Bidirectional Joint Representation Learning with Symmetrical Deep Neural Networks for Multimodal and Crossmodal Applications

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

Common approaches to problems involving multiple modalities (classification, retrieval, hyperlinking, etc.) 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. In this work, we propose a flexible crossmodal deep neural network architecture for multimodal and crossmodal representation. By tying the weights of two deep neural networks, symmetry is enforced in central hidden layers thus yielding a multimodal representation space common to the two original representation spaces. The proposed architecture is evaluated in multimodal query expansion and multimodal retrieval tasks within the context of video hyperlinking. Our method demonstrates improved crossmodal translation capabilities and produces a multimodal embedding that significantly outperforms multimodal embeddings obtained by deep autoencoders, resulting in an absolute increase of 14.14 in precision at 10 on a video hyperlinking task ($\alpha = 10^{-4}$).

Keywords

neural networks; deep learning; representation; embedding; multimodal; crossmodal; retrieval; video retrieval; video hyperlinking; image and text; autoencoder; bidirectional learning; tied weights; shared weights

1. INTRODUCTION

Deep neural networks have been long know to produce meaningful data representations [5], either as deep belief networks, autoencoders or 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 [4, 9, 6], ranking [10] and classification tasks [2, 8] to generative tasks [9]. Data often consist of bimodal pairs such as images and tags [2], images and speech [4, 6], audio and video [8], 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 [4], 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 classification tasks.

In this work, we present a novel deep neural network architecture for crossmodal mapping and multimodal embedding. The seminal idea of the approach is to keep separate deep neural networks for each modality 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 common multimodal representation space that is attainable from either one of the modalities and from which we can attain either one of the modalities. As a proof of concept, experimental evidence of the benefits of the architecture is given by multimodal query expansion and multimodal retrieval tasks within the context of video hyperlinking. Extensions to other multimodal and crossmodal tasks remains straightforward.

2. METHODOLOGY

In this section, we analyze two methods for creating joint multimodal representations: multimodal autoencoders and

ICMR 2016, June 6, 2016, New York, USA.

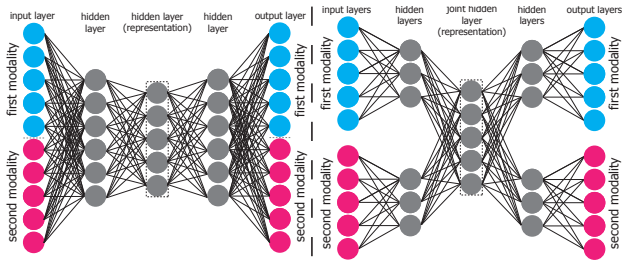

Figure 1: Two typical autoencoder architectures: left - concatenated representations at input and output, all hidden layers are joint; right - separated inputs, outputs and hidden layers, one hidden layer in common

our proposed method, i.e., bidirectional learning with deep neural networks with forced symmetry. Both approaches can do both 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. Our method is 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.

2.1 Multimodal Autoencoders

Two typical autoencoders are shown in Figure 1. The first (left) one illustrates a common approach that consists in concatenating the representations [8, 6] of the two modalities and training the autoencoder to reconstruct the data presented as input. The hidden layer in the middle is then used to obtain a joint multimodal representation (multimodal embedding).

The second (right) architecture is quite similar but has separate inputs and outputs (one for each modality) and separate hidden layers. One hidden layer in common 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 [8] in addition to a joint representation.

2.2 Bidirectional Representation Learning - Deep Neural Networks with Tied Weights

Previously described multimodal autoencoders include fully connected hidden layers that interact with both modalities. In other terms, the two modalities are available on the input and/or the two are reconstructed. The middle hidden layer represents a multimodal embedding and crossmodal translation is possible through this layer. However, mixing modalities is less optimal than directly mapping from one modality to another and back.

We propose a variation of the second autoencoder that implements a deep neural network translating from one modality to the other and back. 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

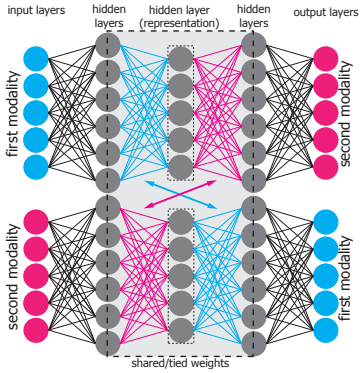

Figure 2: Proposed architecture: training is done in both directions; a shared representation is created by tying the weights (sharing the variables) and enforcing symmetry in the central part

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 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:

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

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

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

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

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

Table 1: Available modalities for anchor/target pairs

		Targets			
		Both	Transcripts	Visual c.	None
Anchors	Both	10,525	236	528	11
	Transcripts	0	0	0	0
	Visual c.	957	15	68	0
	None	0	0	0	0

$\mathbf{h_1^{(2)}}$ and/or $\mathbf{h_1^{(2)}}$. If both modalities are available, a concatenation of their embedding is used. If only one modality is available its embedding is duplicated. This allows for easy comparison, independently of modality availability.

3. EXPERIMENTS

3.1 Dataset and Initial Representations

We tested the symmetric bidirectional architecture within the framework of the video hyperlinking task using the MediaEval 2014 dataset and the respective groundtruth that was collected as part of the challenge [3]. 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 two modalities: the automatic transcripts of the audio track and KU Leuven visual concepts. 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 the 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. Table 1 illustrates the different cases of modality availability within the dataset. 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 chose to represent the transcripts and visual concepts of each anchor and target with a *Word2Vec* skip-gram model with hierarchical sampling [7], 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 (see Table 2) and additionally, our work is focused on cross-modality and joint multimodal representations, thus the choice of the input representations is not crucial. A standard cosine distance is used in all the experiments.

We implemented the two autoencoders described in Section 2.1 in *Keras*[1] and our proposed bidirectional symmetrical deep neural network in *Lasagne*[2]. All embeddings have a dimension of 100 as larger dimensions did not bring any significant improvement. The autoencoders we tested had architectures with 200-100-200 hidden layers, as had the symmetrical bidirectional networks. Other sizes were also tested but performed worse or not better and were not included in the comparison. Since this is unsupervised learning, the learning was performed on the part of the dataset that contains both transcripts and visual concepts and tested on the whole dataset.

[1]http://keras.io

[2]https://github.com/Lasagne/Lasagne

Table 2: Comparison of initial (single-modal) representations. Precision at 10 (%) obtained with TF-IDF, Word2Vec and Paragraph Vectors

Representation	Transcripts	Visual concepts
LSI & TF-IDF	23.67	18.33
Avg. Word2Vec	**58.67**	**50.00**
PV-DM	45.00	45.33
PV-DBOW	41.67	48.33

Table 3: Comparison of the tested methods: precision at 10 (%) and standard deviation

Method	P@10	σ
Baseline		
Only transcripts	58.67	-
Only visual concepts	50.00	-
Linear combination	61.32	3.1
Autoencoders		
Embedded tran. and v. c. with AE #1	57.40	1.24
Embedded tran. and v. c. with AE #2	59.60	0.65
Bidirectional DNNs with tied weights		
Embedded transcripts	70.43	0.46
Embedded visual concepts	54.92	0.99
Embedded transcripts and visual c.	**73.74**	**0.82**
Expanded transcripts	58.16	0.24
Expanded visual concepts	55.75	0.13
Expanded transcripts and visual c.	62.35	0.25

Table 3 reports the performance of the different methods. Using only transcripts yields a precision at 10 (P@10) of 58.67 % while using only visual concepts yields 50.00 %. Combining the two modalities in different linear combinations yields an improved P@10 to 61.32 %.

3.2 Crossmodal Query Expansion

Bidirectional symmetrical architectures enable crossmodal expansion where a missing modality is filled in by translating from the other one. If the transcript is not available for a segment, it is generated from the visual concepts and conversely. Using query expansion so that all segments have all modalities, we obtain 55.75 % for the visual concepts and 58.16 % for the transcripts. The first difference is significant ($\alpha = 0.001$) improvement, while the second is not. This is due to the relatively small number of samples with one missing modality, so filling the missing modalities does not have a big influence.

3.3 Multimodal Embedding

A multimodal representation created with the two autoencoders (AE) of Sec. 2.1 yields 57.40 % (left part of Fig. 1) and 59.60 % (right part of Fig. 1). With symmetrical bidirectional networks, the two modalities are handled in the

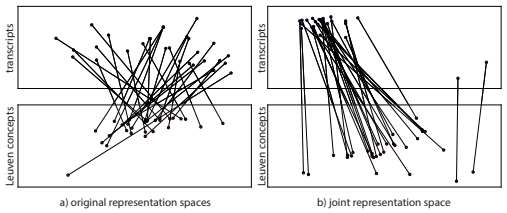

Figure 3: Mapping the two modalities of the 30 anchors (first two components): a) in the original representation spaces b) after embedding them

following manner to generate a multimodal embedding: 1) the modalities are both present for both the target and the anchor: all four representations are projected into the new multimodal representation space 2) one segment (anchor or target) has just one modality while the other has both: everything is projected into the new multimodal representation space, the one available modality of one segment being compared against the two modalities of the other 3) segments have only one modality: both modalities are projected into the new multimodal representation space.

When embedding only the transcripts to the new joint multimodal representation space with our proposed architecture, we obtain a P@10 of 70.43 %. Embedding only the visual concepts yields 54.92 %. The biggest improvement is achieved when embedding the two modalities, which yields 73.74 %, a significant ($\alpha = 10^{-4}$) improvement over autoencoders of 14.14 in P@10.

Figure 3 illustrates the first two principal components of the two modalities, in their original representation spaces (left) and the first two principal components of their representation in the joint embedding space. The 30 anchors that are shown display a more ordered structure in the embedding space then in their original space, indicating that projections to a common representation space are successful.

Figure 4 shows a comparison of the two possible usage of bidirectional symmetrical architectures: for multimodal expansion (as a crossmodal translator) and as a multimodal embedder that translates the two modalities to a joint representation space that has been trained in a bidirectional way and that represents a common space for the crossmodal translation. The graph shows P@10 values for 10k training iterations. For reference, also the three baseline values are included (original transcripts, visual concepts and their linear combination). Multimodal query expansion improves P@10, not by a large margin due to the small number of samples missing one modality but it proves successful crossmodal translation. Embedding into the new joint representation space improves P@10 with a bigger margin.

4. CONCLUSIONS

Our proposed deep neural network architecture with tied weights is trained bidirectionally from the first modality to the second and from the second modality to the first and creates a crossmodal mapping between the two representation spaces that can be successfully used in multimodal query expansion. Due to its enforced symmetry, a joint multimodal embedding is also created that further improves multimodal data representation. We have shown that projecting into this space is significantly better than translating between the initial spaces, even when using only one modality. Projecting both modalities into the joint multimodal representation space yields the best results on our task. We expect that this architecture may provide similar improvements for

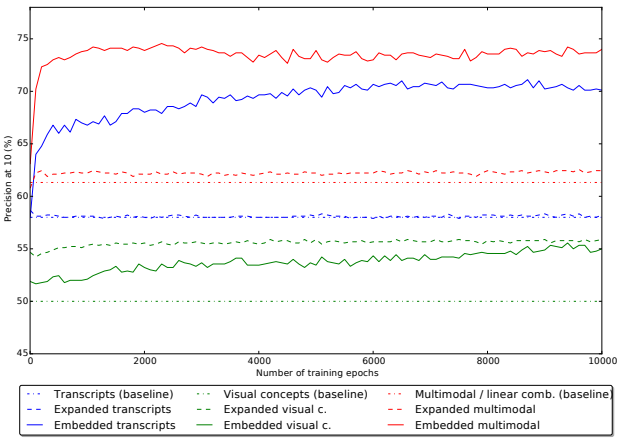

Figure 4: Comparison of P@10 for different training epochs (averages across multiple runs) and the baseline (original representations)

other tasks that would be the subject of future works.

5. ACKNOWLEDGMENTS

Partially funded via the CominLabs excellence laboratory financed by the National Research Agency under reference ANR-10-LABX-07-01.

6. 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] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. Jones. The search and hyperlinking task at MediaEval 2014. In *Working Notes MediaEval Workshop*, 2014.

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

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

[6] 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.

[7] 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.

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

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

[10] 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.