A Crossmodal Approach to Multimodal Fusion in Video Hyperlinking

Vedran Vukotić, Christian Raymond, Guillaume Gravier

A Crossmodal Approach to Multimodal Fusion in Video Hyperlinking

Vedran Vukotić, Christian Raymond, Guillaume Gravier

Abstract—With the recent resurgence of neural networks and the proliferation of massive amounts of unlabeled data, unsupervised learning algorithms became very popular for organizing and retrieving large video collections in a task defined as video hyperlinking. Information stored as videos typically contain two modalities, namely an audio and a visual one, that are used conjointly in multimodal systems by undergoing fusion. Multimodal autoencoders have been long used for performing multimodal fusion.
In this work, we start by evaluating different initial, single-modal representations for automatic speech transcripts and for video keyframes. We progress to evaluating different autoencoding methods of performing multimodal fusion in an offline setup. The best performing setup is then evaluated in a live setup at TRECVID's 2016 video hyperlinking task. As in offline evaluations, we show that focusing on crossmodal translations as a way of performing multimodal fusion yields improved multimodal representations and that our simple system, trained in an unsupervised manner, with no external information information, defines the new state of the art in a live video hyperlinking setup. We conclude by performing an analysis on data gathered after the live evaluations at TRECVID 2016 and express our thoughts on the overall performance of our proposed system.

Index Terms—multimodal fusion, video hyperlinking, multimodal autoencoders, multimodal retrieval, crossmodal, neural networks, deep learning, unsupervised representation learning, video retrieval, bidirectional learning, tied weights, shared weights.

1 Introduction

The seminal idea of video hyperlinking is to create hyperlinks between different videos and/or video segments based on their data. In video hyperlinking, there are two main concepts: anchors and targets. Anchors represent segments of interest within videos that a user would like to know more about. Targets represent potential segments of interests that might or might not be related with a specific anchor. The goal is to hyperlink relevant targets for each anchor by using multimodal approaches. Each video consists of at least two data streams: a visual stream and an audio stream. A visual stream is typically represented by frames, where only keyframes are stored completely and other frames are encoded as a difference from a keyframe. Audio streams also provide information - most often (but not limited to) as speech and thus, after automatic transcription, a sequence of words. Data from an audio source does not have to correlate with data from the corresponding video source but it certainly can. Given this nature of videos, it is necessary to perform content analysis and comparison of both visual information and spoken information both in a crossmodal and in a multimodal fashion (e.g., a link between two video segments can reflect a connection between a concept being discussed in the first video segment and a location being displayed in the second video segment). State-of-the-art continuous representation spaces exists for both visual and audio modalities, as well as for different levels of embedding of each modality (e.g., visual embedding vs semantic embedding and visual concept recognition). However, to fully describe a video segment, it is necessary to combine both modalities, as they typically do not provide the same information and suplement each other. In all generality, methods for fusing modalities are often required when working with multimodal data. The most common approach consists in creating a joint multimodal representation by embedding every single-modal representations into a common representation space.

There are two main groups of such approaches:

1) *Multimodal approaches* create a joint representation of the initially disjoint modalities or otherwise merge the initial modalities without necessarily providing a bidirectional mapping of the initial representation spaces to the new representation space and back. These approaches are typically used in retrieval and classification tasks where translating back from the multimodal representation to the single-modal ones is not required.
2) *Crossmodal approaches* focus on bidirectional mapping of the initial representations [1], often by also creating a joint representation space in the process of doing so. They are able to map from one modality to another and back, as well as representing them in a joint representation space. These approaches can be used where crossmodal translation is required (e.g., multimodal query expansion, crossmodal retrieval) in addition to multimodal fusion.

In this work, we start by analyzing different methods for multimodal embedding and crossmodal mapping, as well as different different single-modal representations to jointly embed descriptors in a new multimedia representation for the task of video hyperlinking. We evaluate multimodal fusion methods in an offline setup with a fixed groundtruth. After that, we progress to evaluating the two initially disjoint modalities and propose to perform multimodal fusion by focusing on crossmodal translations, with bidirectional deep neural networks. We perform evaluations offline and in a live video hyperlinking setup where we show that the proposed method defines the new state of the art. We then

conclude by a further post-hoc analysis and and further discussion.

2 Content Representation

Content representation with appropriate embedding is key to video hyperlinking. We discuss speech and visual embedding, before presenting multimodal embedding techniques.

2.1 Initial Single-modal Representations

All methods presented in this work utilize two data modalities: i) automatic audio transcripts and ii) video keyframes. Automatic audio transcripts are used instead of subtitles which are not always available and would include a human component in the system. Video keyframes are considered in two different settings: using *ImageNet* concepts [2] or directly describing images with features obtained with state-of-the-art convolutional neural networks.

2.1.1 Representing Automatic Transcripts

Automatic transcripts of a video segment consist of one or more sentences, each with multiple words. This makes sentence/paragraph/document representation methods suitable for the task. Two methods were evaluated (each in different settings): paragraph vectors [3] and Word2Vec [4]. Contrary to paragraph vectors, Word2Vec is not specifically designed for embedding bigger blocks of text. However, it was shown that Word2Vec can perform quite well [5] and can be suitable when combined with an aggregation of the embedded words.

2.1.2 Representing Visual Information with Concepts

For each keyframe of each video segment, a set of top scoring visual concepts is used as information indicating what's visible in the image. Visual concepts describe a class of objects or entities and includes all related subcategories. We treat each visual concept as a word and utilize it to obtain word embeddings representing the visual information of a video segment provided by its visual concepts in a continuous representation space.

2.1.3 Representing Visual Information with CNN Features

Convolutional Neural Networks provide state-of-the art visual descriptors [6] that have been shown to perform well in computer vision applications [7], [8] and in video summarization tasks [9]. In this work, we test three different state-of-the art deep convolutional neural network architectures, namely AlexNet [10], VGG-16 and VGG-19 [11].

2.2 Multimodal and Crossmodal Approaches

There is a multitude of ways to combine and fuse multiple modalities. After exploring basic methods for combining different modalities and classical ways to perform multimodal fusion, we elaborate the feasibility of using crossmodal translations as a way to perform multimodal fusion.

2.2.1 Simple Methods for Combining Multiple Modalities

A simple way to perform multimodal early fusion is by simply concatenating single-modal representations. This does not provide the best results, as each representation still belongs to its own representation space. It is also possible to utilize two separate modalities by performing a linear combination [12] of the similarities obtained by comparing each of the two modalities. We use these two methods as a baseline to compare standard autoencoders and bidirectional deep neural networks against.

2.2.2 Multimodal/Crossmodal Autoencoders

When using autoencoders for multimodal embedding, a classical approach is to concatenate the modalities at the inputs and outputs of a network [13], [14]. This approach, however, does not offer crossmodal translation. A better approach is to have autoencoders with separate inputs and separate outputs for each modality, often with additional separate fully connected layers attached to each input and output layer, as illustrated in Figure 1. One common hidden layer is used for creating a joint multimodal representation. One modality might be sporadically removed from the input to make the autoencoder learn to represent both modalities from one. This enables autoencoders to also provide crossmodal mappings [13] in addition to a joint representation.

Autoencoders however have some downsides which slightly deteriorate performance:

- Both modalities influence the same central layer(s), either directly or indirectly, through other modality-specific fully connected layers. Even when translating from one modality to the other, the input modality is either mixed with the other or with a zeroed input.
- Autoencoders need to learn to reconstruct the same output both when one modality is marked missing (e.g., zeroed) and when both modalities are presented as input.
- Classical autoencoders are primarily made for multimodal embedding while crossmodal translation is offered as a secondary function.

These issues are addressed by bidirectional (symmetrical) deep neural networks [15], which we discuss next.

2.2.3 Bidirectional Deep Neural Networks

In BiDNNs, learning is performed in both directions: one modality is presented as an input and the other as the expected output while at the same time the second one is presented as input and the first one as expected output. This is equivalent to using two separate deep neural networks and tying them (sharing specific weight variables) to make them symmetrical, as illustrated in the bottom part of Figure 1. Implementation-wise the variables representing the weights are shared across the two networks. Learning of the two crossmodal mappings is then performed simultaneously and they are forced to be as close as possible to each other's inverses by the symmetric architecture in the middle. Symmetry is enforced only on the central layers, as enforcing complete symmetry would prevent the architecture to converge. A joint representation in the middle of the two crossmodal mappings is also formed while learning.

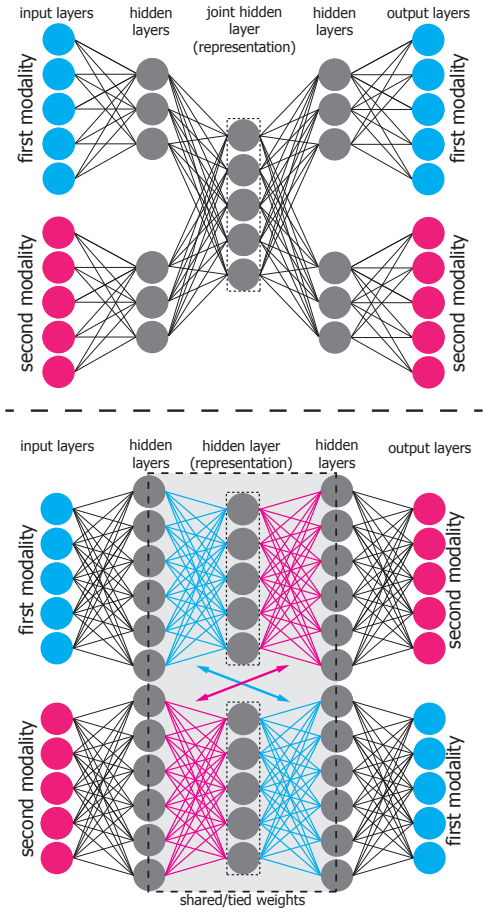

Fig. 1. Illustration of the architectures of a classical multimodal autoencoder (top) and a bidirectional symmetrical deep neural network (bottom)

Formally, let $\mathbf{h_i^{(j)}}$ denote (the activation of) a hidden layer at depth j in network i ($i = 1, 2$; one for each modality), $\mathbf{x_i}$ the feature vector for modality i and $\mathbf{o_i}$ the output of the network for modality i. In turn, for each network, $\mathbf{W_i^{(j)}}$ denotes the weight matrix of layer j and $\mathbf{b_i^{(j)}}$ the bias vector. Finally, we assume that each layer admits f as an activation function. The architecture is then defined by:

$$\mathbf{h_1^{(1)}} = f(\mathbf{W_1^{(1)}} \times \mathbf{x_1} + \mathbf{b_1^{(1)}})$$
$$\mathbf{h_2^{(1)}} = f(\mathbf{W_2^{(1)}} \times \mathbf{x_2} + \mathbf{b_2^{(1)}})$$

$$\mathbf{h_1^{(2)}} = f(\mathbf{W^{(2)}} \times \mathbf{h_1^{(1)}} + \mathbf{b_1^{(2)}})$$
$$\mathbf{h_2^{(2)}} = f(\mathbf{W^{(3)^T}} \times \mathbf{h_2^{(1)}} + \mathbf{b_2^{(2)}})$$

$$\mathbf{h_1^{(3)}} = f(\mathbf{W^{(3)}} \times \mathbf{h_1^{(2)}} + \mathbf{b_1^{(3)}})$$
$$\mathbf{h_2^{(3)}} = f(\mathbf{W^{(2)^T}} \times \mathbf{h_2^{(2)}} + \mathbf{b_2^{(3)}})$$

$$\mathbf{o_1} = f(\mathbf{W_1^{(4)}} \times \mathbf{h_1^{(3)}} + \mathbf{b_1^{(4)}})$$
$$\mathbf{o_2} = f(\mathbf{W_2^{(4)}} \times \mathbf{h_2^{(3)}} + \mathbf{b_2^{(4)}})$$

It is important to note that in the above equations, the weight matrices $\mathbf{W^{(3)}}$ and $\mathbf{W^{(2)}}$ are used twice due to weight tying, for computing $\mathbf{h_1^{(2)}}$, $\mathbf{h_2^{(3)}}$ and $\mathbf{h_2^{(2)}}$, $\mathbf{h_1^{(3)}}$ respectively. Training is performed by applying gradient descent to minimize the mean squared error of $(\mathbf{o_1}, \mathbf{x_{m_2}})$ and $(\mathbf{o_2}, \mathbf{x_{m_1}})$ thus effectively minimizing the reconstruction error in both directions and creating a joint representation in the middle, where both representations are projected.

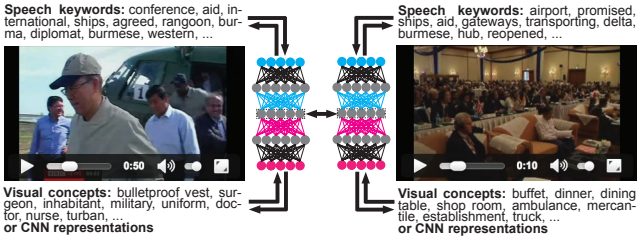

Fig. 2. Video hyperlinking with BiDNNs: two video segments, both with two modalities are compared after their multimodal embeddings are computed

Given such an architecture, crossmodal translation is done straightforwardly by presenting the first modality as $\mathbf{x_{m_i}}$ and obtaining the output in the representation space of the second modality as $\mathbf{o_i}$. A multimodal embedding is obtained by presenting one modality or both modalities ($\mathbf{x_{m_1}}$ and/or $\mathbf{x_{m_2}}$) at their respective inputs and reading the central hidden layers $\mathbf{h_1^{(2)}}$ and/or $\mathbf{h_1^{(2)}}$.

Multimodal embeddings are obtained in the following manner:

- When the two modalities are available, both are presented at their respective inputs and the activations are propagated through the network. The multimodal embedding is then obtained by concatenating the outputs of the middle layer.
- When one modality is available and the other is not, the available modality is presented to its respective input of the network and the activations are propagated. The central layer is then used to generate an embedding by being duplicated, thus still generating an embedding of the same size while allowing to transparently compare video segments regardless of modality availability.

Finally, segments are then compared as illustrated in Figure 2: for each video segment, the two modalities are taken and a multimodal embedding is created with a BiDNN. The two multimodal embeddings are then simply compared with a cosine similarity measure.

3 Experiments

We first describe the datasets used for evaluation. After that, we proceed to evaluating the different methods described in Section 2.1 for creating single-modal representation and we evaluate methods for performing fusion, described in Section 2.2 in both offline and live evaluations.

3.1 Dataset

3.1.1 Post MediaEval 2014 Dataset

The generation of hyperlinks within video segments is the focus of the "Search and Hyperlinking" evaluation at MediaEval and more recently at TRECVid [16]. All the methods

were evaluated within the task of video hyperlinking using the MediaEval 2014 dataset and the respective groundtruth that was collected as part of the challenge [17]. The videos for the task were provided by BCC. For each video, multiple data and modalities are available. In this work, we used a combination of two modalities: for the speech modality we use automatic transcripts embedded in different ways, while for the the visual modality we either use embedded visual concepts (ImageNet [2] classes) provided by KU Leuven [18] or CNN features.

During the challenge, targets are not given and are suggested for each anchor by the systems of the participants. Their relevance is then judged post-hoc by human evaluated on Amazon Mechanical Turk (AMT). After the challenge, a groundtruth is form containing the top-10 targets that each that each participating team proposed for each anchor, along with the relevance judgments from AMT. Using this groundtruth, we formulate the problem as a reranking problem, where we rerank the targets from the groundtruth and evaluate relevance against it.

In total, the dataset consists of 30 anchors, 10,809 targets and a ground truth with 12,340 anchor-target pairs (either related or unrelated). Interestingly, among the anchor and target segments, not all have both transcripts and visual concepts available. On average, there are 34.3 keyframes per video segment. The task consists of using multimodal information to rank the targets by relevance for each anchor and comparing their relevance with the previously established groundtruth.

3.1.2 TRECVID 2016 Dataset

In the previous dataset, that is formed post-hoc, the segmentation is already given and defined by the limited set of video segments the participants suggested as possible targets. For a live setup, segmentation is not given and should be computed a priori. Video segmentation must result in a set of 10 to 120 seconds long potential targets, as a limit imposed by the benchmark organizers. We chose to segment videos by taking only 30 seconds of contiguous speech and cut at the following speech pause, as detected by the speech transcription system. Additionally, we run this segmentation another time, using an offset of one speech segment at the second pass, in order to obtain an overlapping segmentation. This results in 307,403 video segments with an average duration of 45 seconds.

Contrary to the previous dataset that contained videos provided by BCC, this dataset is formed with videos from the BlipTV video sharing platform and contains different user-submitted videos of with different topics, styles and languages. The dataset contained 14,838 videos an average duration of 13 minute. We used automatic speech transcripts provided by LIMSI and video keyframes provided with the dataset. We did not make use of visual concepts and metadata, although they were provided with the dataset.

3.2 Initial Representations

We represent transcripts and visual concepts of each anchor and target with a *Word2Vec* skip-gram model with hierarchical sampling [4], a representation size of 100 and a window size of 5. The visual concepts were sorted previous to learning and the representations of the words and concepts found within a segment were averaged [5]. This option worked best for our task.

Convolutional neural network representations were obtained by using the output of the last fully connected layers of AlexNet, VGG-16 and VGG-19, respectively. The deep convolutional neural networks are based on *Keras-Convnets*[1], a framework based on *Keras*, offering models already pretrained on *ImageNet*. All three convolutional neural network architectures yield a representation of size 4096. Since there are multiple keyframes in each video segment, aggregation done by averaging [19]. Averaged VGG-16 provide the best visual embedding, yielding a result of 70.67 % in precision at 10. A standard cosine similarity is used in all the experiments. The performance of the different methods is shown in Table 1.

3.3 Multimodal Embedding

Multimodal embedding with classical autoencoders and BiDNNs are tested. For a fair comparison, the sizes of the layers and, concordly, the representation dimensionality are the same for both architectures. Initial, single-modal representations are of size 100 for automatic transcripts and visual concepts. For representations obtained with convolutional neural network, the size is 4096. In case of simple concatenation, the multimodal representation sizes are clearly of 200 and 4196, respectively.

Multimodal autoencoders and BiDNNs were configured to yield a representation of size 1000. Bigger representation sizes (up to 4196) did not improve performance, while smaller representation sizes resulted in deteriorated results. All systems were trained with stochastic gradient descent (SGD) with Nesterov momentum, dropout of 20 %, in minibatches of 100 samples, for 1000 epochs (although convergence was achieved quite earlier). Each system had its weights randomly initialized by sampling from an appropriate uniform distribution [20] and was run five times. The average scores and their respective standard deviations due to random initialization are shown in Table 1.

3.3.1 Simple Multimodal Approaches

For a fair and complete comparison, we test two simple ways of combining multiple modalities: concatenation and linear combination of similarity scores [12]. There is no significant improvement when concatenating embedded transcripts and visual concepts. However, a simple concatenation of embedded transcripts and embeddings obtained with convolutional neural networks improves over each single-modal representation alone. For instance, combining VGG-16 embeddings with embedded transcripts yields 75.33 % (precision at 10) over the initial performance of 70.67 % and 58.67 % respectively.

3.3.2 Multimodal Embedding with Autoencoders

Multimodal autoencoders are the most common current method for obtaining multimodal embeddings. We implemented a model as illustrated in Figure 1 in the top part: a multimodal autoencoder with separate inputs and outputs and separate fully connected layers assigned to each

1. https://github.com/heuritech/convnets-keras

input/output. The two modalities are then merged in a central fully connected layer where the multimodal embedding is obtained. Since autoencoders with separate modalities perform better than simple autoencoders where the modalities are concatenated and used as one input/output pair [15], we didn't test the classical simple version but only the best performing one. This autoencoder architecture offers crossmodal translation by being additionally trained with one zeroed modality while asked to reconstruct both modalities. We implemented this architecture in *Keras*[2], with a central layer of size 1000. Bigger sizes did not improve the results but smaller ones did deteriorate them. The inputs, outputs and their associated fully connected layers were sized accordingly with the dimensionality of the input data.

Table 1 reports the results. It can be clearly seen that multimodal embedding performs better than each single modality by itself; e.g., combining embedded transcripts and VGG-19 features yields 74.73 %, compared to 58.67 % and 68.07 % respectively. However, in some cases, it seems that embeddings obtained in such a way do not yield significantly better results than simple methods. We believe this to be caused by the already good single representations and the fact that autoencoders have to train to represent the correct output with both modalities being present at their input and with one zeroed modality. In cases where the initial embeddings perform less (e.g., embedded visual concepts combined with embedded transcripts), autoencoders seem to improve in a more significant way.

3.3.3 BiDNN Multimodal Embedding

As explained in Section 2.2.3, BiDNNs try to address the problems found in classical multimodal autoencoders. We implemented a BiDNN comparable with the previously described autoencoder: a central fully connected layer yielding a representation of size 1000 and inputs/outputs dependent on the modalities used. BiDNNs behaved similarly to autoencoders as representation sizes bigger than 1000 did not bring any significant improvement while smaller ones deteriorated the performance. Each model was trained with five independent runs of 1000 epochs each, well after achieving convergence, the results were averaged and, together with their respective standard deviations, are reported in Table 1. We use this to report significance levels, computed with single-tailed t-tests, where comparing different methods.

Our implementation of BiDNNs was implemented in *Lasagne*[3] and is available[4] as an open source command-line tool that can be used both independently and as a *Python* module. In both cases it can be used to perform multimodal embedding and multimodal query expansion (filling of missing modalities with crossmodal translation) with a multitude of additional options.

Multimodal embedding with BiDNNs creates a common joint representation space where both modalities are projected from their initial representation spaces. This provides multimodal embedding superior to multimodal autoencoders and brings significant improvement. For instance, combining embedded transcripts with VGG-19 embeddings yields a precision at 10 of 80.00 %, compared to 58.67 % and 68.67 % respectively. All the other tested combinations also yielded better results and high quality multimodal embeddings.

3.3.4 BiDNN Single Modality Embedding

Although BiDNNs are trained in a multimodal setup, it is possible to embed only one modality by presenting in to the respective input and propagating the activations forward until the central representation layer. Results clearly show that each newly formed common representation space is significantly better than its respective original representation space. Automatic transcripts improve from 58.67 % to 66.78 %, visual concepts from 50.00 % to 54.92 % and VGG-19 embeddings from 68.67 % to 70.81 %. These results are obviously not as good as multimodal embeddings obtained by combining two modalities but they clearly show the improvement that BiDNNs bring even when used in a single-modal fashion and not only as a common space where representations from originally different representation spaces are projected.

3.3.5 BiDNN Crossmodal Query Expansion

BiDNNs naturally enable crossmodal expansion where a missing modality is filled in by translating from the other one. If a transcript is not available for a video segment, it is generated from the visual concepts and conversely. Using crossmodal query expansion so that all segments have all modalities, and concatenating them in their original spaces, we obtain, e.g., 62.35 % when combining transcripts and visual concepts (originally 58.00 %) and no significant improvement for pairs computed with high-performing deep convolutional neural networks and automatic transcripts. This is due to the relatively small number of samples with one missing modality, so filling the missing modalities does not have a big impact and the fact that the original representation spaces are not as good-performing as the new multimodal spaces.

3.4 Live Evaluation at TRECVID 2016

In the previous part, we evaluated different initial single-modal representations and different autoencoding methods and we have shown that focusing at crossmodal translations to perform multimodal fusion yields improved results over classical multimodal autoencoders, regardless of the initial representations. However, all the evaluations were done with a fixed groundtruth of limited size. Evaluation of arbitrary video segments was not possible in such a setup. To further analyze BiDNNs and evaluate whether they also work in a live setup, we decided to evaluate them through the TRECVID's 2016 video hyperlinking benchmark. To see not only how well the proposed system performs in regards of the systems proposed by other participants but also how much does it improve over the initial, disjoint single-modal representations, we decided to evaluate the best performing single-modal representations, according to our previous offline evaluation, together with the two modalities fused in a crossmodal fashion by BiDNNs. Each submission was made by finding the top 10 most similar video segments, for each given anchor, out of the 307,403 overlapping video

2. http://keras.io
3. https://github.com/Lasagne/Lasagne
4. https://github.com/v-v/BiDNN

TABLE 1
Comparison of the tested methods: precision at 10 (%) and standard deviation

Modalities	Method	P@10 (%)	σ (%)
Single-modal speech representations			
Word2Vec	**average**	**58.67**	-
PV-DM	-	45.00	-
PV-DBOW	-	41.67	-
Single-modal visual representations			
KU Leuven c., W2V	average	50.00	-
KU Leuven c., PV-DM	-	45.33	-
KU Leuven c., PV-DBOW	-	48.33	-
AlexNet	average	63.00	-
VGG-16	**average**	**70.67**	-
VGG-19	average	68.67	-
Simple multimodal approaches			
Transcripts, v.c.	concat	58.00	-
Transcripts, AlexNet	concat	70.00	-
Transcripts, VGG-16	concat	75.33	-
Transcripts, VGG-19	concat	74.33	-
Transcripts, v.c.	lin. comb.	61.32	3.10
Transcripts, AlexNet	lin. comb.	67.38	2.66
Transcripts, VGG-16	lin. comb.	71.86	4.11
Transcripts, VGG-19	lin. comb.	71.78	3.90
Multimodal autoencoders			
Transcripts, visual concepts		59.60	0.65
Transcripts, AlexNet		69.87	1.64
Transcripts, VGG-16		74.53	1.52
Transcripts, VGG-19		75.73	1.79
BiDNN single modality embedding			
Transcripts		66.78	1.05
Visual concepts		54.92	0.99
AlexNet		66.33	0.58
VGG-16		68.70	1.98
VGG-19		70.81	1.08
BiDNN multimodal embedding			
Transcripts, visual concepts		73.74	0.46
Transcripts, AlexNet		73.41	1.08
Transcripts, VGG-16		76.33	1.60
Transcripts, VGG-19		**80.00**	**0.80**
BiDNN query expansion			
Transcripts, visual concepts		62.35	0.25
Transcripts, AlexNet		70.11	1.25
Transcripts, VGG-16		75.33	0.10
Transcripts, VGG-19		74.33	0.10

segments. In case two proposed segments overlapped, only the most similar one was kept and the other was removed from the stack. Relatnedness was, as in the previous offline evaluation, determined by the cosine similarity of the evaluated video segment and the given target, using the chosen representation for that run (either one of the single-modal ones or the fused representation obtained with BiDNNs).

The results of the submitted runs, as well as the statistics of all the team that participated are given in Table 2. The official scores at TRECID's 2016 video hyperlinking task were given in precision at 5.

The live evaluation confirmed what we have previously shown in offline evaluations: focusing on crossmodal translations with added restrictions (namely enforced symmetry) yields improved multimodal embeddings. Not only did BiDNN improve the two initially disjoint modalities (40 % and 45 % into a representation that yields 52 % in terms of precision at 5) but it also proved that it obtaining

TABLE 2
Results of our submitted runs containing solely the initial modalities, fused modalities with BiDNN and overall statistics for all the participants of TRECVID's 2016 video hyperlinking benchmark.

Evaluation	P@5 (%)
Speech transcripts only	40
VGG-19 features only	45
BiDNN - transcripts and VGG-19	**52**
Min (all teams)	24
1st quartile (all teams)	32
Mean (all teams)	35
3rd quartile (all teams)	41
Max (all teams)	**52**

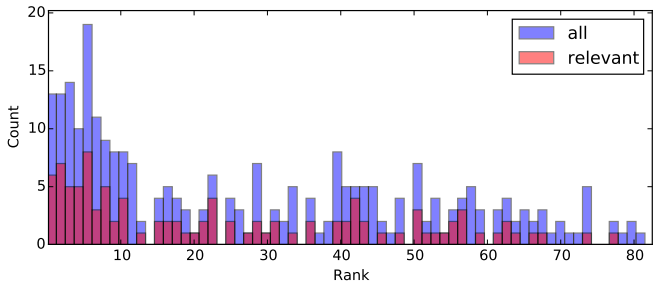

Fig. 3. Histogram illustrating the distribution of ranks of our proposed targets in the collection of targets proposed by all the participants, ordered by similarity according to BiDNN fused representations. The blue histogram represents all our proposed targets while the one in red solely those that we proposed and were judged relevant.

multimodal fusion in a crossmodal fashion outperformed multimodal fusion methods of the other teams and it defined the new state of the art and performed best in the overall submissions. This shows how a simple, unsupervised learning system, with no additional information and no hand crafted features can compete with more complex multimodal systems. It also proves that crossmodal translations with enforced symmetry improve over classic ways of multimodal fusion and define the new state of the art.

3.5 Post TRECVID 2016 Evaluation

While we have seen that BiDNNs define the new state of the art in a live evaluation, it is still interesting to further evaluate the overall system and see how would the different parts impact the evaluation if altered. After the TRECVID 2016 live evaluation was performed, a new groundtruth was again collected by the organizers, containing the targets proposed by all their participants and their relevance, as judged by AMT human evaluators. These targets are not consistent with our evaluated video segments as every participant had their own segmentation. We thus use our proposed system (multimodal representations of each video segment, fused with BiDNNs) to analyze the similarity of the targets proposed by all the participants for each anchor and the rank (determined by similarity) of the targets we proposed. On average, there are 79.19 proposed targets for each anchor.

Figure 3 illustrates the distribution of our proposed targets between all the targets that were proposed for a specific anchor by all the participants, according to the their

similarity given by multimodal embeddings obtained with a BiDNN model. While most of the targets are within the top 10 by similarity to the given anchor, we proposed many targets that are quite low ranking according to our own system. The average rank for our proposed targets is 27.98 and furthermore, 61.17 % of our proposed targets are not within the top 10 video segments when evaluating all the video segments proposed by all participating teams. According to BiDNN similarity, within the top 10 best targets for any given anchor, there are on average only 3.02 targets that we proposed. This indicates that, had we evaluated the same video segments as some other participants, the system could further be improved, at least in regards of similarity between targets and anchors given by BiDNN based representations.

Additionally, by performing an offline analysis on all the proposed targets (without retraining the model) and basically rerank the targets proposed by all the participants, in the same way we did in Section 3.3.3, we obtain a result of 49.56% in precision at 5, compared to 30.8% for the live evaluation, which strongly indicates that although offline evaluations are acceptable for choosing the best performing method, they largely differ from live evaluations in terms of performance. This could however change if large enough groundtruths are formed that contain relevance information between almost all the video segments, which is practically not feasible to obtain.

4 Conclusions

We started by analyzing different methods for obtaining continuous representations for the task of video hyperlinking by using automatic transcripts and visual information and methods to fuse them into a multimodal representation. Expectedly, visual embeddings obtained with deep convolutional neural networks outperformed embedded visual concepts and proved to be more relevant than automatic transcripts. Very deep VGG convolutional neural network architectures significantly outperformed the less deep AlexNet architecture. VGG-16 performed best and produced a single-modal visual embedding that yields 70.68 % in precision at 10. We have shown that the few downsides of autoencoders can affect their results and that BiDNNs successfully tackle these problems and clearly outperform multimodal autoencoders by a significant margin. Although VGG-16 performed better than VGG-19 in a single modal setup, the best performance was obtained by multimodal fusion of embedded automatic transcripts and embedded VGG-19 features, yielding a precision at 10 of 80.00 %.

We then chose the best performing setup and evaluated it in a live setup at TRECVID's 2016 video hyperlinking task. The live evaluation confirmed that focusing on crossmodal translations with added restrictions, namely in the form of symmetry, is a feasible method to perform multimodal fusion and BiDNNs defined the new state of the art by achieving the best performance at the challenge. After the post-evaluation groundtruth was formed, we used it to further analyze our model and found that, although our submissions were most often ranked between the top 10 ones, multimodal representations obtained with BiDNNs often favored other targets proposed by other participants. This indicates that further improvements can be made solely by changing the initial segmentation.

BiDNNs have already been shown to successfully tackle the downsides of multimodal autoencoders and to provide superior multimodal embeddings. In this work, we extensively tested BiDNNs both in offline and live evaluations in the context of video hyperlinking, which further reinforces the points already made in [15] and defines the new state of the art at the last TRECVID 2016 video hyperlinking benchmarking initiative.

References

[1] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *ACM Intl. Conf. on Multimedia*, 2014, pp. 7–16.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[3] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents." in *ICML*, vol. 14, 2014, pp. 1188–1196.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013.

[5] M. Campr and K. Ježek, "Comparing semantic models for evaluating automatic document summarization," in *Text, Speech, and Dialogue*, 2015.

[6] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[8] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang, "Eventnet: A large scale structured concept library for complex event detection in video," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 471–480.

[9] L. Jiang, S.-I. Yu, D. Meng, Y. Yang, T. Mitamura, and A. G. Hauptmann, "Fast and accurate content-based semantic search in 100m internet videos," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 49–58.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[12] C. Guinaudeau, A. R. Simon, G. Gravier, and P. Sébillot, "HITS and IRISA at MediaEval 2013: Search and hyperlinking task," in *Working Notes MediaEval Workshop*, 2013.

[13] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Intl. Conf. on Machine Learning*, 2011.

[14] H. Lu, Y. Liou, H. Lee, and L. Lee, "Semantic retrieval of personal photos using a deep autoencoder fusing visual features with speech annotations represented as word/paragraph vectors," in *Annual Conf. of the Intl. Speech Communication Association*, 2015.

[15] V. Vukotić, C. Raymond, and G. Gravier, "Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 2016, pp. 343–346.

[16] P. Over, J. Fiscus, G. Sanders, D. Joy, M. Michel, G. Awad, A. Smeaton, W. Kraaij, and G. Quénot, "Trecvid 2014–an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID*, 2014, p. 52.

[17] T. Search and H. T. at MediaEval 2014, "Maria eskevich and robin aly and david n. racca and roeland ordelman and shu chen and gareth j.f. jones," in *Working Notes MediaEval Workshop*, 2014.

[18] T. Tommasi, T. Tuytelaars, and B. Caputo, "A testbed for cross-dataset analysis," *CoRR*, vol. abs/1402.5923, 2014.

[19] V. Vukotic, C. Raymond, and G. Gravier, "Multimodal and Crossmodal Representation Learning from Textual and Visual Features with Bidirectional Deep Neural Networks for Video Hyperlinking," in *ACM Multimedia 2016 Workshop: Vision and Language Integration Meets Multimedia Fusion (iV&L-MM'16)*. Amsterdam, Netherlands: ACM Multimedia, Oct. 2016. [Online]. Available: https://hal.inria.fr/hal-01374727

[20] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in *Aistats*, vol. 9, 2010, pp. 249–256.

Vedran Vukotić Vedran Vukotić is a PhD student at IRISA and INSA Rennes. He received his B.Sc. and M.Sc. in computer science in 2012 and 2014 and his M.Sc. in nautical sciences in 2010. His main area of interests are unsupervised methods of obtaining multimodal representations with deep learning architectures and in particular multimodal fusion of speech and vision in video hyperlinking.

Christian Raymond Guillaume Gravier is senior research scientist at CNRS, France. He leads the Linkmedia research group focusing on content-based media analysis, indexing and linking with the ultimate goal of enabling better multimedia applications and new innovative services. His current research interests are in media analytics, multimedia collection modeling, deep learning and multimodality, graph-based methods for multimedia content representation.

Guillaume Gravier Christian Raymond received the M.S. in 2000 and the Ph.D. in 2005, both in computer science, from the University of Avignon, France.He was appointed in September 2009 associate professor at the INSA Rennes. He's member of LinkMedia team, devoted to multimedia document analysis. His research activities focus on speech understanding, machine learning for natural language processing, data driven stochastic approaches.