Learning to detect visual relations

Julia Peyre

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

Learning to detect visual relations

Soutenue par

Julia PEYRE

Le 29 Août 2019

École doctorale n°386

Sciences Mathématiques de Paris Centre

Spécialité

Informatique

Composition du jury :

Frédéric JURIE University of Caen Normandie	*Président du jury*
Matthieu CORD Sorbonne University	*Rapporteur*
Svetlana LAZEBNIK University of Illinois at Urbana-Champaign	*Rapporteur*
Ivan LAPTEV Inria	*Directeur de thèse*
Cordelia SCHMID Inria	*Directrice de thèse*
Josef SIVIC Inria	*Directeur de thèse*

ENS | PSL

Abstract

We study the problem of visual relation detection in images. We call visual relation a triplet of the form $(subject, predicate, object)$ where the predicate is typically a preposition (e.g. "under", "in front of") or a verb (e.g. "hold", "repair") that links a pair of objects $(subject, object)$. As visual relations are semantic units of intermediate granularity between objects and scenes, better recognizing and localizing visual relations could in turn help develop more accurate and interpretable models for higher-level tasks in image understanding such as image retrieval, captioning or visual question answering.

Yet, detecting visual relations is a challenging task. Learning detectors for visual relations in a fully-supervised set-up requires pairs of bounding boxes to be annotated for all the predicates in the vocabulary. This is extremely costly, especially when the vocabulary of objects and predicates is large. Also, in such set-up, the triplets typically follow a long tail distribution, raising the issue of learning detectors when there is few or no training data. Second, visual relations are subject to important internal variability: in the visual world, the same relation can have different visual appearance, while the same visual entity can be described by different words.

In the first part of this thesis, we address the problem of learning with less supervision by developping a weakly-supervised model to detect visual relations using only image-level annotations. Our model is compositional and can be used to predict unseen visual relations. We also develop a new visual representation of a relation that generalizes well to unseen relations. Finally, we introduce an evaluation dataset, UnRel, for Unusual Relations, that enables us to evaluate without the problem of missing annotations. Our experiments show that, given pre-trained object detectors, object relations can be learnt from weak image-level annotations without a significant loss of recognition performance.

In the second part of this thesis, we tackle the issue of variability of visual appearance of a relation. For this, we complement the compositional module with visual phrase detectors, that are more robust to change of appearance. We also change our formulation from a classification into a fixed number of categories of relations to learning joint visual semantic embeddings. This allows us to generalize to open vocabulary of predicates, through the use of pre-trained word embeddings. Finally, a central contribution of this part is the introduction of analogy reasoning to transfer visual phrase embeddings from seen to unseen visual relations. Experimental results demonstrate the improvement brought by visual phrase embeddings over a purely compositional model and validate the benefits of our transfer by analogy to retrieve unseen triplets.

Résumé

Nous nous intéressons au problème de détection de relations visuelles dans les images. Une relation visuelle peut être formulée par un triplet de la forme *(sujet, prédicat, objet)*, où le prédicat est typiquement une préposition (par exemple "au-dessous de", "devant") ou un verbe (par exemple "tenir", "réparer") qui décrit le lien entre une paire d'objets *(sujet, objet)*. Les relations visuelles constituent des unités sémantiques intermédiaires entre les objets pris de manière isolée et les scènes visuelles complexes où plusieurs objets interagissent. En cela, l'amélioration des modèles de détection de relations visuelles pourrait à son tour aider à développer des modèles plus précis et interprétables pour des tâches complexes de compréhension d'image, telle que la recherche automatique d'image ou la description de contenu visuel.

Cependant, la détection des relations visuelles est une tâche difficile. Premièrement, l'entrainement de modèles fortement supervisés nécessite l'annotation exhaustive des relations visuelles pour l'ensemble des paires d'objets dans les images d'entrainement ; ce type d'annotations est très coûteux, d'autant plus dans le cas où le vocabulaire d'objets et de prédicats est étendu. Par ailleurs, de nombreuses relations visuelles sont peu fréquentes, posant la question de la généralisation des détecteurs aux triplets peu voire pas observés dans les images d'entrainement. Deuxièmement, les modèles développés doivent être robustes face aux variations d'apparence visuelle au sein d'une même relation et à la diversité des descriptions textuelles.

Dans la première partie de cette thèse, nous développons un modèle d'apprentissage de relations visuelles faiblement supervisé, utilisant seulement des annotations au niveau de l'image. Notre modèle, compositionnel, sépare la détection des objets de la prédiction des prédicats, ce qui permet de généraliser à des relations visuelles non observées à l'entrainement. De plus, nous introduisons une base de données d'évaluation, UnRel, constituée de relations rares, qui nous permet de contourner le problème d'annotations manquantes source de bruit au moment de l'évaluation. Nous montrons expérimentalement que, étant donné des détecteurs d'objets pré-entrainés, il est possible d'apprendre des détecteurs de relations visuelles en utilisant des annotations faiblement supervisées au niveau de l'image, avec une précision proche de celle de modèles fortement supervisés.

Dans la seconde partie de cette thèse, nous nous intéressons à la question de variation d'apparence visuelle d'une relation. Pour cela, nous proposons un module holistique, complémentaire au modèle purement compositionnel, qui permet une robustesse supplémentaire face aux changements d'apparence visuelle. Contrairement à la première partie où nous apprenons un nombre fini de classifieurs distincts, notre modèle est formulé en termes d'apprentissage d'espaces visuels et textuels communs. Cela permet en particulier d'utiliser les similarités textuelles pour généraliser à un plus grand nombre de relations. Enfin, nous nous réattaquons au problème de détection de relations visuelles non observées à l'entrainement en proposant un modèle de raisonnement par analogie entre triplets source observés et triplets cibles non observés.

Nos résultats expérimentaux confirment le bénéfice apporté par le module holistique en comparaison d'un modèle purement compositionnel et valident notre modèle de transfert par analogie pour la détection de triplets non observés.

Acknowledgments

I would like to thank Josef, Ivan and Cordelia for giving me the opportunity to work within the Willow department. Thank you for sharing with me your knowledge, your enthusiasm, your efforts and your time. I am grateful for everything I have learnt with you during these last four years.

I would like to thank Svetlana Lazebnik and Matthieu Cord for accepting the role of rapporteurs of my thesis, as well as Frédéric Jurie for agreeing to be part of my jury.

I would like to thank Jean Ponce for critical and encouraging feedbacks about my research. Thank you also for kindly introducing me to many researchers and entrepreneurs.

I would like to thank Francis Bach for giving me the opportunity to do teaching assistance and offering technical guidance.

I would like to thank Eric de la Clergerie and Sylvain Arlot for taking the time to be part of my comité de suivi doctoral.

I would like to thank Alain Marchand for his efficient and friendly technical assistance.

I would like to thank David Dinis, Sabrine Boumizy, Hélène Bessin-Rousseau, Hélène Milome and Mathieu Mourey for helping with administrative procedures.

I would like to thank Guillaume, Jean-Baptiste, Antoine and Yana for their help on the cluster.

I would like to thank all my colleagues from Willow and Sierra for their good mood and very enjoyable discussions. This has been a great support. In particular, I would like to thank Guilhem, Matthew and Margaux for having maintained a supportive and friendly atmosphere in office C314, which greatly contributed to the achievement of this PhD.

I would like to thank my family : Michael, Georges, Catie and Henri, whose constant support and sound advice have been priceless.

Contents

Chapter 1

Introduction

1.1 Goals

What is a visual relation? Our goal is to develop methods for automatic detection of visual relations in images. We call visual relation a relation that holds between objects in an image. For example, in Figure 1-1, there are two objects in the image - a "person" and a "surfboard" - which interact in a way that can be described by the predicate "hold". Such visual relation can thus be formulated as a triplet of the form $(subject, predicate, object)$ where the predicate is typically a preposition (eg. "under", "in front of") or a verb ("hold","ride") that links a pair of objects $(subject, object)$ in an image. In this thesis, we are interested in making the link between pairs of objects in the visual world and triplets in language, i.e. we wish to develop algorithms that automatically map pairs of objects to language triplets and vice-versa. The types of visual relations that we handle are varied: we do not impose any lexical constraints on subject, object or predicate (e.g. a "person" can both intervene as *subject* or *object*, and a *subject* is not necessarily a "person"). The only constraint we impose is grammatical: our algorithms are applicable to relations between two objects that can be described as triplets. In particular, handling relations between more than two objects (e.g. "table between chairs") is out of scope of this thesis.

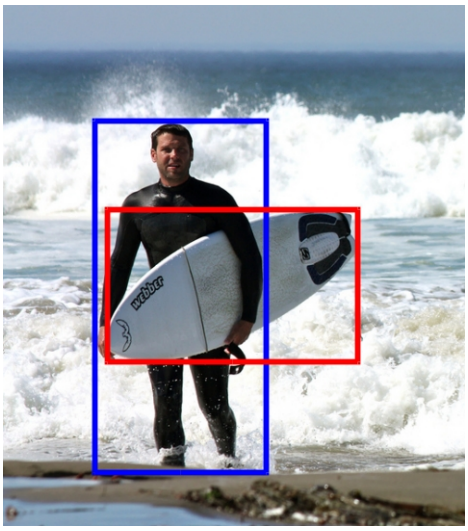


Figure 1-1 – A visual relation is a triplet of the form $(subject, predicate, object)$, here $(person, hold, surfboard)$.

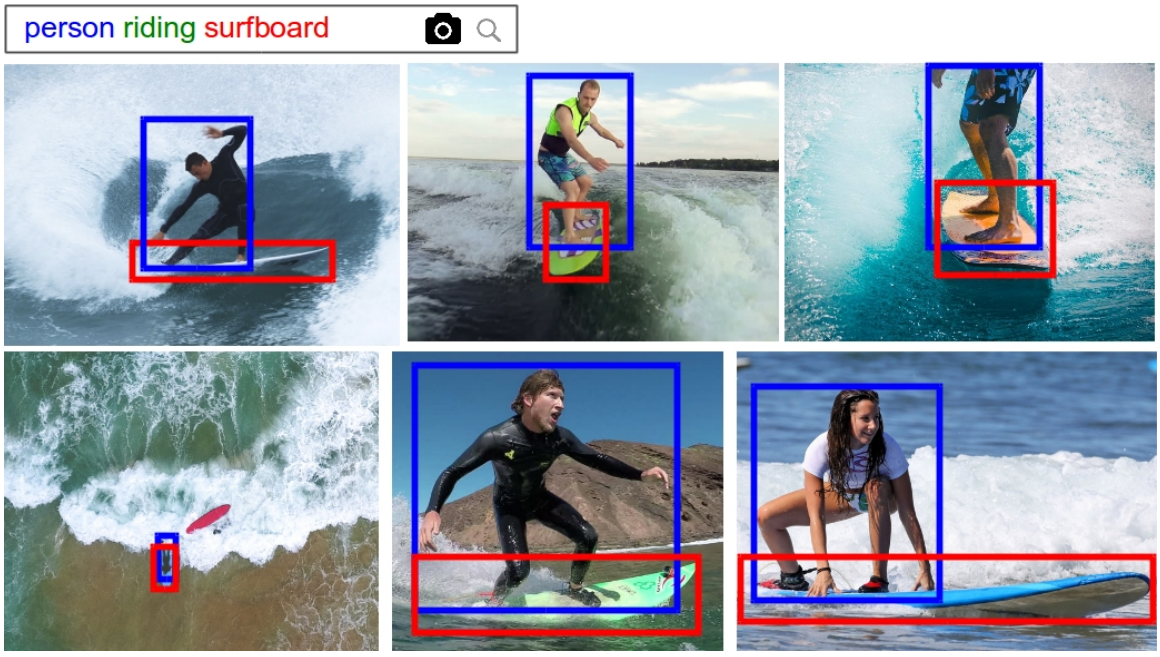


Figure 1-2 – *Text to image, i.e. visual search using text*: for a given triplet e.g. "person riding surfboard", the task is to retrieve images depicting the described interaction and output the location (bounding box) of the corresponding subject and object in the image.

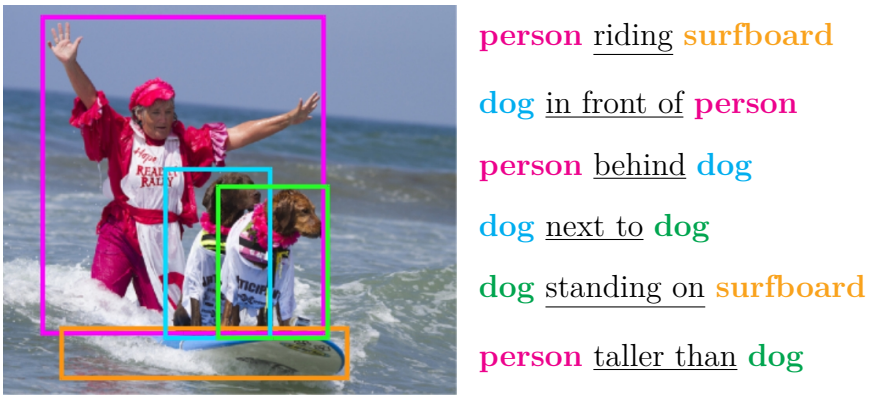


Figure 1-3 – *Image to text*: for all pairs of objects in an image, the task is to output the most likely text descriptions in the form of triplets.

Detecting visual relations. Our goal is to *detect* visual relations in images, i.e. for a given triplet $(subject, predicate, object)$, we aim at localizing the *subject* and the *object* involved by drawing bounding boxes around the corresponding visual entities in the image. We consider two target visual-language tasks. The first one is visual search by text: given a query of the form $(subject, predicate, object)$, the goal is to retrieve images containing objects in the given interaction and return the location (bounding box) of the interacting objects. This task, illustrated in Figure 1-2, typically corresponds to the set-up of image seach with language query constrained in the form of triplet. We will refer to it as the retrieval task. The second task involves producing text from an image: the goal is, for each image and each pair of objects in the image, to predict the most probable language triplet. This process, illustrated in Figure 1-3, corresponds to producing localized descriptions of image content. We will refer to it as the description task. In this thesis, we use both tasks to evaluate our models. Also, when talking about visual relation detection, we do not refer to one task or the other in particular but mean the mapping between the language triplets and visual pairs of objects.

Learning with weaker supervision. The leitmotiv of this thesis is to build models to detect visual relations with limited training data, both in terms of the granu-

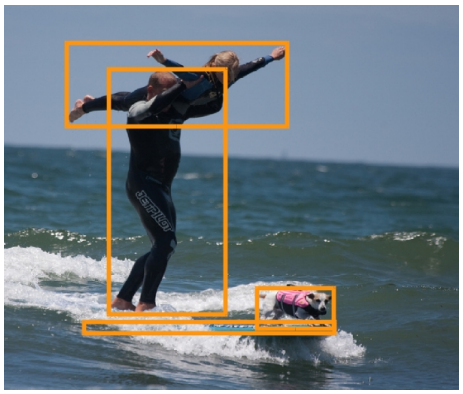


Figure 1-4 – Learning with image-level labels: at training time, we do not know the correspondences between language triplets (on the right) and pairs of candidate object bounding boxes in the image (on the left, represented in orange).

larity of annotations (image-level vs. bounding boxes) and in terms of the number of examples. These two aspects can be re-formulated as: (1) learning with weak supervision, i.e. using only image-level annotations, (2) generalizing to unseen triplets, i.e. transferring our models to triplets with no training annotations.

The first aspect is illustrated in Figure 1-4. We wish to develop a model that learns only from image-level labels, i.e. without knowing the correspondence of the language triplets to their visual entities in the image at training, while still being able to output the location of the triplets in the image at test time. In other words, our model should compensate the mismatch between the level of supervision at training (image-level) and the granularity of the desired output outlining the bounding boxes of objects participating in the relation in the test image (box-level).

The second aspect is to develop models that can generalize to *unseen triplets*, which are triplets of the form $(subject, predicate, object)$ whose components have been seen independently but not in the specific combination in training images. For instance, we might have seen examples of "person ride horse", and "person pet cow" at training but no example of "person ride cow". We also wish to explore generalization to triplets involving totally unseen predicates through the use of pre-learnt language models.

1.2 Motivation

Automatic scene understanding. One of the objectives of computer vision is automatic scene understanding, where the goal is to build models which can automatically analyze and interpret visual content. Recent technical advances such as digital cameras, reduction of storage cost and content sharing through Internet have lead to dramatical amount of visual data, making automatic scene understanding not just interesting but very desirable. So how exactly can automatic scene understanding help us? In a historical perspective, the two Industrial Revolutions allowed us to automatize tasks in industry and manufacturing that were once laboriously done by humans. The result was an increase in productivity, enabling the spread consumer goods in everyday life that make our lifes easier (e.g. washing machines, kitchen appliances, cars). The Digital Revolution we are living now tries to analyze human behaviors and habits to make these tools "smarter", i.e. more interactive, more personalized, more precise. In other words, it attempts to increase the quality of the objects we use everyday. This process is enabled by analyzing large quantity of data to understand the environment and user habits, and in particular visual data. For instance, new smart home devices more and more rely on visual sensors to better serve the user (e.g. indicating whether the refrigerator is empty, or whether an intruder entered the house), industry chains use visual inputs to detect anomalities in production lines, and the development of autonomous vehicles heavily relies on processing large amount of visual content.

Whatever the specific application, visual relations naturally appear as a core building block in scene understanding, as one of the key questions in scene understanding is: what are the objects in the scene and how do they relate to each other?

Visual relations: between objects and scene. Visual relations naturally appear as intermediate entities between objects and scenes as we illustrate in Figure 1-5. On the right hand side of the picture are scenes, which are complex compositions of many

Figure 1-5 – Sadeghi and Farhadi [2011] introduces *visual phrases*, as intermediate visual composites between objects and scenes.

entities that are not easily automatically interpretable, while on the other hand are objects that are isolated scene components, whose appearance may vary greatly as they enter in interaction. In between the simplest semantic units - objects - and the more complex ones - scenes - are intermediate semantic units that are visual relations. Sadeghi and Farhadi [2011] call *visual phrases* these intermediate "chunks of meaning bigger than object and smaller than scene". In a similar spirit, Lan et al. [2013] introduce the notion of *visual composites* which are groups of two or more objects that closely interact. These higher-order composites, that are groups of interacting objects which exhibit spatial and appearance regularities, can be used to reduce the complexity of analyzing a scene.

This idea of intermediate semantic unit also has groundings in psychology and neuroscience. In early 20th century, psychologists have been interested in understanding how humans group visual elements such as lines or points to recognize objects. The results of these studies lead to the Gestalt principles of grouping. For instance, one of these principles, named law of proximity, suggests that elements that are close to each other are perceived as a group. In neuroscience, there is evidence that our brain benefits from spatial regularities to group objects that tend to co-occur [Kaiser et al., 2014; Stein et al., 2015] enabling to spare attentional ressources for efficient

image processing. Visual relations, as intermediate units between objects and complex scenes, have drawn a lot of attention among neuroscientists who tried to discover when and in which region of the brain visual relation recognition occurs [Harel et al., 2013; Stansbury et al., 2013]. Interestingly, a study carried by [Kim and Biederman, 2010] suggests that identifying scene-like relations (i.e. relations involving multiple objects) do not occur after object identification but rather simultaneously with the specification of object shape. This means that, in the human brain, the process of analyzing a scene is not strictly hierarchical from objects to a scene, but rather involves shortcuts for complex, recurrent groups of objects. This question of whether there exist neurons that directly encode higher-level concepts also appears in the conflict between localist and distributed models of the brain, also known in neuroscience as the "grandmother cell" debate [Bowers, 2009].

While our purpose is not to build machines that reproduce the human brain, we can still take inspiration from these studies and speculate that modeling groups of interacting objects is a promising step for scene understanding. The semantic unit of interest in this thesis are visual relations, which are special cases of visual composites that involve interactions between exactly two objects. While other visual composites are definitely worth studying (e.g. object-attribute or higher-order interactions), visual relations are especially interesting in that most scenes can be decomposed into pairwise interactions between objects.

Why visual relation detection? The task of visual relation detection as introduced in Section 1.1 is thus an interesting intermediate goal on the way to scene understanding. In particular, grounding of visual relations, i.e. spatial localization of entities interacting in the image, is desirable both for fine-grained reasoning and interpretability. We illustrate this point in Figure 1-6 by showing examples of concrete applications of scene understanding that require reasoning with visual relations.

- In *image search with a natural language query* (Figure 1-6 (a)), the user wishes

Figure 1-6 – Concrete applications of visual relation detection for scene understanding: (a) image search with natural language queries, (b) image captioning, (c) human-robot interaction and (d) visual question answering.

to retrieve images corresponding to free-form textual queries such as "children playing chess under a tree", often involving the understanding of visual relations (here: "children playing chess" and "children under a tree").

- In *image captioning* (Figure 1-6 (b)), the machine is asked to produce a description of the content of an image in natural language. Informative content involves the objects present (here: "group of people", "swimming pool") in the image and the relations between them (here: people are located *in* the swimming pool, and they are *having dinner*).

- *Human-robot interaction* (Figure 1-6 (c)) requires the ability of robots to interpret natural language instructions from humans. Such instructions often involve localizing objects in the visual world and understanding their interactions. For instance, here, understanding the instruction "put the coffee on the table", requires detecting the objects involved ("coffee", "table") and understanding the desired interactive state between them ("on").

- Finally, *visual question answering* (Figure 1-6 (d)) commonly involves reasoning about the spatial configuration and interaction of objects in images. For an abstract question like "is this image funny?", the machine should be able to relate the person who is on the scale, with the person behind who presses his foot on the scale to make a joke, and the smiling person in the background. Fully interpreting this complex scene requires external knowledge (e.g. what is a scale? who is Obama?), but relating visual elements is a necessary step.

The ability to localize visual relations in images is thus a key step for scene understanding. The first focus of this thesis, which is to perform visual grounding using only image-level labels, is practical. If we could easily obtain box-level annotations for triplets, there would be no need to develop weakly-supervised models. The second focus of this thesis, which is the ability to generalize to unseen visual relations, is not only practical but motivated by the evolution of our world. Whatever the time,

(a) (b) (c)

Figure 1-7 – Examples of novel interactions. Novel interactions between objects keep appearing over time: either people find novel use of existing objects such as in (a) and (b), or new objects are invented like in (c).

unseen visual relations are constantly appearing, with people finding new ways to use objects and new objects being invented every day as illustrated in Figure 1-7. From a learning perspective, enhancing the generalization capability of our model is a desirable feature to ensure its robustness when confronted with new situations.

The link to language. One last element to discuss is the link to language. An important question is: why would we need language to learn visual relations? Animals are certainly able to get some understanding of visual relations without having developed a complex and structured language as the human one. Machine learning algorithms can discover higher-order composites automatically by exploiting visual regularities without the need of language [Li et al., 2011a; Singh et al., 2012]. Our first motivation to use language here is driven by application: humans need language to interact with the machine, either at the beginning to formulate our request, or at the end to convert the results of the machine analysis into a human-readable format. Though we can always communicate with the machine by pressing buttons, language is far more convenient. For instance, imagine the convenience brought by home device that you could command through language, or think of a surgeon whose hands would be busy and who would be able to get machine assistance by speaking. Language is

also attractive for non-experts who do not have the time or the will to learn to use a software. For all these tools, developping the ability to interpret natural language and translate it into specific commands would provide a great value. Our second motivation is that language can provide almost free supervision. There is a large amount of paired visual-language data on the web that is waiting to be used: on social media, users often post images with textual comments or videos with narration. Though using natural language as a source of supervision introduces new challenges such as the presence of noise and ambiguities, language provides additional information that complements purely visual understanding.

1.3 Challenges

Detecting visual relations in images is a challenging task. There are two main types of difficulties: (1) the difficulty of obtaining the right kind and amount of annotations, (2) the variability of a relation both in terms of visual appearance and language description. We illustrate below these challenges in details.

Difficulty of getting annotations at box-level. Detecting visual relations in images requires localizing the subject and the object in interaction by drawing bounding boxes around the corresponding visual entities in the image. Desirable training data for fully-supervised models are thus images with visual relations annotated at box-level, i.e. where bounding boxes are drawn around objects and each pair of objects is labeled with its descriptive triplet. However getting such annotations is extremely costly. We illustrate this in Figure 1-8. Already the first step, i.e. drawing a bounding box around all the objects in the image might lead to hundreds of annotations if the image is cluttered, and this number is squared when moving to the second step of annotating the relations between all pairs of objects. Moreover, such annotations heavily depend on the vocabulary of objects and predicates: if predicates in the vocabulary are not mutually exclusive, multiple words can be suitable to label each pair

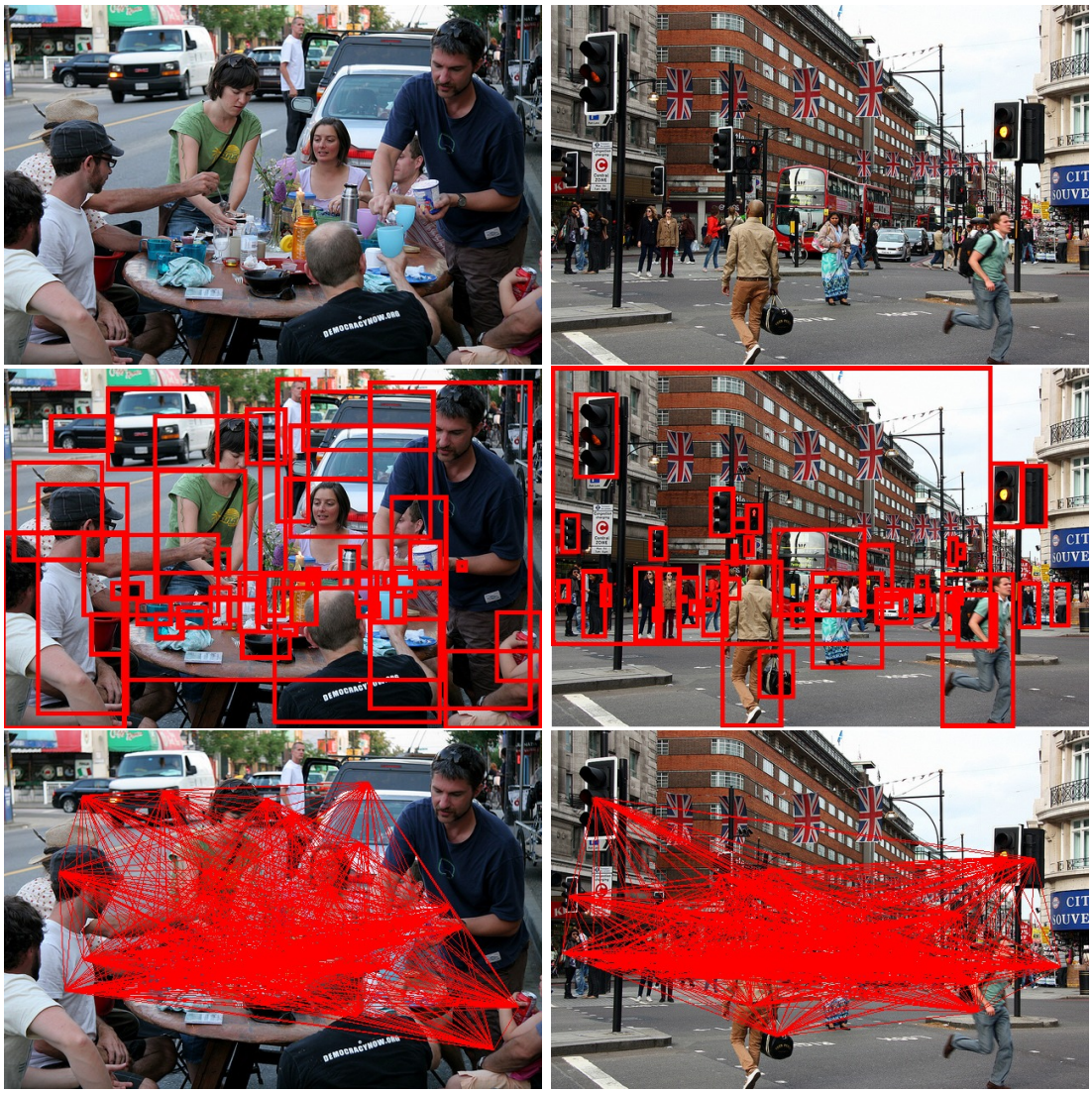


Figure 1-8 – Difficulty of getting exhaustive box-level annotations on two typical images from the COCO dataset: (a) raw images, (b) box-level annotations for all the objects in the image, (c) box-level annotations for all pairs of objects in the image.

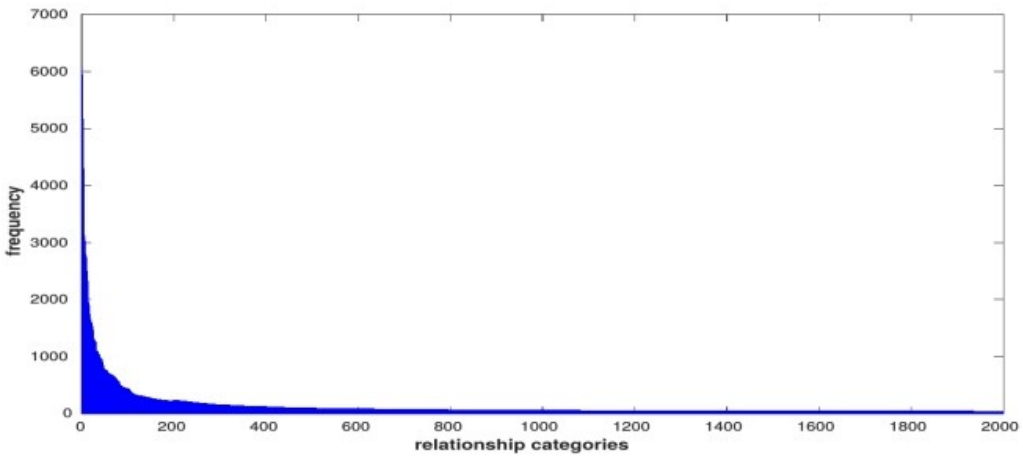


Figure 1-9 – Long-tail label distribution in the Human-Centric Visual Relationship Detection dataset (HCVRD) [Zhuang et al., 2018], a subset of Visual Genome dataset, illustrating the fact that the large majority of relations encountered in current datasets are infrequent.

of objects.

Missing annotations and noisy evaluation. The difficulty of getting annotations at box-level impacts both training, with the limited number of data which is possible to collect to train our algorithm, and evaluation. Indeed, evaluating visual relations consists not only in checking the correctness of the predicted subject, object and predicate categories, but also the accuracy of the localization. Thus, exhaustive box-level annotations are needed at test time. But as for training, the cost of annotation usually leads to datasets with incomplete and missing annotations, a phenomenon that is even more critical when the vocabulary of objects and predicates is large. This leads to noisy evaluation, making both numerical and qualitative results more complicated to interpret.

Long tail distribution of visual relations. Another reason for the difficulty of getting annotations is due to the combinatorial nature of relations. For a vocabulary of N different object classes and K different predicates, the number of possible relations is $N \times N \times K$. For instance, for $N = K = 100$, there are potentially 1 million

possible triplets. As most of these triplets are rare or unseen in the real world, the training data has a long-tail distribution shown in Figure 1-9: the annotations concentrate on very few relations, while most triplets in the vocabulary have few or no training data. Scaling visual relation detectors to a large number of triplets is thus a major challenge.

Variability of visual appearance. The problem of variability of appearance, known as intra-class variability, occurs in many areas of computer vision. It is due to the fact that humans group objects into classes of entities (according to some rules such as their functionality), yet instances of the same class have different visual appearance. In object recognition for instance, cars can have different colors, shapes or can be seen from different viewpoints. The visual variability is tightly linked to the granularity of annotations: the more precise the labeling, the smaller the variability. For instance, there is less variation of appearance between different instances of sports cars, than between instances of motor vehicles. All this creates a challenge for the algorithm that needs to model these variations. Visual relations, which relate two different objects make the problem of modelling intra-class variation even more challenging. Indeed, a visual relation is not just a juxtaposition of objects, but a combination of objects in a certain spatial and appearance arangement which is not unique. The variability of objects is further augmented by the variability of their interactions. We illustrate this in Figure 1-10 where similarly labeled visual relations (e.g. "person ride boat") have yet different visual appearance.

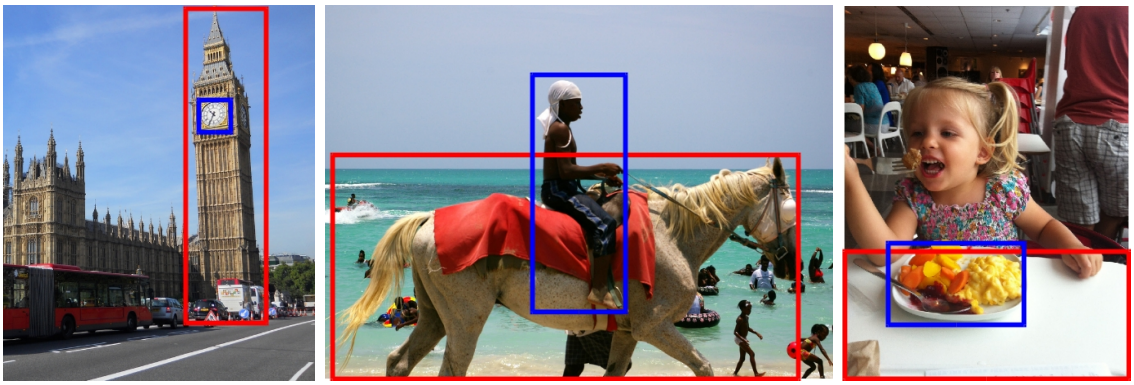
Variability of language descriptions. The problem of intra-class variability is in fact tightly related to the ambiguity of language, where definition and naming of concepts is not rigorous, as it results from a long process of sometimes erratic changes and approximations throughout years of usage. First, the same concept can be described in different manner. For instance in Figure 1-11, the same relation such as "clock on tower" can also be described by other predicates such as "attached to" or

(a)

(b)

(c)

Figure 1-10 – Variability of visual appearance. All the above images are possible answers for the queries (a) “person ride boat”, (b) “person adjust glasses”, (c) “person carry child”.



clock **on** tower
clock **attached to** tower
clock **part of** tower

person **on** horse
person **ride** horse
person **straddle** horse

plate **on** table
plate **on top of** table
plate **lay on** table

Figure 1-11 – Ambiguity of language descriptions. The same word (e.g. "on") can describe different interactions (e.g. "clock on tower", "person on horse" or "plate on table"). Also, the same visual entity can be described in different manners (the triplet "clock attached to tower" is as valid as "clock part of tower" to describe the pair of objects drawn on the left).

"part of". And reciprocally, the same word such as "on" can refer to different visual interactions. Thus, mapping visual and language modalities is difficult as there is no one-to-one correspondence: a given pair of objects in an image could be described by different language triplets, and a given triplet could refer to various visual situations.

1.4 Contributions

1.4.1 Contributions

The contributions in this thesis are described into two chapters.

Weakly-supervised learning of visual relations. In Chapter 3, we focus on two important challenges mentionned in Section 1.3: (i) the difficulty to get annotations at box-level, (ii) the challenge of missing annotations at test time. To address (i), we develop a model, based on discriminative clustering [Bach and Harchaoui, 2007], for detecting visual relations using only image-level labels for relations. For (ii), we propose a novel way to evaluate visual relations without missing annotations by rely-

ing on unusual relations. We introduce UnRel dataset, a dataset of unusual relations, which has two useful properties: first, it allows us to evaluate visual relation detection with reduced level of noise and second to assess the generalization capability of the model on these unusual relations. We use UnRel dataset for evaluation in Chapters 3 and 4. We further explore the visual representation of a relation and introduces a new spatial representation allowing to account for the multimodality of relations.

Detecting unseen visual relations using analogies. In Chapter 4, we attack the three other major challenges described in Section 1.3: (i) the variability of appearance of interactions, (ii) the long-tail distribution of training data, (iii) the ambiguity of language. For (i), we propose to learn visual relations at two different granularities: the compositional model where visual relations are grouped according to their predicate as we did in Chapter 3 and the visual phrase model which is more robust to appearance variation but suffers from scarcity of training data. To generalize well to unseen triplets and overcome the challenges of (ii), we propose to link visual relations through analogies, a concept that has been explored for image generation. Inspired by [Reed et al., 2015], we propose a formulation incorporating analogies as arithmetic operations on relation embeddings. Finally, we address (iii) by learning to map visual and language modalities into a common embedding space which can benefit from pre-trained word embedding vectors to find similar concepts and generalize to unseen predicates.

1.4.2 Publications

Our work led to the following publications:

- J. Peyre, I. Laptev, C. Schmid, J. Sivic. Weakly-supervised learning of visual relations. In *ICCV*, 2017. [Peyre et al., 2017] (Chapter 3)
- J. Peyre, I. Laptev, C. Schmid, J. Sivic. Detecting unseen visual relations using analogies. In *ICCV*, 2019. [Peyre et al., 2019] (Chapter 4)

1.4.3 Software and dataset contributions

The code for the two chapters of this thesis has been publicly released, as well as the new UnRel dataset:

- Weakly supervised learning of visual relations (Chapter 3): https://github.com/jpeyre/unrel. The repository contains the code for training and evaluating our model. Pre-computed object detections and features are provided. We also release pre-trained models to exactly reproduce the results in our chapter.
- Detecting unseen visual relations using analogies (Chapter 4). The code with pre-trained models will be released in https://github.com/jpeyre/analogy.
- UnRel evaluation dataset is publicly released in https://www.di.ens.fr/willow/research/unrel/data/. The directory contains the 1071 images collected on the web, that are annotated at box-level for 76 unusual triplet queries. The evaluation code is available in https://github.com/jpeyre/unrel.

1.4.4 Outline

This manuscript is split into five chapters including this introduction.

Literature survey. We review related work in Chapter 2, by first making links with the fields of object detection and scene understanding, then focusing on works that address the problem of visual relation detection from different perspectives.

Weakly-supervised learning of visual relations. In Chapter 3, we focus on the difficulty of obtaining annotations at box-level and propose a weakly-supervised approach which allows us to learn visual relation detectors using image-level labels only. We also address the problem of missing annotations at test time by introducing UnRel, a new evaluation dataset made of unusual relations.

Detecting unseen visual relations using analogies. In Chapter 4, we first propose to better model the variability of appearance of interactions by combining compositional embeddings for subject, object and predicate with holistic visual phrase embeddings representing triplets. We then tackle the challenge of long-tail distribution of relations by performing analogical reasoning on the visual phrase embeddings to better generalize to unseen triplets.

Discussion. We conclude this thesis in Chapter 5 by summarizing our contributions and proposing directions for future work.

Chapter 2

Literature Review

Visual relations, as we have seen in Chapter 1, are semantic units of intermediate granularity between objects and scenes. In this chapter, we review literature about these different semantic units and underline how they relate, beginning by object detection in Section 2.1, continuing with scene understanding in Section 2.2, and finishing by visual relation detection in Section 2.3.

2.1 Object Detection

Object detection and visual relation detection are entangled tasks. The literature on object detection presents three main interests with regard to visual relation detection. First, as objects are the first level semantic units, visual relation detection methods can naturally build on top of object detection models. Therefore, it is instructive to understand how to build object detectors beforehand. Second, visual relations can be envisioned as more complex objects: where an object (e.g. "person") is defined by the spatial arangements and appearance of its parts (e.g. "legs", "arms"), a visual relation (e.g. "person riding horse") is determined by a specific spatial configuration and appearance of the objects that participate in the relation (e.g. the "person" is on top of "horse" with his legs from either side and the "horse" is walking or galloping).

This makes the literature on object detection a relevant source of inspiration when developping models for visual relation detection. Finally, visual relations between objects can in turn provide useful priors for object recognition. In that, the task of object detection can benefit from better understanding of visual relations. In this section, we first describe the different approaches for object detection, then we review the works that leverage visual relations to improve object detection and finally we explain how object detectors have been used in the context of visual relation detection.

Building object detectors. The task of object detection is to determine where are the objects in the image (object localization), and to which category they belong to (object classification). Object detection is a challenging task due to difficulties such as viewpoint variation, illumination, occlusion, scale, deformation, intra-class variation and background clutter; difficulties that are inherited by visual relation detection. Traditional models for object detection are naturally split into different stages: (1) determining the regions of interest (RoI) in the image, (2) extracting meaningful representations for these regions of interest, (3) classifying each region into an object category. The naive approach for selecting regions of interest is to do exhaustive search at every location in the image [Dalal and Triggs, 2005; Harzallah et al., 2009; Felzenszwalb et al.]. However this process is computationally expensive and produces many regions that are not relevant. More efficient sampling methods have later been proposed, either based on segmentation [Uijlings et al., 2013] or based on selection through a measure of objectness [Alexe et al., 2012; Zitnick and Dollár, 2014]. Earliest object detectors relied on handcrafted, low-level descriptors such as SIFT [Lowe, 2004] or HOG [Dalal and Triggs, 2005] on top of which classification was performed, typically with SVM [Cortes and Vapnik, 1995] or Adaboost [Freund and Schapire, 1997]. One particularly successful approach [Felzenszwalb et al.] that stood out represented objects as a collection of parts in a certain deformable arrangement, accounting for the intra-class variability.

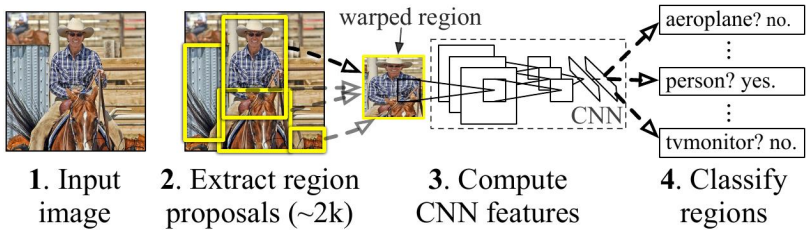
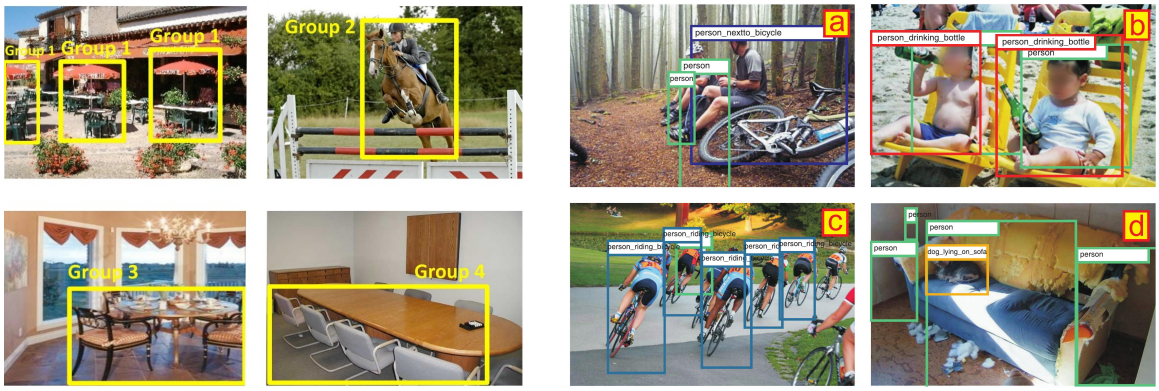


Figure 2-1 – Architecture of R-CNN [Girshick et al., 2014] showing the three typical stages of object detection: (1) region proposals, (2) feature extraction, (3) object classification

The renewed interest in Deep Neural Networks [Krizhevsky et al., 2012] enabled a major improvement in object detection. One main contribution was done by Girshick et al. [2014] who showed that Convolutional Neural Network (CNN) features pre-trained on an auxiliary task could favorably replace handcrafted features for object detection. We display its architecture, R-CNN, in Figure 2-1 as it is a good illustration of a typical object detection pipeline. Further improvements have been added to [Girshick et al., 2014] with SPP-net [He et al., 2014] and Fast-RCNN [Girshick, 2015] that allow to share feature computation across all candidate regions of interest and later with Faster-RCNN [Ren et al., 2015b] which introduces a Region Proposal Network (RPN) to learn object proposals instead of having to rely on external ones such as [Uijlings et al., 2013]. Other approaches such as YOLO [Redmon et al., 2016] and SSD [Liu et al., 2016] adopt a different formalism: instead of viewing object detection as a classification task on top of region proposals, object detection is framed as a regression problem where bounding boxes and class probabilities are predicted at the same time. Such methods allow important speed-up but struggle to precisely localize small objects.

Though object detection has greatly improved in the past years, several challenges remain. One first challenge is related to the detection of small objects, encouraging to explore approaches that can operate at multiple scales such as Feature Pyramid Network (FPN) [Lin et al., 2017a]. A second and complex challenge is to scale object detectors to large number of object classes. This is a major difficulty as labels



(a) (b)

Figure 2-2 – Higher-order composites: (a) groups of objects [Li et al., 2012], automatically discovered, that involve an arbitrary number of objects (e.g. table and chairs around it), (b) visual phrases [Sadeghi and Farhadi, 2011] which are manually defined objects performing an action (e.g. person riding), or a pair of objects interacting with each other (e.g. person drinking bottle).

required for detection are more complicated to collect than labels for classification. YOLO9000 [Redmon and Farhadi, 2017] makes a step in this direction by trying to leverage classification labels to scale to 9000 object categories. Yet, this problem is still largely unexplored.

Modeling contextual information for better object detection. Object detection can be improved by using contextual information, for instance relations between objects. In fact, one of the primary interests for visual relations was motivated by potential improvements in object recognition. Early works incorporate contextual interactions between neighboring objects as an additional cue. In particular, co-occurrence and relative spatial locations between objects have helped to improve object detection in 3D scenes [Hoeim et al., 2006], object categorization [Galleguillos et al., 2008] and multi-class object detection [Desai et al., 2011]. Another line of work changed the reasoning about individual objects, and instead proposed to discover groups of objects that interact in a predictable and structured way. Li et al. [2012], fervent advocates of these *groups of objects*, illustrated in Figure 2-2(a), show

that considering higher-order composites provide a useful prior on object locations with direct benefits to both object detection and scene categorization. Sadeghi and Farhadi [2011] share the same spirit by introducing intermediate composites called *visual phrases*, pictured in Figure 2-2(b), and show that a joint decoding scheme over objects and visual phrases helps object detection. Since Lu et al. [2016a] who framed visual relation detection as a separate task, exploiting contextual cues to improve object detection has not drawn as much attention. We are only aware of the work of Hu et al. [2018] who introduce a differentiable object relation module enabling to reason simultaneously over multiple objects.

Using object detectors for visual relation detection. Detecting objects beforehand is not a pre-requisite for visual relation detection. Indeed, similar to [Sadeghi and Farhadi, 2011], one might directly attempt to learn visual relations as single non-splittable entities. Yet, most of the works on visual relation detection following [Lu et al., 2016a] have chosen to take benefit from object detectors such as [Ren et al., 2015b] to identify regions of interest and extract a generic representation. On one hand, this solution is smart as visual relation detection can thus directly benefit from advances in object detection: higher object recall immediately translates to higher visual relation recall and better representation of objects generally echoes on visual relation recognition performance. On the other hand, building on object detectors draws an upperbound on the performance: when an object participating in a relation is not detected in early step, the error propagates. Such error propagation encourages to keep as many object candidates as possible at the expense of computational complexity. Solutions have been proposed [Zhang et al., 2017c; Yang et al., 2018a] to build relation proposal networks that estimate the degree of relatedness between two objects and discard the unlikely interactions to reduce the complexity. Yet, these attempts still rely on individual proposals for subject and object, ignoring the fact that an object in a group is sometimes easier to detect than in isolation (e.g. if the

object is small or occluded). Finally, we mention an elegant approach by [Kolesnikov et al., 2018] which frames visual relation detection as an object detection problem allowing, with few changes, to directly rely on object detection pipelines.

2.2 Scene understanding

Scenes are the highest level semantic units. Oliva [2009] defines a visual scene as a view of an environment comprised of objects and surfaces organized in a meaningful way. Teaching a computer to understand a scene, i.e. to seize the meaning of the specific organization of objects in a scene, is one of the primarily goals of computer vision. For the neuroscientist David Marr, understanding a scene is "to know what is where by looking. [It is] the process of discovering from images what is present in the world and where it is." [Marr, 1982]. While the task of scene understanding might look intuitive, actually measuring the level of understanding of a computer is not trivial. For many years, object detection and recognition were the most popular metrics to evaluate the capability of an intelligent visual system. Now that significant progress has been made in this direction (see Section 2.1), scene understanding has shifted towards tasks that require higher-level reasoning about object attributes and relationships. In this section, we first give a broad picture of the variety of tasks related to visual scene understanding. We especially focus on scene understanding in 2D static images. Then, we briefly describe the different types of approaches to solve these tasks, and analyze how lower-level semantic units, such as objects and visual relations, have been progressively incorporated to improve scene understanding. We conclude this section by reviewing the current challenges in scene understanding and why visual relations will likely be even more important in this field.

A variety of tasks. Scene understanding materializes in a large growing number of tasks. The reasons for this are twofold. First, scene understanding has a lot of practical applications in diverse domains, demanding specific metrics and databases

to be developed for each so that the performance of an algorithm directly links to a measure of concrete utility. Second, there has always been a desire to build a general multi-task AI that can understand the world around us. As evaluating the capabilities of such machine is hard to capture into a single measure, people have decoupled the measure of intelligence into proxy tasks, each of them enabling to evaluate specific properties of the model in an interpretable manner. We now review some of the practical tasks related to scene understanding, and try to understand, for each of them, what good performance might tell about the capabilities of the algorithm. It is to note that most taks we review here involve a joint understanding of images and text, as natural language provides a convenient way for humans to interact with a computer and interpret the output of an algorithm.

One of the earliest tasks of interest was *image captioning* [Farhadi et al., 2010; Kulkarni et al., 2011; Ordonez et al., 2011], where the goal is, given an image, to produce a description of its content in free-form natural language. Success in image captioning is viewed as a form of understanding, as recognizing the main objects in an image, where they are, and how they interact is a pre-requisite for generating a meaningful description. A closely related task is *image-text matching* [Farhadi et al., 2010; Hodosh et al., 2013] where textual queries should be matched to the most relevant images in a database, and reversely, images should be associated to their most relevant descriptions. This is entirely a retrieval problem, which does not evaluate the ability to generate grammatically correct sentences, contrary to image captioning. In terms of understanding, image-text matching requires at least as much visual reasoning capabilities as image captioning (potentially even more if the images in the database are hard to distinguish). While both tasks have a lot of practical applications, they do not provide easily interpretable information on the properties of the algorithm being evaluated. For many images in current datasets, a rough recognition of the salient entities is enough to achieve good performance.

This has drawn attention to the task of *visual question answering (VQA)* [Antol

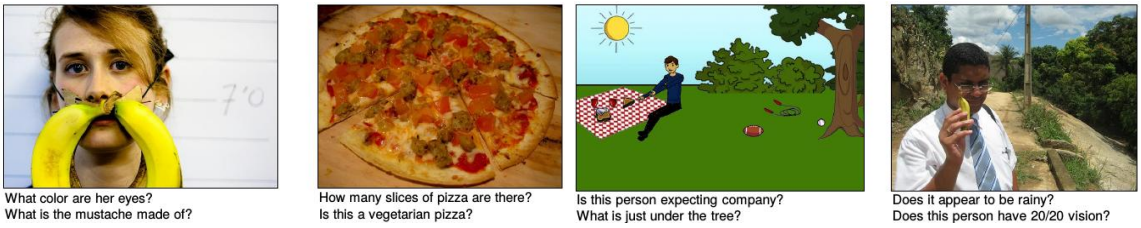


Figure 2-3 – Examples of free-form, open-ended questions from the Visual Question Answering dataset [Antol et al., 2015].

et al., 2015; Ren et al., 2015a; Gao et al., 2015; Zhu et al., 2016; Krishna et al., 2016]. In VQA, the goal is, given an input image and a question in natural language, to output an answer either among a set of pre-defined answers or in free-form text. Many researchers believe that VQA, also sometimes called Visual Turing Test [Geman et al., 2015], is the ideal test for AI systems. For instance, Lehnert [1977] claimed that "when a person understands a story, [they] can demonstrate [their] understanding by answering questions about the story. Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding". In fact, the view of Lehnert [1977] points out two main properties of the VQA: first, the task is flexible in the sense that the granularity of the question can be adjusted throughout time; second, the task is interpretable as the agent should answer in a human-readable format and is forced to give as many details as required to be judged as correct. If we now view image captioning as a specific case of VQA, with the generic question "What happens in this image?", we realize how poor are the requirements of image captioning when compared to the complex questions in current VQA datasets that involve attribute recognition, counting or object localization, as shown in Figure 2-3.

Variants of VQA have been later proposed, complexifying the properties a successful agent should acquire. For instance, in the Visual Dialog task [Das et al., 2017b], illustrated in Figure 2-4(a), a conversational agent has to answer a series of questions about an image, requiring to overcome the challenges of VQA (i.e. reasoning about an image), and the challenges of dialog (i.e. reasoning about past history of a con-

(a) (b)

Figure 2-4 – Extension of VQA tasks: (a) visual dialog [Das et al., 2017b] where a conversational agent answers a series of questions about an image, (b) knowledge-base visual question answering [Wu et al., 2016] which exploits external facts to answer a question about an image.

versation) at the same time. Another interesting variant of VQA is to marry it with the field of *common-sense knowledge* to answer questions that require not only visual reasoning but also reasoning about external facts stored in a Knowledge Base [Chen et al., 2013; Zhu et al., 2014; Sadeghi et al., 2015a]. This variant, explored notably in [Wu et al., 2016; Wang et al., 2017, 2018; Marino et al., 2019], could allow to tackle more complex questions and push further the understanding of visual scene as shown in Figure 2-4(b).

Among all the properties an intelligent agent should have is one that is absolutely crucial. It is the ability to perform visual grounding, i.e. to spatially localize named entities in the image. This property is desirable for fine-grained reasoning, and importantly, for interpretability. The spatial counterpart of image captioning is *dense captioning* [Johnson et al., 2016] whose task is to generate descriptions for salient regions in images. It can also be seen as an extension of object detection where object categories are free-form text. Figure 2-5 provides a good illustration of the link between dense captioning and these two tasks. Many other set-ups have been

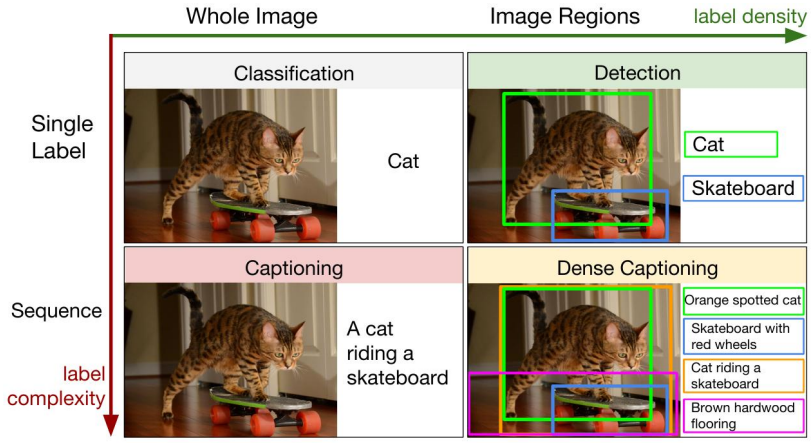


Figure 2-5 – Illustration of the dense captioning task as introduced by [Johnson et al., 2016] that pushes forward image captioning in terms of spatial localization and object detection in terms of label complexity.

invented to evaluate the ability to make correspondences between textual phrases and image regions. For instance in *referring expression comprehension* [Kazemzadeh et al., 2014; Mao et al., 2016; Hu et al., 2016], the agent should localize, by drawing a bounding box, the region in the image that is described in the referring expression. Borrowing the same idea of pointing to image regions, Zhu et al. [2016] extend visual question answering datasets by linking the object mentioned in the question to its corresponding region in the image.

While the set-ups described above ground a textual description to a unique box in the image, *phrase localization* [Plummer et al., 2015, 2017] aims at finding correspondences between regions in an image and all noun phrases in its description. This effort to visually ground entities in a sentence is formalized in a more structured way in *scene graph grounding* [Johnson et al., 2015]. A *scene graph* is a graph-structured representation of a content of a scene where nodes encode objects and edges store the relationships between them. Grounding a scene graph to an image, as illustrated in Figure 2-6, means associating each node of the scene graph with a region in the image that respects the relationships specified in its surrounding edges. By explicitly modeling constraints between entities and imposing visual grounding, scene graph

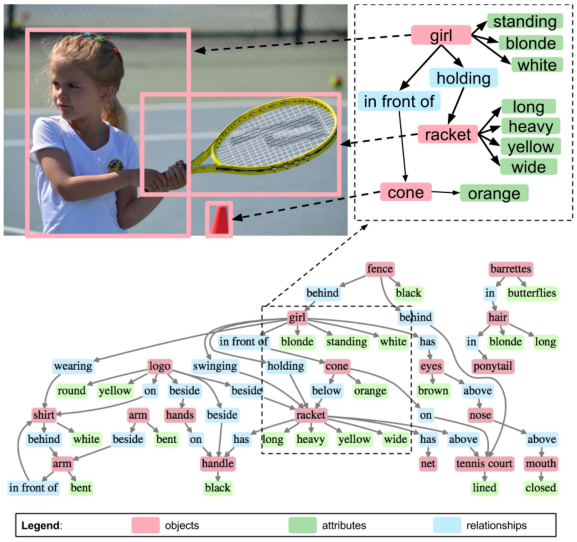


Figure 2-6 – Illustration of scene graph grounding [Johnson et al., 2015]. Each node of the graph is grounded to an object in the image outlined by a bounding box. The relations between objects are encoded through edges in the scene graph. This formulation also allows to add attributes to object nodes.

representations have been benefitting tasks like image retrieval [Johnson et al., 2015], VQA [Teney et al., 2017], image captioning [Yao et al., 2018], image generation [Johnson et al., 2018] and automatic captioning evaluation [Anderson et al., 2016].

In the same way that researchers have been interested in directly evaluating visual relation detection on its own rather than implicitly using it for scene understanding, Xu et al. [2017] introduce *scene graph generation*, consisting in generating a visually-grounded scene graph from an image. In this perspective, the task of visual relation detection that we address in this thesis is a special case of scene grounding/generation, corresponding to the situation where objects and relationships are handled in isolation, ignoring the mutual context between them, just as if each relation was part of a different image.

All these tasks are different ways to evaluate and push forward our understanding of visual scenes. Despite the diversity of these tasks, there are two fundamental properties emerging that seem key for success: (i) the first one is the ability to *reason* about the nature and relations between visual entities, (ii) the second one

is the capacity to *ground* entities in the image, for interpretability and fine-grained reasoning. With this respect, visual relation detection is a key part in the development of intelligent visual systems.

From monolithic to visually grounded approaches. Early approaches in captioning [Kulkarni et al., 2011; Li et al., 2011b; Mitchell et al., 2012; Elliott and Keller, 2013] aimed at filling a pre-defined caption template with objects and attributes detected from visual data. This resulted in captions that were visually grounded and interpretable, but not natural due to the rigid language template. With the success of recurrent neural networks (RNN) in sequence-to-sequence learning for machine translation [Bahdanau et al., 2015], template-based models were replaced by a language generation model based on RNN. The standard architecture takes the form of an encoder/decoder where the entire image is encoded by a convolutional neural network (CNN), and then decoded with a RNN to produce a description [Kiros et al., 2014]. Compared to previous template-based approaches, the generated descriptions are much more natural. Works in image-text matching follow a similar pattern where global representations of images and text are projected into a *joint visual-semantic embedding space*. Two main approaches have been explored: one based on Canonical Correlation Analysis (CCA) [Hardoon et al., 2004] that learns projections maximizing the correlation between the two modalities in the projected space [Andrew et al., 2013; Gong et al., 2014; Klein et al., 2015; Yan and Mikolajczyk, 2015], the other one typically optimizes a ranking loss with Stochastic Gradient Descent (SGD) so that the projected vectors of correct correspondences are closer than those of incorrect matchings in the joint space [Wang et al., 2016a; Faghri et al., 2017]. In VQA, the focus has been to learn efficient and expressive fusion mechanisms to merge global representations of the image and the textual question in a space where the decision is easy [Fukui et al., 2016; Kim et al., 2016; Ben-younes et al., 2017, 2019]. Most of these works are monolithic, in the sense that they reason about global image and

language representations. While they propose simple and effective models, they often lack of visual grounding and interpretability.

Among modern deep neural network approaches, the work of Karpathy et al. [2014] was one of the first efforts to attempt visual grounding of noun phrases. For this, they break down the sentence into multiple fragments using a dependency parser and the image into regions with an object detector and try to infer fragment-region alignment. Karpathy and Fei-Fei [2015] subsequently improved this model in particular by replacing the dependency parser by a Bidirectional Recurrent Neural Network [Schuster and Paliwal, 1997] allowing to encode each word in its context. At about the same time, Fang et al. [2015] proposed to generate image captions based on detected concepts. The novelty of their approach is to train a word detector on image captions in a weakly-supervised way, allowing to scale to larger number of concepts that commonly appear in captions. Yet, one weakness of such models is that salient image regions are pre-defined with respect to a set of visual detectors, as if blinders were put in advance to decide where the model should attend. Also, each grounded region-fragment contributes equally to the final score, which does not reflect the reality of an image where the decision usually depends only on a small subset of informative regions.

Visual attention, a major concept introduced by Itti et al. [1998] and revisited in particular by Xu et al. [2015] for image captioning, was and still is a fertile ground to address part of these weaknesses. In particular, it allows to go beyond pre-defined visual categories by computing an attentional spatial map concentrated on image regions relevant for the final task. This process is illustrated in Figure 2-7. Attention models had a lot of success, especially in the VQA task [Yang et al., 2016; Lu et al., 2016b; Zhu et al., 2016], where being able to choose which regions to attend conditional on the question is a desirable property. Attention mechanisms are great tools to inject flexibility in a model, yet, for visual grounding, attentional maps are not always semantically interpretable, i.e. do not necessarily correspond to visually meaningful regions.

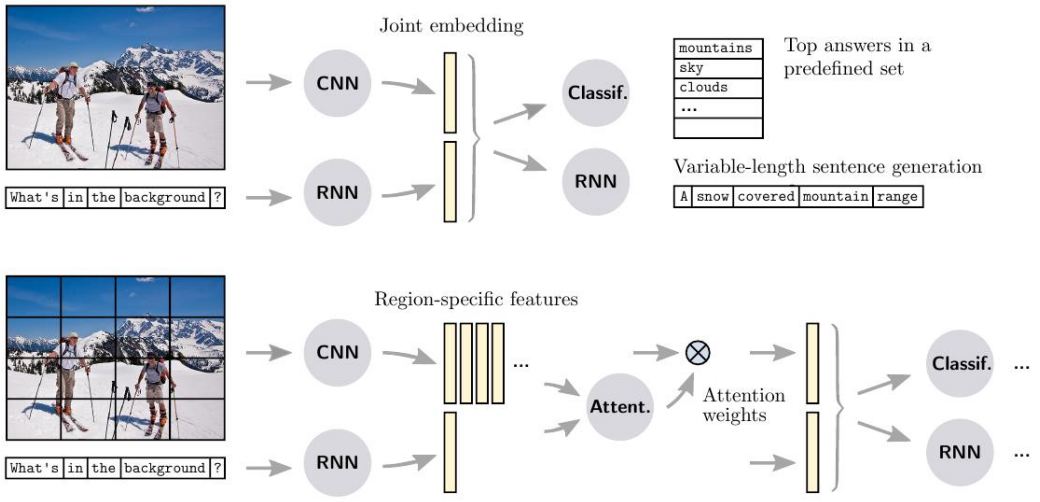


Figure 2-7 – Illustration of the attention mechanism [Wu et al., 2017]: (1) the monolithic approach (top row) which reasons at the level of global image and text features, (2) an approach using attention mechanism (bottom row) which learns to reweight the contributions of each region depending on the task.

Most recent approaches combine a bottom-up and top-down attention mechanism that works as follows: (1) the bottom-up (i.e. objects to scene) mechanism proposes a set of interesting image regions, for instance using an object detector or class-agnostic proposals, (2) the top-down (i.e. scene to objects) is a selection mechanism that reweights the features of the different regions depending on the task. This mechanism lead to improvements in phrase localization [Rohrbach et al., 2016], image-text matching [Lee et al., 2018], image captioning [Lu et al., 2018a] and VQA [Anderson et al., 2018], also allowing better explainability of the algorithm.

While these works enable more interpretable outputs, they perform visual grounding of entities in isolation without explicitly constraining the relationships between them. This leaves behind key structural and semantic information.

Incorporating visual relationships. There is much to gain in understanding entities in combination, that is to say in interpreting the meaning of an entity by looking at the context around. If the first type of context used was global, at the level of the image, the need to build more discriminative models together with the

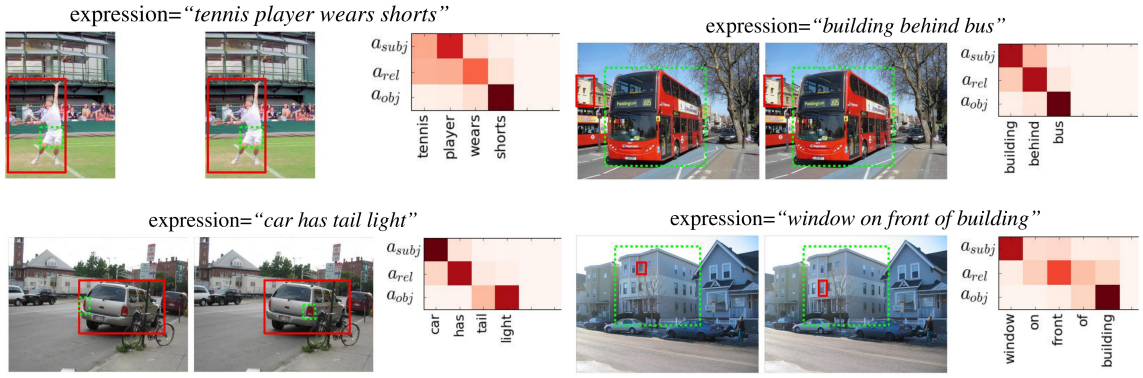


Figure 2-8 – Automatic extraction of (subject, predicate, object) by learning soft attention weights a_{subj}, a_{rel}, a_{obj} activated by different fragments of the description [Hu et al., 2017].

progress in object detection encouraged the use of local context provided by neighboring entities. In an image, local context can be extracted by looking at the mutual configuration and appearance between objects. For instance, in referring expressions, Nagaraja et al. [2016]; Yu et al. [2016] found that an explicit encoding of the visual differences between objects of the same category outperforms global image context. Similarly, encouraging competition between image regions in the objective function also proved useful for generating discriminative captions [Mao et al., 2016; Vedantam et al., 2017]. While these works encourage to reason over mutual context in images, they do not benefit from the rich and informative language structure and semantics. Yet, the grammatical structure of a sentence indicates how noun phrases are linked through verbs and prepositions. One of the simplest versions of this idea is to encourage neighboring words in the sentence to be grounded to the same region in the image [Karpathy and Fei-Fei, 2015]. Other works [Wang et al., 2016b; Xiao et al., 2017] exploit pre-defined relational constraints extracted from the sentence by an external dependency parser to find a coherent alignment between parts of images and parts of sentences. As an alternative to relying on an off-the-shelf parser that might not be well tuned to the specific dataset, one approach is to extract grammatical structure in a soft manner using attention weights. For example, Hu et al. [2017] learn to decompose a sentence into a triplet of the form (subject, predicate, object)

and exploit this structure to ground visual entities. We illustrate some of their results in Figure 2-8. Yu et al. [2018] follow the same spirit by learning to decompose an expression into a (subject, location, relationship) structure which is better suited for referring expressions.

A natural way to encode local context is to represent a scene in the form of a scene graph, where nodes are entities and edges encode their relations. Initially, scene graph representations have been used in computer graphics for scene generation. For instance, Chang et al. [2014, 2015] parse a textual input into an abstract scene template capturing the objects present in the scene and the relations between them, and use this representation to generate a compatible 3D scene. Later, scene graphs have been used as an intermediate representation for image retrieval in abstract [Zitnick et al., 2013] and real-world [Johnson et al., 2015] scenes, video retrieval [Lin et al., 2014a], 3D scene semantic parsing [Kong et al., 2014] and visual question answering [Teney et al., 2017]. This view of translating a scene into an intermediate structured representation where reasoning is more explicit has also been explored through *neural module networks* [Andreas et al., 2016a,b; Johnson et al., 2017b] in VQA. The underlying idea, illustrated in Figure 2-9, is to build, for each question, a specific neural network, made of re-usable modules, that is able to perfectly adapt to the question.

Other types of neural networks have flourished [Scarselli et al., 2009; Battaglia et al., 2016; Santoro et al., 2017], exhibiting various structures designed to capture core properties for *relational reasoning*. These formulations have demonstrated important gain in terms of reasoning capabilities, but many have only been tested on synthetic datasets such as [Johnson et al., 2017a]. Most recent approaches on real-world datasets such as [Norcliffe-Brown et al., 2018; Cadene et al., 2019] represent an image by a set of localized region features, linked in a graph, and use this structured representation to refine each region representation based on its neighbors, producing relation-aware region representation. Yao et al. [2018]; Lu et al. [2018b] adopt a similar formulation and further propose to use semantic knowledge from visual relation

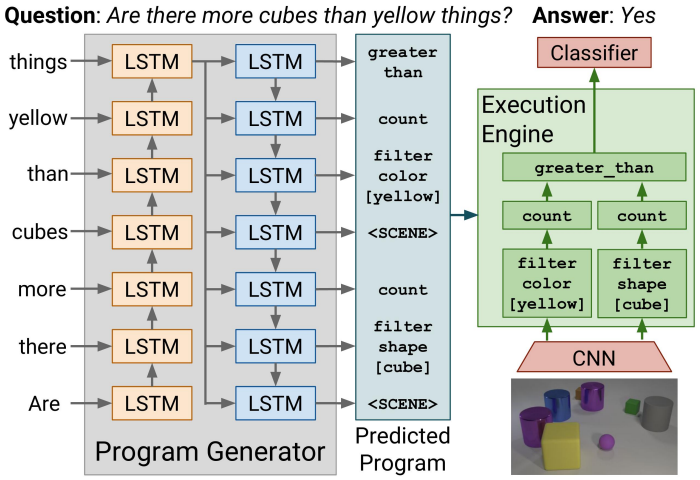


Figure 2-9 – A neural module network [Johnson et al., 2017b] encodes the structure and semantic of a question (e.g. *"Are there more cubes than yellow things"*) into a deep neural network made of re-usable modules (e.g. `filter color`, `filter shape`, `greater_than`). In particular, the relations between objects directly translate into the structure of the network.

detectors learnt on external datasets such as Visual Genome [Krishna et al., 2016].

Improving evaluation of scene understanding. While important numerical improvements have been achieved in scene understanding, future progress in this field is likely to be tightly linked to the capacity to develop new evaluation procedures. Recent studies by Agrawal et al. [2016]; Jabri et al. [2016]; Hendricks et al. [2018] have indeed demonstrated that current algorithms largely benefit from the bias in training data to reach top performance, without deep understanding of the image. Even the most recent state-of-the-art approaches [Anderson et al., 2018; Lu et al., 2018a] in image captioning yet based on top-down and bottom-up approaches have been shown to produce fluent language before using visual information from image regions [Rohrbach et al., 2018]. A typical symptom is object hallucination in a caption that current metrics such as CIDEr [Vedantam et al., 2015], METEOR [Banerjee and Lavie, 2005] or SPICE [Anderson et al., 2016] do not penalize. In this regard, it might be that we learn to give advantage to methods that successfully exploit dataset biases rather than methods that indeed reason about visual content.

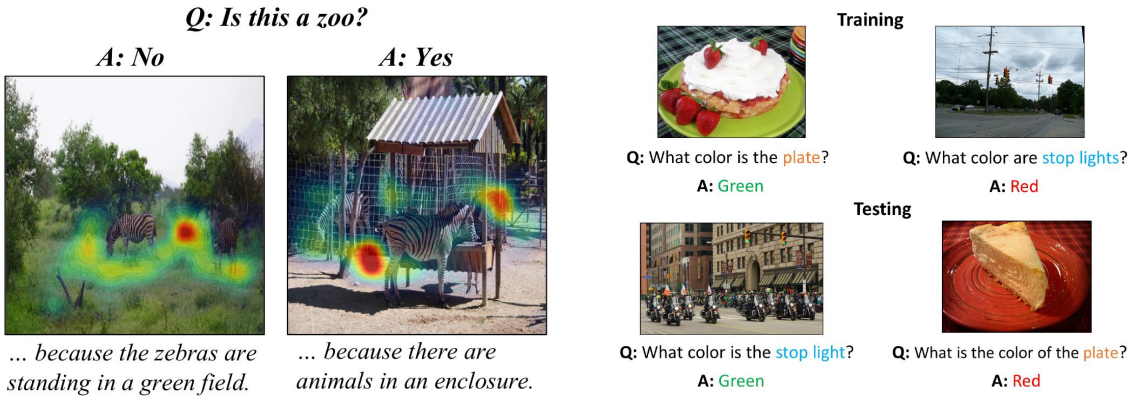


Figure 2-10 – Active areas of research in scene understanding: (a) grounding visual explanations in images [Park et al., 2018], (b) transferring to unseen combinations [Agrawal et al., 2017].

There are two possible answers to this problem. The first one is to constrain the algorithms to produce interpretable, visually grounded outputs that would be explicitly evaluated. For instance, Das et al. [2017a] propose to evaluate attentional maps produced by VQA models by comparing them to human attention. Scene graphs annotations of Visual Genome dataset [Krishna et al., 2016] can also be used to evaluate visual grounding of phrases in image-text matching [Engilberge et al., 2018] or visual question answering [Hudson and Manning, 2019]. Another interesting line of work in this direction is visual explanation [Hendricks et al., 2016a; Park et al., 2018; Zellers et al., 2019], illustrated in Figure 2-10(a), where the machine should also provide a justification, either textual or visual, to its decision.

A second answer can be seen in the development of new tasks [Xie et al., 2018; Hu et al., 2019; Liu et al., 2018a] and datasets [Goyal et al., 2017] less prone to bias, as well as new metrics [Rohrbach et al., 2018] that better correlate with human judgment. In particular, evaluating on new situations is quite interesting as biases do not help anymore. For instance, Hendricks et al. [2016b]; Venugopalan et al. [2017]; Agrawal et al. [2018] have been interested in producing captions that involve novel objects, while Agrawal et al. [2017] have focused on evaluation of unseen compositions of seen

concepts in the context of VQA. We provide an example of such unseen combinations in Figure 2-10(b).

Our work in this manuscript also builds on these ideas as (1) we are interested in grounding visual relations to image regions, (2) we aim to transfer to unseen combinations.

2.3 Visual Relationship Detection

Visual relations are intermediate visual composites that allow to connect objects to scenes. In this part, we review the literature on visual relation detection using different perspectives. We first begin by a brief history of the related tasks and datasets. Second, we study how visual relations have been represented in the literature. Third, we analyze related work in the light of two different views, either regarding a visual relation as a combination of objects (compositional view), or as a whole entity (holistic view). We then inspect the literature with respect to two axes developed in this thesis: first, the type of supervision used to learn visual relations; second, the way to generalize to unseen visual relations.

2.3.1 From action recognition to visual relation detection

Visual relation detection takes root in the field of action recognition, whose goal is to predict the activity of a person (e.g. “walking”, “dancing”). Understanding human activities is a more complex task than recognizing objects in isolation as it also involves the analysis of interactions between those entities. For instance, it is not enough to discriminate a “bicycle” from a “motorcycle”, one also needs to disambiguate subtle configurations such as “riding a bicycle” from “pushing a bicycle”. This requires a joint reasoning about the human and the object configurations and appearance as illustrated in Figure 2-11.

Thus, earliest works in action recognition had to find strong visual cues on which to

Figure 2-11 – Examples of diverse actions involving the object "bike" from Lu et al. [2016a]. Action recognition is one level more difficult than object recognition as it also requires disambiguating subtle configurations between objects.

rely. There were two main approaches: one on videos relying on motion cues [Schüldt et al., 2004; Blank et al., 2005; Niebles et al., 2006; Laptev et al., 2008], another one, not less challenging, on static images based on human shape or pose [Wang et al., 2006; Ikizler et al., 2008; Maji et al., 2011]. But these works have only modelled the human without taking into account additional context given by the objects in interaction. Gupta et al. [2009] are the first to envision an action as an interaction of a human and an object and exploit spatial and/or functional constraints to complement appearance cues from the human. Other works by Desai et al. [2010]; Yao and Fei-Fei [2010b] followed, confirming that with adequate modeling of mutual context and pose, one could recognize actions without temporal cues, at least to a certain extent. We refer to Guo and Lai [2014] for an exhaustive survey of action recognition in still images.

Advances in computer vision are strongly correlated with datasets that provide new goals and challenges to overcome. Yet, earliest action recognition datasets such as PASCAL VOC [Maji et al., 2011] and Stanford 40 Actions [Yao et al., 2011] were small, with a handcrafted vocabulary of mutually exclusive action classes. A shift was created by Lu et al. [2016a] with the introduction of the Visual Relationship Dataset (VRD) that significantly expands the vocabulary of interactions and offers box-level annotations for relations. Compared to previous datasets, it brought two main novelties: (1) the types of interactions are much more diverse, including actions, spatial relations, comparatives or any interaction that can be formulated as a triplet of the form $(subject, predicate, object)$, (2) the number of possible triplets is large

	# images	# objects	# predicates	# triplets
Visual Phrases	2,769	8	5	13
COCO-a	4,413	80	140	1,681
V-COCO	10,346	80	26	554
HICO-DET	47,774	80	117	600
VRD	5,000	100	70	6,672
UnRel	1,071	41	18	76
HCVRD	52,855	1,824	927	28,323
Visual Genome[1]	107,077	53,304	29,086	572,613
Open Images	100,522	57	10	329

Table 2.1 – Statistics of different datasets for visual relation detection. We indicate the total number of images (train and test), the number of object and predicate categories as well as the number of unique triplets.

($\sim 6K$ triplet types) and includes unseen relations at test time. In comparison, the Visual Phrases dataset of Sadeghi and Farhadi [2011] has only 17 categories (among which 13 triplets), all seen at training. Lu et al. [2016a] also formalized the shift from action recognition by defining the new task of *visual relationship detection* where the goal is to detect pairs of objects in an image and classify the relation between them. This effort has been further extended by larger datasets such as the Visual Genome dataset [Krishna et al., 2016] which contains as much as $\sim 40K$ different types of relationships. An important variant is the task of *human-object interaction* detection which involves visual relations where the subject is a human and the predicate is an action. While action recognition focus on predicting the human action solely, human-object interaction also jointly models the object in interaction. This topic naturally attracts a lot of interest as studying human behavior has many practical applications. Constantly larger datasets, in terms of vocabulary, have been developed such as V-COCO [Gupta and Malik, 2015], COCO-a [Ronchi and Perona, 2015], HICO-DET [Chao et al., 2015, 2018] and HCVRD [Zhuang et al., 2018]. We provide some details on the statistics of these datasets in Table 2.1.

[1]Split VG80K, after the cleaning process of Zhang et al. [2019]. Other cleaning processes have been used, resulting in different number of categories.

These datasets have pros and cons. On one hand, they encourage research towards methods that can better generalize, as the number of different triplets is large and visual training data cannot cover all possibilities. Yet, on the other hand, these datasets introduce new issues during evaluation, as many of them have missing and/or ambiguous annotations due to a less controlled data collection process. Recently, new datasets on visual relations tend to adopt a more controlled annotation set-up. For instance, the Open Images dataset [Kuznetsova et al., 2018] proposes only 329 different types of triplets, carefully chosen such that: (1) they are not obvious, i.e. they cannot be simply deduced from object co-occurrences and spatial proximity, (2) they are well defined in advance, contrary to previous datasets which authorize free-form annotation, (3) they are exhaustively annotated in the image.

We also share this spirit that seeks to build a more controlled evaluation set-up. In Chapter 3, we introduce a new dataset, UnRel, for evaluating visual relation detection without missing annotations and further evaluate our model in Chapter 4 on it.

2.3.2 Representing a visual relation

Learning visual relations has relied on many different cues such as human body, object appearance, image context and even external cues from language or knowledge bases. We review how these cues have been represented and used in the literature.

Human body and parts. When a human is involved in a visual relation, as in the case of human-object interaction, informative cues can be learnt by observing the human body such as its shape [Wang et al., 2006], appearance [Delaitre et al., 2010], pose [Thurau and Hlavác, 2008; Ikizler et al., 2008] or body parts [Maji et al., 2011; Raja et al., 2011; Yao and Fei-Fei, 2012]. While many efforts have been invested in the development of structured mid-level representations for action recognition [Yao and Fei-Fei, 2010a; Delaitre et al., 2011; Desai and Ramanan, 2012], today's most used representations are based on high-level CNN features extracted from the region

enclosing the human. There have been recent improvements in the field of human understanding such as segmentation [He et al., 2017], pose [Cao et al., 2017] or 3D surface-based representation [Güler et al., 2018], yet only few works [Chao et al., 2015; Shen et al., 2018] have tried to exploit these cues for human-object interaction.

Object appearance. Objects provide strong cues for understanding an interaction. The appearance, shape, size and functionality of an object indeed put important constraints on the authorized interactions. In a recent study, Zellers et al. [2018] have performed a detailed analysis on the Visual Genome dataset [Krishna et al., 2016] and have found that, given the categories of subject and object involved in a relation, predicting the predicate that most frequently co-occurs with the object categories already gives 70% chance of success. According to them, such astonishing proportions occur because most relations in this dataset are geometric and possessive (such as clothing, part entities) which are much less ambiguous than semantic relations (such as activities). Rather than relying on the object categories, most works in visual relation detection use object appearance which contains richer information on the state of the specific object instance in the image (e.g. viewpoint, attributes). This choice could be related to the Visual Memex model [Malisiewicz and Efros, 2009] which advocates to go beyond categories and encode the local appearance. In recent works [Chao et al., 2018; Gkioxari et al., 2018; Shen et al., 2018; Gao et al., 2018], the object appearance is encoded by a high-level feature extracted from the object region with a CNN pre-trained on ImageNet [Deng et al., 2009] and often fine-tuned for object detection. Other works such as [Lu et al., 2016a; Dai et al., 2017; Li et al., 2017a; Zhang et al., 2019; Qi et al., 2018] extract a feature from the box enclosing the union of subject and object to capture the joint appearance of the interacting objects. A recent interesting work by Yang et al. [2018b] remarked that appearance features extracted in this manner are too much biased towards object category which is harmful for generalization, and instead proposed to learn object-agnostic features

using a strategy inspired by self-supervised learning methods [Zhu et al., 2017].

Subject-object spatial configuration. Additionally modeling the spatial configuration between the subject and the object boxes has proved beneficial. Desai et al. [2010] represent the relative location of subject and object into a sparse binary vector, where each bin encodes a canonical configuration such as "next to" or "above" (Figure 2-12(a)). For each object category, Maji et al. [2011] formulate the relative spatial location as a mixture model in order to capture the variability of spatial configurations (e.g. depending on the view: front or side) (Figure 2-12(b)). In addition to the relative location, Prest et al. [2011] include other cues such as the relative scale, overlap and the euclidean distance between subject and object boxes. Interesting works in human-object interaction such as [Delaitre et al., 2010] have also attempted to model spatial relations between object and human parts (Figure 2-12(c)). Most recent works [Plummer et al., 2017; Zhuang et al., 2017; Zhang et al., 2017a] typically rely on a small dimensional vector encoding the relative location between subject and object boxes, as well as other handcrafted cues such as aspect ratio or relative scale, that is fed into several non-linearities of a neural network. An alternative approach explored by Dai et al. [2017]; Chao et al. [2018] is to explicitly encode the location of subject and object boxes in the form of binary masks (Figure 2-12(d)).

Related types of cues such as 3D, surface normals or segmentation might be advantageously incorporated to help reason about spatial relations and disambiguate occlusions. To the best of our knowledge, such cues have not been explored yet in the context of visual relation detection. This might change in the future as tools to estimate these cues from 2D images are rapidly evolving [Hoeim et al., 2006; Wang et al., 2015; He et al., 2017].

Image context. Detecting visual relations consists in making a local prediction about the relation between two objects in an image, in isolation. Yet, undeniably, using context of what is around can help. Plummer et al. [2017] have found that

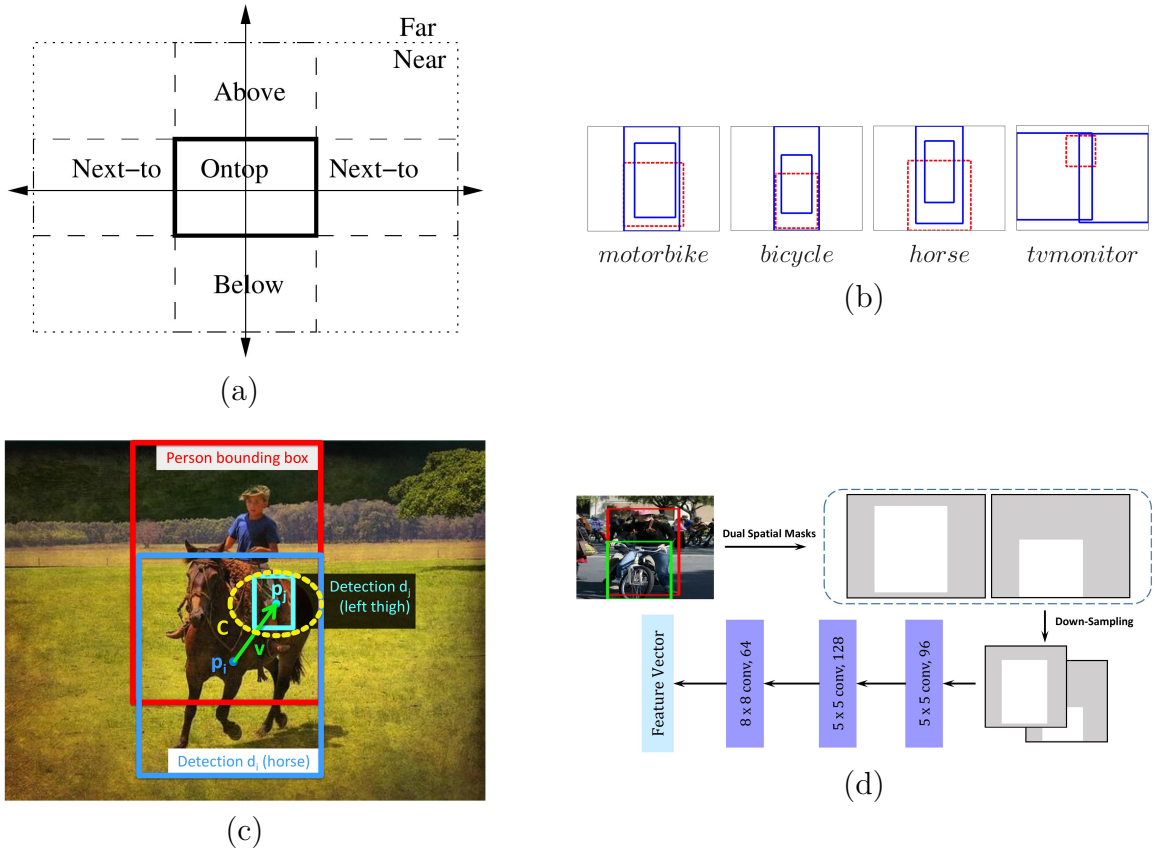


Figure 2-12 – Different ways to encode the spatial configuration between subject and object boxes: (a) the spatial feature of Desai et al. [2010] bins the relative location between subject and object into one of the 7 canonical relations, (b) for each of the 4 object types, Maji et al. [2011] fit a two component mixture model that can capture the possible spatial configurations between the person (blue) and the object (red), (c) [Delaitre et al., 2010] model interactions between object and human body parts (e.g. leg of a person) by measuring their relative scale-space displacement, (d) Dai et al. [2017] encode the spatial configuration as a binary mask with 2 channels, one for the subject and one for the object, that is processed by a few convolutional layers.

the absolute location and size of a region in the image are already informative, as people tend to describe big, central objects in the image. The appearance of the whole scene can be used as an additional context feature [Delaitre et al., 2010], as well as the appearance of neighboring objects [Gupta et al., 2009]. Even only adding immediate surrounding context has been shown useful by Dai et al. [2017] who extract the appearance of objects from its enclosing box augmented by a small margin. One step further are the works on scene graph generation [Xu et al., 2017; Li et al., 2017b; Newell and Deng, 2017; Liang et al., 2017; Yang et al., 2018a; Woo et al., 2018] that jointly reason over all relations in the image at the same time. These works show that exploiting this kind of global context can help resolve ambiguities that occur when relations are considered in isolation and is beneficial for visual relation detection.

Language. One can learn to recognize visual relations without using language, but language might be used as a guide to orient our attention on what and where to look for. In particular, language comes with a compositional structure where small language chunks usually correspond to informative visual entities. In visual relation detection, the triplet formulation as $(subject, predicate, object)$ provides a strong cue on the entities to be grounded and how they relate. Both structural and semantic information can thus be exploited. In action recognition, semantic cues have first been used by Yao et al. [2011] who decompose an interaction into a sparse vector on attribute action bases extracted from image descriptions (e.g. verbs describing the human actions such as "riding" or "sitting"). As recent datasets are moving towards larger vocabulary of objects and predicates, scarcity of training data is becoming one of the main challenges, and leveraging linguistic regularities is even more relevant. Also, the evaluation of visual relation detection as a description task involves, just like image captioning, to not only recognize the interaction but also find the right words to express it. Lu et al. [2016a] regularize the predictions of the visual model with a language prior indicating the likelihood of a triplet based

on a pre-trained Word2vec language representation [Mikolov et al., 2013]. Yu et al. [2017]; Plesse et al. learn to distill linguistic knowledge into a deep network for visual relation detection and confirm that using language cues improves the predictions. Liao et al. [2017b] even predict the relation using only semantic cues of object categories and spatial cues, without relying on appearance. Linguistic cues have also been exploited in joint visual-semantic embedding frameworks, widely used in image-text matching. In visual relation detection, such approaches have been recently explored by [Plummer et al., 2017; Zhang et al., 2019] who learn to map visual features for different combinations of subject, object and union regions with linguistic features from pre-trained word embeddings. There are in fact two components behind the success of integrating language priors in visual relation detection: (1) the ability to exploit statistical dependencies between object and predicate categories, (2) the use of external knowledge. One alternative way to exploit statistical dependencies between objects and predicates explored by [Xu et al., 2017; Li et al., 2017a; Yin et al., 2018] consists in learning correlations between objects and predicates features through message passing. Similar to the use of external knowledge from linguistic corpus, this provides priors on the set of plausible relations, for instance telling us that the triplet *"cat eat fish"* is common while the triplet *"fish eat cat"* is very unlikely. We detail this point in the next paragraph.

Knowledge base (KB). Using pre-trained word embeddings learnt on large text corpora such as Wikipedia is one way to exploit common-sense knowledge about facts that are likely to happen in everyday world. Another way to represent and use such information in a more structured way is through knowledge base graph [Suchanek et al., 2007; Carlson et al., 2010; Movshovitz-Attias and Cohen, 2015]. Knowledge base graphs are repositories of entities and rules which can be used for problem solving. We illustrate this form of knowledge representation in Figure 2-13. Researchers have explored various types of entities (i.e. what a node should represent), rules (i.e.

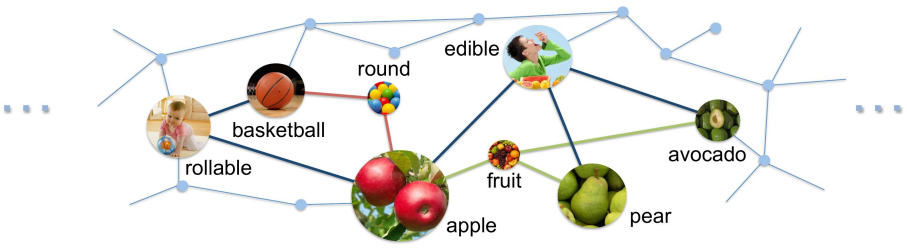
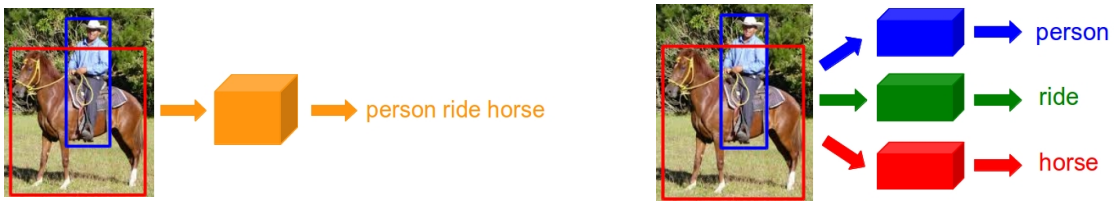


Figure 2-13 – Example of knowledge structure provided by [Zhu et al., 2014]: different concepts (e.g. objects, attributes, affordances) represented by nodes are linked by relations of various nature, shown by edges of different colors.

what types of relations between nodes should be learnt) and knowledge sources (i.e. on which type of data should it be learnt, e.g. text corpora, visual data). Extracting knowledge from visual sources has been first attempted with NEIL [Chen et al., 2013], a computer program that automatically discovers common-sense relationships from web images. It uses object detectors to improve semantic understanding that in turn allows to learn detectors for new object categories. By repeating this process, the program can constantly increase its knowledge of new concepts. Sadeghi et al. [2015a] further extend this work by considering more general relationships and apply it to the task of visual relation verification. Tightly linked is also the work of Zhu et al. [2014] who demonstrate benefits of using a KB to reason about object functionalities. For visual relation detection, KB approaches have recently been explored by Gu et al. [2019] which use common-sense facts from ConceptNet [Speer and Havasi, 2013] to refine visual features.

In both Chapters 3 and 4, our visual representation is a concatenation of individual CNN appearance features of subject and object complemented by spatial configuration features. Chapter 3 investigates a quantized spatial representation based on Gaussian Mixture, while Chapter 4 adopts a simpler version. In Chapter 4, we exploit language cues to learn joint visual-semantic embeddings.



(a) holistic (b) compositional

Figure 2-14 – Illustration of holistic versus compositional approaches: (a) in holistic approaches, a visual phrase detector is learnt for each triplet (e.g. "person ride horse"), (b) in compositional approaches, separate detectors are learnt for the subject (e.g. "person"), the predicate (e.g. "ride") and the object (e.g. "horse") whose confidence scores are combined.

2.3.3 Learning triplets: from holistic to compositional approaches

Earliest works in visual relation recognition [Sadeghi and Farhadi, 2011; Prest et al., 2011; Hu et al., 2013] learn holistic detectors for visual relations, which consists in training separate detectors for each category of relation as we illustrate in Figure 2-14(a). Such approaches are suited for small vocabulary datasets with many training data available for each class. Yet, with the recent development of datasets such as Visual Relationship Detection [Lu et al., 2016a] built on open vocabulary rather than on a pre-defined set of categories, the number of triplets in vocabulary is so large that most of them have little, if any, training examples. In this context, previous holistic approaches are not scalable anymore. This encourages the development of compositional models, which, instead of learning a separate detector for each visual relation, compose it from simpler visual primitives that can be shared across multiple visual relations. We illustrate this approach in Figure 2-14(b). Such change of view is naturally guided by the structure of language where visual relations are expressed in the compositional form of a triplet, whose individual components can be observed in isolation, or as part of different interactions. In visual relation detection, this approach was popularized by Lu et al. [2016a] who break down the task in two stages: first detecting the objects, second predicting the predicates. Since then, this

generic view has been largely adopted and refined [Li et al., 2017a; Dai et al., 2017; Zhang et al., 2019]. One immediate advantage of compositional approaches is that computational complexity is reduced from learning $N \times N \times K$ different triplet detectors to learning only $N \times K$ detectors, with N being the number of objects and K the number of predicates in the vocabulary. Another benefit is that detectors for individual components can be combined to recognize unseen combinations. However, compositional models often sin by their lack of expressiveness: the meaning of a word (e.g. "on") indeed changes when entering in a combination (e.g. the interaction in "glass on table" is different from the one in "painting on wall"). In other words, the *context* of an interaction, provided by the subject and object, modifies the interpretation of the relation. Visually, this translates into different appearances and spatial configurations that generic compositional detectors might fail to capture.

This tradeoff between expressiveness and complexity is one of the major challenges when modeling relational data in large vocabulary setting [Sutskever et al., 2009; Bordes et al., 2013; Socher et al., 2013a]. We refer to Nickel et al. [2015] for an exhaustive review of relational modeling, especially in the context of knowledge graphs where relational modeling is crucial. The question at stake is: how can we *scale* to large number of categories while maintaining a good *accuracy*? The work of Jenatton et al. [2012] is a good illustration of this dilemna. In this work, the authors try to scale well to a large number of relations by: (1) capturing expressivity through various orders of interaction with n-grams, (2) sharing parameters across relations through matrix factorization. The first point, i.e. modeling different orders of interaction through n-grams, has been explored in visual relation recognition as well [Sadeghi et al., 2015a; Elhoseiny et al., 2016; Plummer et al., 2017]. The higher the n in n-grams, the more expressive it is, but also the less training data gets available. As learning a detector for each possible composite is computationally expensive, interesting works such as [Rosenthal et al., 2017] have proposed to learn a strategy to determine in advance, for each relation, the right type of detectors that

Figure 2-15 – The work of Divvala et al. [2014] is a good illustration of the trade-off between expressiveness and complexity. The question is about how to scale to a large number of categories (*anything*) while still accounting for the appearance variation within each category (*everything*).

should be learnt. Also relevant is the work of Divvala et al. [2014], illustrated in Figure 2-15, who propose to automatically discover the vocabulary of n-grams that cover appearance variations for a large number of concepts. The second point, i.e. sharing parameters through tensor factorization, has also been applied in visual relation detection, notably by Hwang et al. [2018] to model relationship priors. Another elegant solution to reduce the computational complexity without sacrificing expressiveness is the use of translation embedding first introduced in KB by Bordes et al. [2013] then applied in visual relation detection by Zhang et al. [2017a]. This approach, called VTransE, proposes to learn a low-dimensional space shared by objects and predicates, where predicates could be interpreted as vector translation between objects, i.e. $predicate \approx subject - object$. A recent work by Hung et al. [2019] extends VTransE to better generalize to rare visual relations by expressing a predicate embedding as $predicate \approx union(subject, object) - subject - object$. An alternative approach adopted by Misra et al. [2017]; Zhuang et al. [2017] is based on the idea that the same interaction (e.g. "riding") with different context (e.g. "horse", "snowboard") should result in different classifiers (e.g. "person riding horse", "person riding snowboard"). In this view, generic classifiers for objects and predicates are combined

to produce the final triplet classifier which is specific to a predicate among its context objects. It is also interesting to relate these types of approaches to the works of Mitchell and Lapata [2008]; Dinu and Baroni [2014] in distributional semantics which propose different ways to form the representation of textual phrases from the representations of individual words.

In Chapter 3, we adopt the view of generic, compositional detectors for visual relations. In Chapter 4, we come back to this dilemna and propose a hybrid model that combines both compositional and holistic views.

2.3.4 Learning with less supervision

One important question in visual relation detection is about the level of supervision. Detecting visual relations in images requires grounding a triplet of the form $(subject, predicate, object)$ into visual entities, i.e. delimiting the regions in the image that correspond to the subject and the object in that interaction. Training fully-supervised models requires two types of supervision: (1) from language side, the text description should be split into a triplet, (2) from visual side, the *subject* and *object* in the text should be mapped explicitly to image regions. Most works in visual relation detection use this type of supervision. But is this a realistic level of supervision to ask? If not, can we learn to detect visual relations with less supervision?

The right kind of supervision. This question can be reformulated as follows: which types of annotations are easy to collect? On the web, we could exploit the pairing of images with their description. Yet, most of the descriptions are written in free-form text, with no pre-defined structure, and thus asking for triplet annotations might look like a strong request. In reality, parallel advances in natural language processing, have led to important improvements in dependency tree parsers such as [Manning et al., 2014], allowing to extract subject-predicate-object triplets from natural language sentences. This has been shown in the context of scene graph

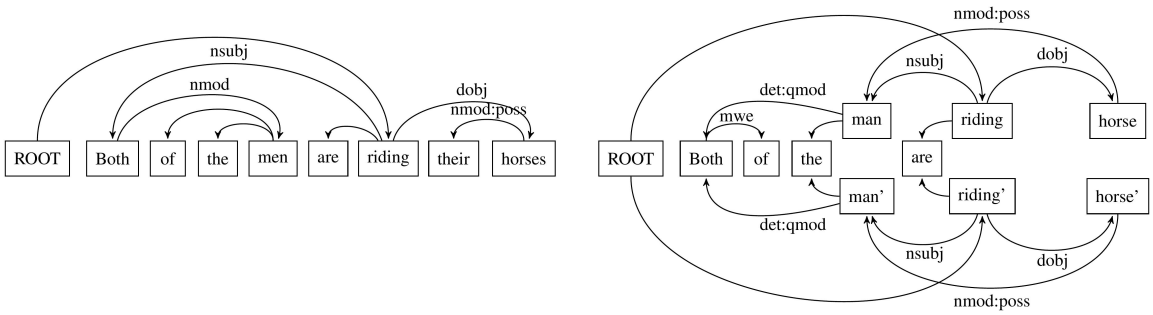


Figure 2-16 – Schuster et al. [2015] show that a scene graph representation (right) can be created automatically from a natural language description by using dependency parsers (left).

extraction by Schuster et al. [2015]. An example is provided in Figure 2-16. Aligned with most works in visual relation detection, we thus suppose in this thesis that full supervision in the form of triplets is given from language side, and focus on discussing the different levels of supervision for visual data. From visual side, the assumption that we have access to annotations that map triplets to pair of bounding boxes in the image seems too strong as already discussed in Chapter 1.3. It seems more reasonable to only require image-level supervision that would indicate the presence of a visual relation in an image without its localization. The setup we explore in this thesis is semi-supervised. We ground on recent state-of-the-art object detectors that provide us with a set of candidate objects. Our interest focus on learning the pairs of objects that interact and the nature of their interaction using only image-level labels for relations. This level of supervision has also been adopted in [Mallya and Lazebnik, 2016] who use a pre-trained object detector to propose candidate human regions and learn actions using image-level labels only. It is also frequently used in recent works on image captioning [Anderson et al., 2018; Lu et al., 2018a] and image-text matching [Lee et al., 2018] where bottom-up attention mechanism are based on pre-trained fully-supervised object detectors.

Weakly-supervised learning of visual relations. Exploiting image-level labels for localization tasks has been mostly explored for object detection [Song et al., 2014;

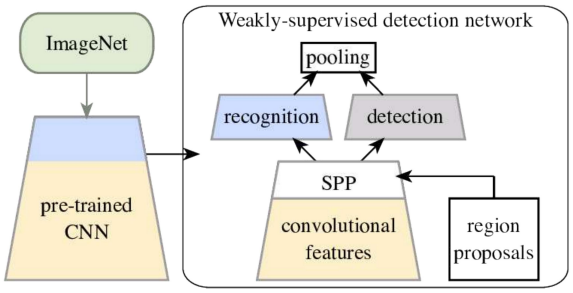


Figure 2-17 – Bilen and Vedaldi [2016] propose a weakly-supervised model for object detection based on two-branch: the first one is responsible for selecting the objects of interest, the second one for classifying them. Both classification and detection scores are aggregated to produce image-level scores.

Bilen et al., 2015; Oquab et al., 2015; Cinbis et al., 2017]. Most methods rely on a Multiple Instance Learning (MIL) framework introduced by [Dietterich et al., 1997] where an image is interpreted as a bag of regions annotated with a binary label. In visual relation detection, very few works have addressed weakly-supervised learning so far. To the best of our knowledge, Prest et al. [2011] were the first to tackle this problem for human-object interaction. Yet, their model rely on reasoning about the single highest scoring human detection in the image which might not work well in current datasets with multiple people in the same image. Gkioxari et al. [2015] built on the same idea of treating the object as an unknown latent variable, relying only on the supervision provided by the human box. This type of formulation, which uses full supervision for the target object and weak supervision for the context object, has also been explored in referring expressions [Yu et al., 2016; Nagaraja et al., 2016; Hu et al., 2017]. Inspired by recent advances in weakly-supervised object detection such as [Bilen and Vedaldi, 2016; Kantorov et al., 2016], a complete weakly-supervised model for object and visual relation detection has been recently proposed by Zhang et al. [2017b]. Their architecture build on the two-branch network of Bilen and Vedaldi [2016], illustrated in Figure 2-17, that performs simultaneously region selection and classification.

Discriminative clustering. In Chapter 3, we propose a novel way to detect visual relations based on a discriminative clustering framework [Xu et al., 2004; Bach and Harchaoui, 2007]. Discriminative clustering aims at separating data into clusters that can be recovered by a discriminative classifier. More precisely, for a set of N data points represented by d-dimensional features $X \in \mathbb{R}^{N \times d}$, the problem is to recover the assignments $Z \in \{0,1\}^{N \times K}$ into K clusters and a set of classifiers $W \in \mathbb{R}^{d \times K}$ that minimize the discriminative clustering cost:

$$\min_{Z \in \mathcal{Z}, W} \ell(Z, XW) + \Omega(W) \tag{2.1}$$

where $\mathcal{Z}$ defines the set of possible assignments, ℓ is the loss function and Ω is the regularization. In this thesis, we use DIFFRAC [Bach and Harchaoui, 2007], a special case of discriminative clustering where ℓ is the square loss and Ω is the Tikhonov regularization. In this framework, it is easy to incorporate weak supervision in the form of constraints on latent assignment variables $Z \in \mathcal{Z}$. The flexibility for incorporating constraints that are specific to each problem makes it an interesting approach for many computer vision applications such as image cosegmentation [Joulin et al., 2010], localizing actions in videos [Bojanowski et al., 2013, 2014, 2015] or learning from narrated instruction videos [Alayrac et al., 2016]. In particular, our work in Chapter 3 uses the online algorithm proposed by Miech et al. based on Block Coordinate Frank-Wolfe [Osokin et al., 2016] which allows us to scale discriminative clustering to large number of training instances.

Learning without annotations? Using methods such as discriminative clustering enables to learn from classification labels instead of detection ones. But collecting classification labels in a clean, large-scale set-up still has a cost [Deng et al., 2009]. To break the deadlock, some works such as [Divvala et al., 2014; Chen and Gupta, 2015] have proposed to learn from web images, by exploiting the performance of web image search engines. The idea is to collect positive examples for a given category

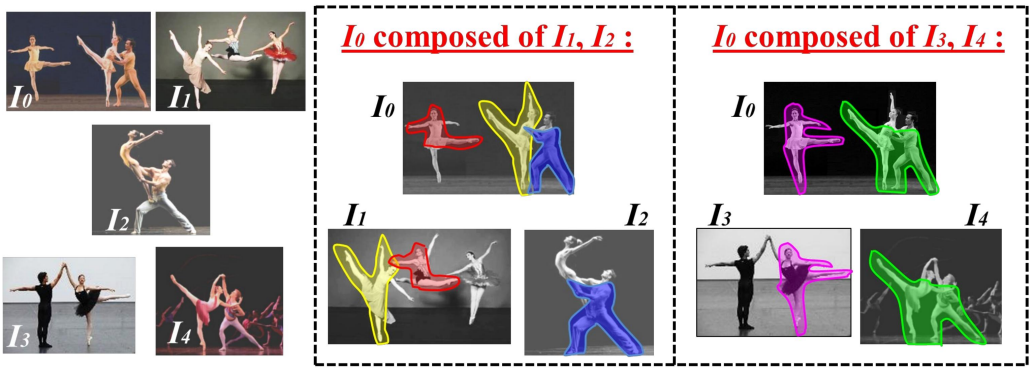


Figure 2-18 – [Faktor and Irani, 2012] propose the model of "Clustering by Composition" for unsupervised discovery of visual categories. The idea is to define a cluster as a group of images where one image can be easily composed from pieces of other images in the cluster, but difficult to compose from images outside.

by entering it as a query in the search engine and taking the top returned images as positives. *Webly-supervised* models enable to learn from a large amount of data at no annotation cost, but face other issues such as noisy and incomplete labeling, which increase as the query becomes more complex. To the best of our knowledge, webly-supervised models have not been used for visual relation detection in images but have been applied for related tasks involving visual relations such as automatic discovery of common-sense relationships [Chen et al., 2013; Sadeghi et al., 2015a]. Yet, all the types of supervision described above suppose paired image-sentence, either manually annotated or collected from the web. Such alignment between visual and text data already provides supervision. Can we learn visual relations in a fully unsupervised setting, i.e learning only from images? Exploiting the regularities in visual data to discover meaningful patterns has been attempted for automatic discovery of actions [Niebles et al., 2006; Wang et al., 2006] and object classes [Sivic et al., 2005; Faktor and Irani, 2012; Cho et al., 2015]. We illustrate one of these approaches in Figure 2-18. In this case, unsupervised discovery of visual categories is performed by searching for statistically significant regions that co-occur between images. For visual relation detection, current works still rely on paired image-text corpus.

In Chapter 3, we propose a novel way to learn visual relations using only weak

supervision for relations. In Chapter 4, our setup is fully-supervised.

2.3.5 Generalizing to unseen visual relations

The question of supervision discussed above is tightly linked to the problem of recognizing entities for which there is no training data, also called *zero-shot learning*. In visual relation detection, this problem occurs as soon as using datasets with a large vocabulary of objects and predicates, which highly increases the chance of encountering unseen combinations at test time. Being able to transfer models from seen relations at training to unseen relations at test time is thus extremely desirable. Here we review some of the methods for transfer learning under four different angles: (1) sharing knowledge through attributes, (2) exploiting relations between labels, (3) using distributional semantics, (4) transferring with analogy.

Sharing knowledge through attributes. A core component in zero-shot learning is to define a measure of similarity between entities. Such measure allows to transfer knowledge from one seen entity to a similar but unseen one. One way to do this is to have an intermediate representation of an entity into a set of shared properties. The best illustration of this is found in object recognition literature, where *attributes* have served as intermediate units to measure similarities between objects. One important work has been done by Farhadi et al. [2009] who have proposed to view object recognition not anymore as the task of naming the object category but as the task of describing it in terms of attributes (e.g. shape, materials, parts). Such view allows to describe even an unseen object, whose name is unknown, but whose attributes have been encountered in other objects. Tightly related is the work of Lampert et al. [2009], illustrated in Figure 2-19, that uses semantic attributes as coupling between seen and unseen object classes. Farhadi et al. go one step further by localizing attributes in the image and reasoning about their spatial arrangement. The shared spirit of these works is to build a compositional representation of an object in terms

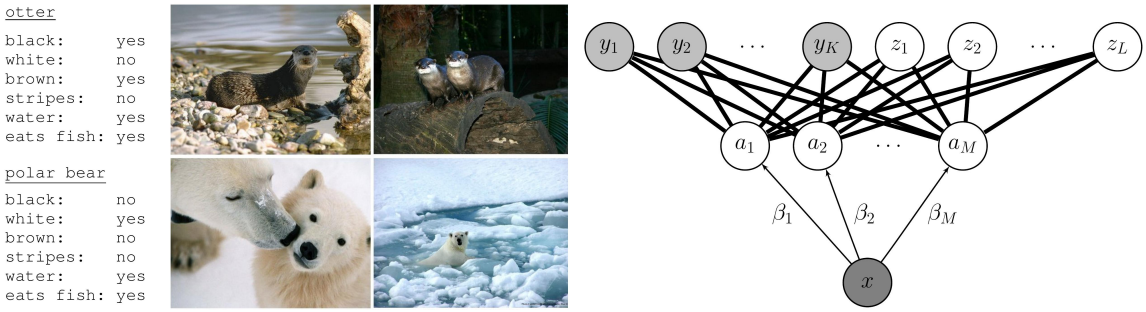


Figure 2-19 – Illustration of transfer by attributes in [Lampert et al., 2009]: the attributes classifiers (a_m) learnt on seen classes (y_k) enables to recognize unseen classes (z_l).

of intermediate semantics whose basic components are generic enough to be shared across different objects. It is interesting to note that recent works in visual relation detection have implicitly adopted this view, as they typically decompose a visual relation into three sub-parts: the subject, the object and the predicate, where each part can be shared across different relations. This decomposition is highly motivated by the triplet structure in language, that allows to express a complex entity in terms of its components. It is thus important to realize that the triplet structure of language orients our understanding of a relation as it explicitly provides the attributes on which the relation should be decomposed.

Exploiting relations between labels. Another way to assess the degree of similarity between two entities is to exploit the structure of real-world labels given by lexical databases such as WordNet [Miller, 1992]. These databases group words into sets of synonyms - called synsets - and define hierarchical relations between them. In WordNet, the similarity between two entities can thus be easily measured by counting the number of edges that separate the two representative synsets. Moreover, the hierarchy between two entities can be advantageously exploited. This has been done in large-scale object classification by Deng et al. [2014] who encode semantic relations between labels into a structured Hierarchy and Exclusion (HEX) graph.

This formulation has been later used by Ramanathan et al. [2015] to address ac-

(a) (b)

Figure 2-20 – Enforcing structure between labels by: (a) exploiting hierarchy between actions [Ramanathan et al., 2015], (b) using relations between different levels of abstraction [Vendrov et al., 2016]

tion retrieval in a large-scale set-up, where the lack of supervision is compensated by semantic relationships between actions. For instance, as shown in Figure 2-20(a), instead of treating each action (e.g. "person interact with panda") independently, the model of Ramanathan et al. [2015] identifies the related relations and treats them either as positives (e.g. "person feeding panda") or negatives (e.g. "person feeding calf"). Another interesting approach by Vendrov et al. [2016] encodes hierarchy between concepts by imposing ordering constraints on the visual-semantic embedding space. This process, illustrated in 2-20(b), is slightly different from Ramanathan et al. [2015] as it uses different levels of abstraction of the same concept (e.g. "woman walking her dog", "woman walking", "person walking", "person" are all valid descriptions for the entity). Finally, the recent work of Kato et al. [2018] uses external knowledge bases such as WordNet or NEIL [Chen et al., 2013] to connect objects (through hypernym-hyponym links) and actions (through affordances) in a graph to better transfer to unseen human-object interactions.

Using distributional semantics. The approaches described previously moves the problem of supervision from annotating visual data to annotating semantic data (ei-

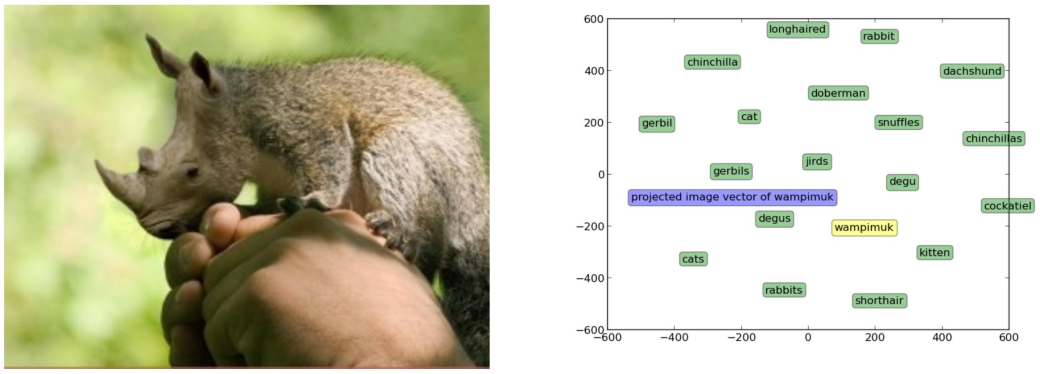


Figure 2-21 – Lazaridou et al. [2014] introduce this imaginary entity named a wampimuk, only known through the sentence *"We found a cute, hairy wampimuk sleeping behind the tree"*. The sentence is projected into the linguistic space and associated to its nearest neighbor, a degus, whose visual detector can be used to retrieve candidate images of wampimuk.

ther parts, attributes or hierarchical relations) which is tedious. Instead, *distributional semantics* [Mikolov et al., 2013] automatically leverages the co-occurrences of words in text corpora without requiring any annotation. The underlying idea is to represent a word according to its context. Words that appear in similar context have similar representation. Distributional language representations have opened up many possibilities in image-text mapping, in particular for the generalization to unseen entities. Notably, Socher et al. [2013b]; Frome et al. [2013] have proposed to learn a mapping of the visual representation of an object into the semantic embedding space of words learnt from large, unsupervised text corpora. The joint visual-semantic embedding space that results provides a simple and efficient framework to recognize unseen visual objects, by using the knowledge about their relations with other concepts in language. A nice illustration of this is given by Lazaridou et al. [2014] who introduce an unseen imaginary entity, a "wampimuk", that could be potentially retrieved by processing its textual description with distributional semantics. We detail this process in Figure 2-21. Beyond mapping objects and words into a joint space, joint mapping of more complex entities such as images and textual descriptions has been studied quite a lot as reviewed in Section 2.2. Still, performing zero-shot learning in such joint visual-semantic embedding spaces remains a difficult problem.

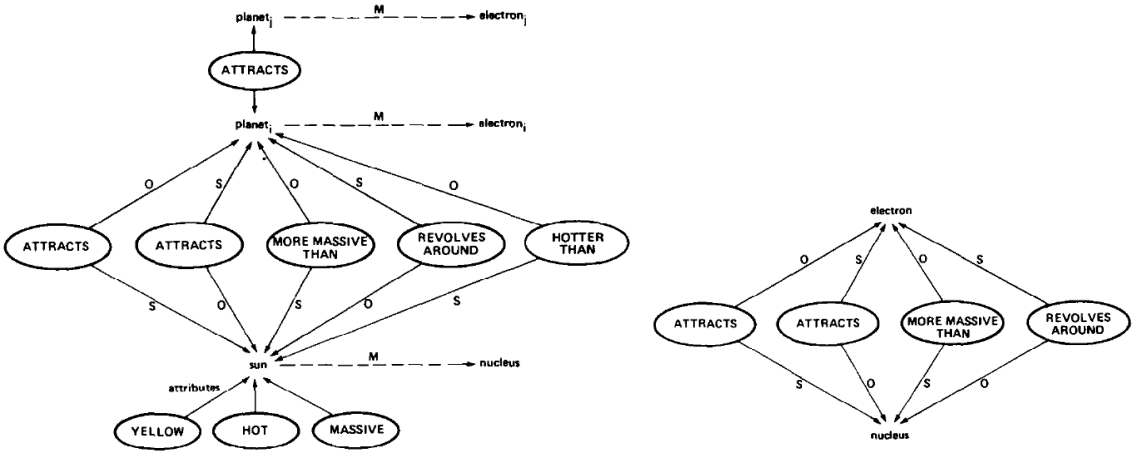


Figure 2-22 – Structure mapping for the Rutherford analogy: ‘‘*The atom is like the solar system*” [Gentner, 2003]. The source domain is the solar system (a). The target domain is the atom system (b). Entities that occupy the same role in a common system of relations (e.g. “sun” and “nucleus” exhibit similar relations to “planet” and “electron” respectively) can be mapped together, even if they have different attributes.

Transferring with analogy. In this manuscript, we propose to explore the idea of analogy to transfer knowledge from familiar, seen, visual relations to unseen ones. In the most generic form, an analogy is given by a statement such as: *“A is like B”*, where A is called the source, and B is called the target. Learning to map A to B is called analogy mapping and defined by Gentner [2003] as the process by which “a familiar situation, the base or source description, is matched with a less familiar situation, the target description”.

In the last thirty years, analogy has been much studied by psychologists [Gentner, 1983; Holyoak, 1985; Ross, 1989] who exhibited some of its core properties. One major theory introduced by Gentner [1983] is *structural mapping* which views an analogy as a similarity of relational structure between two domains. In other words, analogical reasoning is about mapping a system of relations from base objects to target objects, rather than mapping attributes or object properties. An example of analogy is given in Figure 2-22 that highlights the structural mapping between the source and target domains. Once a mapping between base and target has been found, analogical

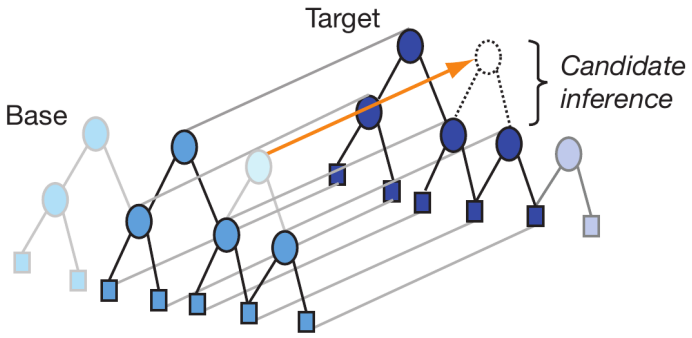


Figure 2-23 – Analogy completion illustrated by [Gentner and Smith, 2012]: first, initial alignment of common relational structure is made between source and target domain (grey lines), second a new candidate inference can be generated by completing the missing relational pattern in the target (orange arrow).

reasoning enables to make new inferences by transferring information that hold in the source domain to the target domain. This is sometimes referred to as analogy completion, because generating a candidate inference consists in completing a missing relational pattern in the target domain as illustrated in Figure 2-23. Analogical reasoning is influenced by a few key principles. Among them is *systematicity* which is our tendency to align domains that share a large number of relations. In Figure 2-23, it means that the more grey lines between source and target domains, the more likely we will align them. Another important factor of influence is *transparency*: we tend to align objects that are similar, i.e. that share some attributes and properties. Conversely, performing analogical reasoning where similar objects have different roles is difficult. We refer to [Gentner, 2003] for an exhaustive discussion about these factors of influence.

In computer vision, the concept of analogy has been used mainly in computer graphics, as a way to learn and apply image transformations. Hertzmann et al. [2001] formulate the problem of image analogies as follows: given a pair of images A and A' (the unfiltered and filtered source images, respectively), along with some additional unfiltered target image B, the goal is to synthesize a new filtered target image B'

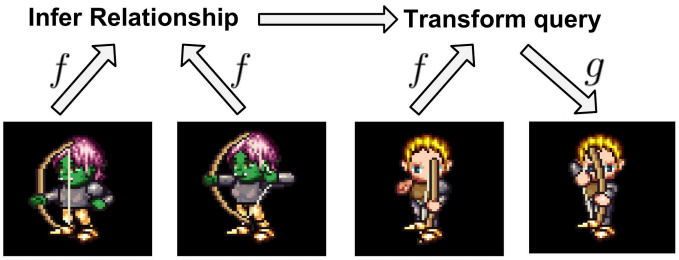


Figure 2-24 – Reed et al. [2015] transfer animations by learning an encoder function f that maps the images into a space where the analogy (e.g. transformation by rotation) can be performed, and a decoder function g that maps back the transformed embedding into the image space.

such that

$$A : A' :: B : B' \tag{2.2}$$

meaning that B' should relate to B in a similar manner than A' relates to A. This formulation, which supports a large number of applications such as texture transfer, super resolution or artistic filters, has been extended in various ways [Cheng et al., 2008; Bénard et al., 2013; Barnes et al., 2015; Reed et al., 2015; Liao et al., 2017a]. In Figure 2-24, we illustrate the analogical reasoning model of Reed et al. [2015] that transfers animations on 2D characters. The idea of analogy also outreached outside computer graphics to applications like object categorization [Hwang et al., 2013] and visual analogy question [Sadeghi et al., 2015b]. Both works present two crucial steps to perform analogical reasoning: (1) building two sets of quadruplets (A, A', B, B') suited for analogy: a valid set where the analogy holds and a negative set where it does not, (2) encoding analogies as vector transformations into an embedding space.

Our work is closely related to theses formulations. In Chapter 4, we explain how we use analogies to detect unseen visual relations. For (1) we exploit the triplet structure into three parts $(subject, predicate, object)$ to easily compute similarities between triplets in terms of visual and semantic cues, enabling to generate a set of valid quadruplets. For (2), our work is especially related to [Reed et al., 2015] who propose an objective function that forces analogy completion by arithmetic operations

on vectors in the embedding space. We adapt this framework to our problem of visual relation detection in a multimodal setting.

Chapter 3

Weakly-supervised learning of visual relations

In this chapter, we introduce a novel approach for modeling visual relations between pairs of objects. We explicitly address the difficulty to get annotations, especially at box-level, for all possible triplets, which makes both learning and evaluation difficult. The contributions are threefold. First, we design strong yet flexible visual features that encode the appearance and spatial configuration for pairs of objects. Second, we propose a weakly-supervised discriminative clustering model to learn relations from image-level labels only. Third we introduce a new challenging dataset of unusual relations (UnRel) together with an exhaustive annotation, that enables accurate evaluation of visual relation retrieval. We show experimentally that our model results in state-of-the-art results on the visual relationship dataset [Lu et al., 2016a] significantly improving performance on previously unseen relations (zero-shot learning), and confirm this observation on our newly introduced UnRel dataset.

3.1 Introduction

While a great progress has been made on the detection and localization of individual objects [Ren et al., 2015b; Zagoruyko et al., 2016], it is now time to move one step forward towards understanding complete scenes. For example, if we want to localize "a person sitting on a chair under an umbrella", we not only need to detect the objects involved: "person", "chair", "umbrella", but also need to find the correspondence of the semantic relations "sitting on" and "under" with the correct pairs of objects in the image. Thus, an important challenge is automatic understanding of how entities in an image interact with each other.

This task presents two main challenges. First, the appearance of objects can change significantly due to interactions with other objects (person cycling, person driving). This visual complexity can be tackled by learning "visual phrases" [Sadeghi and Farhadi, 2011] capturing the pair of objects in a relation as one entity, as opposed to first detecting individual entities in an image and then modeling their relations. This approach, however, does not scale to the large number of relations as the number of such visual phrases grows combinatorially, requiring large amounts of training data. To address this challenge, we need a method that can share knowledge among similar relations. Intuitively, it seems possible to generalize frequent relations to unseen triplets like those depicted in Figure 3-1: for example having seen "person ride horse" at training could help recognizing "person ride dog" at test time.

The second main challenge comes from the difficulty to provide exhaustive annotations on the object level for relations that are by their nature non mutually-exclusive (i.e. "on the left of" is also "next to"). A complete labeling of R relations for all pairs of N objects in an image would indeed require $\mathcal{O}(N^2R)$ annotations *for each image*. Such difficulty makes both learning and evaluation very challenging. For learning, it would be desirable to learn relations from image-level annotations only. For evaluation, current large-scale datasets [Krishna et al., 2016; Lu et al., 2016a] make retrieval evaluation difficult due to large amount of missing annotations. This leads to correct

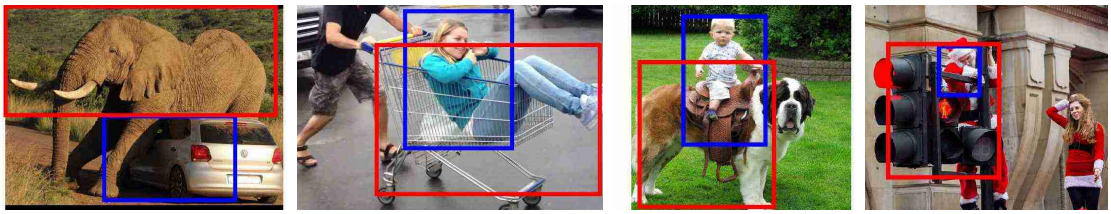


Figure 3-1 – Examples of top retrieved pairs of boxes in UnRel dataset for unusual queries (indicated below each image) with our weakly-supervised model described in 3.3.2.

predictions being penalized making (i) results less interpretable, (ii) numerical performances decrease potentially differently across models (e.g. if missing annotations are not uniformly spread across relations, some methods could be better at modeling the noise due to missing annotations, but worst at predicting visual relations).

Contributions. The contributions of this chapter are three-fold. First, to address the combinatorial challenge, we develop a method that can handle a large number of relations by sharing parameters among them. For example, we learn a single "on" classifier that can recognize both "person on bike" and "dog on bike", even when "dog on bike" has not been seen in training. The main innovation is a new model of an object relation that represents a pair of boxes by explicitly incorporating their spatial configuration as well as the appearance of individual objects. Our model relies on a multimodal representation of object configurations for each relation to handle the variability of relations. Second, to address the challenge of missing training data, we develop a model that, given pre-trained object detectors, is able to learn classifiers for object relations from image-level supervision only. It is, thus, sufficient to provide an image-level annotation, such as "person on bike", without annotating the objects involved in the relation. Finally, to address the issue of missing annotations in test data, we introduce a new dataset of unusual relations (UnRel), with exhaustive annotation for a set of unusual triplet queries, that enables to evaluate retrieval on rare triplets and validate the generalization capabilities the learned model.

3.2 Related Work

Alignment of images with language. Learning correspondences between fragments of sentences and image regions has been addressed by the visual-semantic alignment which has been used for applications in image retrieval and caption generation [Chang et al., 2015; Karpathy and Fei-Fei, 2015; Karpathy et al., 2014]. With the appearance of new datasets providing box-level natural language annotations [Kazemzadeh et al., 2014; Krishna et al., 2016; Mao et al., 2016; Plummer et al., 2015], recent works have also investigated caption generation at the level of image regions for the tasks of natural language object retrieval [Hu et al., 2016; Mao et al., 2016; Rohrbach et al., 2016] or dense captioning [Johnson et al., 2016]. Our approach is similar in the sense that we aim at aligning a language triplet with a pair of boxes in the image. Typically, existing approaches do not explicitly represent relations between noun phrases in a sentence to improve visual-semantic alignment. We believe that understanding these relations is the next step towards image understanding with potential applications in tasks such as Visual Question Answering [Andreas et al., 2016a].

Learning triplets. Triplet learning has been addressed in various tasks such as mining typical relations (knowledge extraction) [Chen et al., 2013; Sadeghi et al., 2015a; Yatskar et al., 2016; Zhu et al., 2014], reasoning [Jenatton et al., 2012; Movshovitz-Attias and Cohen, 2015; Socher et al., 2013a], object detection [Gupta and Davis, 2008; Sadeghi and Farhadi, 2011], image retrieval [Johnson et al., 2015] or fact retrieval [Elhoseiny et al., 2016]. In this chapter, we address the task of relationship detection in images. This task was studied for the special case of human-object interactions [Delaitre et al., 2011; Desai et al., 2010; Gupta et al., 2009; Prest et al., 2011; Ramanathan et al., 2015; Yao and Fei-Fei, 2010a,b; Yao et al., 2011], where the triplet is in the form $(person, action, object)$. Contrary to these approaches, we do not restrict the *subject* to be a person and we cover a broader class of predicates

that includes prepositions and comparatives. Moreover, most of the previous work in human-object interaction was tested on small datasets only and does not explicitly address the combinatorial challenge in modeling relations [Sadeghi and Farhadi, 2011]. Recently, [Lu et al., 2016a] tried to generalize this setup to non-human subjects and scale to a larger vocabulary of objects and relations by developing a language model sharing knowledge among relations for visual relation detection. In our work we address this combinatorial challenge by developing a new visual representation that generalizes better to unseen triplets without the need for a strong language model. This visual representation shares the spirit of [Galleguillos et al., 2008; Johnson et al., 2015; Li et al., 2012] and we show it can handle multimodal relations and generalizes well to unseen triplets. Our model also handles a weakly-supervised set-up when only image-level annotations for object relations are available. It can thus learn from complex scenes with many objects participating in different relations, whereas previous work either uses full supervision or typically assumes only one object relation per image, for example, in images returned by a web search engine. Finally, we also address the problem to evaluate accurately due to missing annotations also pointed out in [Elhoseiny et al., 2016; Lu et al., 2016a]. We introduce a new dataset of unusual relations exhaustively labeled for a set of triplet queries, the UnRel dataset. This dataset enables the evaluation of relation retrieval and localization. Our dataset is related to the *"Out of context"* dataset of [Choi et al., 2012] which also involves objects in unusual configurations, as shown in Figure 3-2. However, this dataset is not annotated with relation labels.

Weak supervision. Most of the work on weakly-supervised learning for visual recognition has focused on learning objects [Bilen and Vedaldi, 2016; Fang et al., 2015; Oquab et al., 2015]. Here, we want to tackle the task of weakly-supervised detection of relations. This task is more complex as we need to detect the individual objects that satisfy the specific relation. We assume that pre-trained detectors for

Figure 3-2 – Images from Out-of-Context dataset [Choi et al., 2012]

individual objects are available and learn relations among objects with image-level labels. Our work uses a discriminative clustering objective [Bach and Harchaoui, 2007], that has been successful in several computer vision tasks [Bojanowski et al., 2014; Joulin et al., 2014], but has not been so far, to the best of our knowledge, used for modeling relations. We refer to Chapter 2.3.4 for a discussion on the different levels of supervision used in visual relation detection.

Zero-shot learning. Zero-shot learning has been mostly explored for object classification [Frome et al., 2013; Lazaridou et al., 2014; Socher et al., 2013b; Xian et al., 2016] and recently for the task of describing images with novel objects [Hendricks et al., 2016b; Venugopalan et al., 2017]. We refer to Chapter 2.3.5 for a more complete review. In our work, we address zero-shot learning of relations in the form of triplets $(subject, predicate, object)$, where each term has already been seen independently during training, but not in that specific combination. We develop a model to detect and localize such zero-shot relations.

3.3 Representing and learning visual relations

We want to represent triplets $t = (s, r, o)$ where s is the subject, o the object and r is the predicate. s and o are nouns and can be objects like "person", "horse", "car" or regions such as "sky", "street", "mountain". The predicate r is a term that links the subject and the object in a sentence and can be a preposition ("in front of", "under"), a verb ("ride", "hold") or a comparative adjective ("taller than"). To detect and localize such triplets in test images, we assume that the candidate object detections for s and o are given by a detector trained with full supervision. Here we use the object detector [Girshick, 2015] trained on the Visual Relationship Detection training set [Lu et al., 2016a]. In 3.3.1, we will explain our representation of a triplet $t = (s, r, o)$ and show in 3.3.2 how we can learn to detect triplets in images given weak image-level supervision for relations.

3.3.1 Visual representation of relations

A triplet $t = (s, r, o)$ such as "person next to surfboard" in Figure 3-3 visually corresponds to a pair of objects (s, o) in a certain configuration. We represent such pairs by the spatial configuration between object bounding boxes $(\boldsymbol{o_s}, \boldsymbol{o_o})$ and the individual appearance of each object.

Representing spatial configurations of objects. Given two boxes $\boldsymbol{o_s} = [x_s, y_s, w_s, h_s]$, $\boldsymbol{o_o} = [x_o, y_o, w_o, h_o]$, where (x, y) are the coordinates of the center of the box, and (w, h) are the width and height of the box, we encode the spatial configuration with a 6-dimensional vector:

$$\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o}) = [\underbrace{\frac{x_o - x_s}{\sqrt{w_s h_s}}}_{r_1}, \underbrace{\frac{y_o - y_s}{\sqrt{w_s h_s}}}_{r_2}, \underbrace{\sqrt{\frac{w_o h_o}{w_s h_s}}}_{r_3}, \underbrace{\frac{\boldsymbol{o_s} \cap \boldsymbol{o_o}}{\boldsymbol{o_s} \cup \boldsymbol{o_o}}}_{r_4}, \underbrace{\frac{w_s}{h_s}}_{r_5}, \underbrace{\frac{w_o}{h_o}}_{r_6}] \tag{3.1}$$

where r_1 and r_2 represent the renormalized translation between the two boxes, r_3 is the

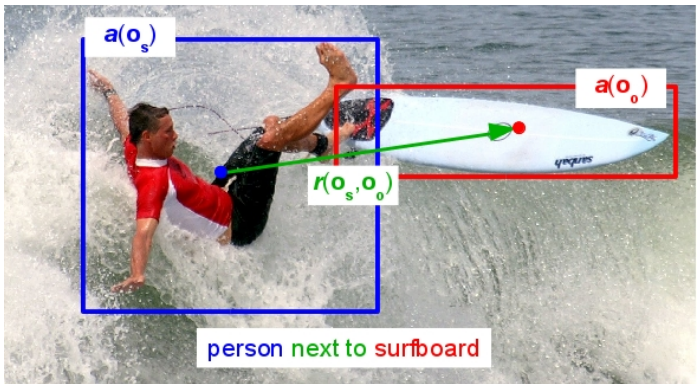


Figure 3-3 – Our visual representation is the composition of appearance features for each object $[\boldsymbol{a}(\boldsymbol{o_s}), \boldsymbol{a}(\boldsymbol{o_o})]$ and their spatial configuration $\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})$ represented by the green arrow.

ratio of box sizes, r_4 is the overlap between boxes, and r_5, r_6 encode the aspect ratio of each box respectively. Directly adopting this feature as our representation might not be well suited for some spatial relations like "next to" which are multimodal. Indeed, "s next to o" can either correspond to the spatial configuration "s left of o" or "s right of o". Instead, we propose to discretize the feature vector in Eq. (3.1) into k bins. For this, we assume that the spatial configurations $\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})$ are generated by a mixture of k Gaussians and we fit the parameters of the Gaussian Mixture Model to the training pairs of boxes. We take the scores representing the probability of assignment to each of the k clusters as our spatial representation. In our experiments, we use $k = 400$, thus the spatial representation is a 400-dimensional vector. In Figure 3-4, we show examples of pairs of boxes for the most populated components of the trained GMM. We can observe that our spatial representation can capture subtle differences between configurations of boxes, see row 1 and row 2 of Figure 3-4, where "person on board" and "person carry board" are in different clusters.

Representing appearance of objects. Our appearance features are given by the fc7 output of a Fast-RCNN [Girshick, 2015] trained to detect individual objects. In our experiments, we use Fast-RCNN with VGG16 pre-trained on ImageNet. As the

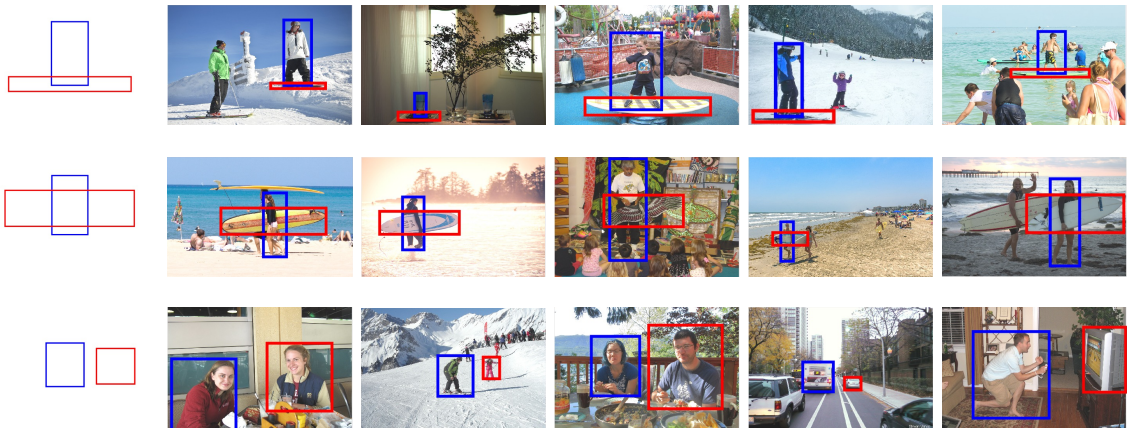


Figure 3-4 – Examples for different GMM components of our spatial configuration model (one per row). In the first column we show the spatial configuration corresponding to the mean of the pairs of boxes per component. Note that our representation can capture subtle differences between spatial configurations, see e.g., row 1 and 2.

extracted features have high dimensionality, we perform PCA on the L2-normalized features to reduce the number of dimensions from 4096 to 300. We concatenate the appearance features of the subject and object and apply L2-normalization again.

Our final visual feature is a concatenation of the spatial configuration $\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})$ and the appearance of objects $[\boldsymbol{a}(\boldsymbol{o_s}), \boldsymbol{a}(\boldsymbol{o_o})]$. In our experiments, it has a dimensionality of $d = 1000$. In the fully supervised setup, where each relation annotation is associated with a pair of object boxes in the image, we use ridge regression to train a multi-way relation classifier to assign a relation to a given visual feature. Training is performed jointly on all relation examples of the training set.

In the next section, we describe how we learn relation classifiers with only weak, image-level, annotations.

3.3.2 Weakly-supervised learning of relations

Equipped with pre-trained detectors for individual objects, our goal here is to learn to detect and localize relations between objects, given image-level supervision only. For example, for a relation "person falling off horse" we are given (multiple) object

detections for "person" and "horse", but do not know which objects participate in the relation, as illustrated in Figure 3-5. Our model is based on a weakly-supervised discriminative clustering objective [Bach and Harchaoui, 2007], where we introduce latent variables to model which pairs of objects participate in the relation. We train a classifier for each predicate r and incorporate weak annotations in the form of constraints on latent variables. Note that the relation classifiers are shared across object categories (eg. the relations "person on horse" and "cat on table" share the same classifier "on") and can thus be used to predict unseen triplets.

Discriminative clustering of relations. Our goal is to learn a set of classifiers $W = [\mathbf{w_1}, ..., \mathbf{w_R}] \in \mathbb{R}^{d \times R}$ where each classifier $\mathbf{w_r}$ predicts the likelihood of a pair of objects (s, o) to belong to the r^{th} predicate in a vocabulary of R predicates. If strong supervision was provided for each pair of objects, we could learn W by solving a ridge regression:

$$\min_{W \in \mathbb{R}^{d \times R}} \frac{1}{N} \|Z - XW\|_F^2 + \lambda \|W\|_F^2 \tag{3.2}$$

where $Z \in \{0, 1\}^{N \times R}$ are the ground truth labels for each of the N pairs of objects across all training images, and $X = [\mathbf{x_1}, ..., \mathbf{x_N}]^T$ is a $N \times d$ matrix where each row $\mathbf{x_k}$ is the visual feature corresponding to the k^{th} pair of objects. However, in a weakly-supervised setup the entire matrix Z is unknown. Building on DIFFRAC [Bach and Harchaoui, 2007], our approach is to optimize the cost:

$$\min_{Z \in \mathcal{Z}} \min_{W \in \mathbb{R}^{d \times R}} \frac{1}{N} \|Z - XW\|_F^2 + \lambda \|W\|_F^2 \tag{3.3}$$

which treats Z as a latent assignment matrix to be learnt together with the classifiers $W \in \mathbb{R}^{d \times R}$. Minimizing the first term encourages the predictions made by W to match the latent assignments Z, while the second term is a L2-regularization on the classifiers W. This framework enables to incorporate weak annotations by constraining the

space of valid assignment matrices $Z \in \mathcal{Z}$. The valid matrices $Z \in \{0,1\}^{N \times R}$ satisfy the multiclass constraint $Z\mathbf{1}_R = \mathbf{1}_N$ which assigns a pair of objects to one and only one predicate r. We describe in the next paragraph how to incorporate the weak annotations as constraints.

Weak annotations as constraints. For an image, we are given weak annotations in the form of triplets $t = (s, r, o) \in \mathcal{T}$. Having such triplet (s, r, o) indicates that at least one of the pairs of objects with object categories (s, o) is in relation r. Let us call $\mathcal{N}_t$ the subset of pairs of objects in the image that correspond to object categories (s, o). The rows of Z that are in subset $\mathcal{N}_t$ should satisfy the constraint:

$$\sum_{n \in \mathcal{N}_t} Z_{nr} \geq 1 \tag{3.4}$$

This constraint ensures that at least one of the pairs of objects in the subset $\mathcal{N}_t$ is assigned to predicate r. For instance, in case of the image in Figure 3-5 that contains 12 candidate pairs of objects that potentially match the triplet $t = (person, falling\ off, horse)$, the constraint in Eq. (3.4) imposes that at least one of them is in relation $falling\ off$.

Triplet score. At test time, we can compute a score for a pair of boxes $(\boldsymbol{o}_s, \boldsymbol{o}_o)$ relative to a triplet $t = (s, r, o)$ as

$$\begin{aligned} S((\boldsymbol{o}_s, \boldsymbol{o}_o) \mid t) = v_{rel}((\boldsymbol{o}_s, \boldsymbol{o}_o) \mid r) + \alpha_{sub} v_{sub}(\boldsymbol{o}_s \mid s) \\ + \alpha_{obj} v_{obj}(\boldsymbol{o}_o \mid o) + \alpha_{lang} \ell((s, o) \mid r), \end{aligned} \tag{3.5}$$

where $v_{rel}((\boldsymbol{o}_s, \boldsymbol{o}_o)|r) = \mathbf{x}_{(\boldsymbol{o}_s, \boldsymbol{o}_o)} \mathbf{w_r}$ is the score returned by the classifier $\mathbf{w}_r$ for predicate r (learnt by optimizing Eq. (3.3)) for the visual representation $\mathbf{x}_{(\boldsymbol{o}_s, \boldsymbol{o}_o)}$ of the pair of boxes. $v_{sub}(\boldsymbol{o}_s|s)$ and $v_{obj}(\boldsymbol{o}_o|o)$ are the object class likelihoods returned by the object detector. $\ell((s, o)|r)$ is a score of a language model that we can optionally combine with our visual model.

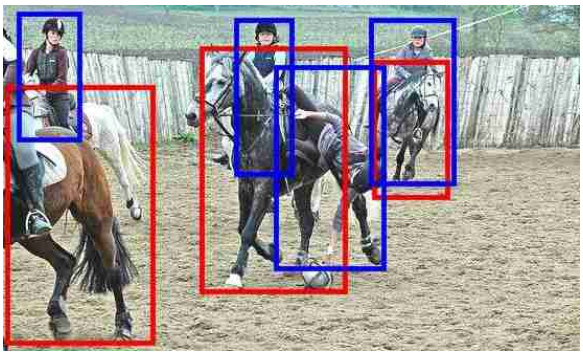


Figure 3-5 – Image from the COCO dataset [Lin et al., 2014b] associated with caption: *"A person falling off the side of a horse as it rides"*. The boxes correspond to the possible candidates for object category *person* (blue) and *horse* (red). Our model has to disambiguate the right pair for the relation "falling off" among 12 candidate pairs.

Optimization. We optimize the cost in Eq. (3.3) on pairs of objects in the training set using a variant of the Frank-Wolfe algorithm [Miech et al.; Osokin et al., 2016]. The hyperparameters $(\alpha_{sub}, \alpha_{obj}, \alpha_{lang})$ are optimized on an held-out fully-annotated validation set which has no overlap with our training and test sets. In our experiments we use the validation split of [Johnson et al., 2016] of the Visual Genome dataset [Krishna et al., 2016]. Unless otherwise specified, the candidate pairs, both at training and test time, are the outputs of the object detector [Girshick, 2015] that we fine-tuned on the Visual Relationship Detection dataset [Lu et al., 2016a]. For each image, we keep the object candidates whose confidence scores is above 0.3 among the top 100 detections. Non-maximum suppression with threshold 0.3 is applied to handle multiple detections. This results in an average of 18 object detections per image, i.e. around 300 pairs of boxes.

3.4 Experiments

In this section, we evaluate the performance of our model on two datasets for different evaluation setups. First, we evaluate our new visual representation for relations on the Visual Relationship Detection dataset [Lu et al., 2016a]. We show results with

our weakly-supervised model learned from image-level supervision and present large improvements over the state of the art for detecting unseen triplets (zero-shot detection). Second, we evaluate our model for the task of unusual triplets retrieval and localization on our new UnRel dataset.

3.4.1 Recall on Visual Relationship Detection dataset

Dataset. We evaluate our method on the Visual Relationship Detection dataset [Lu et al., 2016a] following the original experimental setup. This dataset contains 4000 training and 1000 test images with ground truth annotations for relations between pairs of objects. Due to the specific train/test split provided by [Lu et al., 2016a], 10% of test triplets are not seen at training and allow for evaluation of zero-shot learning. Some of these triplets are rare in the linguistic and visual world (e.g. "laptop on stove"), but most of them are only infrequent in the training set or have not been annotated (e.g. "van on the left of car"). Around 30K triplets are annotated in the training set, with an average of 7.5 relations per image. The dataset contains 100 objects and 70 predicates, i.e. $100 \times 100 \times 70$ possible triplets. However there are only 6672 different annotated triplets.

Evaluation set-up. Following [Lu et al., 2016a], we compute `recall@x` which corresponds to the proportion of ground truth pairs among the `x` top scored candidate pairs in each image. We use the scores returned by Eq. (3.5) to sort the candidate pairs of boxes. The evaluation is reported for three setups. In **predicate detection**, candidate pairs of boxes are ground truth boxes, and the evaluation only focuses on the quality of the predicate classifier. In the other two more realistic setups, the subject and object confidence scores are provided by an object detector and we also check whether the candidate boxes intersect the ground truth boxes: either both subject and object boxes should match (**relationship detection**), or the union of them should match (**phrase detection**). For a fair comparison with [Lu et al., 2016a], we

report results using the same set of R-CNN [Girshick et al., 2014] object detections as them. The localization is evaluated with IoU = 0.5.

Benefits of our visual representation. We first evaluate the quality of our visual representation in a fully supervised setup where the ground truth spatial localization for each relation is known, i.e. we know which objects in the image are involved in each relation at training time. For this, we solve the multi-label ridge regression in Eq. (3.2). Training with one-vs-rest SVMs gives similar results. We compare three types of features described in Section 3.3.1 in Table 3.1: [S] the spatial representation (f.), [A] the appearance representation (g.) and [S+A] the concatenation of the two (h.). We compare with the Visual Phrases model [Sadeghi and Farhadi, 2011] and several variants of [Lu et al., 2016a] [1] : Visual model alone (b.), Language (likelihood of a relationship) (c.), combined Visual+Language model (d.). In row (e.) we also report the performance of the full language model of [Lu et al., 2016a], that scores the candidate pairs of boxes based on their predicted object categories, that we computed using the model and word embeddings provided by the authors. Because their language model is orthogonal to our visual model, we can combine them together (i.). The results are presented on the complete test set (column All) and on the zero-shot learning split (column Unseen). Table 3.1 shows that our combined visual features [S+A] improve over the visual features of [Lu et al., 2016a] by 40% on the task of predicate detection and more than 10% on the hardest task of relationship detection. Furthermore, our purely visual features without any use of language (h.) reach comparable performance to the combined Visual+Language features of [Lu et al., 2016a] and reach state-of-the-art performance (i.) when combined with the language scores of [Lu et al., 2016a]. The good performance of our spatial features [S] alone (f.) confirms the observation we made in Figure 3-4 that our spatial clusters group pairs of objects in similar relations. That could partly explain why the visual model of [Lu

[1] When running the evaluation code of [Lu et al., 2016a], we found slighlty better performance than what is reported in their paper. See Annex A for more details.

	Predicate Det.		Phrase Det.		Relationship Det.	
	All	Unseen	All	Unseen	All	Unseen
Full sup.						
a. Visual Phrases [Sadeghi and Farhadi, 2011]	0.9	-	0.04	-	-	-
b. Visual [Lu et al., 2016a]	7.1	3.5	2.2	1.0	1.6	0.7
c. Language (likelihood) [Lu et al., 2016a]	18.2	5.1	0.08	0.00	0.08	0.00
d. Visual + Language [Lu et al., 2016a]	47.9	8.5	16.2	3.4	13.9	3.1
e. Language (full) [Lu et al., 2016a]	48.4	12.9	15.8	4.6	13.9	4.3
f. Ours [S]	42.2	22.2	13.8	7.4	12.4	7.0
g. Ours [A]	46.3	16.1	14.9	5.6	12.9	5.0
h. Ours [S+A]	50.4	**23.6**	16.7	**7.4**	14.9	**7.1**
i. Ours [S+A] + Language [Lu et al., 2016a]	**52.6**	21.6	**17.9**	6.8	**15.8**	6.4
Weak sup.						
j. Ours [S+A]	46.8	19.0	16.0	6.9	14.1	6.7
k. Ours [S+A] - Noisy	46.4	17.6	15.1	6.0	13.4	5.6

Table 3.1 – Results on Visual Relationship Detection dataset [Lu et al., 2016a] for R@50. See Table A.2 for results with R@100.

et al., 2016a] has low performance. Their model learns a classifier only based on the appearance of the union of the two object boxes and lacks information about their spatial configuration.

Weak supervision. We evaluate our weakly-supervised classifiers W learned on image-level labels as described in Section 3.3.2. We use the ground truth annotations of the Visual Relationship Detection dataset as image-level labels. We report the results using our combined spatial and appearance features (j.) in Table 3.1. We see that when switching to weak supervision the recall@50 only drops from 50.4% to 46.8% for predicate detection and has limited influence on the other tasks. This is an interesting result as it suggests that, given pre-trained object detectors, weak image-level annotations are enough to learn good classifiers for relations. To investigate this further we have also tried to learn relation classifiers directly from noisy image-level labels without inferring at training time which objects participate in which relation. For each triplet $t = (s, r, o)$ in an image containing candidate pairs of boxes $(\boldsymbol{o_s}, \boldsymbol{o_o})$ we randomly select one of the pairs as being in relation r and discard the other

object pairs. This is equivalent to training in a fully-supervised setup but with noisy labels. The performance obtained by this classifier (k.) is below our weakly-supervised learning set-up but is surprisingly high. We believe that this is related to a particular bias present in the Visual Relationship Detection dataset [Lu et al., 2016a], which contains many images with only two prominent objects involved in a specific relation (more than half of the triplets fall into this category). To underline the ability of the weakly-supervised model to disambiguate the correct bounding boxes, we evaluate in a more difficult setup where we replace the candidate test pairs of [Lu et al., 2016a] by all candidate pairs formed by objects of confidence scores above 0.3. This multiplies by 5 the number of candidate pairs, resulting in an increased level of ambiguity. In this more challenging setup, our approach obtains a recall@50 for Phrase Detection (resp. Relationship Detection) of 17.9% (resp. 12.0%) compared to the "Ours [S+A] Noisy" baseline which drops to 15.3% (resp. 10.1%).

Unseen triplets. Following [Lu et al., 2016a] we report results on the "zero-shot split" of the test set containing only the test triplets not seen in training. Results for both of our fully-supervised and weakly-supervised methods are shown in Table 3.1 (column Unseen). Interestingly, our fully supervised model almost triples the performance on the unseen triplets compared to the Visual+Language model of [Lu et al., 2016a]. Even using weak supervision, our recall of 19.0% is significantly better than their fully supervised method. We believe that this improvement is due to the strength of our visual features that generalize well to unseen triplets.

Figure 3-6 shows examples of predictions of both seen and unseen triplets (last row) by our model [S+A] trained with weak-supervision. We note that many of the misclassified relations are in fact due to missing annotations in the dataset (yellow column). First, not all pairs of objects in the image are labeled; second, the pairs that are labeled are not labelled exhaustively, i.e. "person riding horse" can be labelled as "person on horse" and predicting "riding" for this pair of objects is considered

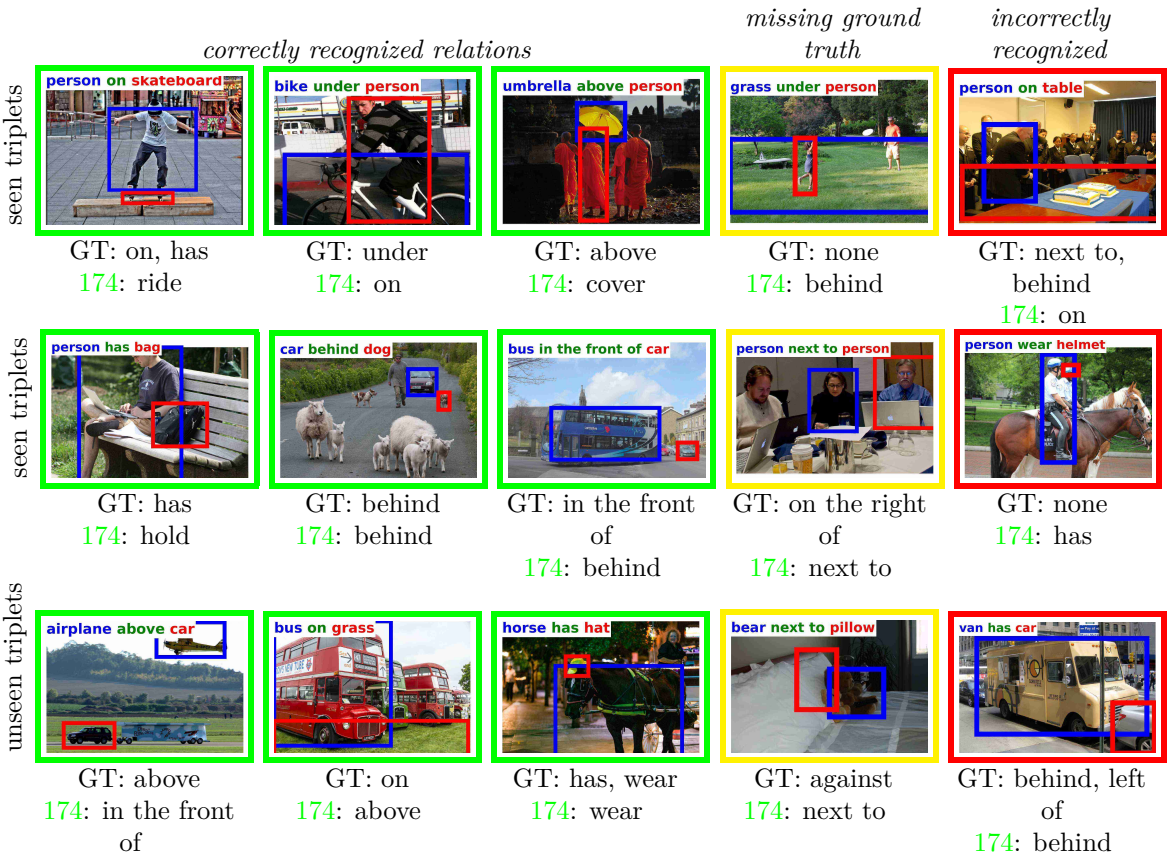


Figure 3-6 – Relationship detections on the test set of [Lu et al., 2016a]. We show examples among the top scored triplets detected for each relation by our weakly-supervised model described in 3.3.2. The triplet is correctly recognized if both the object detections and the relation match ground truth (in green), else the triplet is incorrect (in red). We also show examples of correctly predicted relations where the ground truth is erroneous: either missing or incomplete (in yellow). The last row shows zero-shot triplets that are not in the training set. See Figure 3-10 and Figure 3-11 for additional qualitative results.

as an error. Not having exhaustive annotation per object pair is therefore an issue as predicates are not necessary mutually exclusive. We tackle this problem in the next section by introducing a new exhaustively labeled dataset that enables retrieval evaluation. Our real errors (red column) are mostly due to two reasons: either the spatial configuration is challenging (e.g.“person on table”), or the spatial configuration is roughly correct but the output predicate is incorrect (e.g. “van has car” has similar configuration to ”person has bag”).

3.4.2 Retrieval of rare relations on UnRel Dataset

Dataset. To address the problem of missing annotations, we introduce a new challenging dataset of unusual relations, UnRel, that contains images collected from the web with unusual language triplet queries such as “person ride giraffe”. We exhaustively annotate these images at box-level for the given triplet queries. UnRel dataset has three main advantages. First, it is now possible to evaluate retrieval and localization of triplet queries in a clean setup without problems posed by missing annotations. Second, as the triplet queries of UnRel are rare (and thus likely not seen at training), it enables evaluating the generalization performance of the algorithm. Third, other datasets can be easily added to act as confusers to further increase the difficulty of the retrieval set-up. Currently, UnRel dataset contains more than 1000 images queried with 76 triplet queries. As it is small scale, this dataset is mainly suited for testing.

Setup. We use our UnRel dataset as a set of positive pairs to be retrieved among all the test pairs of the Visual Relationship Dataset. We evaluate retrieval and localization with mean average precision (mAP) over triplet queries $t = (s, r, o)$ of UnRel in two different setups. In the first setup (with GT) we rank the manually provided ground truth pairs of boxes $(\boldsymbol{o}_s, \boldsymbol{o}_o)$ according to their predicate scores $v_{rel}((\boldsymbol{o}_s, \boldsymbol{o}_o) \mid r)$ to evaluate relation prediction without the difficulty of object detection. In the second setup (with candidates) we rank candidate pairs of boxes $(\boldsymbol{o}_s, \boldsymbol{o}_o)$ provided by the ob-

	With GT	With candidates		
	-	union	subj	subj/obj
Chance	38.4	8.6	6.6	4.2
Full sup.				
DenseCap [Johnson et al., 2016]	-	6.2	6.8	-
Reproduce [Lu et al., 2016a]	50.6	12.0	10.0	7.2
Ours [S+A]	**62.6**	**14.1**	**12.1**	**9.9**
Weak sup.				
Ours [S+A]	58.5	13.4	11.0	8.7
Ours [S+A] - Noisy	55.0	13.0	10.6	8.5

Table 3.2 – Retrieval on UnRel (mAP) with IoU=0.3

ject detector according to predicate scores $v_{rel}((\boldsymbol{o}_s, \boldsymbol{o}_o) \mid r)$. For this second setup we also evaluate the accuracy of localization: a candidate pair of boxes is positive if its IoU with one ground truth pair is above 0.3. We compute different localization metrics: *mAP-subj* computes the overlap of the predicted subject box with the ground truth subject box, *mAP-union* computes the overlap of the predicted union of subject and object box with the union of ground truth boxes and *mAP-subj/obj* computes the overlap of both the subject and object boxes with their respective ground truth. Like in the previous section, we form candidate pairs of boxes by taking the top-scored object detections given by [Girshick, 2015]. We keep at most 100 candidate objects per image, and retain at most 500 candidate pairs per image. For this retrieval task where it is important to discriminate the positive from negative pairs, we found it is important to learn an additional "no relation" class by adding an extra column to W in Eq. (3.3). The negative pairs are sampled at random among the candidates that do not match the image-level annotations.

Results. Retrieval results are shown in Table 3.2. Our classifiers are trained on the training subset of the Visual Relationship Dataset. We compare with two strong baselines. The first baseline is our implementation of [Lu et al., 2016a] (their trained models are not available online). For this, we trained a classifier [Ren et al., 2015b] to

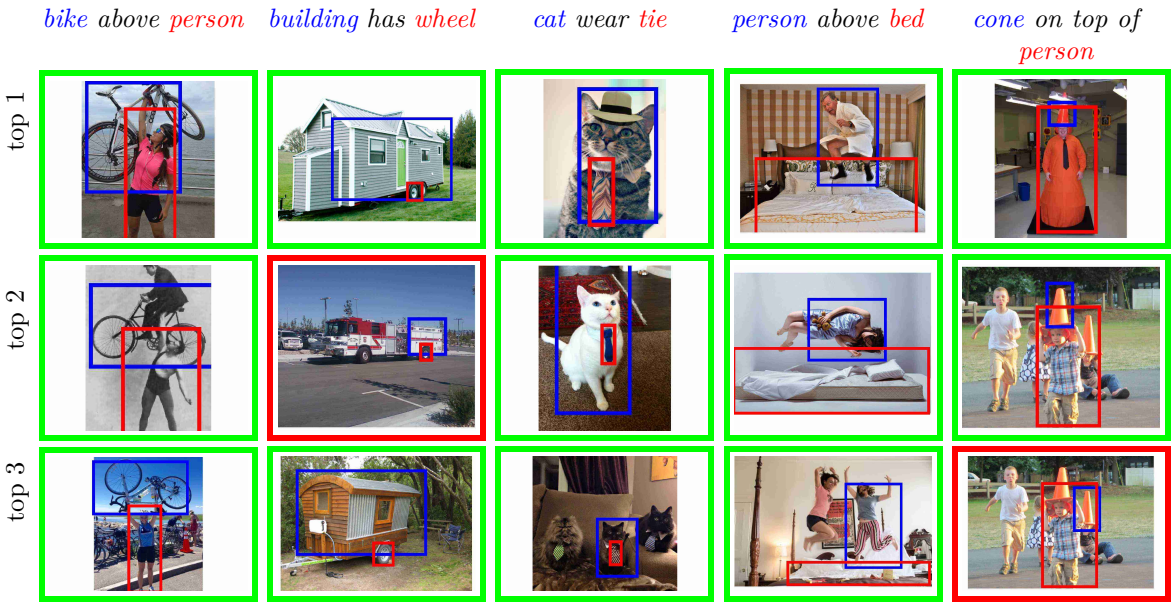


Figure 3-7 – Top 3 retrieved pairs of boxes for a set of UnRel triplet queries (first line is best) with our weakly-supervised model. The pair is marked as positive (green) if the candidate subject and object boxes coincide with a ground truth subject and object boxes with $IoU \geq 0.3$. We provide more qualitative results in Figure 3-9.

output predicates given visual features extracted from the union of subject and object bounding boxes. We do not use the language model as its score does not affect the retrieval results (only adding a constant offset to all retrieved images). We verified our implementation on the Visual Relationship Dataset where results of [Lu et al., 2016a] are available. As the second baseline, we use the DenseCap [Johnson et al., 2016] model to generate region candidates for each image and sort them according to the score of the given triplet query. Note that this is a strong baseline as we use the pre-trained model released by the authors which has been trained on 77K images of [Krishna et al., 2016] in a fully supervised manner using localized language descriptions, compared to our model trained on only 4K training images of [Lu et al., 2016a]. DenseCap outputs only a single bounding box (not a pair of boxes) but we interpret its output as either a subject box or a union of boxes. We cannot compare with the Visual Phrases [Sadeghi and Farhadi, 2011] approach as it requires training data for each triplet, which is not available for these rare queries. We report the

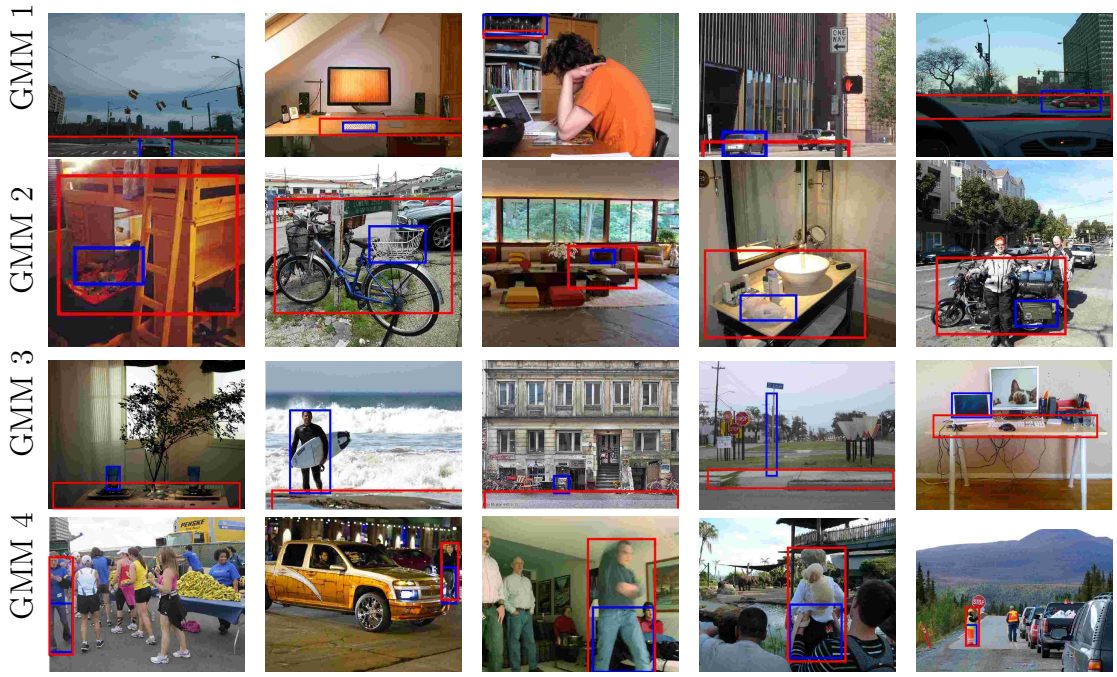


Figure 3-8 – High-scoring GMM components for the relation "on" learned by our weakly-supervised model. Each row shows examples of pairs of boxes (blue on red) for one GMM component. Note that our GMM-based spatial model can capture different configurations for highly multimodal relations such as "on".

chance as the performance given by random ordering of the proposals. Results in Table 3.2 show significant improvements of our method over the baselines. Note that our weakly-supervised method outperforms these strong baselines that are fully supervised. This confirms our results from the previous section that (i) our visual features are well suited to model relations, (ii) they generalize well to unseen triplets, and (iii) training from weak image-level supervision is possible and results only in a small loss of accuracy compared to training from fully supervised data. Examples of retrieved unusual relations are shown in Figure 3-7.

3.5 Qualitative Analysis

3.5.1 Handling multimodal relations

In this section, we provide visualization of learned components of our spatial model to illustrate that our spatial features can handle ambiguous multimodal relations. In Figure 3-8 we show examples of pairs of boxes belonging to the top-4 scoring GMM components for the ambiguous relation "on". Each row corresponds to a different learned mode of "on". In particular, components 1 and 3 represent an object being on top of another object, component 4 corresponds to a garment worn by a person, which is often described by an "on" relation, such as "pants on person". Component 2 often corresponds to the "on top" configuration, where the two objects are captured from an elevated viewpoint.

3.5.2 Qualitative results on UnRel dataset

In Figure 3-9 we show additional qualitative results for triplet retrieval on the UnRel dataset using our weakly-supervised model. Each line corresponds to one unusual triplet query and we plot examples of top-scoring retrieved pairs of boxes (green), top-scoring incorrect pairs (red) and missed detections (blue). A pair of boxes is considered as positive if both subject and object candidates overlap with the corresponding subject and object ground truth with IoU ≥ 0.3. The false positives (red) are either due to incorrect object detection/localization (e.g. the dog in "person ride giraffe" is confused with a giraffe) or failures of the relation classifier in challenging configurations (e.g. "person hold car", "person stand on bench"). The missed detections are often due to the failure of the object detector, which is by itself challenging, as the UnRel dataset contains images of objects in unusual contexts (e.g. "dog ride bike").

3.5.3 Qualitative results for Visual Relationship Detection

What is learnt by action predicates? In Figure 3-10, we show examples of predictions with our weakly-supervised model (Section 3.3.2) for the task of predicate detection. In this task, candidate object boxes are fixed to ground truth boxes and the goal is to predict the relation between a pair of objects. We perform retrieval per class, i.e. for each predicate (one row in Figure 3-10) we show examples of top-scoring object pairs for this relation. This allows us to visualize what our model has learnt for less frequent predicates such as "ride", "carry" or "drive", which are less frequently predicted by our model, as biases in the dataset favor predicates such as "on", "has" or "wear". Similar to prepositions, we see that the spatial configuration of object boxes plays a key role in the prediction of verbs. Indeed, the top-ranked pairs in each row share similar spatial patterns. The top-ranked negatives (in red) demonstrate that it is still challenging to disambiguate subtle differences between relations (e.g. "person ride horse" versus "person on horse", or "person hold watch" versus "person wear watch"). Ground truth can also be incomplete or ambiguous (in yellow), i.e. "person ride bike" is predicted correctly, whereas the ground truth "sit on" is less relevant for this example.

Predicting unseen triplets. In Figure 3-11 we provide additional examples for retrieval of zero-shot triplets. Similar to the previous set-up, we assume the ground truth object boxes to be given and focus on predicting relations. We compare predictions of our weakly-supervised model with the fully supervised Visual+Language model of [Lu et al., 2016a]. For each pair of boxes, we indicate below each image the output of [Lu et al., 2016a]. We also report the ground truth predicates as 'GT'. These examples demonstrate the benefit of our visual features for predicting zero-shot triplets. In some cases, the Visual+Language model of [Lu et al., 2016a] appears to heavily rely on language (e.g. "elephant feed elephant" instead of "elephant next to elephant", which is transferred from "person feed elephant" via language) where

our spatial features predict the correct relation. In other cases the language model suppresses incorrect relations such as "surfboard wear hand" as well as disambiguates subtly different spatial configurations ("kite on street" instead of "kite above street").

3.6 Conclusion and future work

We have developed a new powerful visual descriptor for representing object relations in images achieving state-of-the-art performance on the Visual Relationship Detection dataset [Lu et al., 2016a], and in particular significantly improving the current results on unseen object relations. We have also developed a weakly-supervised model for learning object relations and have demonstrated that, given pre-trained object detectors, object relations can be learnt from weak image-level annotations without a significant loss of recognition performance. Finally, we introduced, UnRel, a new evaluation dataset for visual relation detection that enables to evaluate retrieval without missing annotations and assess generalization to unseen triplets.

The work in this chapter opens-up the possibility of learning a large vocabulary of visual relations directly from large-scale Internet collections annotated with image-level natural language captions. A concrete step in this direction would be to extend our model from fixed to open vocabulary of objects and predicates, transforming the problem of learning visual relation detectors into learning a mapping between visual and textual modalities as in [Bojanowski et al., 2015]. Another extension would be to augment the model capacity by replacing the linear classifier with a deep neural network and allow fine-tuning of visual features. This could be done by iteratively estimating the latent assignments and the weights of the neural network as in [Caron et al., 2018].

Figure 3-9 – Examples of retrieved results for triplet queries on UnRel with our weakly-supervised method using candidate proposals.

Figure 3-10 – Predicate detections on the test set of [Lu et al., 2016a]. We show examples among the top 100 scored triplets for some action relations retrieved by our weakly-supervised model described in Section 3.3.2. In this task, the candidate object boxes are the ground truth boxes. The triplet is correctly recognized if the relation matches ground truth (in green), else the triplet is incorrect (in red). We also show examples of correctly predicted relations where the ground truth is erroneous, either missing or incomplete (in yellow). Below each image, we indicate the ground truth predicates ('GT') for the pair of boxes shown.

Figure 3-11 – Examples of predicate detections on the unseen test triplets of [Lu et al., 2016a] by our weakly-supervised model described in Section 3.3.2 using ground truth object boxes. The triplet is correctly recognized if the relation matches ground truth (in green), else the triplet is incorrect (in red). We also show examples where the ground truth is missing or ambiguous (in yellow). Below each image, we report the prediction of the Visual+Language model of [Lu et al., 2016a], as well as the correct ground truth predicates ('GT') for the pair of boxes.

Chapter 4

Detecting unseen visual relations using analogies

In the previous chapter, we have introduced a new dataset, UnRel, to benchmark visual relation detection models with respect to rare visual relations. We have also developed a compositional approach to recognize unseen triplets, for which training examples of the individual entities are available but their combinations are unseen at training. In this chapter, we seek to further explore the modeling of unseen visual relations. This is an important set-up due to the combinatorial nature of visual relations: collecting sufficient training data for all possible triplets would be very hard. We also propose to go beyond the previously introduced compositional approach and combine it with a holistic visual phrase approach that is more robust the variation of appearance. Contrary to Chapter 3, we adopt a fully-supervised set-up to concentrate on the problem of unseen relations. The contributions in this chapter are three-fold. First, we learn a representation of visual relations that combines (i) individual embeddings for subject, object and predicate together with (ii) a visual phrase embedding that represents the relation triplet. Second, we learn how to transfer visual phrase embeddings from existing training triplets to unseen test triplets using analogies between relations that involve similar objects. Third, we demonstrate the

benefits of our approach on three challenging datasets : on HICO-DET, our model achieves significant improvement over a strong baseline for both frequent and unseen triplets, and we observe similar improvement for the retrieval of unseen triplets with out-of-vocabulary predicates on the COCO-a dataset as well as the challenging unusual triplets in the UnRel dataset.

4.1 Introduction

Understanding interactions between objects is one of the fundamental problems in visual recognition. To retrieve images given a complex language query such as "a woman sitting on top of a pile of books" we need to recognize individual entities "woman" and "a pile of books" in the scene, as well as understand what it means to "sit on top of something". In this chapter we aim to recognize and localize unseen interactions in images, as shown in Figure 4-1, where the individual entities ("person", "dog", "ride") are available at training, but not in this specific combination. Such ability is important in practice given the combinatorial nature of visual relations where we are unlikely to obtain sufficient training data for all possible relation triplets.

Existing methods [Dai et al., 2017; Li et al., 2017a; Lu et al., 2016a] to detect visual relations in the form of triplets $t = (subject, predicate, object)$ typically learn generic detectors for each of the entities, i.e. a separate detector is learnt for subject (e.g. "person"), object (e.g. "horse") and predicate (e.g. "ride"). The outputs of the individual detectors are then aggregated at test time. This *compositional approach* can detect unseen triplets, where subject, predicate and object are observed separately but not in the specific combination. However, it often fails in practice [Zhang et al., 2017a; Peyre et al., 2017], due to the large variability in appearance of the visual interaction that often heavily depends on the objects involved; it is indeed difficult for a single "ride" detector to capture visually different relations such as "person ride horse" and "person ride bus".

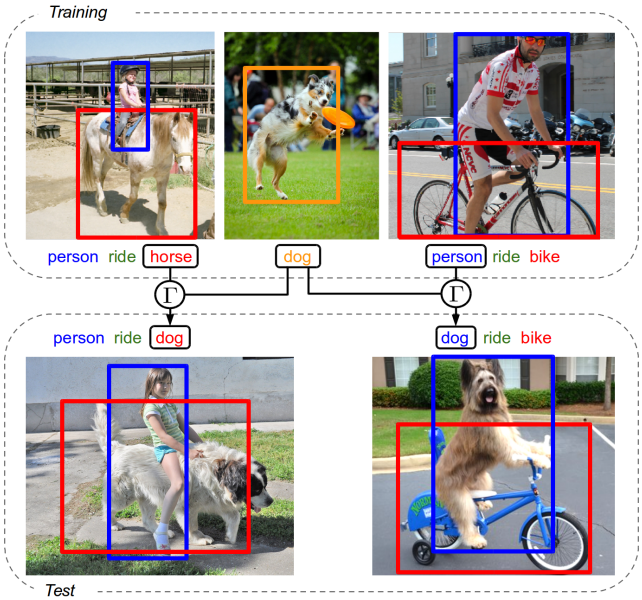


Figure 4-1 – Illustration of transfer by analogy with our model described in 4.3.2. We transfer visual representations of relations seen in the training set such as "person ride horse" to represent new unseen relations in the test set such as "person ride dog".

An alternative approach [Sadeghi and Farhadi, 2011] is to treat the whole triplet as a single entity, called a visual phrase, and learn a separate detector for each of the visual phrases. For instance, separate detectors would be learnt for relations "person ride horse" and "person ride surfboard". While this approach better handles the large variability of visual relations, it requires training data for each triplet, which is hard to obtain as visual relations are combinatorial in their nature and many relations are unseen in the real world.

In this chapter we address these two key limitations. First, what is the right representation of visual relations to handle the large variability in their appearance, which depends on the entities involved? Second, how can we handle the scarcity of training data for unseen visual relation triplets?

To address the first challenge, we develop a hybrid model that combines the compositional and visual phrase representations. More precisely, we learn a compositional representation for subject, object and predicate by learning separate visual-language embedding spaces where each of these entities is mapped close to the language em-

bedding of its associated annotation. In addition, we also learn a relation triplet embedding space where visual phrase representations are mapped close to the language embedding of their corresponding triplet annotations. At test time, we aggregate outputs of both compositional and visual phrase models.

To address the second challenge, we learn how to transfer visual phrase embeddings from existing training triplets to unseen test triplets using analogies between relations that involve similar objects. For instance, as illustrated in Figure 4-1, we recognize the unseen triplet "person ride dog" by using the visual phrase embedding for triplet "person ride horse" after a transformation that depends on the object embedding "dog" and "horse". Because we transfer training data only from triplets that are visually similar, we expect transferred visual phrase detectors to better represent the target relations compared to a generic detector for a relation "ride" that may involve also examples of "person ride train" and "person ride surfboard".

Contributions. Our contributions are three fold. First, we take advantage of both the compositional and visual phrase representations by learning complementary visual-language embeddings for subject, object, predicate and the visual phrase. Second, we develop a model for transfer by analogy to obtain visual-phrase embeddings of never seen before relations. Third, we perform experimental evaluation on three challenging datasets where we demonstrate the benefits of our approach on both frequent and unseen relations.

4.2 Related work

Visual relation detection. Learning visual relations belongs to a general class of problems on relational reasoning Bansal et al. [2017]; Battaglia et al. [2016]; Jenatton et al. [2012]; Kipf and Welling [2016]; Santoro et al. [2017] that aim to understand how entities interact. In the more specific set-up of visual relation detection, the approaches can be divided into two main groups: (i) compositional models, which learn

detectors for subject, object and predicates separately and aggregate their outputs; (ii) and visual phrase models, which learn a separate detector for each visual relation. Visual phrase models such as Sadeghi and Farhadi [2011] have demonstrated better robustness to the visual diversity of relations than compositional models. However, with the introduction of datasets with a larger vocabulary of objects and predicates Chao et al. [2015]; Krishna et al. [2016], visual phrase approaches have been facing severe difficulties as most relations have very few training examples. Compositional methods Gao et al. [2018]; Gkioxari et al. [2018]; Johnson et al. [2015]; Lu et al. [2016a]; Peyre et al. [2017]; Qi et al. [2018]; Shen et al. [2018], which allow sharing knowledge across triplets, have scaled better but do not cope well with unseen relations. To increase the expressiveness of the generic compositional detectors, recent works have developed models of statistical dependencies between the subject, object and predicate, using, for example, graphical models Dai et al. [2017]; Li et al. [2017a], language distillation Yu et al. [2017], or semantic context Zhuang et al. [2017]. Others Atzmon et al. [2016]; Divvala et al. [2014]; Plummer et al. [2017]; Sadeghi et al. [2015a] have proposed to combine unigram detectors with higher-order composites such as bigrams (subject-predicate, predicate-object). In contrast to the above methods that model a discrete vocabulary of labels, we learn visual-semantic (language) embeddings able to scale to out-of-vocabulary relations and to benefit from powerful pre-learnt language models.

Visual-semantic embeddings. Visual-semantic embeddings have been successfully used for image captioning and retrieval Karpathy and Fei-Fei [2015]; Karpathy et al. [2014]. With the introduction of datasets annotated at the region level Krishna et al. [2016]; Plummer et al. [2015], similar models have been applied to align image regions to fragments of sentences Izadinia et al. [2015]; Wang et al. [2016a]. In contrast, learning embeddings for visual relations still remains largely an open research problem with recent work exploring, for example, relation representations using de-

formations between subject and object embeddings Zhang et al. [2017a]. Our work is, in particular, related to models Zhang et al. [2019] learning separate visual-semantic spaces for subject, object and predicate. However, in contrast to Zhang et al. [2019], we additionally learn a visual phrase embedding space to better deal with appearance variation of visual relations, and develop a model for analogy reasoning to infer embeddings of unseen triplets.

Unseen relations and transfer learning. Learning visual phrase embeddings suffers from the problem of lack of training data for unseen relations. This has been addressed by learning factorized object and predicate representations Hwang et al. [2018] or by composing classifiers for relations from simpler concepts Kato et al. [2018]; Misra et al. [2017]. In contrast, our approach transfers visual relation representations from seen examples to unseen ones in a similar spirit to how previous work dealt with inferring classifiers for rare objects Aytar and Zisserman [2011]. The idea of sharing knowledge from seen to unseen triplets to compensate for the scarcity of training data has been also addressed in Ramanathan et al. [2015] by imposing constraints on embeddings of actions. Different from this work, we formulate the transfer as an analogy between relation triplets. To achieve that, we build on the computational model of analogies developed in Reed et al. [2015] but extend it to representations of visual relations. This is related to Sadeghi et al. [2015b] who also learn visual analogies as vector operations in an embedding space, but only consider visual inputs while we learn analogy models for joint image-language embeddings.

4.3 Model

In this section we describe our model for recognizing and localizing visual relations in images. As illustrated in Figure 4-2, our model consists of two parts. First, we learn different visual-language embedding spaces for the subject (s), the object (o), the predicate (p) and the visual phrase (vp), as shown in Figure 4-2(a). We explain

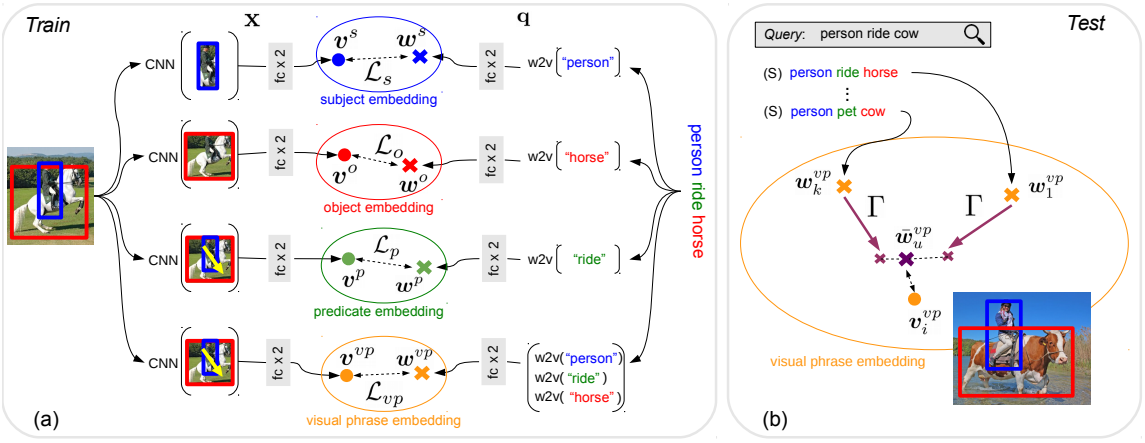


Figure 4-2 – **Model overview.** Our model consists of two parts: (a) learning embedding spaces for subject, object, predicate and visual phrase by optimizing the joint loss $\mathcal{L}_{joint} = \mathcal{L}_s + \mathcal{L}_o + \mathcal{L}_p + \mathcal{L}_{vp}$ combining the input visual $\mathbf{x}$ and language $\mathbf{q}$ representations; (b) inferring the visual phrase embedding $\bar{\boldsymbol{w}}_u^{vp}$ of a new unseen triplet (e.g. "person ride cow") by aggregating the visual phrase embeddings $\boldsymbol{w}_k^{vp}$ of seen triplets (e.g. "person ride horse", "person pet cow") transformed by analogy transformation Γ.

how to train these embeddings in Section 4.3.1. Second, we transfer visual phrase embeddings of seen triplets to unseen ones with analogy transformations, as shown in Figure 4-2(b). In Section 4.3.2 we explain how to train the analogy transformations and form visual phrase embeddings of new unseen triplets at test time.

Notation for relation triplets. The training dataset consists of N candidate pairs of bounding boxes, each formed by a subject candidate bounding box proposal and object candidate bounding box proposal. Let $\mathcal{V}_s$, $\mathcal{V}_o$ and $\mathcal{V}_p$ be the vocabulary of subjects, objects and predicates, respectively. We call $\mathcal{V}_{vp} = \mathcal{V}_s \times \mathcal{V}_p \times \mathcal{V}_o$ the vocabulary of triplets. A triplet t is of the form $t = (s, p, o)$, e.g. $t = (person, ride, horse)$. Each pair of candidate subject and object bounding boxes, $i \in \{1, ..., N\}$, is labeled by a vector $(y_t^i)_{t \in \mathcal{V}_{vp}}$ where $y_t^i = 1$ if the i^{th} pair of boxes could be described by relation triplet t, otherwise $y_t^i = 0$. The labels for subject, object and predicate naturally derive from the triplet label.

4.3.1 Learning representations of visual relations

We represent visual relations in joint visual-semantic embedding spaces at different levels of granularity: (i) at the unigram level, where we use separate subject, object and predicate embeddings, and (ii) at the trigram level using an a visual phrase embedding of the whole triplet. Combining the different types of embeddings results in a more powerful representation of visual relations as will be shown in section 4.4. In detail, as shown in Figure 4-2(a), the input to visual embedding functions (left) is a candidate pair of objects i encoded by its visual representation $\mathbf{x}_i \in \mathbb{R}^{d_v}$. This representation is built from (i) pre-computed appearance features obtained from a CNN trained for object detection and (ii) a representation of the relative spatial configuration of the object candidates. The language embeddings (right in Figure 4-2(a)) take as input a triplet t encoded by its language representation $\mathbf{q}_t \in \mathbb{R}^{d_q}$ obtained from pre-trained word embeddings. We provide more details about these representations in 4.4.2. Next we give details of the embedding functions.

Embedding functions. Our network projects the visual features $\mathbf{x}_i$ and language features $\mathbf{q}_t$ into separate spaces for the subject (s), the object (o), the predicate (p) and the visual phrase (vp). For each input type $b \in \{s, o, p, vp\}$, we embed the visual features and language features into a common space of dimensionality d using projection functions

$$\boldsymbol{v}_i^b = f_v^b(\mathbf{x}_i), \tag{4.1}$$

$$\boldsymbol{w}_t^b = f_w^b(\mathbf{q}_t), \tag{4.2}$$

where $\boldsymbol{v}_i^b$ and $\boldsymbol{w}_t^b$ are the output visual and language representations, and the projection functions $f_v^b : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^d$ and $f_w^b : \mathbb{R}^{d_q} \rightarrow \mathbb{R}^d$ are 2-layer perceptrons, with ReLU non linearities and Dropout, inspired by [Wang et al., 2016a]. Additionally, we L2 normalize the output language features while the output visual features are not

normalized, which we found to work well in practice.

Training loss. We train parameters of the embedding functions (f_v^b, f_w^b) for each type of input b (i.e subject, object, predicate and visual phrase) by maximizing log-likelihood

$$\begin{aligned}\mathcal{L}_b = &\sum_{i=1}^{N}\sum_{t\in\mathcal{V}_b} \mathbb{1}_{y_t^i=1} \log\left(\frac{1}{1+e^{-\boldsymbol{w}_t^{bT}\boldsymbol{v}_i^b}}\right)\\ &+\sum_{i=1}^{N}\sum_{t\in\mathcal{V}_b} \mathbb{1}_{y_t^i=0} \log\left(\frac{1}{1+e^{\boldsymbol{w}_t^{bT}\boldsymbol{v}_i^b}}\right),\end{aligned} \tag{4.3}$$

where the first attraction term pushes closer visual representation $\boldsymbol{v}_i^b$ to its correct language representation $\boldsymbol{w}_t^b$ and the second repulsive term pushes apart visual-language pairs that do not match. As illustrated in Figure 4-2, we have one such loss for each input type and optimize the joint loss that sums the individual loss functions $\mathcal{L}_{joint} = \mathcal{L}_s + \mathcal{L}_o + \mathcal{L}_p + \mathcal{L}_{vp}$. A similar loss function has been used in [Mikolov et al., 2013] to learn word representations, while visual-semantic embedding models [Karpathy et al., 2014; Wang et al., 2016a] typically use triplet ranking losses. Both loss functions work well, but we found embeddings trained with log-loss (4.3) easier to combine across different input types as their outputs are better calibrated.

Inference. At test time, we have a language query in the form of triplet t that we embed as $(\boldsymbol{w}_t^b)_b$ using Eq. (4.2). Similarly, pairs i of candidate object boxes in the test images are embedded as $(\boldsymbol{v}_i^b)_b$ with Eq. (4.1). Then we compute a similarity score $S_{t,i}$ between the triplet query t and the candidate object pair i by aggregating predictions over the different embedding types $b \in \{s, p, o, vp\}$ as

$$S_{t,i} = \prod_{b\in\{s,p,o,vp\}} \frac{1}{1+e^{-\boldsymbol{w}_t^{bT}\boldsymbol{v}_i^b}}. \tag{4.4}$$

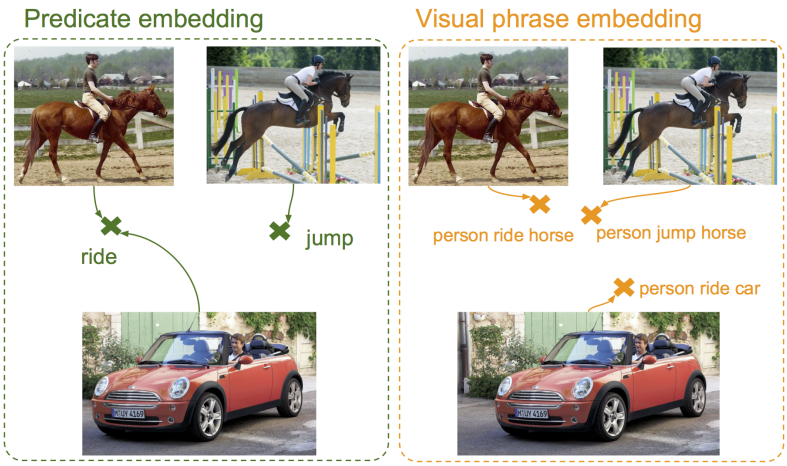


Figure 4-3 – Illustration of the differences between predicate (p) (left) and visual phrase (vp) (right) embeddings. In the p space, visually different relations such as "person ride horse" and "person ride car" map to the same location defined by the predicate "ride". In contrast, they are mapped to distinct locations in the visual phrase space that considers the entire relation triplet.

Interpretation of embedding spaces. The choice of learning different embedding spaces for subject, object, predicate and visual phrase is motivated by the observation that each type of embedding captures different information about the observed visual entity. In Figure 4-3 we illustrate the advantage of learning separate predicate (p) and visual-phrase (vp) embedding spaces. In the p space, visual entities corresponding to "person ride horse" and "person ride car" are mapped to the same point, as they share the same predicate "ride". In contrast, in the vp space, the same visual entities are mapped to two distinct points. This property of the vp space is desirable to handle both language polysemy (i.e., "ride" has different visual appearance depending on the objects involved and thus should not be mapped into a single point) and synonyms (i.e., "person jump horse" and "person ride horse" projections should be close even if they do not share the same predicate).

4.3.2 Transferring embeddings to unseen triplets by analogy transformations

We propose to explicitly transfer knowledge from seen triplets at training to new unseen triplets at test time by analogy reasoning. The underlying intuition is that if we have seen examples of "person ride horse", it might be possible to use this knowledge to recognize the relation "person ride cow", as "horse" and "cow" have similar visual appearance. As illustrated in Figure 4-2(b), this is implemented as an *analogy transformation* in the visual phrase embedding space, where a representation of the source triplet (e.g. "person ride horse") is transformed to form a representation of target triplet (e.g. "person ride cow"). There are two main steps in this process. First, we need to learn how to perform the analogy transformation of one visual phrase embedding (e.g. "person ride horse") to another (e.g. "person ride cow"). Second, we need to identify which visual phrases are suitable for such transfer by analogy. For example, to form a representation of a new relation "person ride cow" we want to transform the representation of "person ride horse" but not "person ride bus". We describe the two steps next.

Transfer by analogy. To transform the visual phrase embedding $\boldsymbol{w}_t^{vp}$ of a source triplet $t = (s, p, o)$ to the visual phrase embedding $\boldsymbol{w}_{t'}^{vp}$ of a target triplet $t' = (s', p', o')$ we learn a transformation Γ such that

$$\boldsymbol{w}_{t'}^{vp} = \boldsymbol{w}_t^{vp} + \Gamma(t, t'). \tag{4.5}$$

Here, Γ could be interpreted as a correction term that indicates how to transform $\boldsymbol{w}_t^{vp}$ to $\boldsymbol{w}_{t'}^{vp}$ in the joint visual-semantic space vp to compute a target relation triplet t' that is analogous to source triplet t. This relates to neural word representations such as [Mikolov et al., 2013] where word embeddings of similar concepts can be linked by arithmetic operations such as "$king$" − "man" + "$woman$" = "$queen$". Here, we

would like to perform operations such as "*person ride horse*" − "*horse*" + "*cow*" = "*person ride cow*".

Form of Γ. To relate the visual phrase embeddings of t and t' through Γ we take advantage of the decomposition of the triplet into subject, predicate and object. In detail, we use the visual phrase embeddings of individual subject, predicate and object to learn how to relate the visual phrase embeddings of triplets. Using this structure, we redefine the analogy transformation given by Eq. (4.5) as

$$\boldsymbol{w}_{t'}^{vp} = \boldsymbol{w}_{t}^{vp} + \Gamma \begin{bmatrix} \boldsymbol{w}_{s'}^{vp} - \boldsymbol{w}_{s}^{vp} \\ \boldsymbol{w}_{p'}^{vp} - \boldsymbol{w}_{p}^{vp} \\ \boldsymbol{w}_{o'}^{vp} - \boldsymbol{w}_{o}^{vp} \end{bmatrix}, \tag{4.6}$$

where $t = (s, p, o)$ and $t' = (s', p', o')$ denote the source and target triplet, and $\boldsymbol{w}_{s}^{vp}$, $\boldsymbol{w}_{p}^{vp}$, $\boldsymbol{w}_{o}^{vp}$ are visual phrase embeddings of subject, predicate and object, respectively, constructed using Eq. (4.2) as $\boldsymbol{w}_{s}^{vp} = f_{w}^{vp}(\mathbf{q}_{[s,0,0]})$, $\boldsymbol{w}_{p}^{vp} = f_{w}^{vp}(\mathbf{q}_{[0,p,0]})$, $\boldsymbol{w}_{o}^{vp} = f_{w}^{vp}(\mathbf{q}_{[0,0,o]})$. Here $[s, 0, 0]$ denotes the concatenation of word2vec embeddings of subject s with two vectors of zeros of size d. For example, the analogy transformation of $t = (person, ride, horse)$ to $t' = (person, ride, camel)$ using Eq. (4.6) would result in

$$\boldsymbol{w}_{t'}^{vp} = \boldsymbol{w}_{t}^{vp} + \Gamma \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \boldsymbol{w}_{camel}^{vp} - \boldsymbol{w}_{horse}^{vp} \end{bmatrix}. \tag{4.7}$$

Intuitively, we would like Γ to encode how the change of objects, observable through the embeddings of source and target objects, $\boldsymbol{w}_{o}^{vp}$, $\boldsymbol{w}_{o'}^{vp}$, influences the source and target triplet embeddings $\boldsymbol{w}_{t}^{vp}$, $\boldsymbol{w}_{t'}^{vp}$. Please note that here we have shown an example of a transformation resulting from a change of object, but our formulation, given by Eq. (4.6), allows for changes of subject or predicate in a similar manner.

While different choices for Γ are certainly possible, we opt for

$$\Gamma(t, t') = MLP \begin{bmatrix} \boldsymbol{w}_{s'}^{vp} - \boldsymbol{w}_{s}^{vp} \\ \boldsymbol{w}_{p'}^{vp} - \boldsymbol{w}_{p}^{vp} \\ \boldsymbol{w}_{o'}^{vp} - \boldsymbol{w}_{o}^{vp} \end{bmatrix}, \tag{4.8}$$

where MLP is a 2-layer perceptron without bias. We also compare different forms of Γ in Section 4.4.

Which triplets to transfer from? We wish to apply the transformation by analogy Γ only between triplets that are similar. The intuition is that to obtain representation of an unseen target triplet $t' = (person, ride, camel)$, we wish to use only similar triplets such as $t = (person, ride, horse)$ but not triplets such as $t = (person, ride, skateboard)$. For this, we propose to decompose the similarity between triplets t and t' by looking at the similarities between their subjects, predicates and objects measured by the dot-product of their representations in the corresponding individual embedding spaces. The motivation is that the subject, object and predicate spaces do not suffer as much from the limited training data compared to the visual phrase space. In detail, we define a weighting function G as:

$$G(t, t') = \sum_{b \in \{s,p,o\}} \alpha_b \boldsymbol{w}_t^{b^T} \boldsymbol{w}_{t'}^{b}, \tag{4.9}$$

where $\boldsymbol{w}_t^{b^T} \boldsymbol{w}_{t'}^{b}$ measures similarity between embedded representations $\boldsymbol{w}_{\cdot}^{b}$ and scalars α_b are hyperparameters that reweight the relative contribution of subject, object and predicate similarities. As we constrain $\sum_b \alpha_b = 1$ the output of $G(t, t') \in [0, 1]$. For a target triplet t', we define as $\mathcal{N}_{t'}$ the set of k most similar source triplets according to G.

Sampling source triplets. We fit the parameters of Γ by learning analogy transformations between triplets available in the training data. To do this, we generate pairs of source t and target t' triplets as follows. For a target triplet t' in the training data, the source triplets for transfer by analogy are sampled in two steps: (i) for a given target triplet t', we first compute the similarity $G(t, t')$ given by Eq. (4.9) using all triplets t in the training data that occur at least 10 times (i.e. the non-rare triplets according to the definition of Chao et al. [2018]), (ii) we sort this set of candidate source triplets, and retain the top k most similar triplets according to G. The outcome is a set of source triplets $\mathcal{N}_{t'}$, similar to the target triplet t' and hence suitable for learning the analogy transformation. Please note that we do not constrain the source triplets to share words with the target triplet, so all words may differ between source and target triplets. Also note that the procedure described above is similar at training and test time. In practice, we take $k = 5$, $\alpha_r = 0.8$, $\alpha_s = \alpha_o = 0.1$ for all datasets. These hyperparameters are optimized by grid-search on the validation set of HICO-DET.

Learning Γ. Given training data pairs of source t and target t' triplets, we fit parameters of Γ by learning analogy transformations between triplets as follows. For each target triplet t' in the training batch, we randomly sample a relevant source triplet $t \in \mathcal{N}_{t'}$ as described above. We call $\mathcal{Q}$ the set of pairs of related triplets (t, t') formed like this. The parameters of Γ are learnt by maximizing the log-likelihood:

$$\begin{aligned}\mathcal{L}_\Gamma = &\sum_{i=1}^{N} \sum_{(t,t')\in\mathcal{Q}} \mathbb{1}_{y^i_{t'}=1} \log\left(\frac{1}{1+e^{-(\boldsymbol{w}_t^{vp}+\Gamma(t,t'))^T \boldsymbol{v}_i^{vp}}}\right) \\ &+ \sum_{i=1}^{N} \sum_{(t,t')\in\mathcal{Q}} \mathbb{1}_{y^i_{t'}=0} \log\left(\frac{1}{1+e^{(\boldsymbol{w}_t^{vp}+\Gamma(t,t'))^T \boldsymbol{v}_i^{vp}}}\right),\end{aligned} \tag{4.10}$$

where $\boldsymbol{v}_i^{vp}$ are the visual features projected to the visual phrase space and $(\boldsymbol{w}_t^{vp} + \Gamma(t, t'))$ is the transformed visual phrase embedding of the source triplet t to target

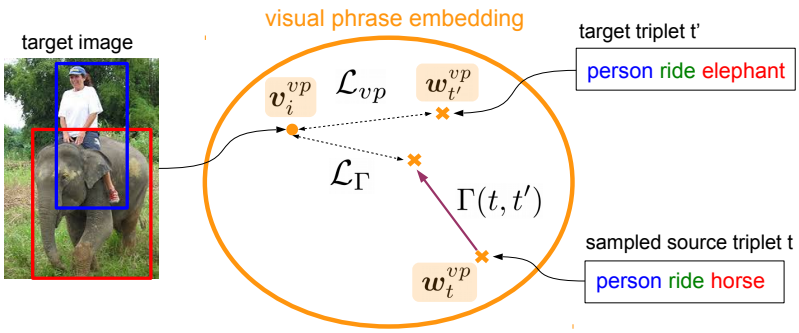


Figure 4-4 – Illustration of training the analogy transformation Γ. For each target triplet t' (e.g. "person ride elephant"), we randomly sample a source triplet t (e.g. "person ride horse"). The first part of the analogy loss $\mathcal{L}_\Gamma$ in Eq (4.10) encourages that the transformed visual phrase embedding $w_t^{vp} + \Gamma(t, t')$ is close to the corresponding visual representation v_i^{vp} of target triplet t.

triplet t' following Eq. (4.6) in Section 4.3. Note that this loss is similar to the loss used for learning embeddings of visual relations, given by Eq. (4.3) in Section 4.3. The first attraction term pulls closer visual representation $\boldsymbol{v}_i^{vp}$ to its corresponding language representation $\boldsymbol{w}_t^{vp}+\Gamma(t,t')$ obtained via analogy transformation, i.e. where the visual representation matches the embedding of the target triplet t' obtained via analogy transformation. We illustrate this term in Figure 4-4. The second repulsive term pushes apart visual-language pairs that do not match, i.e. where the visual representation does not match the target triplet t' obtained via the analogy transformation. The main idea behind Eq. (4.10) is to use the analogy transformation Γ to make the link between the language embedding of a source triplet t and the visual embedding of a target triplet t' in the joint vp space. For example, let us consider source-target pairs of triplets $\mathcal{Q} = \{(t_1, t_1'), (t_2, t_2')\}$ in a mini-batch, where, $t_1 = (person, ride, horse)$, $t_1' = (person, ride, elephant)$, $t_2 = (person, pet, cat)$, $t_2' = (person, pet, sheep)$. The analogy loss in Eq. (4.10) learns Γ that transforms, in the joint vp space, the language embedding of the source triplet $(person, ride, horse)$ such that it is close to the visual embedding of the target triplet $(person, ride, elephant)$ (first term in the loss) but far from the visual embedding of the other target triplet $(person, pet, sheep)$ (second term in the loss).

Aggregating embeddings. At test time, we compute the visual phrase embedding of an unseen triplet u by aggregating embeddings of similar seen triplets $t \in \mathcal{N}_u$ transformed using the analogy transformation:

$$\bar{\boldsymbol{w}}_u^{vp} = \sum_{t \in \mathcal{N}_u} G(t,u)\ (\boldsymbol{w}_t^{vp} + \Gamma(t,u)), \tag{4.11}$$

where $\boldsymbol{w}_t^{vp}$ is the visual phrase embedding of source triplet t obtained with Eq. (4.2), $\Gamma(t,u)$ is the analogy transformation between source triplet t and unseen triplet u computed by Eq. (4.8) and $G(t,u)$ is a scalar weight given by Eq. (4.9) that re-weights the contribution of the different source triplets. This process is illustrated in Figure 4-2(b).

4.4 Experiments

In this section we evaluate the performance of our model for visual relation retrieval on three challenging datasets: HICO-DET [Chao et al., 2018], UnRel [Peyre et al., 2017] and COCO-a [Ronchi and Perona, 2015]. Specifically, we numerically assess the two components of our model: (i) learning the visual phrase embedding together with the unigram embeddings and (ii) transferring embeddings to unseen triplets by analogy transformations.

4.4.1 Datasets and evaluation set-ups

HICO-DET. The HICO-DET [Chao et al., 2015, 2018] dataset contains images of human-object interactions with box-level annotations. The interactions are varied: the vocabulary of objects matches the 80 COCO [Lin et al., 2014b] categories and there are 117 different predicates. The number of all possible triplets is $1 \times 117 \times 80$ but the dataset contains positive examples for only 600 triplets. All triplets are seen at least once in training. The authors separate a set of 138 rare triplets, which are

the triplets that appear fewer than 10 times at training. To conduct further analysis of our model, we also select a set of 25 triplets that we treat as unseen, exclude them completely from the training data in certain experiments, and try to retrieve them at test time using our model. These triplets are randomly selected among the set of non-rare triplets in order to have enough test instances on which to reliably evaluate.

UnRel. UnRel [Peyre et al., 2017] is an evaluation dataset containing visual relations for 76 unusual triplet queries. In contrast to HICO-DET and COCO-a, the interactions do not necessarily involve a human, and the predicate is not necessarily an action (it can be a spatial relation, or comparative). The vocabulary of objects and predicates matches those of Visual Relation Detection Dataset [Lu et al., 2016a]. UnRel is only an evaluation dataset, so similar to [Peyre et al., 2017] we use the training set of Visual Relationship Dataset as training data.

COCO-a. The COCO-a dataset [Ronchi and Perona, 2015] is based on a subset of COCO dataset [Lin et al., 2014b] augmented with annotations of human-object interactions. Similar to HICO-DET, the vocabulary of objects matches the 80 COCO categories. In addition, COCO-a defines 140 predicates resulting in a total of 1681 different triplets. The released version of COCO-a contains 4413 images with no pre-defined train/test splits. Given this relatively small number of images, we use COCO-a as an evaluation dataset for models trained on HICO-DET. This results in an extremely challenging set-up with 1474 unseen triplets among which 1048 involve an out-of-vocabulary predicate that has not been seen at training in HICO-DET.

Evaluation measure. On all datasets, we evaluate our model in a retrieval setup. For each triplet query in the vocabulary, we rank the candidate test pairs of object bounding boxes using our model and compute the performance in terms of Average Precision. Overall, we report mean Average Precision (mAP) over the set of triplet queries computed with the evaluation code released by [Chao et al., 2018] on HICO-

DET and [Peyre et al., 2017] on UnRel. On COCO-a, we use our own implementation as no evaluation code is released.

4.4.2 Implementation details

Candidate pairs. We use pre-extracted candidate pairs of objects from an object detector trained for the vocabulary of objects specific to the dataset. On HICO-DET, we train the object detector on the COCO training data using Detectron [Girshick et al., 2018]. To be comparable to [Gkioxari et al., 2018], we use a Faster-R-CNN [Ren et al., 2015b] with ResNet-50 Feature Pyramid Network [Lin et al., 2017a]. We post-process the candidate detections by removing candidates whose confidence scores are below 0.05 and apply an additional per-class score thresholding to maintain a fixed precision of 0.3 for each object category. For COCO-a, we re-train the object detector excluding images from COCO that intersect with COCO-a. On UnRel, we use the same candidate pairs as [Peyre et al., 2017] to have directly comparable results.

Visual representation. Following [Peyre et al., 2017], we first encode a candidate pair of boxes $(\boldsymbol{o_s}, \boldsymbol{o_o})$ by the appearance of the subject $\boldsymbol{a}(\boldsymbol{o_s})$, the appearance of the object $\boldsymbol{a}(\boldsymbol{o_o})$, and their mutual spatial configuration $\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})$. The appearance features of the subject and object boxes are extracted from the last fully-connected layer of the object detector. The spatial configuration $\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})$ is a 8-dimensional feature that concatenates the subject and object box coordinates renormalized with respect to the union box, i.e. we concatenate $[\frac{x_{min}-T}{A}, \frac{x_{max}-T}{A}, \frac{y_{min}-T}{A}, \frac{y_{max}-T}{A}]$ for subject and object boxes where T and A are the origin and the area of the union box, respectively. The visual representation of a candidate pair is then

$$\mathbf{x}_i = \begin{bmatrix} MLP_s(\boldsymbol{a}(\boldsymbol{o_s})) \\ MLP_o(\boldsymbol{a}(\boldsymbol{o_o})) \\ MLP_r(\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})) \end{bmatrix}, \tag{4.12}$$

where MLP_s, MLP_o contain one layer that projects the appearance features into a vector of dimension 300 and MLP_r is a 2-layer perceptron projecting the spatial features into a vector of dimension 400, making the final dimension of $\mathbf{x}_i$ equal to 1000. For the subject (resp. object) embeddings, we only consider the appearance of the subject (resp. object) without the spatial configuration. Note that both p and vp use the same visual input (including spatial features) while s and o modules only use the appearance features.

Language representation. For a triplet $t = (s, p, o)$, we compute the word embeddings e_s (resp. e_p, e_o) for subject (resp. predicate, object) with a Word2vec [Mikolov et al., 2013] model trained on GoogleNews. The representation of a triplet is taken as the concatenation of the word embeddings $\mathbf{q}_t = [\boldsymbol{e}_s; \boldsymbol{e}_p; \boldsymbol{e}_o] \in \mathbb{R}^{900}$.

Embedding functions. The embedding projection functions are composed of two fully connected layers, with a ReLU non-linearity. For the visual projection functions, we use Dropout. The dimensionality of the joint visual-language spaces is set to $d = 1024$ for HICO-DET and COCO-a. We use $d = 256$ for UnRel as the training set is much smaller.

Training details. First, we learn the parameters of embedding functions by optimizing $\mathcal{L}_{joint} = \mathcal{L}_s + \mathcal{L}_o + \mathcal{L}_p + \mathcal{L}_{vp}$ (Eq. (4.3) in Section 4.3) for 10 epochs with Adam optimizer [Kingma and Ba, 2015] using a learning rate 0.001. Then, we fix parameters of the embedding functions for s, o and p and only finetune parameters of the visual phrase embedding function vp while learning parameters of analogy transformation Γ. This is done by jointly optimizing $\mathcal{L}_{vp} + \lambda\mathcal{L}_\Gamma$ for 5 epochs with Adam optimizer Kingma and Ba [2015] using a learning rate 0.001. In practice, we take $\lambda = 1$. In this joint optimization, we found it helpful to restrict back-propagation of gradients coming from $\mathcal{L}_\Gamma$ only to the parameters of analogy transformation Γ and parameters of the visual embedding functions f_v^b (Eq. (4.1)), i.e. we exclude

back-propagation of gradients coming from $\mathcal{L}_\Gamma$ to parameters of language embedding functions f_w^b. These parameters are finetuned using gradients back-propagated from $\mathcal{L}_{vp}$. The hyperparameters α_s, α_o, α_p and k are optimized by grid-search on the validation set.

Batch sampling. In practice, our model is trained with mini-batches containing 64 candidate object pairs. 25% of the candidate pairs are positive, i.e. the candidate subject and object are interacting. The rest are negative, randomly sampled among candidate pairs involving the same subject and object category (but not interacting). For training, we use candidates from both ground truth and object detector outputs. At test time, we only use candidate pairs from the object detector.

4.4.3 Evaluating visual phrases on seen triplets

We first validate the capacity of our model to detect triplets seen at training and compare with recent state-of-the-art methods. In Table 4.1, we report mAP results on HICO-DET in the Default setting defined by Chao et al. [2018] on the different subsets of triplets (full), (rare), (non rare) as described in 4.4.1. First, we compute three variants of our model : (i) the compositional part using all unigram terms (s+o+p), which can be viewed as a strong fully compositional baseline, (ii) the visual phrase part combined with object scores (s+o+vp), and (iii) our full model (s+o+p+vp) that

	full	rare	non-rare
Chao et al. [2018]	7.8	5.4	8.5
Gupta and Malik [2015]	9.1	7.0	9.7
Gkioxari et al. [2018]	9.9	7.2	10.8
GPNN Qi et al. [2018]	13.1	9.3	14.2
iCAN Gao et al. [2018]	14.8	10.5	16.1
s+o+p	18.7	13.8	20.1
s+o+vp	17.7	11.6	19.5
s+o+p+vp	**19.4**	**14.6**	**20.9**

Table 4.1 – Retrieval results on HICO-DET dataset (mAP).

	Base	With aggregation G			
	-	$\Gamma=\emptyset$	$\Gamma=0$	$\Gamma=linear$	$\Gamma=deep$
s+o+p	23.2	-	-	-	-
s+o+vp+transfer	24.1	9.6	24.8	27.6	**28.6**
s+o+p+vp+transfer	23.6	12.5	24.5	25.4	**25.7**
supervised	33.7	-	-	-	-

Table 4.2 – mAP on the 25 zero-shot test triplets of HICO-DET with variants of our model trained on the *trainval* set excluding the positives for the zero-shot triplets. The first column shows the results without analogy transfer (Section 4.3.1) while the other columns display results with transfer using different forms of analogy transformation Γ (Section 4.3.2). Last line (supervised) is the performance of (s+o+p+vp) trained will all training instances.

corresponds to the addition of the visual phrase representation on top of the compositional baseline (Section 4.3.1). The results show that our visual phrase embedding is beneficial, leading to a consistent improvement over the strong compositional baseline on all sets of triplets, improving the current state-of-the art Gao et al. [2018] by more than 30% in terms of relative gain. We provide ablation studies in Section 4.5 as well as experiments incorporating bigrams modules (sr+ro) leading to improved results.

4.4.4 Transfer by analogy on unseen triplets

Next, we evaluate the benefits of transfer by analogy focusing on the challenging set-up of triplets never seen at training time. While the HICO-DET dataset contains both seen (evaluated in previous section) and manually constructred unseen triplets (evaluated here), in this section we consider additional two datasets that contain only unseen triplets. In particular, we use UnRel to evaluate retrieval of unusual (and unseen) triplets and COCO-a to evaluate retrieval of unseen triplets with out-of-vocabulary predicates.

Evaluating unseen triplets on HICO-DET. First, we evaluate our model of transfer by analogy on the 25 zero-shot triplets of HICO-DET. In Table 4.2, we show

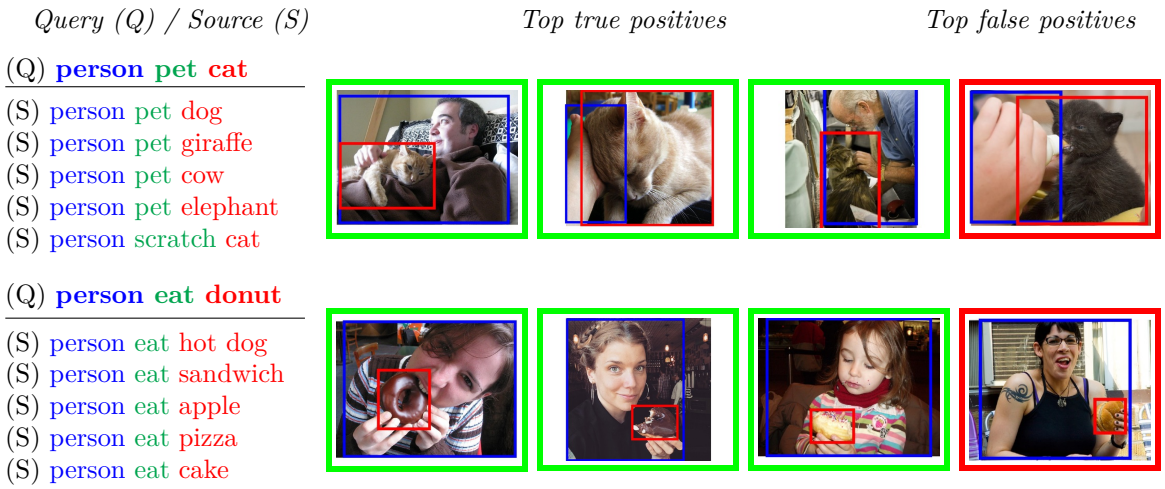


Figure 4-5 – Top retrieved positive (green) and negative (red) detections with our model (s+o+vp+transfer) on unseen triplets excluded from HICO-DET. For a target triplet (Q) (e.g. "person pet cat"), our model automatically learns to select meaningful source triplets (S) involving visually similar objects or predicates ("person pet dog", "person scratch cat") and transfers their visual phrase embeddings by analogy transformation Γ. The top false positives are challenging, either corresponding to a visually related action ("person feed cat" in first row) or to a visually similar object ("donut" confused with "sandwich" in second row). Additional examples are in Figure 4-9.

results for different types of analogy transformations applied to the visual phrase embeddings to be compared with the base model not using analogy (first column). First, $\Gamma=\emptyset$ corresponds to aggregation of visual phrase embeddings of source triplets without analogy transformation. Then, we report three variants of an analogy transformation, where visual phrase embeddings are trained with analogy loss and the embedding of source triplet is either (i) aggregated without transformation ($\Gamma=0$), or transformed with (ii) a linear transformation ($\Gamma=linear$) or (iii) a 2-layer perceptron ($\Gamma=deep$). The results indicate that forming visual phrase embeddings of unseen test triplets by analogy transformations of similar seen triplets, as described in 4.3.2, is beneficial, with the best model (s+o+vp+transfer using $\Gamma=deep$) providing a significant improvement over the compositional baseline (from mAP of 23.2 to 28.6), thus partly filling the gap to the fully supervised setting (mAP of 33.7). It is also interesting to note that, when aggregating visual phrase embeddings of different

source triplets as described in Eq. (4.11), transforming the visual phrase embedding via analogy prior to the aggregation is necessary, as indicated by the significant drop of performance when $\Gamma=\emptyset$. In Figure 4-5 we show qualitative results for retrieval of unseen triplets with the (s+o+vp+transfer) model. For a query triplet (Q) such as "person pet cat" we show the top 3 retrieved candidate pairs (green), and the top 1 false positive (red). Also, for each target triplet, we show the source triplets (S) used in the transfer by analogy (Eq. (4.11)). We note that the source triplets appear relevant to the query.

Evaluating unseen (unusual) triplets on UnRel. Table 4.3 shows numerical results for retrieval on the UnRel dataset. Similar to [Peyre et al., 2017], we also do not use subject and object scores as we found them uninformative on this dataset containing hard to detect objects. For transfer by analogy we use $\Gamma=$*deep*. First, we observe that our (p+vp+transfer) method improves over all other methods, significantly improving the current state-of-the-art [Peyre et al., 2017] on this data, as well as outperforming the image captioning model of [Johnson et al., 2016] trained on a larger corpus. Note that we use the same detections and features as [Peyre et al., 2017], making our results directly comparable. Second, the results confirm the benefits of transfer by analogy (p+vp+transfer) over the fully compositional baseline (p)

	With GT	With candidates		
	-	union	subj	subj/obj
Johnson et al. [2016]	-	6.2	6.8	-
Lu et al. [2016a]	50.6	12.0	10.0	7.2
Peyre et al. [2017] full	62.6	14.1	12.1	9.9
p	62.2	16.8	15.2	12.6
vp	53.4	13.2	11.7	9.4
p+vp	61.7	16.4	14.9	12.6
vp+transfer	53.7	13.7	12	9.7
p+vp+transfer	**63.9**	**17.5**	**15.9**	**13.4**

Table 4.3 – Retrieval on UnRel (mAP) with IoU=0.3.

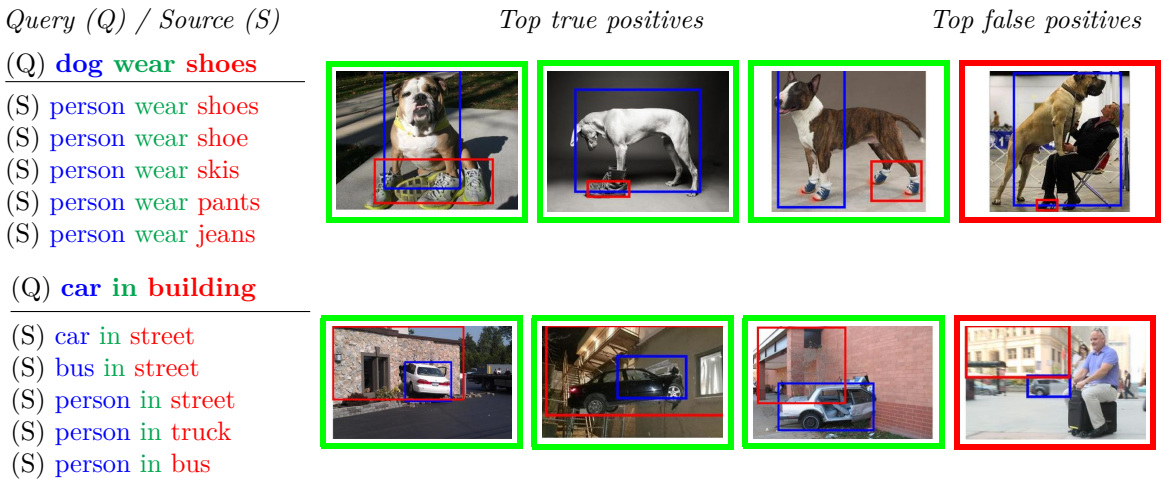


Figure 4-6 – Top retrieved positive (green) and negative (red) detections with our model (p+vp+transfer) on UnRel triplets. The embedding of the unseen query triplet (Q) is formed from the embedding of seen source triplets (S) via analogy transformation. While transfer with analogy on HICO-DET is often done through change of object, here, for retrieving the unseen triplet "dog wear shoes", our model samples source triplets involving a different subject, "person", in interaction with similar objects (e.g. "person wear shoes", "person wear skis"). Additional results are in Figure 4-10.

	all	out of vocabulary
s+o+p	4.3	4.2
s+o+vp	6.0	6.2
s+o+p+vp	5.1	5.1
s+o+vp+transfer	**6.9**	**7.3**
s+o+p+vp+transfer	5.2	5.1

Table 4.4 – Retrieval on unseen triplets of COCO-a (mAP). We show the performance on all unseen triplets (first column) and on unseen triplets involving out-of-vocabulary predicates (second column).

with a consistent improvement in all evaluation metrics. Interestingly, contrary to HICO-DET, using visual phrase embeddings without transfer (p+vp) does not bring significant improvements over (p). This is possibly due to the large mismatch between training and test data as the UnRel dataset used for testing contains unusual relations, as shown in the qualitative examples in Figure 4-6. This underlines the importance of the transfer by analogy model.

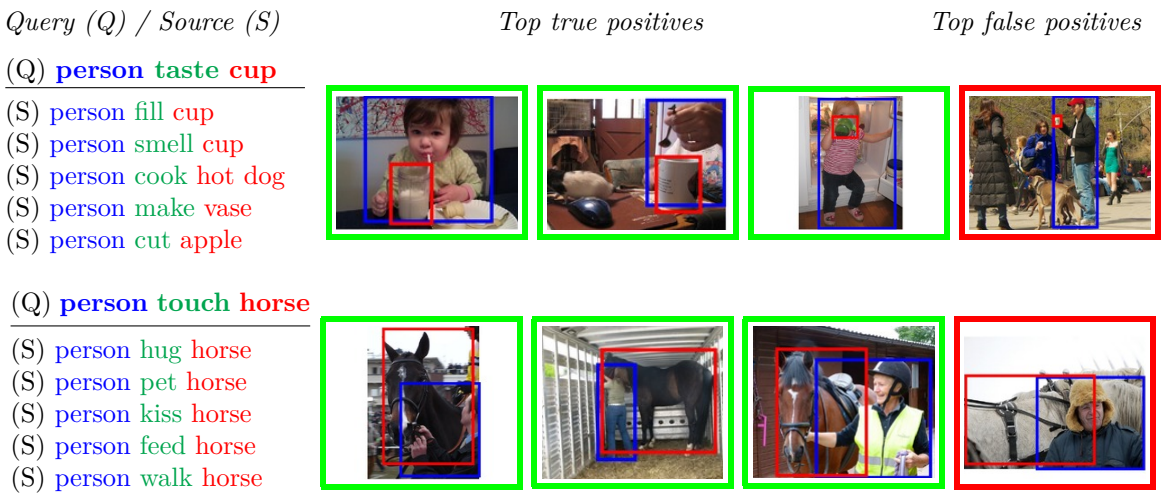


Figure 4-7 – Top retrieved positives (green) and negatives (red) detections with our model (s+o+vp+transfer) of COCO-a triplets. The embedding of the query triplet (Q) to retrieve is formed with the embedding of source triplets (S) by analogy. For retrieving out-of-vocabulary triplets such as "person taste cup", our model of transfer by analogy automatically samples relevant source triplets involving similar predicates and objects (e.g. "person smell cup", "person make vase"). Additional results are in Figure 4-11.

Evaluating unseen (out-of-vocabulary) triplets on COCO-a. Finally, we evaluate our model trained on HICO-DET dataset for retrieval on the unseen triplets of COCO-a dataset. This is an extremely challenging setup as the unseen triplets of COCO-a involve predicates that are out of the vocabulary of the training data. The results shown in Table 4.4 demonstrate the benefits of the visual phrase representation as previously observed on HICO-DET and UnRel datasets. Furthermore, the results also demonstrate the benefits of analogy transfer: compared to the fully compositional baseline (s+o+p) our best analogy model (s+o+vp+transfer) obtains a relative improvement of 60% on all, and more than 70% on the out of vocabulary triplets. Qualitative results are shown in Figure 4-7.

4.5 Ablation studies

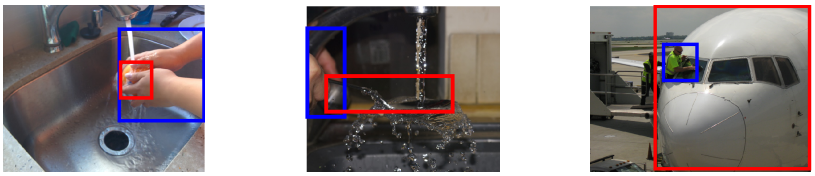
In this section, we perform ablation studies that complement the analysis in Section 4.4. We discuss the benefits of the different components of our model introduced

		full	rare	non-rare
(a)	s+o (obj.det.)	5.6	4.2	6.5
(b)	s+o	10.0	7.6	10.8
(c)	p	14.9	9.4	16.5
(d)	bigrams	14.9	9.6	16.5
(e)	vp	16.5	10.4	18.4
(f)	s+o+vp (*Table 4.1*)	17.7	11.6	19.5
(g)	s+o+p (classifier)	18.0	13.4	19.4
(h)	s+o+p (random words)	18.4	13.7	19.8
(i)	s+o+p (*Table 4.1*)	18.7	13.8	20.1
(j)	s+o+p (finetuned words)	18.8	14.5	20.1
(k)	s+o+p+vp (*Table 4.1*)	19.4	14.6	20.9
(l)	s+o+p+bigrams	19.5	14.6	21.0
(m)	s+o+p+vp+bigrams	**20.0**	**15.0**	**21.5**

Table 4.5 – Ablation study on HICO-DET.

in Section 4.3.1, and in particular the benefits of the visual phrase module. We also analyze the influence of pre-trained word embeddings and the effect of adding bigrams modules.

Benefits of different components of our model. Our contribution is a hybrid model which combines subject (s), object (o), predicate (p) and visual phrase (vp) modules. We show in Table 4.5, which complements Table 4.1, that each of these modules is making a complementary contribution. The performance of our compositional model $s+o+p$ builds on our strong unigram models $s+o$ (row (b)) that already significantly improve over the baseline using only the object scores returned by pre-trained object detectors (row (a)) typically used by other methods Gao et al. [2018]; Gkioxari et al. [2018]. The strength of our modules for representing visual relations is clearly demonstrated by the good performance of our unigram predicate model p (row (c)) and the visual phrase model vp (row (e)) over using objects alone (cf. $s+o$, row (b)). In addition, vp alone performs better than p alone (row (e) > (c)). Importantly, these modules are complementary as clearly shown by the best performance of our combined model (row (k)) that can also easily incorporate bigrams (row (m)), see



(a) person wash orange (b) person wash spoon (c) person wash airplane

Figure 4-8 – Retrieval examples where *s+o+p+vp* is better than *s+o+p*.

below.

Benefits of visual phrase (vp) model. The improvement thanks to the *vp* model is consistent over several datasets. We found qualitatively that the *vp* branch handles important unusual situations where the compositional model (*s+o+p*) fails, which happens when (at least) one of the *s*, *o* and *p* branches has a low score, e.g. due to object occlusion (Figure 4-8(a)), unusual object appearance (Figure 4-8(b)) or unusual spatial configuration (Figure 4-8(c)). The visual phrase model (*vp*) can better handle these situations because it better models the specific appearance and spatial configuration of triplets seen in training.

Benefits of word vectors. In Table 4.5 we (1) show benefits of mapping input triplets to image-language embedding space instead of learning *s*, *o* and *p* classifiers (row (h) > (g)) and (2) confirm that using pre-trained word embeddings helps, but only slightly (row (i) > (h)). Because of the mismatch between word usage in the pre-training text corpus and our dataset, we also found that fine-tuning the pre-trained word embeddings is beneficial (row (j) > row (i)).

Incorporating bigrams. While our primary focus is to marry compositional (unigrams) and visual phrase (trigram) models, we can easily incorporate bigram branches (*sp+po*) in our model. As shown in Table 4.5, bigrams provide an improvement over unigrams modelling subject and object independently *s+o* (row (d) > (b)) and combined with unigrams (row (l)) they reach comparable results to a combination of

unigrams and trigram (row (k)). Interestingly, bigrams and trigram are complementary. Their combination leads to the overall best results (row (m)).

Alternative to weighting function G. We tested an alternative to the weighting function G (Eq. (4.9) taking as input word2vec embeddings instead of joint visual-semantic embeddings in s, o and p spaces. This lead to a slight performance drop (28.3 vs. 28.6 for our analogy transfer in Table 4.2). This result suggests that while pre-trained language embeddings are core ingredients to establish similarities between concepts, they can be further strengthen by using visual appearance.

4.6 Qualitative analysis

4.6.1 Qualitative results on HICO-DET dataset

In Figure 4-9 we show additional examples of retrieved detections for unseen triplets that supplement Figure 4-5. These qualitative examples confirm that our model for transfer by analogy (s+o+vp+transfer) (Section 4.3.2) automatically selects relevant source triplets (S) given an unseen triplet query (Q). For instance, for the query triplet "person throw frisbee" (first row), our model selects (1) a source triplet that involves the same action, with a different, but similar, object "person throw sports ball", (2) two source triplets with the same object, and different, but related, actions "person catch frisbee", "person block frisbee" and (3) two other source triplets with different, but related, object and actions "person hit sports ball", "person serve sports ball". Similar conclusions hold for the other examples displayed. The top false positives indicate that the main failure mode is the confusion with another similar interaction (e.g. "lie on" is confused with "sit on" in row 3 or "inspect" is confused with "hold" in row 4. Some detections are also incorrectly classified as failure, as they are still some missing ground truth annotations (e.g. row 2, row 6).

Figure 4-9 – **Retrieval examples on the HICO-DET dataset.** Top retrieved positives (green) and negatives (red) using our model (s+o+vp+transfer) for unseen triplet queries. The query is marked as (Q). The source triplets automatically selected by our model are marked as (S). For instance, for the query triplet "person throw frisbee" (first row), our model selects (1) a source triplet that involves the same action, with a different, but similar, object "person throw sports ball", (2) two source triplets with the same object, and different, but related, actions "person catch frisbee", "person block frisbee" and (3) two other source triplets with different, but related, object and actions "person hit sports ball", "person serve sports ball". The top false positives show the main failure mode: the interaction is confused with another similar interaction (e.g. "lie on" is confused with "sit on" in row 3 or "inspect" is confused with "hold" in row 4). Also, we note that some mistakes among the top false positives are due to missing ground truth annotations.

4.6.2 Qualitative results on UnRel dataset

In Figure 4-10 we show additional qualitative results for our model (p+vp+transfer) for retrieval of unseen (unusual) triplets on the UnRel dataset supplementing results shown in Figure 4-6. We show the source triplets (S) automatically sampled by our analogy model that are used to form the visual phrase embedding of the target query (Q). The top true positive retrievals are shown in green, the top false positive retrieval is shown in red. The automatically sampled source triplets all appear relevant. Our method samples source triplets involving (1) a different subject ("dog ride bike" is transferred from "person ride bike", "building has wheel" is transferred from "truck has wheel"), (2) a different object ("person stand on horse" is transferred from "person stand on sand"), or (3) a different predicate ("cone on the top of person" is transferred from "sky over person"). The results confirm that our model works well not only for human-object interactions but also for more general interactions involving spatial relations (e.g. "in", "on the top of") or a subject different from a person (e.g. "cone", "car", "building", "dog"). There are two main failure modes illustrated by the top false positive detections. The first one is an incorrect object detection (e.g. "train" is confused with "building" in row 3, or "motorcycle" is confused with "bike" in row 2). The second failure mode is due to the confusion with another similar triplet, possibly due to the unusual character of UnRel queries which sometimes make it difficult to sample close enough source triplets for the transfer by analogy. For instance, it is hard to form a good embedding for "car in building" from source triplets "car in street", "bus in street", "person in street" as these source triplets have fairly different visual appearance (row 5).

4.6.3 Qualitative results on COCO-a dataset

In Figure 4-11, we show additional qualitative results of our model for transfer by analogy (s+o+vp+transfer) on retrieval of unseen (out of vocabulary) triplets in the

Figure 4-10 – **Querying for unseen (unusual) triplets on the UnRel dataset.** Examples of retrieval using our model (p+vp+transfer). The query triplet is marked as (Q). The source triplets (S) seen in training are automatically selected by our model described in Section 4.3.2 and used to transfer the visual phrase embedding using the analogy transformation. The automatically selected source triplets all appear relevant. Our method selects source triplets involving (1) a different subject ("dog ride bike" is transferred from "person ride bike", "building has wheel" is transferred from "truck has wheel"), (2) different object ("person stand on horse" is transferred from "person stand on sand"), or (3) different predicate ("cone on the top of person" is transferred from "sky over person").

COCO-a dataset, complementing results in Figure 4-7. We display the source triplets (S) automatically sampled by our model for a target query (Q). Despite the fact that the target predicates are not seen in training, our model manages, most of the time, to sample relevant source triplets for transfer. For instance, our model would link the unseen triplet "person use laptop", involving the unseen predicate "use" (row 2) to source triplets such as "person type on laptop", "person read laptop" or "person text on phone", all involving a predicate that is relevant to the unseen target predicate "use". The same holds for the unseen triplet "person touch horse" (row 3) for which our model samples source triplets involving contact interaction such as "person hug horse", "person pet horse" or "person kiss horse". The top false detections are informative: (i) either they correspond to interactions involving related triplets, which are likely to be sampled as source triplets (e.g. "person shear sheep" confused with "person caress sheep" in row 1), (ii) or they correspond to interactions with ambiguous semantics (e.g. "person get frisbee" or "person prepare kite" that involve ambiguous predicates that could correspond to a large variety of spatial configurations).

4.6.4 Visualization of joint embedding spaces

Here, we provide additional insights about the embedding spaces learnt on the HICO-DET dataset and UnRel dataset using the t-sne visualization Van der Maaten and Hinton [2008] of the final learnt joint embedding. First, we show t-sne visualization Van der Maaten and Hinton [2008] of joint embedding spaces learnt for objects and predicates on HICO-DET to better understand which concepts are close together in the learnt space. For the object embedding, as shown in Figure 4-12, objects are grouped according to their visual and semantic similarity. The same holds for predicate embeddings shown in Figure 4-13. We draw similar plots for UnRel dataset, showing the object embedding in Figure 4-14 and the predicate embedding in Figure 4-15. The visualization of predicate embedding on UnRel dataset in Figure 4-15 is especially interesting as it involves spatial relations. We remark that our model is

Figure 4-11 – **Querying for unseen (out of vocabulary) triplets on the COCO-a dataset.** Examples of retrieval using our model (s+o+vp+transfer). The query triplet is marked as (Q). The source triplets (S) seen in training are automatically selected by our model described in Section 4.3.2 and used to transfer the visual phrase embedding using the analogy transformation. The automatically selected source triplets all appear relevant despite the difficulty that all predicates involved in the shown triplet queries are unseen at training time. The transfer to unseen predicates is made possible by the use of pre-trained word2vec embeddings. Given out-of-vocabulary triplets such as "person use laptop" (row 2), our model automatically samples source triplets involving a relevant predicate such as "person type on laptop". However, we also observe that sometimes the out-of-vocabulary predicate is ambiguous (e.g. "prepare" or "get"), which makes it challenging to identify relevant source triplets among the set of available training triplets (e.g. "person launch kite", "person catch frisbee").

able to separate spatial relations such as “under” from “above” which are semantically very similar. Learning good embedding for unigrams is crucial in our model for transfer by analogy, as unigram embeddings directly influence the analogy transformation from the seen visual phrases to the unseen ones.

4.7 Conclusion and future work

In this chapter, we have developed a new approach for visual relation detection that combines complementary compositional and visual phrase representations. Furthermore, we have proposed a model for transfer by analogy able to compute visual phrase embeddings of never seen before relations. We have demonstrated benefits of our approach on three challenging datasets involving unseen triplets.

In the future, we hope to further improve the generalization to unseen combinations involving unseen entities. In particular, in this chapter we have seen that it is possible to transfer to triplets involving unseen predicates (results on COCO-a in Table 4.4). Yet the results we obtained are still far from those on the seen predicates of HICO-DET (Table 4.2). We thus wish to better exploit semantic cues to learn relations between predicate categories, possibly enforcing logical consistencies such as in [Deng et al., 2014; Ramanathan et al., 2015]. Also, a future direction not addressed in this chapter involves the generalization to unseen objects. Evaluation could be done on an extension of UnRel dataset with images from Open Images dataset [Kuznetsova et al., 2018] involving out-of-vocabulary objects. The work presented in this chapter also opens-up possibilities to learn from data annotated at different levels of granularities such as unigram, bigram and trigram terms. An interesting extension could be to replace the trigram branch by a module able to map image regions with short unstructured text descriptions such as [Johnson et al., 2016]. This would enable us to train visual phrase embeddings with more data, augmenting the diversity of source triplets, and is thus likely to improve the benefits of transfer by analogy.

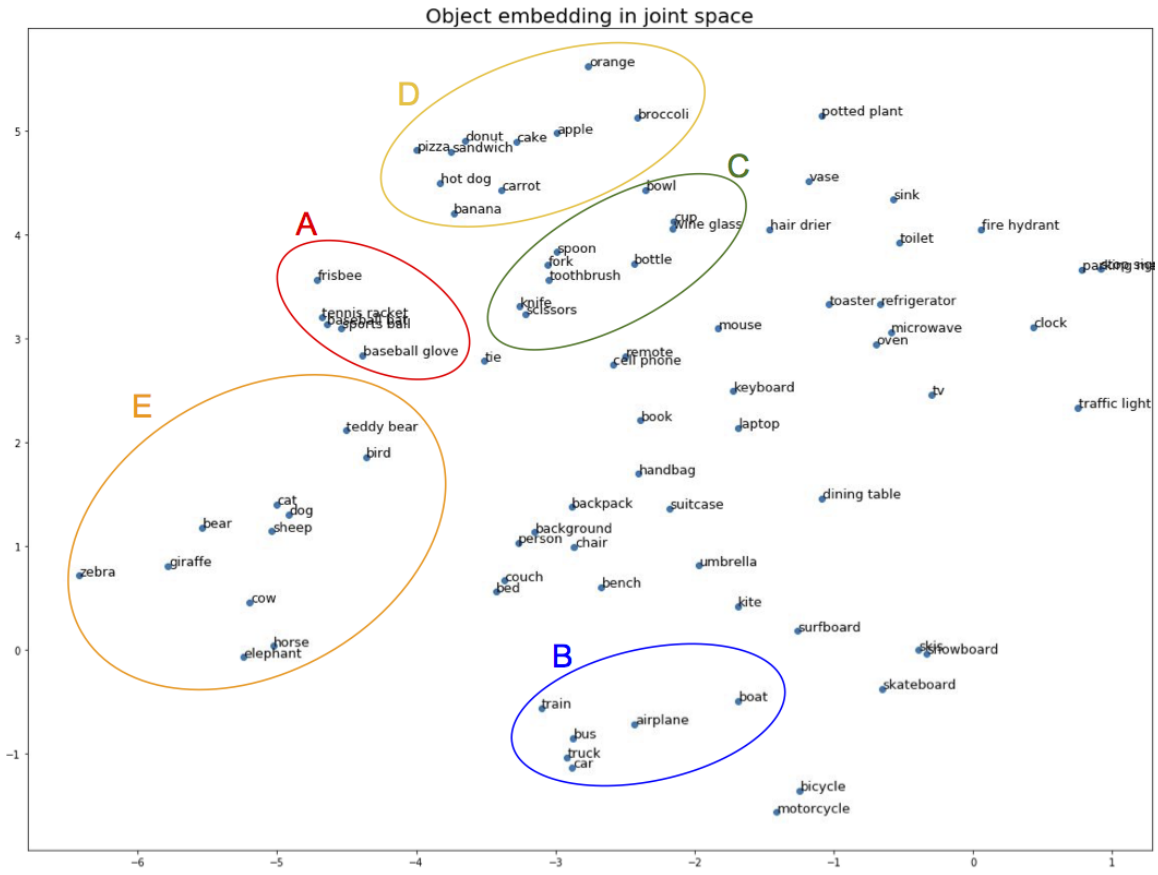


Figure 4-12 – Object embedding on HICO-DET visualized using T-sne [Van der Maaten and Hinton, 2008]. Objects appear to be grouped according to their visual and semantic similarity. For example, we highlight regions corresponding to: (A) sports instruments (e.g. "tennis racket", "frisbee"), (B) big transportation (e.g. "bus", "train"), (C) eating utensils (e.g. "fork", "cup"), (D) food (e.g. "pizza", apple"), (E) animals (e.g. "giraffe", "bird"). Learning a good embedding for unigrams (here objects) is crucial in our model that uses the transfer by analogy, as unigram embeddings directly influence the analogy transformation from the seen visual phrases to the unseen ones.

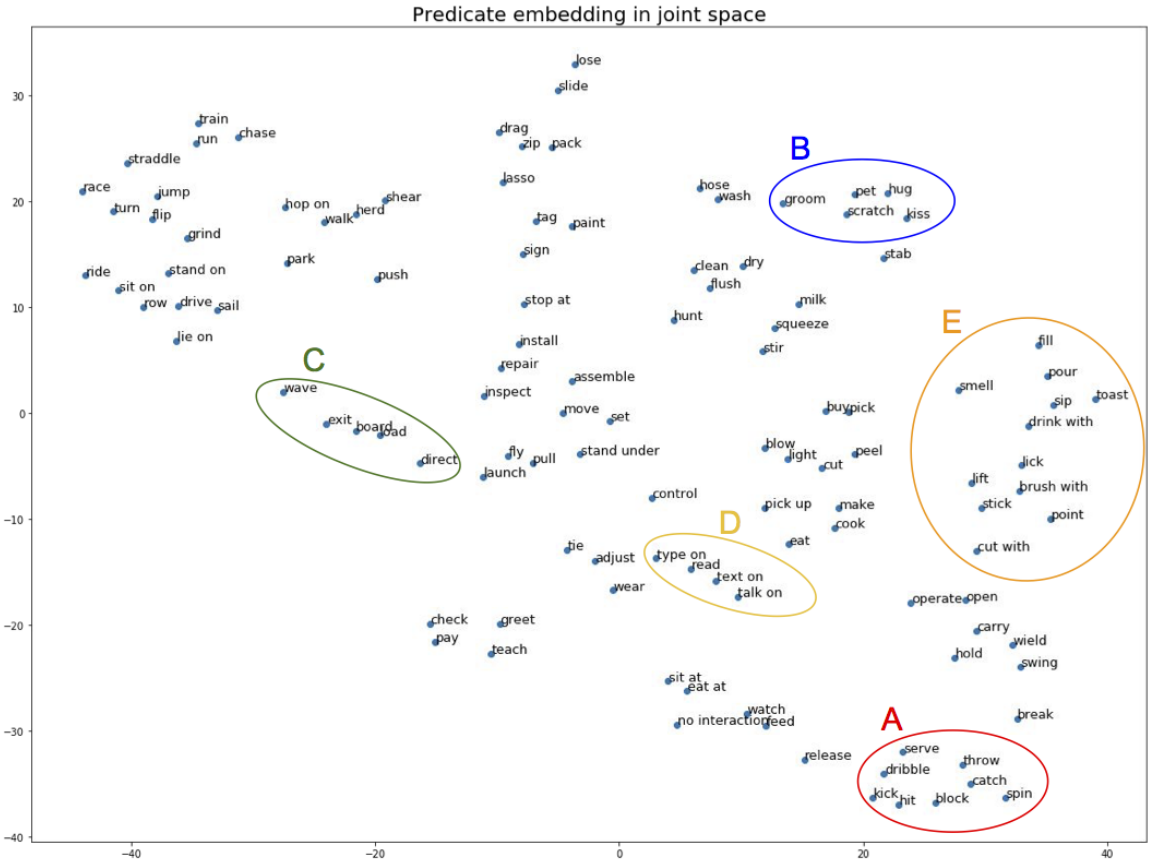


Figure 4-13 – Predicate embedding on HICO-DET visualized with T-sne [Van der Maaten and Hinton, 2008]. The predicates are grouped according to their visual and semantic similarity. For example, we highlight regions corresponding to: (A) interactions related to sports (e.g. “throw”, “dribble”), (B) gentle interactions with an animal/person (e.g. “hug”, “kiss”), (C) interactions with transportation vehicles (e.g. “board”, “exit”), (D) interactions with (electronic) devices (e.g. “text on”, “read”), (E) interactions with food (e.g. “smell”, “lick”). Learning a good embedding for unigrams (here predicates) is crucial in our model that uses transfer by analogy, as unigram embeddings directly influence the analogy transformation from the seen visual phrases to the unseen ones.

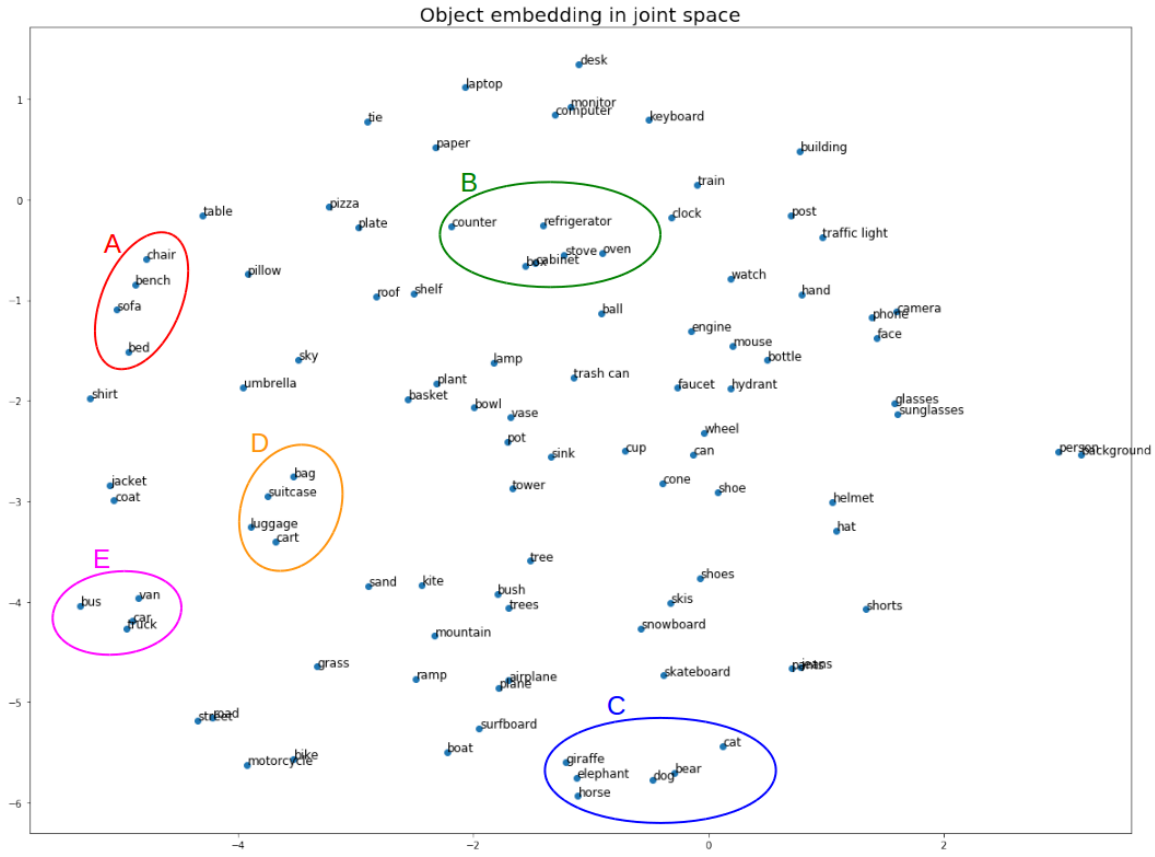


Figure 4-14 – Object embedding on the UnRel dataset visualized using T-sne [Van der Maaten and Hinton, 2008]. Objects appear to be grouped according to their visual and semantic similarity. For example, we highlight regions corresponding to: (A) piece of furniture on which to sit (e.g. “chair”, “bench”), (B) kitchen furniture (e.g. “refrigerator”, “stove”), (C) animals (e.g. “giraffe”, “cat”), (D) bags and containers (e.g. “suitcase”, cart”), (E) motorized transportation (e.g. “bus”, “car”). Learning a good embedding for unigrams (here objects) is crucial in our model that uses the transfer by analogy, as unigram embeddings directly influence the analogy transformation from the seen visual phrases to the unseen ones.

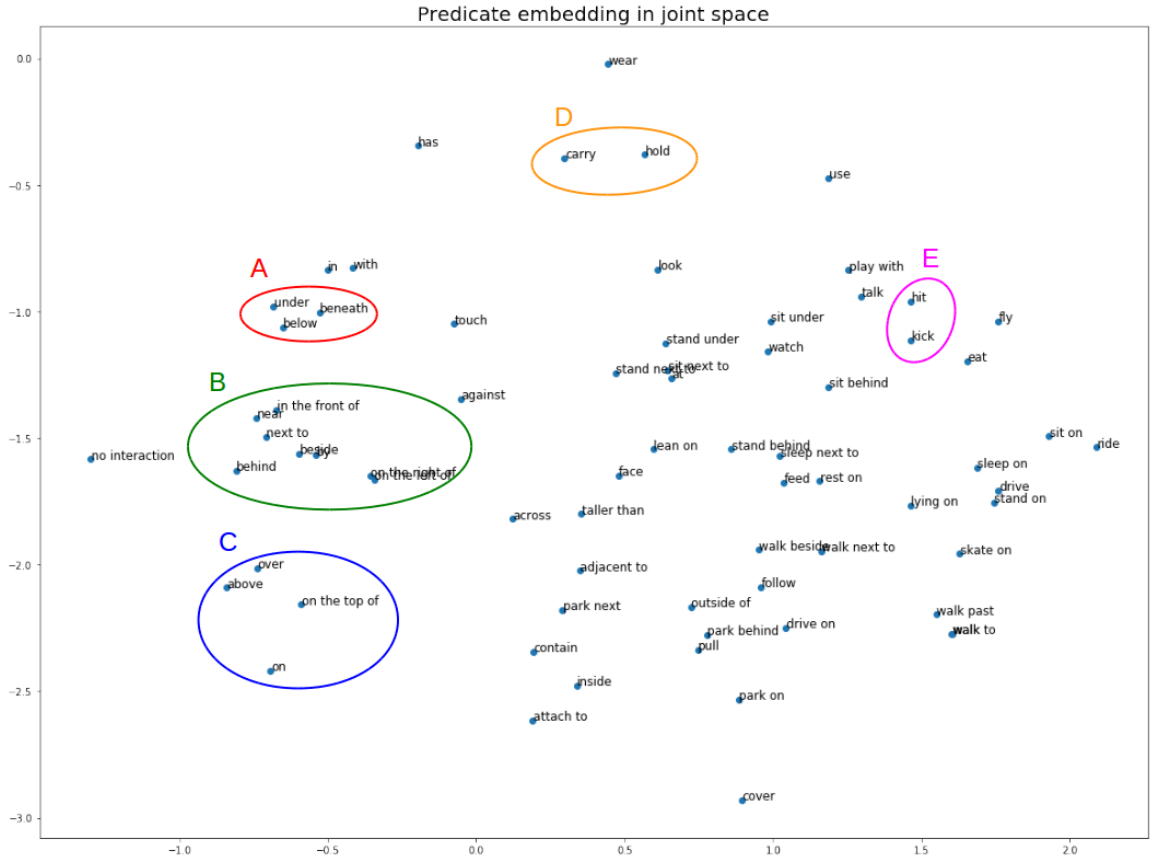


Figure 4-15 – Predicate embedding on the UnRel dataset visualized with T-sne [Van der Maaten and Hinton, 2008]. The predicates are grouped according to their visual and semantic similarity. For example, we highlight regions corresponding to: (A) spatial relations related to "under" (e.g. "below", "beneath"), (B) spatial relations related to "next to" (e.g. "near", "beside"), (C) spatial relations related to "above" (e.g. "over", "on the top of"), or similar actions (D) and (E). Note that it is remarkable that our visual-semantic embedding separates relations such as those in (A) from those in (C) while they are very similar from a strictly semantic point of view (in pre-trained word2vec embeddings). Learning a good embedding for unigrams (here predicates) is crucial in our model that uses transfer by analogy, as unigram embeddings directly influence the analogy transformation from the seen visual phrases to the unseen ones.

Chapter 5

Discussion and perspectives

In this chapter, we summarize our contributions and propose directions for future research.

5.1 Summary of contributions

In Section 1.3, we have highlighted core challenges in the task of visual relation detection. Here, we review the contributions in this manuscript in the light of these different questions:

1. **detecting visual relations with less supervision:** in Chapter 3, we introduce a model based on discriminative clustering that enables us to learn detectors for visual relations using only image-level labels, provided pre-trained detectors for objects are already available. Our experiments show that this results in a small loss in performance compared to a similar model trained with full supervision.

2. **generalizing to unseen relations:** our compositional model in Chapter 3 generalizes well to unseen visual relations, in particular thanks to a powerful visual representation based on individual object appearance and quantized spatial representation of the relation. In Chapter 4, we investigated the question

of transfer of visual phrase embeddings to unseen triplets based on analogical reasoning.

3. **evaluating visual relations without missing annotations:** an important contribution of this thesis is the introduction in Chapter 3 of an evaluation dataset, UnRel, as a new way to measure the performance of visual relation models (1) in a retrieval set-up without missing annotations, (2) on unusual triplets, for testing the generalization capabilities.

4. **variability of visual appearance:** we address this question in Chapter 4 by introducing an hybrid model that unites compositional and visual phrase modules in a single framework. We experimentally show that the visual phrase module complements the compositional branch on both seen and unseen triplets.

5. **variability of language descriptions:** we make two attempts in this direction. First, in Chapter 3 by explicitly modelling a visual relation in terms of its modes of variation of the spatial configuration, we capture the fact that one word can refer to different concepts, especially different spatial configurations. Second, in Chapter 4, we learn visual relations in a joint visual-language embedding where different triplets can be mapped to close locations if they describe similar entities. Also we benefit from semantic similarities encoded in pre-trained word embeddings to identify related concepts.

5.2 Future work

In this section, we propose possible directions for future research. We first suggest some ideas to improve visual relation detectors. Second, we discuss possible ways to use visual relation detectors to solve other problems.

5.2.1 Improving visual relation detection

Fine-grained modeling. One possible way to improve the performance of visual relation detectors is to improve the visual representation. In this thesis, our representation was motivated by the triplet structure, framing a visual relation as two parts: a subject and an object. Thus, we naturally decomposed a relation into these two shareable components. Yet, it might be worth to explore finer-grained decompositions into smaller components that could be more discriminative and helpful for ambiguous cases. For instance, earliest works such as [Yao and Fei-Fei, 2010b; Maji et al., 2011; Delaitre et al., 2011] used human pose and body parts for action recognition. But this tradition of fine-grained modeling of humans in action recognition has slowed down with the success of deep neural networks that provide a powerful encoding of image regions. As models for human pose and body part estimation have improved, can we now exploit them to improve human-object interaction? Or do CNN features already contain all the useful information? It is not clear whether CNN implicitly captures intrinsic spatial relationships between the different parts of an object. For instance, the widely used max-pooling method throws away positional information [Sabour et al., 2017]. An explicit modeling of the arangement between the parts of a visual relation could thus be beneficial. While pose or keypoint information could serve as guidance for extracting human parts as done in Chéron et al. [2015], for general visual relations, one can try to automatically discover the relevant parts. Interesting ideas can be taken from works performing weakly-supervised discovery of parts for scene categorization [Pandey and Lazebnik, 2011; Sharma et al., 2013] as well as fine-grained recognition of objects [Yang et al., 2012; Krause et al., 2015; Simon and Rodner, 2015]. We also feel that a finer-grained modeling of a relation could be especially useful for recognition of unseen relations. Just like decomposing an object in its attributes has helped zero-shot object recognition [Farhadi et al., 2009], identifying sub-parts of a relation looks like a very promising direction to generalize to new unseen combinations.

Other visual cues. In this thesis, we have solely relied on appearance and 2D spatial cues between image regions. However, other types of cues could be advantageously exploited. For instance, depth information would be useful to disambiguate spatial relations such as "in front of" or "behind". Similar to the work of Hoeim et al. [2006] in object detection, one could model contextual relations between objects within the 3D world. Using segmentation masks complementary to object bounding boxes could also help disambiguate subtle occlusions. This could be studied in the set-up proposed by Liu et al. [2018b] who recently introduced a new dataset, Person In Context (PIC), that provides segmentation masks for both subject and object in interaction.

Adaptive object proposals. Our models in Chapters 3 and 4 decompose in two steps: (1) extracting candidate object bounding boxes using a pre-trained object detector, (2) training a visual relation detection on the pairs of extracted objects. Thus, the performance of visual relation detection is upperbounded by the recall of object detection. As we cannot retain all possible pairs of objects due to computational complexity, the recall can be low. Some efforts such as [Zhang et al., 2017c] have been done to develop relation proposal network selecting relevant candidate pairs of objects that are potentially interacting. Yet these proposals, in the same manner as candidate boxes from Region Proposal Network (RPN) in object detectors, are class-agnostic. Instead, an interesting approach to explore could be to sample candidate pairs non-uniformly in image regions depending on the triplet query. For instance, one might use an attentional mechanism as in VQA to roughly determine the regions of interest in the image (conditional on the query), and retain more candidate pairs of objects in these relevant regions. This would provide higher precision without increasing the complexity, that could be extremely useful to improve recall for small or occluded objects.

Learning joint embeddings with image-level labels. In Chapter 4, our analogy model allowed us to transfer embeddings from seen triplets to unseen ones. We have

shown that this model works well to retrieve unseen target triplets, providing that we can sample similar enough source triplets. This means that to be able to transfer successfully, we still need to train models for a large number of triplets. And for this, we need annotations. We thus come back to the problem of Chapter 3 where we proposed to learn visual relation detectors using only image-level annotations. A desirable extension would be to marry Chapters 3 and 4, i.e. to learn joint visual-semantic embeddings in a weakly-supervised manner. This would allow us to leverage a large number of image-level annotations to learn more source triplets. Learning from image-level annotations in the form of text descriptions could also provide us with additional useful information to localize visual relations in images. For instance, in an image description such as "A girl wearing a life vest floats in water", we can extract the triplets $(girl, wearing, life\ vest)$ and $(girl, floats\ in, water)$ but we also have the additional information that the subject ("girl") in the two extracted triplets refers to the same entity. This provides additional constraints at the level of images that could be exploited to better localize visual relations.

Active learning. Another way to alleviate the problem of annotation is to concentrate the annotation budget on the visual relations that are most informative. This is the framework of active learning. In this set-up, the central question is to identify the visual relations that are worthy of this extra labeling effort. For instance, we do not want to annotate additional examples of "person ride horse" if the corresponding embeddings are already well learnt. Instead, we wish to expand embeddings in direction that we do not know well. To identify the relations where to spend the annotation effort, it could be interesting to explore an analogy model as in Chapter 4. One possible way to proceed to extend visual relation embeddings to a larger number of triplets could be as follows: (1) initialize the embeddings with our model in Chapter 4 on seed annotated data, (2) sample an unseen triplet that is close to already seen triplets, (3) use a web search engine to query images for the unseen

triplet, (4) use transfer by analogy as in Chapter 4 to rank the web images, (5) give the badly ranked images to a human annotator, (6) update the embeddings with the new annotations and add the query to the set of seen triplets. Repeat the process.

5.2.2 Beyond visual relations

Here, we review some tasks towards which our work on visual relation detection could be extended.

Using visual relations to improve image retrieval. One natural and important extension of our work goes in the direction of image retrieval with natural language queries. Current state-of-the-art pipelines [Lee et al., 2018] for image retrieval still do not model relations between objects. Yet, visual relations carry a lot of useful semantic information. In Figure 5-1, we display image retrieval results by a top web search engine for different types of language queries. Our observation is that while retrieval of short text works well in general (except for rare situations), the performance on longer text queries rapidly degrades. A future promising line of research is to learn models that can combine these well learnt short chunks into bigger ones. An extension of our work towards this direction would proceed in three steps:

1. *decomposing a textual query into short language chunks*: in Chapters 3 and 4, we have assumed that language triplets were provided, but in reality we will have to process totally unstructured text. We thus need a method to break down long textual queries into shorter chunks that we can visually ground. For visual side, there is a Region Proposal Network (RPN) that proposes candidate regions of interest. We can imagine a similar mechanism for language side, i.e. a language chunks proposals that would pre-select interesting groups of words in the sentence. One way to do this is through dependency parser [Manning et al., 2014] from which different grammatical structures can be extracted. Such methods can provide a good starting point but will soon suffer from the dis-

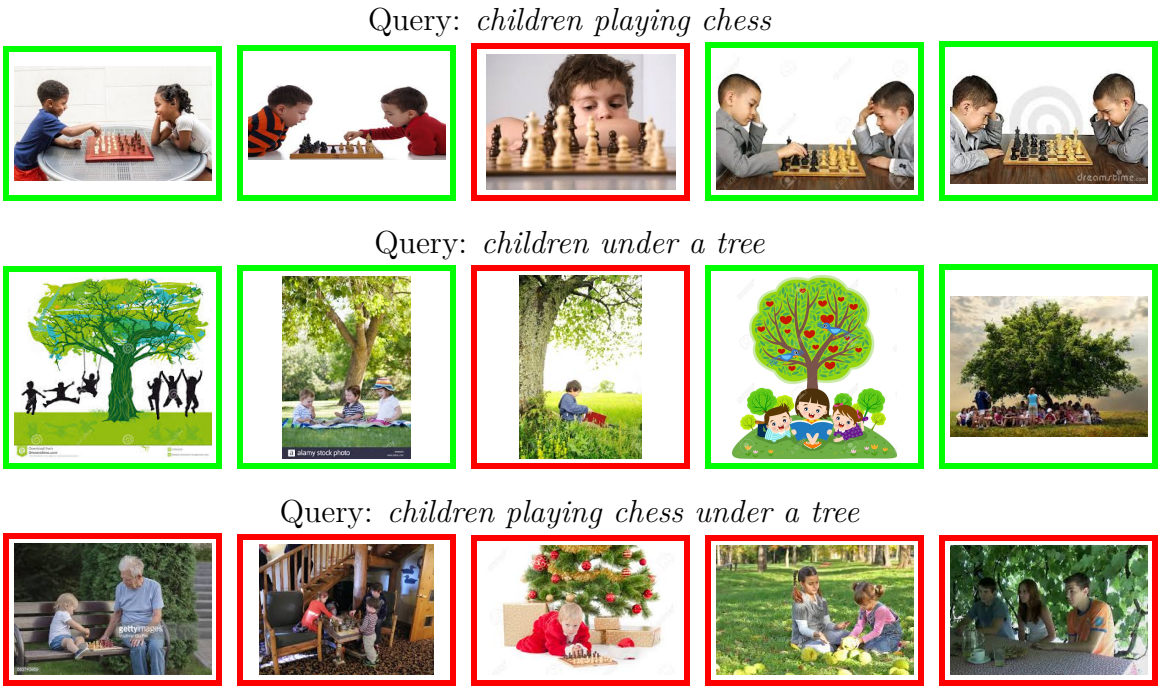


Figure 5-1 – Top retrieved images of a commercial image search engine for natural language queries of different complexity

tribution gap between the corpus where the parser was pre-trained, and the image-sentence dataset on which it will be applied. Another interesting direction would be to use an attentional mechanism to identify candidate groups of words in language that could be visually grounded. Hu et al. [2017] use attention to extract a triplet of the form $(subject, predicate, object)$ from sentences. To extract multiple triplets, or different types of chunks, an idea could be to adapt the work of Lin et al. [2017b] who represent a sentence by a 2-D matrix where each row corresponds to a different part of the sentence.

2. *grounding short textual chunks in images*: the second extension would generalize our work on triplets to handle other types of short textual chunks such as attributes. Our model in Chapter 4 can be easily extended to bigrams. One possibility could be to add another branch to our model that would learn joint embedding space between image regions and short unstructured text in the man-

ner of DenseCap [Johnson et al., 2016]. There is some recent relevant literature on this that addresses the task of open-vocabulary phrase detection [Hinami and Satoh, 2018; Plummer et al., 2019]. The major challenge here is to develop a model that can generalize well to unseen chunks of text. In particular, it would be interesting to see how to generalize our analogy model to unstructured chunks.

3. *aggregating local predictions*: the last step concerns the aggregation of these short textual chunks into the final prediction for the full query. In particular, a desirable aggregation operator should be able to reason at different levels of granularities (unigram, trigram, whole sentence), e.g. by enforcing constraints between chunks at different levels of hierarchy.

We detailed this process for image retrieval, but other related tasks such as visual question answering or image captioning could also benefit from visual relations. In general, incorporating visual relations could be useful when analyzing complex or unusual scenes, where adding more structure could help compensating the lack of training data.

Extending UnRel dataset. An important element to discuss towards the extension of visual relations to scene understanding is the dataset question. In this thesis, we have been especially interested in unseen visual relations as they constitute one of the main challenges in visual relation detection. Similarly, one of the main challenges in scene understanding is to generalize towards unseen situations. If the dataset is not specifically built to evaluate unseen situations, then frequent seen situations dominate at test time, and the top performing methods in terms of numerical results are not necessarily the ones which generalize better. Thus, we believe that efforts should be first made to develop datasets and/or evaluation protocols that target unseen situations. In Chapter 3, we introduced UnRel, an evaluation dataset made of unusual relations. It could be advantageous to extend UnRel with language descriptions. One

interesting feature of this dataset would be that both triplets and textual descriptions would be available, enabling to evaluate both scene understanding (either cast as captioning, image retrieval or a VQA problem) and visual grounding. An extension of UnRel to include never-seen objects (for instance using images and annotations from Open Images dataset [Kuznetsova et al., 2018]) could be of interest.

Visual relations in videos. Our manuscript has addressed the task of detecting visual relations in static images. While already interesting for many applications, there is an inherent limitation that some relations need temporal extent do be disambiguated. For instance, it is difficult to know whether one person is "opening" or "closing" something from a single frame. In videos, most works have been interested in spatio-temporal localization of human actions. The works that try to jointly model the action and the object in interaction are still rare [Kato et al., 2018; Kalogeiton et al., 2017]. Recently, Shang et al. [2017] introduced a Video Visual Relation Detection (VidVRD) dataset containing 1000 videos labeled at box-level with 35 object categories and 132 predicate categories for a total of 3219 different triplets, 258 of them being unseen in training. This task thus inherits the problem of long-tail distribution that we have addressed in Chapter 4. One interesting extension would be to marry our model with action recognition pipelines in videos such as [Chéron et al., 2018].

Image generation. Other applications of visual relations involve image generation conditional on langage query. There have been many improvements in this domain in particular thanks to Generative Adversarial Networks [Goodfellow et al., 2014; Reed et al., 2016b,a]. Recently, Johnson et al. [2018] have tried to further condition image generation on an input scene graph, encoding the objects in the image and the relationships between them. Also the work of Zhu et al. [2017] on image-to-image translation has enabled significant improvement in style transfer but fails when presented with unusual situations such as in Figure 5-2. Image generation or editing

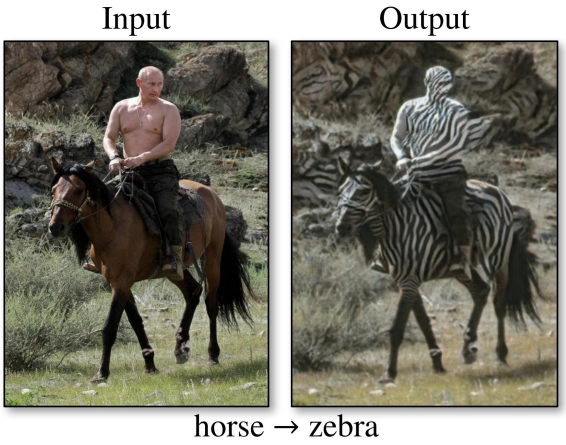


Figure 5-2 – Style transfer from horse → zebra by Zhu et al. [2017]: the transfer also applies on the "person" and not just on the "horse" as there was no examples of a person riding a horse or zebra in training. Training with images of "person riding horse" and re-using our analogy reasoning in Chapter 4 could be helpful to successfully transfer to this unusual case.

is an interesting field of research. In Chapter 4 we applied analogical reasoning for retrieval. A potential extension could be to use similar analogy to generate images of unseen visual relations. For instance, given images of "person ride horse", "horse" and "cow" at training, can we synthesize an image of "person ride cow"? This task could be an intermediate step towards the generation of complex images corresponding to unseen situations.

Language disambiguation. Visual relations have potential applications in natural language processing, for instance in sentence disambiguation. Current syntactic parsers indeed show important weaknesses in these two following situations: (1) *coreference resolution* is the task of finding all the expressions that refer to same entity. For instance, in the sentence: "A table is in the room. Next to it is a chair.", one should understand that the pronoun "it" refers to the "table", (2) *prepositional-phrase attachment resolution* is the problem of finding the phrase a preposition is attached to. For instance, in the sentence "I shot an elephant in my pajamas", one should understand that "I" is wearing the pajamas, not the "elephant". There has been

some research on these problems that tried to jointly reason over visual and language modalities [Kong et al., 2014; Christie et al., 2016; Kottur et al., 2018], but these are still open research problems. Any improvement in this field would benefit both language and vision communities.

Annex A

Additional experiments

In this chapter, we provide additional results on our weakly-supervised model described in Chapter 3.

A.1 Evaluation on Visual Genome Dataset

Here we show results for the new challenging dataset of [Krishna et al., 2016] that depicts complex scenes (21 objects and 17 relationships per image on average). Table A.1 compares the accuracy of our method with the method of [Krishna et al., 2016]. Since not all details are given in the paper, we have reproduced the experimental setting as well as we could based on a communication with the authors. In particular, we keep a vocabulary corresponding to the 100 most frequent relations and nouns that occur at least 50 times in one of these relations (we end up with 2618 nouns). We use the

Relationship	Top-1	Top-5
Krishna et al. [2016] full	8.7	26.6
Ours [S+A] full	**46.5**	**76.4**
Ours [S+A] weak	35.5	70.1

Table A.1 – Results on the Visual Genome dataset [Krishna et al., 2016] for the Relationship recognition task.

training/validation/test splits given by [Johnson et al., 2016] and retain around 27K unique triplets for training and validation as in [Krishna et al., 2016] while testing on the whole test split. We compare with the fully-supervised baseline experiment *Relationship* in [Krishna et al., 2016] which trains a VGG16 network to predict the predicate for a pair of boxes given the appearance of the union of the boxes. We train our model described in Section 3.3.2 first with full supervision (Ours [S+A] full) then with weak supervision (Ours [S+A] weak). Our appearance features are extracted from a VGG16 network trained on COCO (we do not fine-tune). For the weak supervision we use ground truth object boxes to form the candidate pairs of boxes in the image. This would correspond to the case where a perfect object detector is given and we only have to disambiguate the correct configuration. The evaluation measure checks the per-instance accuracy to predict the right predicate for each pair of boxes.

A.2 Varying evaluation parameters

In this part, we want to test the robustness of our model when varying parameters in evaluation measures such as the number of candidate boxes retained in recall and the IoU threshold to evaluate localization.

R@100 on Visual Relationship Dataset [Lu et al., 2016a]. Complementary to results with R@50 provided in Table 3.1, we show results with R@100 in Table A.2. Similar to previous observations, our method outperforms [Lu et al., 2016a; Sadeghi and Farhadi, 2011], in particular on the zero-shot split. Also, the relatively high performance of appearance features alone (g.), which can incorporate only limited context around objects, and the language model only (e.), which ignores image information, reveals a bias in this dataset: knowing object categories already provides a strong prior on the set of possible relations. This emphasizes the value of our UnRel dataset which is created to remove such bias by considering unusual relations among objects.

	Predicate Det.		Phrase Det.		Relationship Det.	
	All	Unseen	All	Unseen	All	Unseen
Full sup.						
a. Visual Phrases [Sadeghi and Farhadi, 2011]	1.9	-	0.07	-	-	-
b. Visual [Lu et al., 2016a]	7.1	3.5	2.6	1.1	1.8	0.7
c. Language (likelihood) [Lu et al., 2016a]	18.2	5.1	0.08	0.01	0.08	0.01
d. Visual + Language [Lu et al., 2016a]	47.8	8.4	17.0	3.7	14.7	3.5
e. Language (full) [Lu et al., 2016a]	48.4	12.9	17.3	5.5	15.3	5.1
f. Ours [S]	42.2	22.2	15.0	8.7	13.5	8.1
g. Ours [A]	46.3	16.1	16.4	6.0	14.4	5.4
h. Ours [S+A]	50.4	**23.6**	18.1	**8.7**	16.1	**8.2**
i. Ours [S+A] + Language [Lu et al., 2016a]	**52.6**	21.6	**19.5**	7.8	**17.1**	7.4
Weak sup.						
j. Ours [S+A]	46.8	19.0	17.4	7.4	15.3	7.1
k. Ours [S+A] - Noisy	46.4	17.6	16.9	6.7	15.0	6.4

Table A.2 – Results for Visual Relationship Detection on the dataset of [Lu et al., 2016a] for R@100.

Retrieval on UnRel with IoU=0.5. In addition to results on the UnRel dataset presented in Table 3.2 for IoU=0.3, Table A.3 shows UnRel results for IoU=0.5. Our results show similar patterns for both IoU thresholds.

	With GT	With candidates		
	-	union	subj	subj/obj
Chance	38.4	6.7	4.9	2.8
Full sup.				
DenseCap [Johnson et al., 2016]	-	2.9	2.3	-
Reproduce [Lu et al., 2016a]	50.6	10.4	7.8	5.1
Ours [S+A]	**62.6**	**11.8**	**9.2**	**6.7**
Weak sup.				
Ours [S+A]	58.5	11.2	8.5	5.9
Ours [S+A] - Noisy	55.5	11.0	8.2	5.7

Table A.3 – Retrieval on UnRel (mAP) with IoU=0.5

		Predicate Det.		Phrase Det.		Relationship Det.	
		All	Unseen	All	Unseen	All	Unseen
	Recall@50						
b.	Visual	7.2	5.4	2.3	1.5	1.6	1.0
e.	Visual + Language	48.7	12.9	16.5	5.1	14.1	4.8
	Recall@100						
b.	Visual	7.2	5.4	2.7	1.7	1.9	1.2
e.	Visual + Language	48.7	12.9	17.3	5.7	15.0	5.4

Table A.4 – Results on the Visual Relationship Detection dataset recomputed for the approach of [Lu et al., 2016a]. Despite using the evaluation code of [Lu et al., 2016a] we have obtained slightly higher results than they report.

A.3 Reproducing results of [Lu et al., 2016a]

When reproducing results of [Lu et al., 2016a] for Visual Relationship Detection task using their evaluation code we obtained slightly higher performance than reported in [Lu et al., 2016a]. Hence we report the full set of obtained results in Table A.4.

Bibliography

Agrawal, A., Batra, D., and Parikh, D. Analyzing the Behavior of Visual Question Answering Models. In *EMNLP*, 2016. 36

Agrawal, A., Kembhavi, A., Batra, D., and Parikh, D. C-VQA: A Compositional Split of the Visual Question Answering (VQA) v1.0 Dataset. *arXiv:1704.08243*, 2017. 37

Agrawal, H., Desai, K., Chen, X., Jain, R., Batra, D., Parikh, D., Lee, S., and Anderson, P. Nocaps: Novel Object Captioning at Scale. *arXiv:1812.08658*, 2018. 37

Alayrac, J.-B., Bojanowski, P., Agrawal, N., Laptev, I., Sivic, J., and Lacoste-Julien, S. Unsupervised learning from Narrated Instruction Videos. In *CVPR*, 2016. 54

Alexe, B., Deselaers, T., and Ferrari, V. Measuring the Objectness of Image Windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 21

Anderson, P., Fernando, B., Johnson, M., and Gould, S. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*, 2016. 30, 36

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*, 2018. 33, 36, 52

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. Deep Compositional Question Answering with Neural Module Networks. In *CVPR*, 2016a. 35, 67

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. Learning to Compose Neural Networks for Question Answering. In *HLT-NAACL*, 2016b. 35

Andrew, G., Arora, R., Bilmes, J. A., and Livescu, K. Deep Canonical Correlation Analysis. In *ICML*, 2013. 31

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: Visual Question Answering. In *ICCV*, 2015. 26, 27

Atzmon, Y., Berant, J., Kezami, V., Globerson, A., and Chechik, G. Learning to generalize to new compositions in image understanding. *arXiv:1608.07639*, 2016. 95

Aytar, Y. and Zisserman, A. Tabula Rasa: Model Transfer for Object Category Detection. In *ICCV*, 2011. 96

Bach, F. R. and Harchaoui, Z. Diffrac: a discriminative and flexible framework for clustering. In *NIPS*, 2007. 16, 54, 69, 73

Bahdanau, D., Cho, K., and Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*, 2015. 31

Banerjee, S. and Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005. 36

Bansal, T., Neelakantan, A., and McCallum, A. Relnet: End-to-end modeling of entities & relations. *arXiv:1706.07179*, 2017. 94

Barnes, C., Zhang, F.-L., Lou, L.-m., Wu, X., and Hu, S. Patchtable: efficient patch queries for large datasets and applications. *ACM Trans. Graph.*, 2015. 62

Battaglia, P. W., Pascanu, R., Lai, M., Rezende, D. J., and Kavukcuoglu, K. Interaction networks for learning about objects, relations and physics. In *NIPS*, 2016. 35, 94

Ben-younes, H., Cadène, R., Cord, M., and Thome, N. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *ICCV*, 2017. 31

Ben-younes, H., Cadene, R., Thome, N., and Cord, M. BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection. In *AAAI*, 2019. 31

Bénard, P., Cole, F., Kass, M., Mordatch, I., Hegarty, J., Senn, M. S., Fleischer, K. W., Pesare, D., and Breeden, K. Stylizing animation by example. *ACM Trans. Graph.*, 2013. 62

Bilen, H. and Vedaldi, A. Weakly Supervised Deep Detection Networks. In *CVPR*, 2016. 53, 68

Bilen, H., Pedersoli, M., and Tuytelaars, T. Weakly supervised object detection with convex clustering. In *CVPR*, 2015. 53

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. Actions as space-time shapes. In *ICCV*, 2005. 39

Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., and Sivic, J. Finding actors and actions in movies. In *ICCV*, 2013. 54

Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., and Sivic, J. Weakly supervised action labeling in videos under ordering constraints. In *ECCV*, 2014. 54, 69

Bojanowski, P., Lajugie, R., Grave, E., Bach, F., Laptev, I., Ponce, J., and Schmid, C. Weakly-supervised alignment of video with text. In *ICCV*, 2015. 54, 87

Bordes, A., Glorot, X., Weston, J., and Bengio, Y. A Semantic Matching Energy Function for Learning with Multi-Relational Data. *Machine Learning*, 2013. 49, 50

Bowers, J. S. On the Biological Plausibility of Grandmother Cells: Implications for Neural Network Theories in Psychology and Neuroscience. *Psychological Review*, 2009. 7

Cadene, R., Ben-Younes, H., Thome, N., and Cord, M. MUREL: Multimodal Relational Reasoning for Visual Question Answering. In *CVPR*, 2019. 35

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *CVPR*, 2017. 42

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., and Mitchell, T. M. Toward an Architecture for Never-Ending Language Learning. In *AAAI*, 2010. 46

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. Deep Clustering for Unsupervised Learning of Visual Features. In *ECCV*, 2018. 87

Chang, A., Monroe, W., Savva, M., Potts, C., and Manning, C. D. Text to 3D Scene Generation with Rich Lexical Grounding. In *ACL*, 2015. 35, 67

Chang, A. X., Savva, M., and Manning, C. D. Learning Spatial Knowledge for Text to 3D Scene Generation. In *EMNLP*, 2014. 35

Chao, Y.-W., Wang, Z., He, Y., Wang, J., and Deng, J. HICO: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 40, 42, 95, 106

Chao, Y.-W., Liu, Y., Liu, X., Zeng, H., and Deng, J. Learning to Detect Human-Object Interactions. In *WACV*, 2018. 40, 42, 43, 104, 106, 107, 110

Chen, X. and Gupta, A. Webly Supervised Learning of Convolutional Networks. In *ICCV*, 2015. 54

Chen, X., Shrivastava, A., and Gupta, A. NEIL: Extracting Visual Knowledge from Web Data. In *ICCV*, 2013. 28, 47, 55, 58, 67

Cheng, L., Vishwanathan, S. V. N., and Zhang, X. Consistent Image Analogies using Semi-supervised Learning. In *CVPR*, 2008. 62

Chéron, G., Laptev, I., and Schmid, C. P-CNN: Pose-based CNN Features for Action Recognition. In *ICCV*, 2015. 131

Chéron, G., Alayrac, J.-B., Laptev, I., and Schmid, C. A flexible model for training action localization with varying levels of supervision. In *NIPS*, 2018. 137

Cho, M., Kwak, S., Schmid, C., and Ponce, J. Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals. In *CVPR*, 2015. 55

Choi, M. J., Torralba, A., and Willsky, A. S. Context models and out-of-context objects. *Pattern Recognition Letters*, 2012. 68, 69

Christie, G., Laddha, A., Agrawal, A., Antol, S., Goyal, Y., Kochersberger, K., and Batra, D. Resolving Language and Vision Ambiguities Together: Joint Segmentation & Prepositional Attachment Resolution in Captioned Scenes. In *EMNLP*, 2016. 139

Cinbis, R. G., Verbeek, J. J., and Schmid, C. Weakly Supervised Object Localization with Multi-Fold Multiple Instance Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 53

Cortes, C. and Vapnik, V. Support-Vector Networks. *Machine Learning*, 1995. 21

Dai, B., Zhang, Y., and Lin, D. Detecting Visual Relationships with Deep Relational Networks. In *CVPR*, 2017. 42, 43, 44, 45, 49, 92, 95

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 21

Das, A., Agrawal, H., Zitnick, L., Parikh, D., and Batra, D. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *CVIU*, 2017a. 37

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J., Parikh, D., and Batra, D. Visual Dialog. In *CVPR*, 2017b. 27, 28

Delaitre, V., Sivic, J., and Laptev, I. Learning person-object interactions for action recognition in still images. In *NIPS*, 2011. 41, 67, 131

Delaitre, V., Laptev, I., and Sivic, J. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010. 41, 43, 44, 45

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 42, 54

Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, C. Y., Neven, H., and Adam, H. Large-Scale Object Classification Using Label Relation Graphs. In *ECCV*, 2014. 57, 124

Desai, C. and Ramanan, D. Detecting Actions, Poses, and Objects with Relational Phraselets. In *ECCV*, 2012. 41

Desai, C., Ramanan, D., and Fowlkes, C. Discriminative models for static human-object interactions. In *(SMiCV) CVPR Workshops*, 2010. 39, 43, 44, 67

Desai, C., Ramanan, D., and Fowlkes, C. C. Discriminative Models for Multi-Class Object Layout. In *International Journal of Computer Vision*, 2011. 23

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. Solving the Multiple Instance Problem with Axis-Parallel Rectangles. *Artificial Intelligence*, 1997. 53

Dinu, G. and Baroni, M. How to make words with vectors: Phrase generation in distributional semantics. In *ACL*, 2014. 51

Divvala, S. K., Farhadi, A., and Guestrin, C. Learning Everything about Anything: Webly-Supervised Visual Concept Learning. In *CVPR*, 2014. 50, 54, 95

Elhoseiny, M., Cohen, S., Chang, W., Price, B., and Elgammal, A. Sherlock: Scalable fact learning in images. In *AAAI*, 2016. 49, 67, 68

Elliott, D. and Keller, F. Image Description using Visual Dependency Representations. In *EMNLP*, 2013. 31

Engilberge, M., Chevallier, L., Pérez, P., and Cord, M. Finding Beans in Burgers: Deep Semantic-Visual Embedding with Localization. In *CVPR*, 2018. 37

Faghri, F., Fleet, D. J., Kiros, J., and Fidler, S. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *BMVC*, 2017. 31

Faktor, A. and Irani, M. "Clustering by Composition" - Unsupervised Discovery of Image Categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 55

Fang, H., Gupta, S., Iandola, F. N., Srivasta, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., Zitnick, C. L., and Zweig, G. From Captions to Visual Concepts and Back. In *CVPR*, 2015. 32, 68

Farhadi, A., Endres, I., and Hoiem, D. Attribute-Centric Recognition for Cross-category Generalization. 56

Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. A. Describing Objects by their Attributes. In *CVPR*, 2009. 56, 131

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. Every picture tells a story: Generating sentences from images. In *ECCV*, 2010. 26

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. A., and Ramanan, D. Object Detection with Discriminatively Trained Part Based Models. 21

Freund, Y. and Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 1997. 21

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., and Mikolov, T. DeViSE: A Deep Visual-Semantic Embedding Model. In *NIPS*, 2013. 59, 69

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *arXiv:1606.01847*, 2016. 31

Galleguillos, C., Rabinovich, A., and Belongie, S. Object Categorization using Co-occurrence, Location and Appearance. In *CVPR*, 2008. 23, 68

Gao, C., Zou, Y., and Huang, J.-B. ICAN: Instance-Centric Attention Network for Human-Object Interaction Detection. In *BMVC*, 2018. 42, 95, 110, 111, 116

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., and Xu, W. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question. In *NIPS*, 2015. 27

Geman, D., Geman, S., Hallonquist, N., and Younes, L. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences of the United States of America*, 2015. 27

Gentner, D. Structure Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 1983. 60

Gentner, D. *Analogical Reasoning, Psychology Of*. 2003. 60, 61

Gentner, D. and Smith, L. A. Analogical Reasoning. *Encyclopedia of Human Behavior*, 2012. 61

Girshick, R. Fast R-CNN. In *ICCV*, 2015. 22, 70, 71, 75, 82

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 22, 77

Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., and He, K. Detectron. https://github.com/facebookresearch/detectron, 2018. 108

Gkioxari, G., Girshick, R. B., and Malik, J. Contextual Action Recognition with R*CNN. In *ICCV*, 2015. 53

Gkioxari, G., Girshick, R., and He, K. Detecting and Recognizing Human-Object Interactions. 2018. 42, 95, 108, 110, 116

Gong, Y., Ke, Q., Isard, M., and Lazebnik, S. A multi-view embedding space for modeling internet images, tags, and their semantics. 2014. 31

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative Adversarial Nets. In *NIPS*, 2014. 137

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*, 2017. 37

Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., and Ling, M. Scene graph generation with external knowledge and image reconstruction. In *CVPR*, 2019. 47

Güler, R. A., Neverova, N., and Kokkinos, I. DensePose: Dense Human Pose Estimation in the Wild. In *CVPR*, 2018. 42

Guo, G. and Lai, A. A survey on still image based human action recognition. *Pattern Recognition*, 2014. 39

Gupta, A. and Davis, L. S. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008. 67

Gupta, A., Kembhavi, A., and Davis, L. S. Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009. 39, 45, 67

Gupta, S. and Malik, J. Visual Role Semantic Labeling. *arXiv:1505.04474*, 2015. 40, 110

Hardoon, D. R., Szedmák, S., and Shawe-Taylor, J. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*, 2004. 31

Harel, A., Kravitz, D. J., and Baker, C. I. Deconstructing Visual Scenes in Cortex: Gradients of Object and Spatial Layout Information. *Cerebral cortex*, 2013. 7

Harzallah, H., Jurie, F., and Schmid, C. Combining efficient object localization and image classification. In *ICCV*, 2009. 21

He, K., Zhang, X., Ren, S., and Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 22

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. Mask R-CNN. In *ICCV*, 2017. 42, 43

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. Generating Visual Explanations. In *ECCV*, 2016a. 37

Hendricks, L. A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., and Darrell, T. Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data. In *CVPR*, 2016b. 37, 69

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., and Rohrbach, A. Women Also Snowboard: Overcoming Bias in Captioning Models. In *ECCV*, 2018. 36

Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B., and Salesin, D. Image analogies. In *SIGGRAPH*, 2001. 61

Hinami, R. and Satoh, S. Discriminative Learning of Open-Vocabulary Object Retrieval and Localization by Negative Phrase Augmentation. In *EMNLP*, 2018. 136

Hodosh, M., Young, P., and Hockenmaier, J. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. In *Journal of Artificial Intelligence Research*, 2013. 26

Hoeim, D., Efros, A. A., and Hebert, M. Putting Objects in Perspective. In *CVPR*, 2006. 23, 43, 132

Holyoak, K. The Pragmatics of Analogical Transfer. *The Psychology of Learning and Motivation*, 1985. 60

Hu, H., Gu, J., Zhang, Z., Dai, J., and Wei, Y. Relation Networks for Object Detection. In *CVPR*, 2018. 24

Hu, H., Misra, I., and van der Maaten, L. Binary Image Selection (BISON): Interpretable Evaluation of Visual Grounding. *arXiv:1901.06595*, 2019. 37

Hu, J., Zheng, W.-S., Lai, J., Gong, S., and Xiang, T. Recognising Human-Object Interaction via Exemplar Based Modelling. In *ICCV*, 2013. 48

Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., and Darrell, T. Natural Language Object Retrieval. In *CVPR*, 2016. 29, 67

Hu, R., Rohrbach, M., Andreas, J., Darrell, T., and Saenko, K. Modeling Relationships in Referential Expressions with Compositional Modular Networks. In *CVPR*, 2017. 34, 53, 135

Hudson, D. A. and Manning, C. D. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *CVPR*, 2019. 37

Hung, Z.-S., Mallya, A., and Lazebnik, S. Union Visual Translation Embedding for Visual Relationship Detection and Scene Graph Generation. *arXiv:1905.11624*, 2019. 50

Hwang, S. J., Ravi, S. N., Tao, Z., Kim, H. J., Collins, M. D., and Singh, V. Tensorize, Factorize and Regularize: Robust Visual Relationship Learning. In *CVPR*, 2018. 50, 96

Hwang, S. J., Grauman, K., and Sha, F. Analogy-preserving Semantic Embedding for Visual Object Categorization. In *ICML*, 2013. 62

Ikizler, N., Cinbis, R. G., Pehlivan, S., and Duygulu, P. Recognizing actions from still images. In *ICPR*, 2008. 39, 41

Itti, L., Koch, C., and Niebur, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998. 32

Izadinia, H., Sadeghi, F., Divvala, S. K., Choi, Y., and Farhadi, A. Segment-phrase table for semantic segmentation, visual entailment and paraphrasing. In *ICCV*, 2015. 95

Jabri, A., Joulin, A., and van der Maaten, L. Revisiting Visual Question Answering Baselines. In *ECCV*, 2016. 36

Jenatton, R., Roux, N. L., Bordes, A., and Obozinski, G. R. A latent factor model for highly multi-relational data. In *NIPS*, 2012. 49, 67, 94

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. Image Retrieval using Scene Graphs. In *CVPR*, 2015. 29, 30, 35, 67, 68, 95

Johnson, J., Karpathy, A., and Fei-Fei, L. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *CVPR*, 2016. 28, 29, 67, 75, 82, 83, 113, 124, 136, 141, 142

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. B. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017a. 35

Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Zitnick, C. L., and Girshick, R. B. Inferring and Executing Programs for Visual Reasoning. In *ICCV*, 2017b. 35, 36

Johnson, J., Gupta, A., and Fei-Fei, L. Image Generation from Scene Graphs. In *CVPR*, 2018. 30, 137

Joulin, A., Bach, F., and Ponce, J. Discriminative clustering for image co-segmentation. In *CVPR*, 2010. 54

Joulin, A., Tang, K., and Fei-Fei, L. Efficient image and video co-localization with frank-wolfe algorithm. In *ECCV*, 2014. 69

Kaiser, D., Stein, T., and Peelen, M. V. Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 2014. 6

Kalogeiton, V., Weinzaepfel, P., Ferrari, V., and Schmid, C. Joint Learning of Object and Action Detectors. In *ICCV*, 2017. 137

Kantorov, V., Oquab, M., Cho, M., and Laptev, I. ContextLocNet: Context-Aware Deep Network Models for Weakly Supervised Localization. In *ECCV*, 2016. 53

Karpathy, A. and Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*, 2015. 32, 34, 67, 95

Karpathy, A., Joulin, A., and Fei-Fei, L. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *NIPS*, 2014. 32, 67, 95, 99

Kato, K., Li, Y., and Gupta, A. Compositional Learning for Human Object Interaction. In *ECCV*, 2018. 58, 96, 137

Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. L. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*, 2014. 29, 67

Kim, J.-H., Lee, S.-W., Kwak, D.-H., Heo, M.-O., Kim, J., Ha, J.-W., and Zhang, B.-T. Multimodal Residual Learning for Visual QA. In *NIPS*, 2016. 31

Kim, J. G. and Biederman, I. Where Do Objects Become Scenes? *Cerebral Cortex*, 2010. 7

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 109

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2016. 94

Kiros, R., Salakhutdinov, R., and Zemel, R. S. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv:1411.2539*, 2014. 31

Klein, B. E., Lev, G., Sadeh, G., and Wolf, L. Fisher Vectors Derived from Hybrid Gaussian-Laplacian Mixture Models for Image Annotation. In *CVPR*, 2015. 31

Kolesnikov, A., Lampert, C. H., and Ferrari, V. Detecting Visual Relationships Using Box Attention. *arXiv:1807.02136*, 2018. 25

Kong, C., Lin, D., Bansal, M., Urtasun, R., and Fidler, S. What Are You Talking About? Text-to-Image Coreference. In *CVPR*, 2014. 35, 139

Kottur, S., Moura, J. M. F., Parikh, D., Batra, D., and Rohrbach, M. Visual Coreference Resolution in Visual Dialog using Neural Module Networks. In *ECCV*, 2018. 139

Krause, J., Jin, H., Yang, J., and Fei-Fei, L. Fine-grained recognition without part annotations. In *CVPR*, 2015. 131

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 2016. 27, 36, 37, 40, 42, 65, 67, 75, 83, 95, 140, 141

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012. 22

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. Babytalk: Understanding and generating simple image descriptions. In *CVPR*, 2011. 26, 31

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., and Ferrari, V. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018. 41, 124, 137

Lampert, C. H., Nickisch, H., and Harmeling, S. Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer. In *CVPR*, 2009. 56, 57

Lan, T., Raptis, M., Sigal, L., and Mori, G. From Subcategories to Visual Composites: A Multi-Level Framework for Object Detection. 2013. 6

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. Learning realistic human actions from movies. In *CVPR*, 2008. 39

Lazaridou, A., Bruni, E., and Baroni, M. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In *ACL*, 2014. 59, 69

Lee, K.-H., Chen, X., Hua, G., Hu, H., and He, X. Stacked Cross Attention for Image-Text Matching. In *ECCV*, 2018. 33, 52, 134

Lehnert, W. G. A conceptual theory of question answering. *International Joint Conferences on Artificial Intelligence Organization*, 1977. 27

Li, C., Parikh, D., and Chen, T. Extracting Adaptive Contextual Cues from Unlabeled Regions. In *ICCV*, 2011a. 10

Li, C., Parikh, D., and Chen, T. Automatic Discovery of Groups of Objects for Scene Understanding. In *CVPR*, 2012. 23, 68

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., and Choi, Y. Composing Simple Image Descriptions using Web-scale N-grams. In *CoNLL*, 2011b. 31

Li, Y., Ouyang, W., and Wang, X. ViP-CNN: A visual Phrase Reasoning Convolutional Neural Network for Visual Relationship Detection. In *CVPR*, 2017a. 42, 46, 49, 92, 95

Li, Y., Ouyang, W., Zhou, B., Wand, K., and Wang, X. Scene Graph Generation from Objects, Phrases and Region Captions. In *ICCV*, 2017b. 45

Liang, X., Lee, L., and Xing, E. P. Deep Variation-Structured Reinforcement Learning for Visual Relationship and Attribute Detection. In *CVPR*, 2017. 45

Liao, J., Yao, Y., Yuan, L., Hua, G., and Kang, S. B. Visual Attribute Transfer through Deep Image Analogy. *ACM Transactions on Graphics*, 2017a. 62

Liao, W., Lin, S., Rosenhahn, B., and Yang, M. Y. Natural Language Guided Visual Relationship Detection. *arXiv:1711.06032*, 2017b. 46

Lin, D., Fidler, S., Kong, C., and Urtasun, R. Visual Semantic Search: Retrieving Videos via Complex Textual Queries. In *CVPR*, 2014a. 35

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *ECCV*, 2014b. 75, 106, 107

Lin, T.-Y., Dollár, P., Girshick, R. B., He, K., Hariharan, B., and Belongie, S. J. Feature Pyramid Networks for Object Detection. In *CVPR*, 2017a. 22, 108

Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. A Structured Self-Attentive Sentence Embedding. In *ICLR*, 2017b. 135

Liu, F., Xiang, T., Hospedales, T. M., Yang, W., and Sun, C. iVQA: Inverse Visual Question Answering. In *CVPR*, 2018a. 37

Liu, S., Feng, J., Han, J., Yan, S., Sun, Y., Liao, Y., Ren, L., and Ren, G. PIC - Person In Context. http://picdataset.com/challenge/index/, 2018b. 132

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C.-Y., and Berg, A. C. SSD: Single Shot MultiBox Detector. In *ECCV*, 2016. 22

Lowe, D. G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 2004. 21

Lu, C., Krishna, R., Bernstein, M., and Fei-Fei, L. Visual Relationship Detection with Language Priors. In *ECCV*, 2016a. vii, 24, 39, 40, 42, 45, 48, 64, 65, 68, 70, 75, 76, 77, 78, 79, 80, 82, 83, 86, 87, 89, 90, 92, 95, 107, 113, 141, 142, 143

Lu, J., Yang, J., Batra, D., and Parikh, D. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *NIPS*, 2016b. 32

Lu, J., Yang, J., Batra, D., and Parikh, D. Neural Baby Talk. In *CVPR*, 2018a. 33, 36, 52

Lu, P., Ji, L., Zhang, W., Duan, N., Zhou, M., and Wang, J. R-VQA: Learning Visual Relation Facts with Semantic Attention for Visual Question Answering. In *KDD*, 2018b. 35

Maji, S., Bourdev, L., and Malik, J. Action Recognition from a Distributed Representation of Pose and Appearance. In *CVPR*, 2011. 39, 41, 43, 44, 131

Malisiewicz, T. and Efros, A. A. Beyond Categories: The Visual Memex Model for Reasoning About Object Relationships. In *NIPS*, 2009. 42

Mallya, A. and Lazebnik, S. Learning models for actions and person-object interactions with transfer to question answering. In *ECCV*, 2016. 52

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL*, 2014. 51, 134

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., and Murphy, K. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 29, 34, 67

Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. *arXiv:1906.00067*, 2019. 28

Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. 1982. 25

Miech, A., Alayrac, J.-B., Bojanowski, P., Laptev, I., and Sivic, J. 54, 75

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In *NIPS*, 2013. 46, 59, 99, 101, 109

Miller, G. A. WORDNET: A Lexical Database for English. *Communications of the ACM*, 1992. 57

Misra, I., Gupta, A., and Hebert, M. From Red Wine to Red Tomato: Composition with Context. In *CVPR*, 2017. 50, 96

Mitchell, J. and Lapata, M. Vector-based models of semantic composition. In *ACL*, 2008. 51

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., C. Mensch, A. C., Berg, A. C., Berg, T. L., and Daumé, H. Midge: Generating Image Descriptions From Computer Vision Detections. In *EACL*, 2012. 31

Movshovitz-Attias, D. and Cohen, W. W. KB-LDA: Jointly Learning a Knowledge Base of Hierarchy, Relations, and Facts. In *ACL*, 2015. 46, 67

Nagaraja, V. K., Morariu, V. I., and Davis, L. S. Modeling Context Between Objects for Referring Expression Understanding. In *ECCV*, 2016. 34, 53

Newell, A. and Deng, J. Pixels to Graphs by Associative Embedding. In *NIPS*, 2017. 45

Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE*, 2015. 49

Niebles, J. C., Wang, H., and Fei-Fei, L. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. In *BMVC*, 2006. 39, 55

Norcliffe-Brown, W., Vafeias, E., and Parisot, S. Learning Conditioned Graph Structures for Interpretable Visual Question Answering. In *NIPS*, 2018. 35

Oliva, A. Visual Scene Perception. *Encyclopaedia of Perception*, 2009. 25

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. Is object localization for free? – Weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 53, 68

Ordonez, V., Kulkarni, G., and Berg, T. L. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NIPS*, 2011. 26

Osokin, A., Alayrac, J.-B., Lukasewitz, I., Dokania, P. K., and Lacoste-Julien, S. Minding the Gaps for Block Frank-Wolfe Optimization of Structured SVMs. In *ICML*, 2016. 54, 75

Pandey, M. and Lazebnik, S. Scene Recognition and Weakly Supervised Object Localization with Deformable Part-Based Models. In *ICCV*, 2011. 131

Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., and Rohrbach, M. Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*, 2018. 37

Peyre, J., Laptev, I., Schmid, C., and Sivic, J. Weakly-Supervised Learning of Visual Relations. In *ICCV*, 2017. 17, 92, 95, 106, 107, 108, 113

Peyre, J., Laptev, I., Schmid, C., and Sivic, J. Detecting Unseen Visual Relations Using Analogies. In *ICCV*, 2019. 17

Plesse, F., Ginsca, A., Delezoide, B., and Prêteux, F. J. Visual Relationship Detection Based on Guided Proposals and Semantic Knowledge Distillation. In *ICME*. 46

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *ICCV*, 2015. 29, 67, 95

Plummer, B. A., Mallya, A., Cervantes, C. M., Hockenmaier, J., and Lazebnik, S. Phrase Localization and Visual Relationship Detection with Comprehensive Linguistic Cues. 2017. 29, 43, 46, 49, 95

Plummer, B. A., Shih, K. J., Li, Y., Xu, K., Lazebnik, S., Sclaroff, S., and Saenko, K. Open-vocabulary Phrase Detection. *arXiv:1811.07212*, 2019. 136

Prest, A., Schmid, C., and Ferrari, V. Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011. 43, 48, 53, 67

Qi, S., Wang, W., Jia, B., Shen, J., and Zhu, S.-C. Learning Human-Object Interactions by Graph Parsing Neural Networks. In *ECCV*, 2018. 42, 95, 110

Raja, K., Laptev, I., Pérez, P., and Oisel, L. Joint pose estimation and action recognition in image graphs. In *ICIP*, 2011. 41

Ramanathan, V., Li, C., Deng, J., Han, W., Li, Z., Gu, K., Song, Y., Bengio, S., Rossenberg, C., and Li, F.-F. Learning Semantic Relationships for Better Action Retrieval in Images. In *CVPR*, 2015. 57, 58, 67, 96, 124

Redmon, J. and Farhadi, A. YOLO9000: Better, Faster, Stronger. In *CVPR*, 2017. 23

Redmon, J., Divvala, S. K., Girshick, R. B., and Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, 2016. 22

Reed, S. E., Zhang, Y., Zhang, Y., and Lee, H. Deep Visual Analogy-Making. In *NIPS*, 2015. 17, 62, 96

Reed, S. E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., and Lee, H. Learning What and Where to Draw. In *NIPS*, 2016a. 137

Reed, S. E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. Generative Adversarial Text to Image Synthesis. In *ICML*, 2016b. 137

Ren, M., Kiros, R., and Zemel, R. S. Exploring Models and Data for Image Question Answering. In *NIPS*, 2015a. 27

Ren, S., He, K., Girshick, R., and Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015b. 22, 24, 65, 82, 108

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., and Schiele, B. Grounding of textual phrases in images by reconstruction. 2016. 33, 67

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object Hallucination in Image Captioning. In *EMNLP*, 2018. 36, 37

Ronchi, M. R. and Perona, P. Describing Common Human Visual Actions in Images. In *BMVC*, 2015. 40, 106, 107

Rosenthal, G., Shamir, A., and Sigal, L. Learn How to Choose: Independent Detectors Versus Composite Visual Phrases. In *WACV*, 2017. 49

Ross, B. H. Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1989. 60

Sabour, S., Frosst, N., and Hinton, G. E. Dynamic Routing Between Capsules. In *NIPS*, 2017. 131

Sadeghi, F., Divvala, S. K., and Farhadi, A. VisKE: Visual Knowledge Extraction and Question Answering by Visual Verification of Relation Phrases. In *CVPR*, 2015a. 28, 47, 49, 55, 67, 95

Sadeghi, F., Zitnick, C. L., and Farhadi, A. VISALOGY: Answering Visual Analogy Questions. In *NIPS*, 2015b. 62, 96

Sadeghi, M. A. and Farhadi, A. Recognition using visual phrases. In *CVPR*, 2011. 6, 23, 24, 40, 48, 65, 67, 68, 77, 78, 83, 93, 95, 141, 142

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. A Simple Neural Network Module for Relational Reasoning. In *NIPS*, 2017. 35, 94

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 2009. 35

Schüldt, C., Laptev, I., and Caputo, B. Recognizing human actions: a local SVM approach. In *ICPR*, 2004. 39

Schuster, M. and Paliwal, K. K. Bidirectional recurrent neural networks. *Signal Processing*, 1997. 32

Schuster, S., Krishna, R., Chang, A. X., Fei-Fei, L., and Manning, C. D. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *VL@EMNLP*, 2015. 52

Shang, X., Ren, T., Guo, J., Zhang, H., and Chua, T.-S. Video Visual Relation Detection. In *ACM International Conference on Multimedia*, 2017. 137

Sharma, G., Jurie, F., and Schmid, C. Expanded Parts Model for Human Attribute and Action Recognition in Still Images. In *CVPR*, 2013. 131

Shen, L., Yeung, S., and Hoffman, J. Scaling Human-Object Interaction Recognition through Zero-Shot Learning. In *WACV*, 2018. 42, 95

Simon, M. and Rodner, E. Neural Activation Constellations: Unsupervised Part Model Discovery with Convolutional Networks. In *ICCV*, 2015. 131

Singh, S., Gupta, A., and Efros, A. A. Unsupervised Discovery of Mid-Level Discriminative Patches. In *ECCV*, 2012. 10

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., and Freeman, W. T. Discovering object categories in image collections. In *ICCV*, 2005. 55

Socher, R., Chen, D., Manning, C. D., and Ng, A. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*, 2013a. 49, 67

Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013b. 59, 69

Song, H. O., Girshick, R. B., Jegelka, S., Mairal, J., Harchaoui, Z., and Darrell, T. One-bit object detection: On learning to localize objects with minimal supervision. In *ICML*, 2014. 52

Speer, R. and Havasi, C. *ConceptNet 5: A Large Semantic Network for Relational Knowledge*. Springer, 2013. 47

Stansbury, D., Naselaris, T., and Gallant, J. L. Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. *Neuron*, 2013. 7

Stein, T., Kaiser, D., and Peelen, M. Interobject grouping facilitates visual awareness. *Journal of Vision*, 2015. 6

Suchanek, F. M., Kasneci, G., and Gerhard Weikum, G. Yago: a core of semantic knowledge. In *WWW*, 2007. 46

Sutskever, I., Salakhutdinov, R. R., and Tenenbaum, J. B. Modelling Relational Data using Bayesian Clustered Tensor Factorization. In *NIPS*, 2009. 49

Teney, D., Liu, L., and van den Hengel, A. Graph-Structured Representations for Visual Question Answering. In *CVPR*, 2017. 30, 35

Thurau, C. and Hlavác, V. Pose primitive based human action recognition in videos or still images. In *CVPR*, 2008. 41

Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., and Smeulders, A. W. M. Selective Search for Object Recognition. *International Journal of Computer Vision*, 2013. 21, 22

Van der Maaten, L. and Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 2008. 122, 125, 126, 127, 128

Vedantam, R., Zitnick, C. L., and Parikh, D. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015. 36

Vedantam, R., Bengio, S., Murphy, K., Parikh, D., and Chechik, G. Context-Aware Captions from Context-Agnostic Supervision. In *CVPR*, 2017. 34

Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. Order-embeddings of images and language. In *ICLR*, 2016. 58

Venugopalan, S., Hendricks, L. A., Rohrbach, M., Mooney, R., Darrell, T., and Saenko, K. Captioning Images with Diverse Objects. In *CVPR*, 2017. 37, 69

Wang, L., Li, Y., and Lazebnik, S. Learning Deep Structure-Preserving Image-Text Embeddings. In *CVPR*, 2016a. 31, 95, 98, 99

Wang, M., Azab, M., Kojima, N., Mihalcea, R., and Deng, J. Structured Matching for Phrase Localization. In *ECCV*, 2016b. 34

Wang, P., Wu, Q., Shen, C., Dick, A. R., and van den Hengel, A. Explicit knowledge-based reasoning for visual question answering. In *IJCAI*, 2017. 28

Wang, P., Wu, Q., Shen, C., Dick, A. R., and van den Hengel, A. FVQA: Fact-Based Visual Question Answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 28

Wang, X., Fouhey, D. F., and Gupta, A. Designing deep networks for surface normal estimation. In *CVPR*, 2015. 43

Wang, Y., Jiang, H., Drew, M. S., Li, Z.-N., and Mori, G. Unsupervised Discovery of Action Classes. In *CVPR*, 2006. 39, 41, 55

Woo, S., Kim, D., Cho, D., and Kweon, I. LinkNet: Relational Embedding for Scene Graph. In *NIPS*, 2018. 45

Wu, Q., Wang, P., Shen, C., Dick, A. R., and van den Hengel, A. Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge from External Sources. In *CVPR*, 2016. 28

Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A. R., and van den Hengel, A. Visual Question Answering: A Survey of Methods and Datasets. *Computer Vision and Image Understanding*, 2017. 33

Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., and Schiele, B. Latent Embeddings for Zero-shot Classification. In *CVPR*, 2016. 69

Xiao, F., Sigal, L., and Lee, Y. J. Weakly-Supervised Visual Grounding of Phrases with Linguistic Structures. In *CVPR*, 2017. 34

Xie, N., Lai, F., Doran, D., and Kadav, A. Visual Entailment Task for Visually-Grounded Language Learning. *arXiv:1811.10582*, 2018. 37

Xu, D., Zhu, Y., Choy, C. B., and Fei-Fei, L. Scene Graph Generation by Iterative Message Passing. In *CVPR*, 2017. 30, 45, 46

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*, 2015. 32

Xu, L., Neufeld, J., Larson, B., and Schuurmans, D. Maximum Margin Clustering. In *NIPS*, 2004. 54

Yan, F. and Mikolajczyk, K. Deep correlation for matching images and text. In *CVPR*, 2015. 31

Yang, J., Lu, J., Lee, S., Batra, D., and Parikh, D. Graph r-cnn for scene graph generation. In *ECCV*, 2018a. 24, 45

Yang, S., Bo, L., Wang, J., and Shapiro, L. G. Unsupervised Template Learning for Fine-Grained Object Recognition. In *NIPS*, 2012. 131

Yang, X., Zhang, H., and Cai, J. Shuffle-Then-Assemble: Learning Object-Agnostic Visual Relationship Features. In *ECCV*, 2018b. 42

Yang, Z., He, X., Gao, J., Deng, L., and Smola, A. J. Stacked Attention Networks for Image Question Answering. In *CVPR*, 2016. 32

Yao, B. and Fei-Fei, L. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010a. 41, 67

Yao, B. and Fei-Fei, L. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010b. 39, 67, 131

Yao, B. and Fei-Fei, L. Action Recognition with Exemplar Based 2.5D Graph Matching. In *ECCV*, 2012. 41

Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L. J., and Fei-Fei, L. Human Action Recognition by Learning Bases of Action Attributes and Parts. In *ICCV*, 2011. 39, 45, 67

Yao, T., Pan, Y., Li, Y., and Mei, T. Exploring Visual Relationship for Image Captioning. In *ECCV*, 2018. 30, 35

Yatskar, M., Ordonez, V., and Farhadi, A. Stating the Obvious: Extracting Visual Common Sense Knowledge. In *NAACL*, 2016. 67

Yin, G., Sheng, L., Liu, B., Yu, N., Wang, X., Shao, J., and Loy, C. C. Zoom-Net: Mining Deep Feature Interactions for Visual Relationship Recognition. In *ECCV*, 2018. 46

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. Modeling Context in Referring Expressions. In *ECCV*, 2016. 34, 53

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., and Berg, T. L. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *CVPR*, 2018. 35

Yu, R., Li, A., Morariu, V. I., and Davis, L. S. Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation. In *ICCV*, 2017. 46, 95

Zagoruyko, S., Lerer, A., Lin, T.-Y., Pinheiro, P. O., Gross, S., Chintala, S., and Dollár, P. A MultiPath Network for Object Detection. In *BMVC*, 2016. 65

Zellers, R., Yatskar, M., Thomson, S., and Choi, Y. Neural Motifs: Scene Graph Parsing with Global Context. In *CVPR*, 2018. 42

Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. From Recognition to Cognition: Visual Commonsense Reasoning. In *CVPR*, 2019. 37

Zhang, H., Kyaw, Z., Chang, S.-F., and Chua, T.-S. Visual Translation Embedding Network for Visual Relation Detection. In *CVPR*, 2017a. 43, 50, 92, 96

Zhang, H., Kyaw, Z., Yu, J., and Chang, S.-F. PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN. In *ICCV*, 2017b. 53

Zhang, J., Elhoseiny, M., Cohen, S., Chang, W., and Elgammal, A. M. Relationship Proposal Networks. In *CVPR*, 2017c. 24, 132

Zhang, J., Kalantidis, Y., Rohrbach, M., Paluri, M., Elgammal, A., and Elhoseiny, M. Large-Scale Visual Relationship Understanding. In *AAAI*, 2019. 40, 42, 46, 49, 96

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *ICCV*, 2017. 43, 137, 138

Zhu, Y., Fathi, A., and Fei-Fei, L. Reasoning about object affordances in a knowledge base representation. In *ECCV*, 2014. 28, 47, 67

Zhu, Y., Groth, O., Bernstein, M. S., and Fei-Fei, L. Visual7W: Grounded Question Answering in Images. In *CVPR*, 2016. 27, 29, 32

Zhuang, B., Liu, L., Shen, C., and Reid, I. Towards Context-aware Interaction Recognition for Visual Relationship Detection. In *ICCV*, 2017. 43, 50, 95

Zhuang, B., Wu, Q., Sehn, C., Reid, I., and van den Hengel, A. HCVRD: a benchmark for large-scale Human-Centered Visual Relationship Detection. In *AAAI*, 2018. 13, 40

Zitnick, C. L. and Dollár, P. Edge Boxes: Locating Object Proposals from Edges. In *ECCV*, 2014. 21

Zitnick, C. L., Parikh, D., and Vanderwende, L. Learning the visual interpretation of sentences. In *ICCV*, 2013. 35

RÉSUMÉ

Nous étudions le problème de détection de relations visuelles de la forme (sujet, prédicat, objet) dans les images, qui sont des entités intermédiaires entre les objets et les scènes visuelles complexes. Cette thèse s'attaque à deux défis majeurs : (1) le problème d'annotations coûteuses pour l'entrainement de modèles fortement supervisés, (2) la variation d'apparence visuelle des relations. Nous proposons un premier modèle de détection de relations visuelles faiblement supervisé, n'utilisant que des annotations au niveau de l'image, qui, étant donné des détecteurs d'objets pré-entrainés, atteint une précision proche de celle de modèles fortement supervisés. Notre second modèle combine des représentations compositionnelles (sujet, objet, prédicat) et holistiques (triplet) afin de mieux modéliser les variations d'apparence visuelle et propose un module de raisonnement par analogie pour généraliser à de nouveaux triplets. Nous validons expérimentalement le bénéfice apporté par chacune de ces composantes sur des bases de données réelles.

MOTS CLÉS

Vision par ordinateur, Détection de relations visuelles, Compréhension d'images, Image et language, Apprentissage faiblement supérvisé, Apprentissage profond.

ABSTRACT

In this thesis, we study the problem of detection of visual relations of the form (subject, predicate, object) in images, which are intermediate level semantic units between objects and complex scenes. Our work addresses two main challenges in visual relation detection: (1) the difficulty of obtaining box-level annotations to train fully-supervised models, (2) the variability of appearance of visual relations. We first propose a weakly-supervised approach which, given pre-trained object detectors, enables us to learn relation detectors using image-level labels only, maintaining a performance close to fully-supervised models. Second, we propose a model that combines different granularities of embeddings (for subject, object, predicate and triplet) to better model appearance variation and introduce an analogical reasoning module to generalize to unseen triplets. Experimental results demonstrate the improvement of our hybrid model over a purely compositional model and validate the benefits of our transfer by analogy to retrieve unseen triplets.

KEYWORDS

Computer vision, Visual relation detection, Scene understanding, Image and language, Weakly-supervised learning, Deep learning.