Joint representation learning from french radiological reports and ultrasound images

Hind Dadoun, Hervé Delingette, Anne-Laure Rousseau, Eric de Kerviler, Nicholas Ayache

JOINT REPRESENTATION LEARNING FROM FRENCH RADIOLOGICAL REPORTS AND ULTRASOUND IMAGES

Hind Dadoun[⋆] *Hervé Delingette*[⋆] *Anne-Laure Rousseau*[†] *Eric de Kerviler*[‡] *Nicholas Ayache*[⋆]

[⋆] Université Côte d'Azur, Inria, Epione Team, Sophia Antipolis, France
[†] Hôpital Européen Georges Pompidou, NHance, Paris, France
[‡] Radiology, Saint Louis Hospital, AP-HP, Paris, France

ABSTRACT

In this study, we explore the value of using a recently proposed multimodal learning method as an initialization for anomaly detection in abdominal ultrasound images. The method efficiently learns visual concepts from radiological reports using natural language supervision and constrastive learning. The underlying requirement of the method is simply the availability of image and textual descriptions pairs. However, in abdominal ultrasound examinations, radiological reports are associated with several images and describe all organs observed during the examination. To address this shortcoming, we automatically construct image and text pairs using 1) deep clustering for abdominal organ classification on ultrasound images and 2) natural language processing tools to extract the corresponding description on the report. We show that pre-training the model with these constructed pairs yields representations that better separate normal classes from abnormal ones on ultrasound images for the kidneys, compared to ImageNet-based representations, with a 10% improvement in macro-average accuracy.

***Index Terms*—** multimodal learning, deep clustering, natural language processing, ultrasound examinations, kidneys

1. INTRODUCTION

In this study, we focus on abnormality detection in abdominal ultrasound images by considering a binary classification task (i.e. normal vs. abnormal organ) with access to limited labeled data. Diseases associated to a given organ may alter its shape, size, contour, position, or textural appearance, resulting in highly variable differences in ultrasound patterns, all grouped into a single "abnormal" category. For this reason, transferring the model weights from ImageNet[1] pre-training can result in poor performance, as the features learned on natural images are not suited to capture the fine-grained visual features necessary to separate the normal class from the abnormal class. Alternatively, in the absence of large number of annotated datasets, the model can be pre-trained on an unlabeled set of ultrasound images using self-supervised learning methods. These image-based self-supervised methods were proven to enhance performance in specific settings [2, 3, 4]. In the case of abdominal ultrasound, we observed that for tasks with high inter-class variability, these methods can be useful. However, on a task such as pathology detection, in which there is greater similarity between classes, and more variability within classes, these methods often fail to achieve better results. More recently, a study [5] proposed the use of multi-modal pre-training to learn fine-grained representations required by medical imaging tasks. They argue that medical reports, as opposed to image labels are often produced by medical experts in their routine workflow and are therefore easily accessible. The approach takes advantage of the medical reports associated to medical images, to learn better latent representations. The underlying assumption of the method is simply the presence of pairs of images and text describing the image. This method was evaluated on four different medical image classification tasks covering 2 different specialties with encouraging results. Yet, the considered hypothesis in which image pairs and a textual description are always available is not consistent with the abdominal ultrasound setting in which a textual radiology report describes a set of ultrasound images. In this work, we present a method to automatically build pairs of text descriptions and ultrasound images using deep clustering for images and named entity recognition for text. We evaluate the pre-trained image encoder on two criteria: its ability to extract discriminative features for the anomaly classification task, and its performance when fine tuning the model on a labeled set. We provide results for the normal/abnormal kidney detection task in an abdominal ultrasound examination.

2. IMAGE AND TEXT PAIRS GENERATION

2.1. Data

During an ultrasound examination, the sonographer performs a complete scan of the area of interest and takes captures, also known as freeze frames, of the standard scanning plane views and potential visible abnormalities. The freeze-frames along with a textual documentation of the examination form the ul-

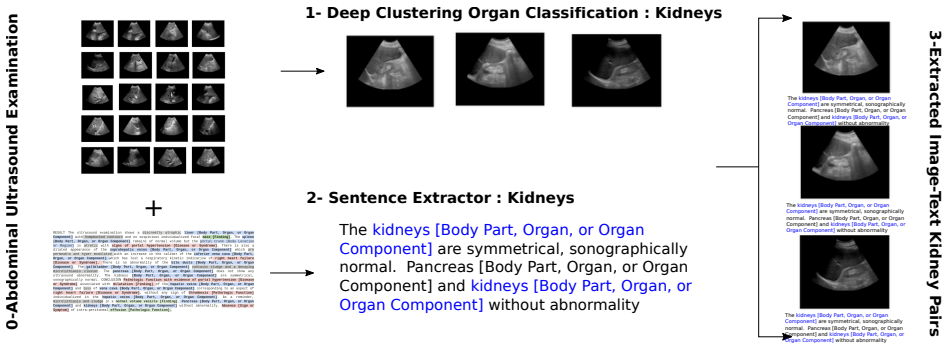

Fig. 1. Generation pipeline of Image-Text Kidney pairs for the pre-training set.

trasound examination report. Our dataset consisted of 8011 abdominal ultrasound examinations ($n_{images} = 120,593$) from 6482 patients with an average of 12.5 images per examination. The images are not restricted to kidneys, and can contain other abdominal organs. We show in the following, how images containing the kidneys are selected among all images. Likewise we show how the sentences in the report that describe the kidneys are detected.

3. IMAGE AND TEXT PAIRS GENERATION

3.1. Data Partition

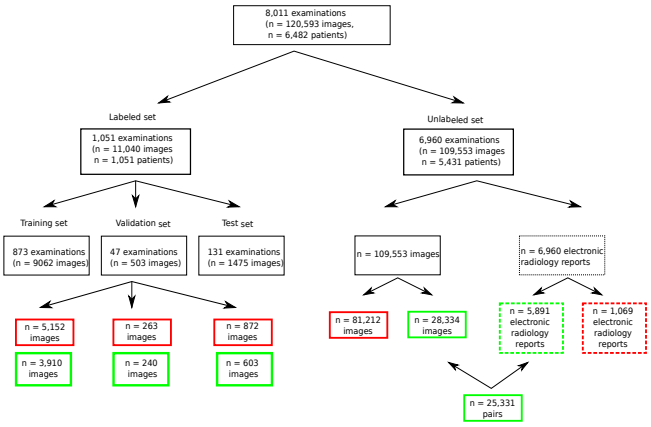

Fig. 2. Data partition. Green boxes are used to indicate that the images or text refer to the kidneys, red boxes are used for all other organs.

A subset of 1,051 ultrasound examinations was randomly selected to constitute the labeled sets of all freeze frames linked to the examinations. For each image the annotators had to assign a label, either normal kidneys, abnormal kidneys or "absent" if the image did not contain any kidneys. 873 of those examinations ($n_{images} = 3,910$ images of kidneys) were assigned to the training set, 47 examinations ($n_{images} = 240$ images of kidneys) to the validation set and 131 examinations ($n_{images} = 603$ images of kidneys) to the test set. The remaining 6,960 unlabeled examinations were used to construct the pre-training set. The image and text data of the pre-training set were processed in three steps which are summarized in Figure 2. First images containing the kidneys are selected, then sentences in the medical report describing the kidneys are selected. Finally, pairs of image and text are constructed for the kidneys.

3.2. Clustering U/S Images

Since we are only interested in kidney images in this study, we use deep clustering to identify a set of clusters linked to this organ. Specifically, we use the method presented in [6] where a framework for abdominal organ clustering using unlabeled ultrasound images is presented. Two different augmentation schemes $(\tilde{\cdot})$ and $(\hat{\cdot})$ are applied to all input images before passing through the CNN. Then a loss term is used to encourage both augmented versions of the image to be assigned to the same class with high probability. Finally, to avoid all images being assigned to the same cluster, an additional loss term is introduced to constrain the distribution of the cluster size to follow a symmetric Dirichlet distribution. This method achieves reasonable performance with an F1-score weighted average of 66.75% and an F1-score of 71.5% for the kidney's class. All images ($n = 87,696$) of our pre-training set are processed by the deep clustering method, and only images assigned to "Kidney" clusters are kept which amounts to a total 28,334 images from 6,531 examinations.

3.3. Text Data

To select the sentences mentioning kidneys in the medical report, we use a tagging tool based on the Unified Medical Language System (UMLS). UMLS is a meta-system that unifies concepts from several dozen terminologies in the biomedical domain. Each UMLS concept is assigned a unique concept identifier (CUI), a set of terms (or synonyms), possibly in multiple languages, and a semantic type. The labeling tool (QuickUMLS) [7], is a fast, unsupervised biomedical concept extraction tool from medical texts that works for multiple languages, including French. Given an unstructured textual med-

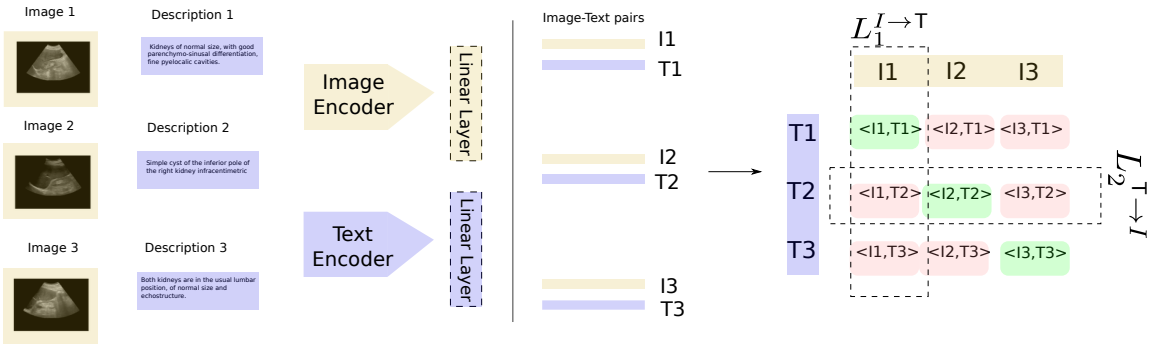

Fig. 3. Pre-training Pipeline: Images and texts are fed to an image encoder and a text encoder respectively. Each input image and text are then converted into d-dimensional vector representation using a Linear Layer. An asymmetric Image-to-text contrastive loss is used for each input modality.

ical report, we first divide it into sentences. For each sentence, we search for a word related to the semantic type "Body part, organ or organ component" and check whether its unique concept identifier refers to kidneys. If so, the sentence is added to the list of sentences assigned to the organ. A total of 5891 examinations, out of the original 6960 examinations, mention kidneys in the report.

3.4. Image-Text Pairing

For each image where the abdominal organ clustering predicted the presence of a kidney, we look if the associated report mentions the kidney. If so, the sentences mentioning the kidney in the report are all associated with the image. This means that multiple images from the same examination will have the same description. Of the 28,334 images, 3003 could not be linked to any textual description (i.e. the associated examination did not mention the kidneys), which leaves us with a final dataset of 25,331 pairs of text and images from 5,513 examinations.

4. JOINT REPRESENTATION LEARNING

Here we detail the model architecture for the renal anomaly detection task. First, the text and image encoders are trained jointly to project the data into the same dimensional space while ensuring coherence between text and image representations. The image encoder is then fine-tuned on a labeled dataset of renal ultrasound images.

4.1. Architecture

We base the pre-training approach on a model (ConVIRT) developed by Zhang *et al* [5]. The model is composed of a text encoder and an image encoder. For the image encoder, we use the ResNet50 architecture. The input images are randomly augmented with different data augmentations: cropping, horizontal flipping, affine transformation, color jittering and Gaussian blur before passing to the encoder. For the text encoder, we use the *CamemBERT* [8] model which is a state-of-the-art language model pre-trained on a French corpus *OSCAR*, based on the *RoBERTa* [9] architecture. Since the text encoder was pre-trained on generic text, it is essential to consider words that are specific to the domain on which we want to refine the model (abdominal ultrasound radiological reports). To do so, we re-train a word-piece Tokenizer to find the set of words that minimize the number of tokens needed to reconstruct the reports in our training set. Each input image and text are then converted into d-dimensional vector representation using a Linear Layer as shown in Figure 3.

4.2. Pre-training Objective Function

The model is trained to predict which image goes with which description and conversely. This is achieved using two InfoNCE [10] losses based on the cosine similarity between the transformed image and text vectors. Let I, T be the d-dimensional vectors of image and text respectively, and $\langle I, T \rangle$ be the cosine similarity between the two, the objective function introduced in [5] is as follows:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (\lambda \cdot L_i^{I \to T} + (1 - \lambda) \cdot L_i^{T \to I})$$

where $L_i^{I \to T}$ is a text-to-image contrastive loss, whose goal is to predict I_i, T_i as the true pair among all possible descriptions and $L_i^{T \to I}$ is an image-to-text contrastive loss, whose goal is to predict I_i, T_i as the true pair among all possible images.

$$L_i^{I \leftarrow T} = -\log \frac{exp(\langle I_i, T_i \rangle / \tau)}{\sum_{j=1}^{n} exp(\langle I_i, T_j \rangle)} \quad \text{and} \quad L_i^{T \leftarrow I} = -\log \frac{exp(\langle I_i, T_i \rangle / \tau)}{\sum_{j=1}^{n} exp(\langle I_j, T_i \rangle)}$$

4.3. Fine-tuning Objective Function

We evaluate our pretrained image encoder on the binary classification task of abnormality detection (normal vs. abnormal

	Precision		Recall		F1-Score		support
	Baseline	**ConVIRT**	**Baseline**	**ConVIRT**	**Baseline**	**ConVIRT**	
Normal Kidneys	0.91	**0.92**	0.63	**0.77**	0.74	**0.84**	427
Abnormal Kidneys	0.49	**0.60**	**0.85**	0.84	0.62	**0.70**	176
accuracy					0.69	**0.79**	603
macro-average	0.70	**0.76**	0.74	**0.80**	0.68	**0.77**	603
weighted average	0.79	**0.83**	0.69	**0.79**	0.71	**0.80**	603

Table 1. Fine-tuning performance on the 603 images of the test set. The baseline initializes the model weights with those of the ImageNet pre-training, whereas ConVIRT initializes them with those of the image-text pre-training.

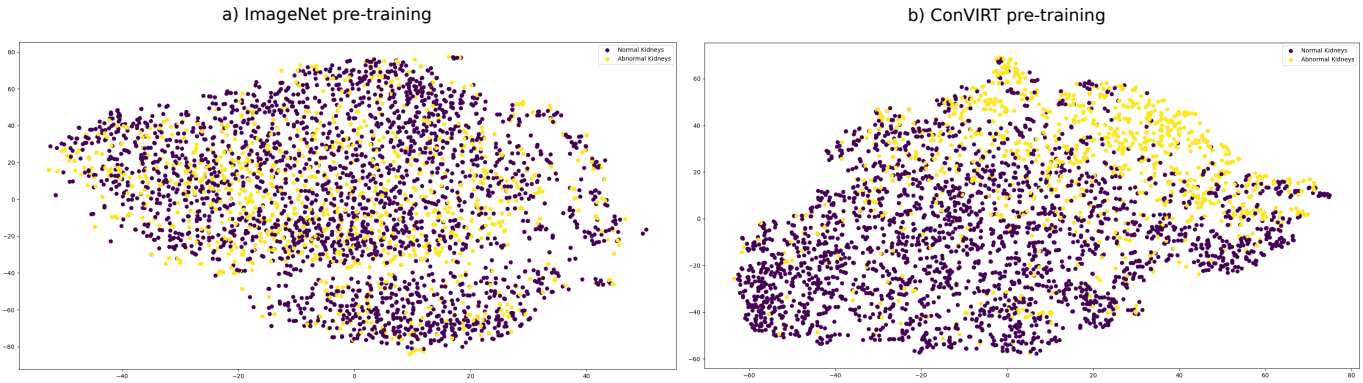

Fig. 4. t-SNE visualizations of encoded image from different pre-training methods. Purple points correspond to images of healthy kidneys, and yellow to images of abnormal kidneys.

kidneys). Both the CNN weights and the linear head are fine-tuned on a labeled training set of 3,910 images. We use the generalized cross entropy loss introduced in [11] to train deep neural networks with noisy labels.

5. RESULTS

5.1. T-SNE Visualization of Extracted Features during Pre-training

First we evaluate how joint representation pre-training impacts the resulting image features, in comparison to the same encoder trained on ImageNet, using t-SNE visualization [12]. t-SNE is a stochastic method for visualizing high-dimensional data. We use the scikit-learn [13] implementation with the default parameters. We can see in Figure 4 that joint representation pre-training helps separate normal classes from abnormal ones in its encoding low-dimensional space.

5.2. Classification Results after Fine-tuning

In order to quantitatively assess the detection performance of this approach after fine-tuning, we evaluated the model on the test cohort presented in Section 3.1. We measured the algorithm's performance using the precision and recall rates as well as the F1-score. One can see in Table 1 that the Image-Text joint pre-traning using ConVIRT yields a 10% improvement in macro-average accuracy compared to the baseline with ImageNet pre-training. Specifically, we found that for ImageNet and ConVIRT pre-training, the negative predictive value (0.91 vs. 0.92) and sensitivity (0.85 vs. 0.84) respectively, were similar. On the other hand, the positive predictive value (0.49 vs. 0.60) and specificity (0.63 vs. 0.77) respectively, were both higher for ConVIRT compared to ImageNet pre-training.

6. CONCLUSION

In this study we explored the value of using unstructured radiological reports to pre-train a model to better separate normal and abnormal kidney ultrasound images. Although a direct link between images and their descriptions is not provided in abdominal ultrasound examinations, we were able to build pairs of images and text using different unsupervised methods. Finally, we showed that this matching strategy, combined with conVIRT pre-training, provided a 10% increase in accuracy during fine-tuning compared to ImageNet pre-training.

7. REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[2] Jianbo Jiao, Richard Droste, Lior Drukker, Aris T Papageorghiou, and J Alison Noble, "Self-supervised representation learning for ultrasound video," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1847–1850.

[3] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical Image Analysis*, vol. 58, pp. 101539, 2019.

[4] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E Petersen, Yike Guo, Paul M Matthews, and Daniel Rueckert, "Self-supervised learning for cardiac mri image segmentation by anatomical position prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 541–549.

[5] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz, "Contrastive learning of medical visual representations from paired images and text," *arXiv preprint arXiv:2010.00747*, 2020.

[6] Hind Dadoun, Hervé Delingette, Anne-Laure Rousseau, Eric de Kerviler, and Nicholas Ayache, "Deep clustering for abdominal organ classification in US imaging," preprint submitted to a journal, 2022.

[7] Luca Soldaini and Nazli Goharian, "QuickUMLS: a fast, unsupervised approach for medical concept extraction," in *MedIR workshop, SIGIR*, 2016, pp. 1–4.

[8] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot, "CamemBERT: a tasty French language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 7203–7219, Association for Computational Linguistics.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[10] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[11] Zhilu Zhang and Mert Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[12] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE.," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[13] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., "Scikit-learn: Machine learning in python," *the Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

8. COMPLIANCE WITH ETHICAL STANDARDS

IRB approval (IRB00011591) was obtained for this multicenter study and informed consent was waived. Data for this pilot study were obtained in collaboration with the Clinical Data Warehouse of Greater Paris University Hospitals registered at the National Commission for Data Protection and Liberties (CNIL-France) under the number 1980120.

9. ACKNOWLEDGMENTS

This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.
The authors are grateful to the OPAL infrastructure from Université Côte d'Azur, the Radiology team from Saint Louis Hospital of Greater Paris University Hospitals and the French Health Data Hub for providing resources and support.
We thank Guillaume Oules, engineer at GoPro, for his help in the design of the NHance project. We thank the Clinical Data Warehouse of Greater Paris University Hospitals and especially the imaging team. The authors declare that they have no conflicts of interest.