Multimodal Learning via Exploring Deep Semantic Similarity

Di Hu
OPTIMAL, NWPU
127 West Youyi Road
Xi'an Shaanxi, China
dtao@mail.nwpu.edu.cn

Xiaoqiang Lu
OPTIMAL, CAS
Xi'an Institute of Optics and Precision Mechanics, CAS
Xi'an Shaanxi, China
luxiaoqiang@opt.ac.cn

Xuelong Li
OPTIMAL, CAS
Xi'an Institute of Optics and Precision Mechanics, CAS
Xi'an Shaanxi, China
xuelong_li@opt.ac.cn

ABSTRACT

Deep learning is skilled at learning representation from raw data, which are embedded in the semantic space. Traditional multimodal networks take advantage of this, and maximize the joint distribution over the representations of different modalities. However, the similarity among the representations are not emphasized, which is an important property for multimodal data. In this paper, we will introduce a novel learning method for multimodal networks, named as *Semantic Similarity Learning* (SSL), which aims at training the model via enhancing the similarity between the high-level features of different modalities. Sets of experiments are conducted for evaluating the method on different multimodal networks and multiple tasks. The experimental results demonstrate the effectiveness of SSL in keeping the shared information and improving the discrimination. Particularly, SSL shows its ability in encouraging each modality to learn transferred knowledge from the other one when faced with missing data.

Keywords

Multimodal learning, semantic similarity, deep learning

1. INTRODUCTION

In the real world, information can be expressed in various kinds of modalities. Typically, images usually come with assigned accompanying tags for describing the same content, regular speech contains audio signal and corresponding lip movements, and environment can be described in both image and 3D depth. These modalities can jointly provide more valuable information than single modality, thus lots of tasks, including image-tag retrieval [14] and speech recognition [6], take consideration of multimodal input.

For each multimodal task, as data modalities generated from the same event are processed in different manners, they own distinct statistical properties which make it difficult to capture patterns across them. Recently, multimodal deep networks have been proposed to learn the shared representation across data modalities after learning each modality with single deep network, which takes advantages of the efficiency of deep network in producing useful representation for image, audio, and text [2]. In such multimodal networks, the generated features from different modalities are considered semantically relevant. Actually, as they share similar semantic information of the same entity, the association among data modalities can be enhanced and provide more valuable information for each one, especially in the case of missing modality. However, this property is not fully explored just by maximizing the joint distribution over modalities in the previous multimodal networks [11, 14, 7], so that the shared representation is easily affected by specific property of single modality and can not be held.

In this paper, we propose a novel learning algorithm for multimodal networks, which encourages the learned representation to keep the shared semantic information across modalities (e.g. audio, video, and text) and reduce the interference from specific modality. Specifically, as the generated high-level features of modalities are usually considered to have similar semantic, they should have similar contribution for the shared hidden layer. Based on this, we compare the hidden units activated by each modality, and enhance the semantic similarity by reducing the difference between them, which is named as *Semantic Similarity Learning* (SSL). To measure the similarity among modalities in different views, several functions are attempted, and the constituted algorithm is performed with multiple multimodal networks in several multimodal tasks. As expected, SSL keeps, even enhances the shared representation, which results in more discriminative representation. Besides, our algorithm shows its effectiveness in dealing with missing modality with the help of learned transferred knowledge.

In the following sections, we first survey the related work about multimodal deep networks in Section 2. In Section 3, we present our proposed learning algorithm for the networks. We then valid the algorithm on different tasks with audio, image, and text in Section 4. Section 5 concludes this paper.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM' 16, October 15-19, 2016, Amsterdam, Netherlands

© 2016 ACM. ISBN 123-4567-24-567/08/06...$15.00

DOI: 10.475/123_4

2. RELATED WORK

In recent years, deep learning methods have performed its effectiveness in feature learning. And the features generated from multimodal networks are considered semantically correlated [14]. In the early work of multimodal deep learning, Ngiam *et al.* [11] proposed a framework that learns the layers of modality-specific network firstly. Then the shared representation across the generated features of modalities are jointly learned using higher networks. Finally, the whole network is fine-tuned to minimize reconstruction errors of both modalities. This kinds of network is called *Multimodal Deep Auto-encoder* (MDAE). Different from the fusion strategy above, Andrew *et al.* [1] proposed *Deep Canonical Correlation Analysis* (DCCA) which measures the linear correlation between generated representations from each single network. They find that deep networks have the ability to transform complex nonlinear data into highly linear correlation and make them share more semantic information. Recently, Wang *et al.* [16] made a comparison between MDAE and DCCA, and attempted to combine them into a kind of *Deep Canonical Correlation Autoencoder* (DCCAE) that is expected to have advantages from both sides. But the simple combination has weakness in emphasizing the shared representation with the help of semantic similarity. Besides, Shu *et al.* [12] presented a multimodal framework to translate cross-domain knowledge based on the generated representations, but they used weakly parameter-shared setting instead of the general joint learning. This framework is more flexible but easier influenced by domain-specific features due to the relaxed constraint for parameters [12].

3. SEMANTIC SIMILARITY LEARNING

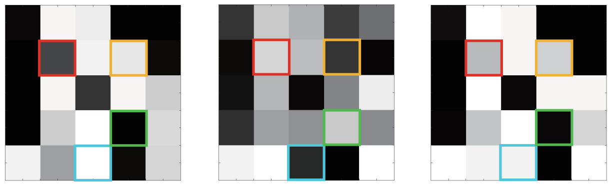
In this section, we first explore the property of the traditional *Maximum Joint Likelihood Learning* (MJLL) on the shared representation learning, then introduce our proposed learning method.

3.1 Multimodal Learning

The standard multimodal network is a modification of *Restricted Boltzmann Machine* (RBM), which is an undirected graphical model that defines a probability distribution of the generated features of different modalities using shared hidden layers [6], named as *Multimodal RBM* (MRBM[1]). The traditional MJLL for MRBM is to maximize the joint distribution over the high-level features of modalities, which contains not only the conditional likelihood across different modalities but also the likelihood of each modality [13]. Although the shared information exists in both modalities is more credible than the single modality, MJLL has weakness in keeping the shared information under the influence of specific modality.

To further explore the effect of single modality on the shared layer of MRBM with the MJLL training, we adopt the general multimodal task, *Audiovisual Speech Recognition* (AVSR), which makes use of the information from both audio and visual modalities for recognition. Inspired by denoising auto-encoders [15], we separately take audio and visual representation as the input to the MRBM and set the other one zero at testing time. Meanwhile, we also take the concatenation of them as the inputs as usual. Fig. 1 shows the activation state of the shared hidden units of MRBM. It's clear to find that both modalities have activated similar hidden units, which confirms the shared information between them, so that they have nearly common influence on the shared representation, as shown in Fig. 1(c). However, there're also some units a bit different from them. To be specific, these units are separated into four groups with different colors. In these groups, different modalities have opposite influences on the shared representation. This is because some modality-specific properties are not relevant. For example, the shape size of mouth when a person is speaking or pronouncing a certain letter has little correlation with the corresponding voice. In fact, these cases may make the shared representation confused as described above, which should be penalized to keep the shared information across modalities. In this paper, we propose to learn the shared representation via reducing the differences between the high-level features based on the MJLL with the perspective of semantic similarity, which actually encourages the shared representation to keep the common information across modalities and reduce the influence from specific modality.

[1]Detailed presentation can be found in the supplementary material.



(a) (b) (c)

Figure 1: Visualization of activation state of the hidden units of MRBM with different inputs to the visible layer after MJLL learning. (a) Only audio modality. (b) Only visual modality. (c) Both audio and visual modality. The unit in complete white means fully activated state. Best viewed in color.

3.2 Objective Function

Suppose that $\mathbf{x}$ and $\mathbf{y}$ are the high-level representations obtained from two modality-specific networks. Let $\mathbf{H}^x \in R^{m\times n}$ and $\mathbf{H}^y \in R^{m\times n}$ denote the activation results of hidden layer $\mathbf{h}$ of the two modalities in MRBM framework, where m is the number of hidden units and n is the size of training set. They are used to evaluate the contribution of each modality to the shared representation, which should be similar. In this case, the goal is to enhance the similarity over $\mathbf{H}^x$ and $\mathbf{H}^y$, therefore, the objective function for MRBM becomes

$$\text{minimize}_{\{\mathbf{W}^x,\mathbf{W}^y,\mathbf{b}^x,\mathbf{b}^y,\mathbf{b}^h\}} \quad -\log p\left(\mathbf{x},\mathbf{y}\right) + \lambda \,\text{sim}(\mathbf{H}^{\text{x}},\mathbf{H}^{\text{y}}), \tag{1}$$

where $\mathbf{W}^x$ is a matrix of pairwise weights between elements of $\mathbf{x}$ and $\mathbf{h}$, and similar for $\mathbf{W}^y$. $\mathbf{b}^x$, $\mathbf{b}^y$, $\mathbf{b}^h$ are bias vectors for $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{h}$, respectively. And λ in Eq. 1 is a regularization constant. The regularization (i.e. $\text{sim}(\mathbf{H}^x, \text{H}^y)$, the similarity function) is a penalty to traditional objective function, which provides convenient fashion to control the similarity between the high-level representations of modalities. As the activation value of hidden units is between 0 and 1 when using the sigmoid activation function, it is intuitive to employ the cross-entropy to measure the similarity as follows,

$$\begin{aligned}\text{sim}\left(\mathbf{H}^x,\mathbf{H}^y\right) &= \mathbb{H}(\mathbf{H}^x \parallel \mathbf{H}^y) \\ &= -\sum_{j,k}\left[\mathbf{H}^x_{jk}\log\mathbf{H}^y_{jk} + (1-\mathbf{H}^x_{jk})\log(1-\mathbf{H}^y_{jk})\right].\end{aligned} \tag{2}$$

Note that the cross-entropy is not symmetry, that is $\mathbb{H}(\mathbf{H}^x \parallel \mathbf{H}^y) \neq \mathbb{H}(\mathbf{H}^y \parallel \mathbf{H}^x)$, so we also take the second form and make comparison between them.

As CCA is a standard statistical method to find a linear correlation of two kinds of data [5], we also employ it to measure the similarity via the following general solution to CCA [10],

$$\mathrm{sim}(\mathbf{H}^x, \mathbf{H}^y) = -\mathrm{corr}(\mathbf{H}^x, \mathbf{H}^y) = -\|\mathbf{T}\|_{tr}, \tag{3}$$

where $\mathbf{T} = \sum_{xx}^{-1/2} \sum_{xy} \sum_{yy}^{-1/2}$, and $\|\cdot\|_{tr}$ is the matrix trace norm. $\sum_{xx}$ and $\sum_{yy}$ denotes the auto-covariance of $\mathbf{H}^x$ and $\mathbf{H}^x$, respectively. And the cross-covariance $\sum_{xy}$ is over $\{\mathbf{H}^x, \mathbf{H}^y\}$. Finally, both the cross-entropy and CCA are treated as the similarity functions and considered in this paper.

For the optimization of Eq. 1, the gradient of the regularization term is performed based on batch-size following the *Contrastive Divergence* (CD) algorithm [4]. Due to space limit, we defer the details about the optimization to supplementary material. And we summarize the learning procedure in Algorithm 1.

Algorithm 1 SSL algorithm for MRBM

1: Initialize model parameters.
2: Update the parameters with respect to each modality $\{\mathbf{x}, \mathbf{y}\}$ using CD learning rule.
3: Update the parameters using the gradient of the regularization term[3].
4: Repeat above steps 2 and 3 until convergence or K steps.

4. EXPERIMENTS

In this section, we conduct sets of experiments for evaluating the SSL. Various kinds of modalities are experimented, including image, audio, and text. And for all the experiments, the parameter λ is fixed at 0.1 that is chosen from $\{10, 1, 0.1, 0.001, 0.0001\}$ and has been validated in the experiments.

4.1 Toy example

We first make a simple experiment on the MNIST handwritten digit recognition dataset [9], which consists of 60,000 train images and 10,000 test images. The 28×28 matrix of pixels in the original image is split into left and right part, treated as two modalities as in [1, 13]. Full image (left + right) is for both the training and testing, but we also test the algorithms with the single part. Table 1 shows the recognition errors obtained with different learning methods. Although the error of SSL is a little higher than MJLL in the case of full image input, it shows noticeable decreases for each single modality input, especially the SSL with cross-entropy function. Specifically, the improvements of error may come from the more exact gradient of MJLL with *Persistent CD* (PCD) compared with SSL that uses CD method. Nevertheless, we consider that the enhanced shared representations in SSL provide each modality more reliable discriminative information via the learned parameters and lead to the above decreases.

[3]The bias term has little influence on the activation of hidden units for MRBM in the experiments, so only the weight terms are considered in this case.

Table 1: The recognition error (%) obtained with different learning methods of MRBM on the MNIST dataset.

Method	Similarity function	Left+Right	Left	Right
MJLL [7]	-	**1.57**	14.98	18.88
SSL	$\mathbb{H}(\mathbf{H}^x \parallel \mathbf{H}^y)$	1.73	**11.93**	**15.87**
	$\mathbb{H}(\mathbf{H}^y \parallel \mathbf{H}^x)$	1.64	11.98	18.31
	$\mathrm{corr}(\mathbf{H}^x, \mathbf{H}^y)$	1.78	13.23	16.20

Table 2: The classification results (MAP) on the MIR-Flickr dataset. Both MJLL and SSL are implemented based on the MDBN.

Method	Similarity function	Text+Image	Image
MJLL [14]	-	59.77	45.78
SSL	$\mathbb{H}(\mathbf{H}^x \parallel \mathbf{H}^y)$	59.84	46.78
	$\mathbb{H}(\mathbf{H}^y \parallel \mathbf{H}^x)$	59.8	47.74
	$\mathrm{corr}(\mathbf{H}^x, \mathbf{H}^y)$	**60.29**	**53.29**

4.2 Image-text

In this experiment, we employ the MIR-Flickr Data set [8] to evaluate our method in the multimodal classification task. The dataset consists of 1 million images and corresponding user assigned tags from the photograpy website Flickr. Among the images, 25,000 are annotated for 24 potential topics and 14 strict category labels. We use the same visual and text features and follow the pre-processing in [14], where the image feature is processed into zero-mean and unit variance for each dimension and the text feature is represented using a word count of the 2000 most frequency tags. We use the random selected 15,000 for training and the rest 10,000 for testing. The performance metric adopts the *Mean Average Precision* (MAP). As for the model architecture, modality-specific deep network of [3857, 1024, 1024] and [2000, 1024, 1024] are built for image and text, respectively. And the shared hidden layer of MRBM consists of 2048 units. Thus, the network is truly a *Multimodal Deep Belief Network* (MDBN) [14].

Table 2 shows the MAP obtained by different learning methods in the cases of different inputs at testing time, i.e., multimodal input (image+text) and unimodal input (image). The proposed SSL outperforms the MJLL in all the settings. There are two points we should pay attention to. First, the improvement in the unimodal case is much higher than the multimodal case, that maybe come from the same reason as the toy example. The added regularization term reduces the uncertain parts and enhances the shared parts at the high-level representation of image modality, even makes the image network learn some transferred knowledge from text modality. Second, the CCA function performs much better than the other two, which may result from the efficient subspace learning of CCA, in which the image and text representations are embedded into highly linear correlation.

4.3 Audio-video

The last sets of experiments are conducted for AVSR, which is an interesting multimodal task. Two datasets are employed for testing, one is AVLetters2 [3] and the other is

Table 3: Audiovisual speech classification performance (accuracy) implemented by different learning methods on two datasets.

Dataset	Method	Similarity function	AV	A
AVLetters2	MJLL[11]	-	59.89	70.88
	SSL	$\mathbb{H}(\mathbf{H}^x \parallel \mathbf{H}^y)$	68.68	74.18
		$\mathbb{H}(\mathbf{H}^y \parallel \mathbf{H}^x)$	64.28	73.62
		$\mathrm{corr}(\mathbf{H}^x, \mathbf{H}^y)$	**70.88**	**75.82**
AVDigits	MJLL[11]	-	58.89	63.89
	SSL	$\mathbb{H}(\mathbf{H}^x \parallel \mathbf{H}^y)$	63.89	71.07
		$\mathbb{H}(\mathbf{H}^y \parallel \mathbf{H}^x)$	60.00	70.56
		$\mathrm{corr}(\mathbf{H}^x, \mathbf{H}^y)$	**70.00**	**71.11**

AVDigits [6]. AVLetters2 is about reading letters from A to Z, spoke by five people, seven times for each letter. Letters spoken by four people are for training and the rest one is for testing. Different from AVLetters2, AVDigits is about speaking digits from 0 to 9, spoke by six people, nine times for each digit. Letters spoken by four people are exploited to train and the rest two are exploited to test. Both datasets provide raw audio record and captured mouth movements.

The audio and visual data are preprocessed, respectively. We extract the spectrogram of the audio signal with 20ms hamming window and 10ms overlap. Then the obtained spectral coefficient is reduced to 50 dimensions using PCA whitening. As for the visual signal, the mouth shape is reshaped to 60×80 pixels and reduced to 100 principal components using PCA whitening as well. 4 audio frames and 1 video frame (almost the same duration) are used as the inputs to the networks. In this task, the MDAE framework [11] is chosen as the experimental network that is commonly used in AVSR.

In Table 3, we show the accuracy of different learning methods applied to the MDAE network, where the multimodal (audiovisual) and unimodal (audio) input setting are employed. There are three points we should pay attention to. First, similar to the above experimental results, the SSL outperforms the MJLL method on all the settings, which confirms the effectiveness of it in multimodal learning. Second, the SSL with cross-entropy of $\mathbb{H}(\mathbf{H}^x \parallel \mathbf{H}^y)$ performs better than $\mathbb{H}(\mathbf{H}^y \parallel \mathbf{H}^x)$, this is because the former is to force the visual modality to fit the expectation of the more reliable audio signal, which makes the shared representation more discriminative. Third, we also note that the audiovisual modality performs worse than single audio information. This is because the visual modality lowers the performance, which is a common situation [11]. But we can significantly reduce the gap between them via emphasizing the similar semantic, as shown in Table 3.

5. CONCLUSION

In this paper, we present a novel learning method for multimodal network in terms of the semantic similarity between the high-level representation of modalities. The conducted sets of experiments show that our method has remarkable improvements over traditional method on various multimodal deep networks. More importantly, the association among different modalities (i.e., image, audio, and text) are all enhanced, which demonstrates the efficient generalization of the method. But the learning is performed just in the M-RBM layers, a better solution may be found by searching the entire deep networks and inferring the modality-specific network with the SSL objective. Thus, a whole learning method of the multimodal deep networks should be derived in the future work.

6. ACKNOWLEDGEMENTS

This work is supported by the Fundamental Research Funds for the Central Universities (Grant no. 3102015BJ(II)JJZ01), the National Basic Research Program of China (973 Program) (Grant No. 2012CB719905), State Key Program of National Natural Science of China (Grant No. 61232010 and 61472413), and by the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences (Grant No. LSIT201408).

7. REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *ICML*, pages 1247–1255, 2013.

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *PAMI, IEEE Transactions on*, 35(8):1798–1828, 2013.

[3] S. J. Cox, R. Harvey, Y. Lan, J. L. Newman, and B.-J. Theobald. The challenge of multispeaker lip-reading. In *AVSP*, pages 179–184, 2008.

[4] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[5] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[6] D. Hu, X. Li, and X. Lu. Temporal multimodal learning in audiovisual speech recognition. In *CVPR*, 2016.

[7] J. Huang and B. Kingsbury. Audio-visual deep learning for noise robust speech recognition. In *ICASSP*, pages 7596–7599, 2013.

[8] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *ACM MIR*, pages 39–43, 2008.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[10] K. V. Mardia, J. T. Kent, and J. M. Bibby. Multivariate analysis. 1980.

[11] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

[12] X. Shu, G.-J. Qi, J. Tang, and J. Wang. Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation. In *ACM Multimedia*, pages 35–44, 2015.

[13] K. Sohn, W. Shang, and H. Lee. Improved multimodal deep learning with variation of information. In *NIPS*, pages 2141–2149, 2014.

[14] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, pages 2222–2230, 2012.

[15] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103, 2008.

[16] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *ICML*, pages 1083–1092, 2015.