Multimodal and Crossmodal Representation Learning from Textual and Visual Features with Bidirectional Deep Neural Networks for Video Hyperlinking

Vedran Vukotić
INSA Rennes
IRISA & INRIA Rennes
Rennes, France
vedran.vukotic@irisa.fr

Christian Raymond
INSA Rennes
IRISA & INRIA Rennes
Rennes, France
christian.raymond@irisa.fr

Guillaume Gravier
CNRS
IRISA & INRIA Rennes
Rennes, France
guillaume.gravier@irisa.fr

ABSTRACT

Video hyperlinking represents a classical example of multimodal problems. Common approaches to such problems are early fusion of the initial modalities and crossmodal translation from one modality to the other. Recently, deep neural networks, especially deep autoencoders, have proven promising both for crossmodal translation and for early fusion via multimodal embedding. A particular architecture, bidirectional symmetrical deep neural networks, have been proven to yield improved multimodal embeddings over classical autoencoders, while also being able to perform crossmodal translation. In this work we focus firstly at evaluating good single-modal continuous representations both for textual and for visual information. Word2Vec and paragraph vectors are evaluated for representing collections of words, such as parts of automatic transcripts and multiple visual concepts, while different deep convolutional neural networks are evaluated for directly embedding visual information, avoiding the creation of visual concepts. Secondly, we evaluate methods for multimodal fusion and crossmodal translation, with different single-modal pairs, in the task of video hyperlinking. Bidirectional (symmetrical) deep neural networks were shown to successfully tackle downsides of multimodal autoencoders and yield a superior multimodal representation. In this work, we extensively tests them in different settings with different single-modal representations, within the context of video hyperlinking. Our novel bidirectional symmetrical deep neural networks are compared to classical autoencoders and are shown to yield significantly improved multimodal embeddings that significantly ($\alpha = 0.0001$) outperform multimodal embeddings obtained by deep autoencoders with an absolute improvement in precision at 10 of 14.1 % when embedding visual concepts and automatic transcripts and an absolute improvement of 4.3 % when embedding automatic transcripts with features obtained with very deep convolutional neural networks, yielding 80 % of precision at 10.

iV&L-MM'16, October 16 2016, Amsterdam, Netherlands

ISBN 978-1-4503-4519-4/16/10…$15.00

DOI: http://dx.doi.org/10.1145/2983563.2983567

Keywords

neural networks; deep learning; convolutional neural networks; CNN; deep neural networks; DNN; representation; embedding; multimodal; crossmodal; retrieval; video retrieval; video hyperlinking; video and text; autoencoder; bidirectional learning; tied weights; shared weights; multimodal fusion

1. INTRODUCTION

The seminal idea of video hyperlinking is to create hyperlinks between different videos and/or video segments based on their data. Each video consists of at least two data streams: a visual stream and an audio stream. A visual stream is represented by a set of consecutive images (frames) of which the most meaningful ones are keyframes. Keyframes (also known as intra-frames) are fully stored frames - frames where the complete information is stored in the video stream. Other frames (known as inter-frames) are expressed as a change from neighbouring keyframes. This is due to the fact that, in most videos, neighbouring frames contain a lot of redundant information. Keyframes provide the whole frame in the beginning, after the accumulated changes from the original previous keyframe are too big and after every scene change. These properties make keyframes a good source of visual information from where visual concept extraction, visual embedding, action or event recognition and other visual content analysis methods can be performed. 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 and/or video segments, 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). Continuous representations are also convenient for crossmodal translation and multimodal fusion with recent deep learning techniques.

Deep neural networks have been long known to produce

meaningful data representations [6], either as deep belief networks, autoencoders or as a combination of both. More recently, deep neural networks have been successfully deployed in tasks requiring consideration of multiple modalities. These tasks vary from retrieval [3, 19, 11], ranking [22] and classification tasks [2, 13] to generative tasks [19]. Data often consist of bimodal pairs such as images and tags [2], images and speech [3, 11], audio and video [13], but the systems exploiting them are not necessary bounded to those pairs.

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 [3], 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 analyze 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 focus on two source modalities: automatic transcription of the audio track and video keyframes. For automatic transcripts, we test different methods to obtain good and meaningful representations. Regarding video keyframes (visual information) we follow two different approaches: i) In the first approach, visual concepts extracted at each keyframe are used, treated as words and then tested with different methods to obtain representations in a continuous space. ii) The second approach utilizes state-of-the-art convolutional neural networks to provide high-quality representations directly from the image [17], without using intermediate interpretable (by a human observer) concepts.

After determining good single-modal representations for the task of video hyperlinking, we progress to analyzing different methods for obtaining multimodal embeddings while also allowing for crossmodal translation. Classical autoencoders and bidirectional symmetrical deep neural networks are evaluated and compared. The focus is put on combining representations obtained from automatic transcripts with embeddings obtained with deep convolutional neural networks with bidirectional symmetrical deep neural networks, which have been shown to outperform classical autoencoders in a multimodal setup with automatic transcripts and visual concepts [21].

The seminal idea of bidirectional symmetrical deep neural networks is to keep separate deep neural networks for each cross-modal translation while tying the weights of the middle layers between the neural networks so as to yield a common multimodal representation. In this setting, the common middle layer acts as a multimodal representation space that is attainable from either one of the modalities and from which we can attain either one of the modalities. This avoids common downsides present in classical autoencoders (see Sec. 2.2.2) and yields a improved multimodal representation. We provide empirical proof of the superiority of such embeddings in different setups, involving different initial single-modal representations.

2. METHODOLOGY

In this section, we analyze methods for obtaining good single modal representations and methods for embedding multiple single modalities into improved multimodal representation spaces. We start with different methods to represent automatic audio transcriptions of the video segments, we progress to methods to represent keyframes of the video segments and we conclude with different methods to create joint multimodal representations, as well as allowing for crossmodal translation. Where appropriate, for some single-modal cases, methods for aggregating multiple embeddings into a single single-modal representation are also tackled.

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 in practice and would include a human component in the system. Video keyframes are considered in two different settings: using *ImageNet* concepts [15] 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 [10] and Word2Vec [12]. 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 [1] 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: e.g., *"n02121808"* indicates *"Any domesticated member of the genus Felis (Domestic cat, house cat, Felis domesticus, Felis catus)"* and includes all related subcategories. We treat each visual concept as a word and utilize it to obtain word embeddings (with Word2Vec or paragraph vectors) representing the visual information of a video segment provided by its visual concepts in a continous representation space.

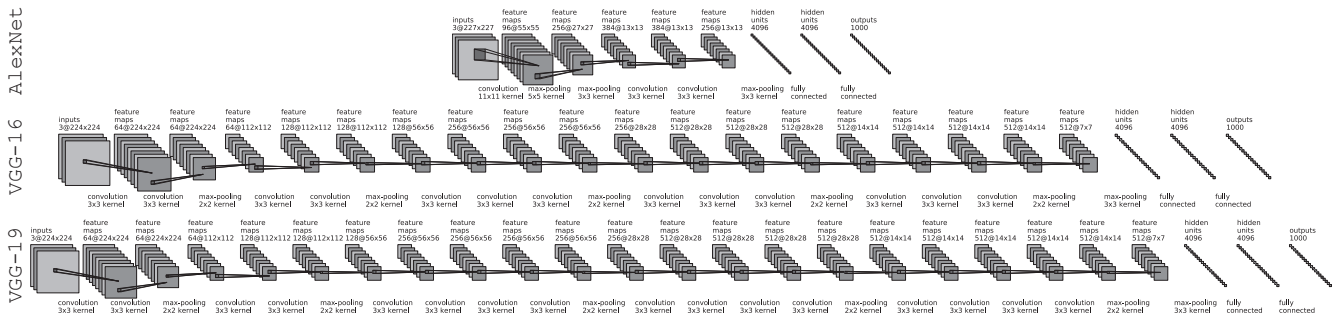

Figure 1: Simplified comparison of the CNN architectures used in this work: AlexNet (top), VGG-16 (middle) and VGG-19 (bottom). For simplicity, only the main layers are shown. Merging, reshaping, padding and other layers are not illustrated.

2.1.3 *Representing Visual Information with CNN Features*

Convolutional Neural Networks (CNNs) provide state-of-the art visual descriptors [17] that have been shown to perform well both in computer vision applications [8, 23] and in video summarization tasks [7]. In this work, we test three different state-of-the art deep convolutional neural network architectures, namely AlexNet, VGG-16 and VGG-19. Figure 1 illustrates, in a simplified manner (only main layers are shown: convolutional, pooling and fully connected layers), the architectures of such networks. AlexNet [9] is deep convolutional neural network of medium depth, with 3 convolutional layers, 3 max-pooling layers and a set of fully connected layers at the end. VGG networks [18] are very deep convolutional neural network architectures defined by the *Visual Geometry Group*. We use two VGG architectures, namely VGG-16 and VGG-19, with 16 and 19 "weight layers" respectively. The VGG-16 architecture consists of 13 convolutional layers, 5 max-pooling layers and 3 fully connected layers. The VGG-19 architecture consists of 16 convolutional layers, 3 max-pooling layers and 3 fully connected layers.

2.1.4 *Aggregation*

Depending on the subtask and the method used, the resulting representations might require aggregation, e.g., to represent all the automatic transcripts of a video segment with Word2Vec or to represent all the keyframes of a video segment. Some methods, on the other side, do not require additional aggregation (e.g., paragraph vectors). In this work, we tested two means of aggregating descriptors: simple averaging and Fisher vectors.

2.2 Multimodal and Crossmodal Approaches

In this section, we analyze methods for creating multimodal embeddings and allowing for crossmodal translation. We compare classical autoencoders and bidirectional symmetrical deep neural networks, where both can do crossmodal translation and provide a joint multimodal embedding. Autoencoders are one of the most commonly used methods for obtaining multimodal representations. Single-modal autoencoders often include forced symmetry and are used for dimensionality reduction. Bidirectional symmetrical DNNs are based on the idea of learning crossmodal mappings in both directions while applying restrictions to force symmetry in deep neural networks in order to form a common multimodal embedding space that is common to the two crossmodal mappings.

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 [5] of the similarities obtained by comparing each of the two modalities. This late fusion avoids multimodal models and might offer slightly better results than simple concatenation (a linear combination can slightly correct the differences by giving more importance to one modality and implicitly reranking similarity scores by different modalities). However, a linear combination requires cross-validation of the parameters, which often might be dependent on the specific dataset and the single modal representations used. 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, 11]. 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 2. One common hidden layer is used for creating a joint multimodal representation. Sometimes, one modality is sporadically removed from the input to make the autoencoder learn to represent both modalities from one. The activations of the hidden layer are used as a multimodal joint representation. This enables autoencoders to also provide crossmodal mapping [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.

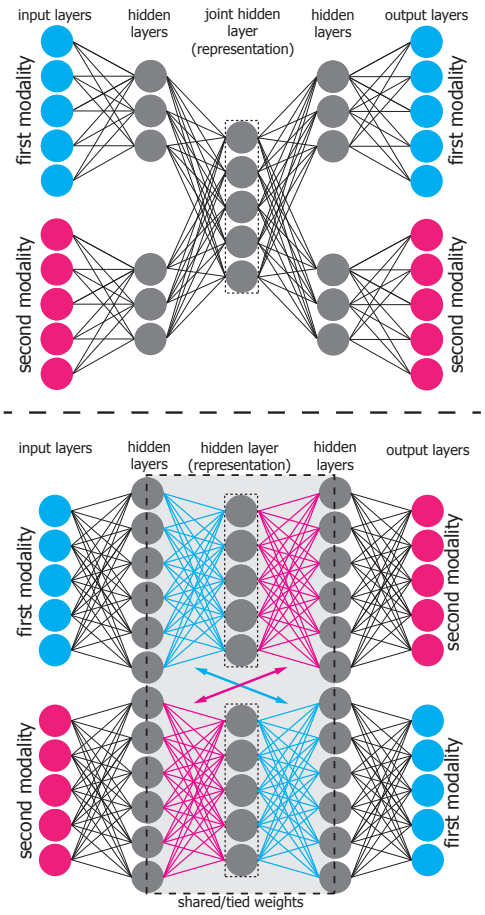

Figure 2: Illustration of the architectures of a classical multimodal autoencoder (top) and a bidirectional symmetrical deep neural network (bottom)

- 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 [21], which we discuss next.

2.2.3 *Bidirectional Deep Neural Networks*

In bidirectional deep neural networks, 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 bottom of Figure 2. Implementation-wise the variables representing the weights are shared across the two networks and are in fact the same variables. 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. A joint representation in the middle of the two crossmodal mappings is also formed while learning.

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), $\mathbf{x_{m_i}}$ 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:

$$\mathbf{h_1^{(1)}} = f(\mathbf{W_1^{(1)}} \times \mathbf{x_{m_1}} + \mathbf{b_1^{(1)}})$$
$$\mathbf{h_2^{(1)}} = f(\mathbf{W_2^{(1)}} \times \mathbf{x_{m_2}} + \mathbf{b_2^{(1)}})$$

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

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

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

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.

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 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 (automatic transcripts and visual concepts or CNN features, depending on the setup), 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 (either only transcripts or only visual information), 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 (either with only one or both modalities).

Finally, segments are then compared as illustrated in Figure 3: for each video segment, the two modalities are taken (embedded automatic transcripts with either embedded vi-

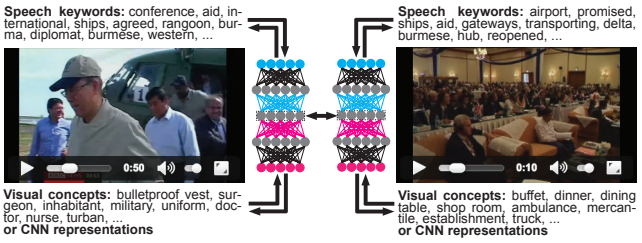

Figure 3: Video hyperlinking with bidirectional deep neural networks: two video segments, both with two modalities (automatic transcripts and either KU Leuven visual concepts or CNN features of each keyframe) are compared after their multimodal embeddings are computed

sual concepts or embedded CNN representations) and a multimodal embedding is created with a bidirectional deep neural network. The two multimodal embeddings are then simply compared with a cosine distance to obtain a similarity measure.

3. EXPERIMENTS

In this section, we first describe the dataset used for evaluation of the previously described methods. After that, we proceed to evaluating the different methods described in Section 2.1 for creating single-modal representation (both of automatic transcripts and visual information). Finally, we describe the details of the multimodal and crossmodal methods described in Section 2.2 and evaluate their performance under different settings.

3.1 Dataset

The generation of hyperlinks within video segments is the focus of the "Search and Hyperlinking" evaluation at MediaEval and more recently at TRECVid [14]. 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 [16]. In this task, 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. For each video, multiple data and modalities are available. In this work, we used a combination of two modalities: either automatic transcripts of the audio track and KU Leuven [20] visual concepts or automatic transcripts of the audio track and descriptors of each keyframe obtained with different convolutional neural network architectures. Both automatic transcripts and KU Leuven visual concepts are provided as part of the dataset. KU Leuven visual concepts consists of multiple ImageNet [15] classes detected in each keyframe with a CNN architecture and provided as a textual description together with each keyframe.

In practice, targets are not given and have to be defined automatically before assessing their relevance to each of the 30 anchors provided. Evaluation of relevance is thus done post-hoc on Amazon Mechanical Turk (AMT). In this paper, we consider a set of targets made of the top-10 targets that each participating team proposed for each anchor, along with the relevance judgments from AMT.

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. Regarding keyframes, there are in total 371,664 keyframes for an average of 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.

We implemented the autoencoder described in Section 2.2.2 in *Keras*[1] and bidirectional symmetrical deep neural networks, described in Section 2.2.3 in *Lasagne*[2]. The deep convolutional neural networks are based on *Keras-Convnets*[3], a framework based on *Keras* offering models already pretrained on *ImageNet*. Our implementation of bidirectional (symmetrical) deep neural networks is now 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.

3.2 Choice of Initial Representations

The performance of the different methods is shown in Table 1. We chose to represent the transcripts and visual concepts of each anchor and target with a *Word2Vec* skip-gram model with hierarchical sampling [12], 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 [1]. 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. All three convolutional neural network architectures yield a representation of size 4096. Since there are multiple keyframes in each video segment, aggregation was either done by averaging or by using Fisher vectors. The average proved to be stable and provide solid representations based on AlexNet, as well as the best representations, based on VGG-16 and VGG-19. For AlexNet, Fisher vectors provided slightly better results (with a previous dimensionality reduction with PCA to a size of 64 and GMM with 64 mixtures). Averaged VGG-16 provide the best visual embedding, yielding a result of 70.67 % in precision at 10. A standard cosine distance is used in all the experiments as a measure of similarity.

3.3 Multimodal Embedding

Multimodal embeddings with both classical autoencoders and bidirectional deep neural networks 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.

[1] http://keras.io
[2] https://github.com/Lasagne/Lasagne
[3] https://github.com/heuritech/convnets-keras
[4] https://github.com/v-v/BiDNN

Table 1: Single modal representations of automatic transcripts and visual information

Representation	Aggregation	P@10 (%)
Automatic transcripts		
Word2Vec	**average**	**58.67**
Word2Vec	Fisher	54.00
PV-DM	-	45.00
PV-DBOW	-	41.67
Visual information		
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
AlexNet	**Fisher**	**65.00**
VGG-16	**average**	**70.67**
VGG-16	Fisher	64.67
VGG-19	**average**	**68.67**
VGG-19	Fisher	66.00

In case of simple concatenation, the multimodal representation sizes are clearly of 200 and 4196, respectively.

Multimodal autoencoders and bidirectional deep neural networks 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 mini-batches 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 [4] and was run five times. The average scores (precision at 10) and their respective standard deviations due to random initialization are shown in Table 2 for all methods. Since there are many different combinations of systems and initial modalities, we report only the best performing initial, single modal representations in combination with the different systems: averaged transcripts, averaged visual concepts and averaged representations obtained with deep convolutional neural networks (even in the case of VGG-16, where Fisher vectors provided a better single-modal score, averaged VGG-16 features performed better when embedded and it is the case we report).

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 [5]. 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. A linear combination of similarities, on the other hand, does not offer a multimodal embedding but might be simpler (often used for relevance reranking) over simple concatenation, at the cost of having to optimize the parameters on another dataset and at the cost of higher variance.

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 2 in the top part: a multimodal autoencoder with separate inputs and outputs and separate fully connected layers assigned to each 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 [21], we didn't test the classical simple version but only the better performing one. This autoencoder architecture offers cross-modal translation by being additionally trained with one zeroed modality while asked to reconstruct both modalities. We implemented the described autoencoder 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 2 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, bidirectional deep neural networks try to address the problems found in classical multimodal autoencoders. We implemented a bidirectional deep neural network 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. Bidirectional deep neural networks behaved similarly to autoencoders as representation sizes bigger than 1000 did not bring any significant improvement while smaller ones deteriorated the performance. This confirms the choice of the dimensionality of the new multimodal representation by two independent methods. Each bidirectional deep neural network was trained with five independent runs of 1000 epochs each, although they converged earlier, the results were averaged and, together with their respective standard deviations, are reported in Table 2. Significance levels of improvements are computed with single-tailed t-tests and reported where appropriate.

Multimodal embedding with bidirectional deep neural networks creates a common joint representation space where both modalities are projected from their initial representation spaces. This provides superior multimodal embeddings that bring 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.

3.4 BiDNN Single Modality Embedding

Although bidirectional deep neural networks 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 forwards until the central representation layer). Doing so might offer an insight about the new representation space, common for both modalities, and its performance compared to the original representation spaces. 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 bidirectional deep neural networks 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.

3.5 BiDNN Crossmodal Query Expansion

Bidirectional deep neural networks 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, 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. Also, the original representation spaces are used and perform less good, as shown in Sections 3.3.3 and 3.4, than the new common representation spaces obtained with bidirectional deep neural networks.

Table 2: Comparison of the tested methods: precision at 10 (%) and standard deviation

Modalities	Method	P@10 (%)	σ (%)
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

4. CONCLUSIONS

In the first part of this work we analyzed different methods for obtaining continuous representations for the task of video hyperlinking by using automatic transcripts and visual information. 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.

The second part of this work was focused on multimodal embedding. Other than simple methods to utilize multimodal information, multimodal autoencoders and novel bidirectional deep neural networks were evaluated. We have shown that the few downsides of autoencoders can affect their results and that bidirectional deep neural networks 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 %.

Bidirectional (symmetrical) deep neural networks have already been shown to successfully tackle the downsides of multimodal autoencoders and to provide superior multimodal embeddings. In this work, we extensively tested bidirectional deep neural networks under different setups and with different single-modal representations in the context of video hyperlinking, which further reinforces the points already made in [21].

Following the results indicating superior new joint representation spaces, we evaluate the representation spaces formed with bidirectional deep neural networks which clearly outperform the original representation spaces even when evaluated solely for each modality separately. Every evaluated single modality (embedded automatic transcripts, embedded visual concepts and embeddings obtained convolutional neural networks) improved when projected to the new representation space obtained with bidirectional deep neural networks, which proves the effectiveness of the symmetrical crossmodal mappings learned by bidirectional deep neural networks and especially of the common representation space formed by the two crossmodal projections with enforced symmetry.

5. REFERENCES

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

[2] M. Cha, Y. Gwon, and H. T. Kung. Multimodal sparse representation learning and applications. *CoRR*, abs/1511.06238, 2015.

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

[4] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256, 2010.

[5] 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.

[6] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[7] 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*, pages 49–58. ACM, 2015.

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

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

[10] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

[11] 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.

[12] 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.

[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] 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*, page 52, 2014.

[15] 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)*, 115(3):211–252, 2015.

[16] 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.

[17] 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*, pages 806–813, 2014.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[19] N. Srivastava and R. Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *Intl. Conf. on Machine Learning*, 2012.

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

[21] 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*, pages 343–346. ACM, 2016.

[22] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.

[23] 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*, pages 471–480. ACM, 2015.