DENSE MULTIMODAL FUSION FOR HIERARCHICALLY JOINT REPRESENTATION

Di Hu, Chengze Wang, Feiping Nie, Xuelong Li

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL)
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China

ABSTRACT

Multiple modalities can provide more valuable information than single one by describing the same contents in various ways. Previous methods mainly focus on fusing the shallow features or high-level representations generated by unimodal deep networks, which only capture part of the hierarchical correlations across modalities. In this paper, we propose to densely integrate the representations by greedily stacking multiple shared layers between different modality-specific networks, which is named as *Dense Multimodal Fusion* (DMF). The joint representations in different shared layers can capture the correlations in different levels, and the connection between shared layers also provides an efficient way to learn the dependence among hierarchical correlations. These two properties jointly contribute to the multiple learning paths in DMF, which results in **faster convergence**, **lower training loss**, and **better performance**. We evaluate our model on audiovisual speech recognition and cross-modal retrieval. The noticeable performance demonstrates that our model can learn more effective joint representation.

Index Terms— Multimodal Learning, Dense Fusion, Hierarchical Correlation

1. INTRODUCTION

The same contents or events can be described in multiple kinds of modalities in the real world. That is, the verbal, vocal, and visual modality can be jointly used for expression in different scenarios. Sometimes, one of them can also provide complementary information for the other one. Considering that visual modality is free of audio noise, it can provide efficient information for speech recognition in the noisy environment [2]. Hence, multiple modalities can jointly provide more valuable information than single one, and there have been many works over the years making use of multimodal data for specific tasks, such as *Audiovisual Speech Recognition* (AVSR) [3], and cross-modal retrieval [5].

Although these works benefit from the valuable multimodal data, different modalities take diverse representations and statistical properties. These different representations make it difficult to capture the complex correlation across modalities [4]. Fortunately, as these modalities are used to describe the same contents, they should share similar patterns to some extent. Recently, deep learning methods have shown their effectiveness in generating useful feature representation [6, 7, 4]. Hence, they propose to learn a kind of joint representation across the top layers of modality-specific networks. The motivation beyond this strategy is that they assume the high-level representations contain sufficient semantic information and the shared patterns across modalities exist in the semantic level. However, there remain two open questions about such strategy. First, if the high-level representations of each modality can provide sufficient information to capture the complex correlation across modalities, especially when the input data are hand-craft features. Second, if the shared patterns only exist in the semantic level or the representation in specific single layer? Actually, the fusion across the high-level representations works like the classical late fusion that fuses the semantic concepts from unimodal features [8]. Compared with other fusion strategies (e.g., early fusion), the late fusion can only capture the correlation in the semantic level but fail to exploit other kinds of correlations, such as the covariation in the early feature level [9], the hierarchical supervision throughout the whole network [10]. Therefore, a kind of hierarchical fusion should be expected for capturing the complex correlations across modalities.

In this paper, to capture the complex correlations across modalities, we propose to densely integrate the representations of different networks, where the higher joint representation not only fuses the modality-specific representations in the same layer but also is conditioned on the lower joint one, which is named as *Dense Multimodal Fusion* (DMF). Different from the traditional fusion scheme based on deep networks, the learned joint representation in the hidden layer can simultaneously capture the covariation in the early fusion and the correlation between the inherent semantic of modalities. More importantly, the dense fusion provides multiple learning paths to enhance the interaction across modalities. For example, when one modality is with high uncertainty or missing, it can be efficiently inferred from the multi-level fused information. To evaluate the proposed DMF scheme, we perform different multimodal tasks on several benchmark datasets, including AVSR, and cross-modal retrieval. Extensive exper-

This work was supported in part by the National Natural Science Foundation of China grant under number 61772427 and 61751202.

iments show that DMF is superior to the traditional fusion schemes in these tasks, not only in the conditions of multimodal inputs but also unimodal input.

2. DENSE MULTIMODAL FUSION

To capture the correlation across modalities, an intuitive way is to directly concatenate the different features of them, then employ multiple layers of nonlinear transformation to generate the high-level joint representation [11], which is named as *Early Multimodal Fusion* (EMF), as shown in Fig. 1. Unfortunately, although such fusion increases the dimensionality, it lacks the ability in capturing more complex correlation across modalities [7]. To tackle the problems of EMF, the general idea is to reduce the influence of individual differences and improve the shared semantic [4]. As the shared layer exists in the middle part of the whole multimodal network, such fusion is named as *Intermediate Multimodal Fusion* (IMF) [12].

Based on the previous multimodal networks, we can easily find that they share the fusion scheme that consists of one shared layer and two modality-specific layers. Such multimodal units have the ability in capturing the correlation between different layers [7, 1]. In this paper, we employ dense multimodal fusion to learn the complex hierarchical correlations between the representations of different modalities.

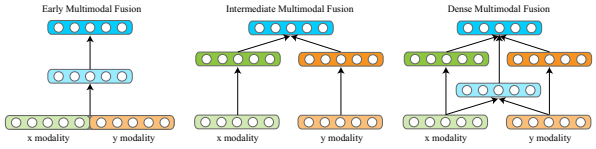


Fig. 1. Different multimodal fusion strategies based on deep networks: early fusion (left), intermediate fusion (middle), and the proposed dense fusion (right).

There are mainly two parts that constitute the proposed DMF, the stacked modality-specific layers and the stacked shared layers. For the former, we can obtain multiple representations in different levels after several stacked layers of nonlinear transformation for each modality, which contain not only the detailed descriptions but also the semantic information. For the latter, to capture the correlations between different modalities in each representation level, several shared layers are densely stacked to correlate modality-specific networks. Each shared layer not only has the ability to capture the correlation in the current level, but also has the capacity to learn the dependency among the correlations. Hence, DMF can capture more complex hierarchical correlation between modalities, and the experimental results also confirm this.

2.1. Multi-paths for Multimodal Learning

To capture the complex hierarchical correlations between modalities, multiple learning paths (in both feed-forward and back-propagation) of the intra-modality and inter-modality should be expected. However, the traditional multimodal networks, i.e., EMF and IMF, contain only one path for learning such correlations. EMF is like a kind of unimodal network but based on the concatenated multimodal features, therefore the top layer just relies on the previous shared layer, and only the error of the shared layer can be propagated to each modality, which is weak in modeling the individual properties of each modality and the correlations in other levels. Concretely, for a IMF of L layers, the top shared layer is obtained by[1],

$$s_L = f\left(W_L^{x\to s} h_L^x + W_L^{y\to s} h_L^y\right) \tag{1}$$

where $f(\cdot)$ is the sigmoid activation function, $W_L^{x\to s}$ is the matrix of pairwise weights between elements of $h_L^x$and s_L, and similarly for $W_L^{y\to s}$. This is a standard multimodal unit. However, the hidden layers, h_L^x and h_L^y, just rely on the modality-specific representations, hence, there exists only one path for learning the correlation across modalities. On the other hand, when backward propagating the error, the gradient to current hidden layer of modality x is written as,

$$\frac{\partial \varepsilon}{\partial h_l^x} = \frac{\partial \varepsilon}{\partial h_{l+1}^x}\frac{\partial h_{l+1}^x}{\partial h_l^x}, \qquad l = 1, 2, ..., L-1 \tag{2}$$

where ε stands for the objective function. The error of the joint representation is propagated via the term of $\partial\varepsilon/\partial h_{l+1}^x$, therefore, the modality-specific weights can be only optimized with respect to the correlation in the top layer.

Compared with the single learning path in EMF and IMF, DMF enjoys multiple paths when feeding the joint representation and propagating the errors, as shown in Fig. 2. The corresponding update and optimization are written as,

Feed-forward:

$$s_l = f\left(W_l^{x\to s} h_l^x + W_l^{y\to s} h_l^y + W_{l-1}^s s_{l-1}\right), \qquad l = 2, ..., L \tag{3}$$

$$h_l^x = f\left(W_{l-1}^x h_{l-1}^x\right), \qquad l = 2, ..., L \tag{4}$$

Back-propagation:

$$\frac{\partial \varepsilon}{\partial h_l^x} = \frac{\partial \varepsilon}{\partial h_{l+1}^x}\frac{\partial h_{l+1}^x}{\partial h_l^x} + \frac{\partial \varepsilon}{\partial s_l}\frac{\partial s_l}{\partial h_l^x}, \qquad l = 1, 2, ..., L-1 \tag{5}$$

$$\frac{\partial \varepsilon}{\partial s_l} = \frac{\partial \varepsilon}{\partial s_{l+1}}\frac{\partial s_{l+1}}{\partial s_l}, \qquad l = 1, 2, ..., L-1 \tag{6}$$

where W_{l-1}^x is the modality-specific weight between layer h_{l-1}^x and h_l^x, while W_{l-1}^s is the weight between adjacent shared layers. These weights of $\{W^x,\ W^{x\to s},\ W^s\}$ jointly model the intra- and inter-modalities correlations, similarly for modality y. In Fig. 2, we can easily find that there are three paths feeding the shared layer s_l from modality x, where the green and yellow one are the same as the paths of IMF and EMF, respectively. These two help to capture the shallow and deep correlation between modalities. The remaining blue path indicates the correlations in the middle layers. When the number of stacked multimodal units increases, there will be

[1]The modality layers and shared layer of one multimodal unit are deemed in the same layer.

more paths connected to higher shared layers. Hence, DMF is more capable of capturing the complex correlation between modalities, not only the ones in the same layer but also in the cross-layers.

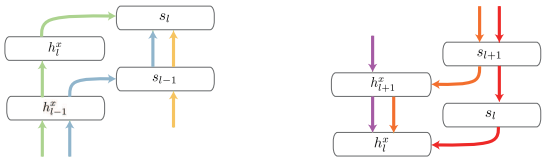
On the other hand, to efficiently optimize the network and infer the shared layers, multiple paths of back-propagation are performed in the DMF. The purple and red path denote the error propagated from the modality-specific network and top shared layer, respectively. They preserve the consistency of intra-modality and inter-modality, which then help to establish the correlations in other layers. Different from EMF and IMF, the distinct orange path is the error propagated via both the shared layer and modality layer.

As the shared layer in each level contains significant representation generated from both modalities, the orange path can provide efficient hierarchical supervision for each modality from the other one. More specifically, let M_l denote $\left(W_l^{x\to s}h_l^x + W_l^{y\to s}h_l^y + W_{l-1}^s s_{l-1}\right)$, then the second term in Eq. 5 can be re-written into

$$\frac{\partial \varepsilon}{\partial s_l}\frac{\partial s_l}{\partial h_l^x} = \left(\frac{\partial \varepsilon}{\partial s_{l+1}}\frac{\partial s_{l+1}}{\partial s_l}\right)\cdot\left(W_l^{s\to x} f\left(M_l\right)\left(1-f\left(M_l\right)\right)\right), \quad (7)$$

$$\frac{\partial s_{l+1}}{\partial s_l} = \left(W_l^s\right)^T f\left(M_{l+1}\right)\left(1-f\left(M_{l+1}\right)\right). \quad (8)$$

Recall that both the term of M_l and M_{l+1} contain the generated information of modality y, hence the error propagated to the current layer of h_l^x contains hierarchical supervision from the other one. Moreover, the multi-level cross-modal supervision can also come from the term of $\partial\varepsilon/\partial h_{l+1}^x$ in Eq. 5. Hence, when one modality is damaged or missing, DMF can still have the ability to provide efficient supervision from the other one and learn effective joint representation, which is also confirmed in the experiments.



(a) Feed forward (b) Back propagation

Fig. 2. An illustration of feed forward and back propagate paths in DMF. Best viewed in color.

2.2. Model Variants

Dense fusion is not one specific network architecture but a novel fusion scheme or mechanism. Hence, it has different model variants for different multimodal learning tasks. One common task is taking advantage of the more valuable information of multiple modalities to perform more exact classification. For this task, DMF can be viewed as a discriminative model, where a regression layer is performed over the top joint representation, as shown in Fig. 3. Such model can be finetuned to minimize the categorical cross-entropy after initializing the feed forward network.

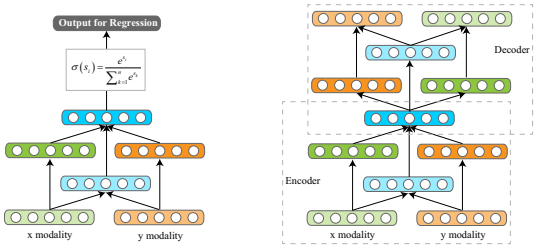


Fig. 3. The DMF variants in the discriminative (left) and generative (right) task, respectively.

Another common task is to infer the robust joint representation with different input modalities, which could be then used for cross-modal retrieval [5] or recognition [2]. To simultaneously preserve the inter- and intra-modal consistency under the unsupervised fashion, we propose to reconstruct the modalities by given the joint representation, which actually treats DMF as an "encoder" and the reversed one as a "decoder", as shown in Fig. 3.

3. EXPERIMENTS

3.1. Toy Example

In this section, following [1], we first evaluate different fusion strategies (i.e., EMF, IMF, and DMF) on the MNIST dataset [13]. As the right and left half of the image jointly describe the same digit, they can be considered as two modalities [1]. To evaluate the ability in learning robust representation when faced with modalities with high uncertainty, we randomly set part of the right half to zeros at different levels, i.e., $\{0,\ 30\%,\ 50\%,\ 70\%\}$. And we also compare these methods under different input conditions, i.e., left+right and right modality. Table. 1 shows the comparison results among EMF, IMF, and DMF. The network for each image pathway is [392, 512, 128] and the shared pathway is [512, 256, 64]. Although EMF and IMF share the same unimodal paths as DMF, the additional shared pathway still leads to the growth of variable number (about 2-3 times larger). The increased variables are correlation learning-related, which means DMF has greater potential in capturing the complex multimodal correlation beyond traditional multimodal network. Hence, DMF does not suffer from the intractable overfitting problem, but significantly outperforms the other two fusion strategies in different input conditions.

Surprisingly, training much more parameters in DMF do not cost more time but less (similar in the following multimodal tasks), compared with classical fusion network, as shown in Fig. 4. The noticeable performance comes from the multiple efficient learning paths of DMF. Specifically, each unimodal layer can receive multiple kinds of error propagated

from different layers (Eq. 5) instead of the single error path of IMF (Eq. 2). Moreover, these learning paths also contribute to the efficient hierarchical supervision in DMF (Eq. 7 and Eq. 8). When one modality is badly damaged, the other reliable one can provide supervision information on different levels for it, while EMF and IMF only provide such information on the bottom and top layer, respectively. Hence, DMF still shows noticeable performance when we destroy one modality. These properties jointly contribute to the lower training loss, higher testing accuracy, and faster convergence of DMF.

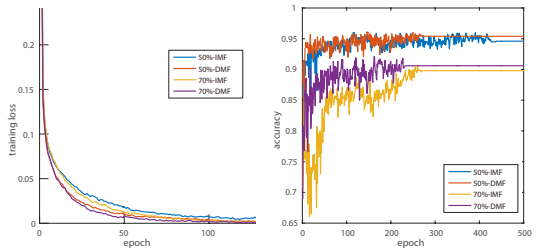


Fig. 4. The training loss and testing accuracy of DMF and IMF on the mnist dataset. The right half of digital image is manually damaged at different levels. Best viewed in color.

Table 1. Recognition performance (in error) for handwritten digit recognition on the MNIST dataset. The percentages indicate the degrees of damage to the right half of images.

Modality	Models	0	30%	50%	70%
Left+Right	EMF	1.90	2.60	4.70	12.20
	IMF	1.60	2.40	5.40	10.20
	DMF	**1.40**	**2.30**	**4.60**	**9.40**
Right	EMF	51.00	53.50	63.40	70.30
	IMF	51.30	54.30	63.30	82.90
	DMF	**39.90**	**47.60**	**57.10**	**67.60**

3.2. Audiovisual Speech Recognition

Audiovisual speech recognition is a classical task that makes use of the information from audio and visual modality to perform robust speech recognition. In this section, we compare DMF with Active Appearance Model (AAM) [14], MDAE [7], MDBN [15], *Recurrent Temporal Multimodal RBM* (RTMRBM) [2], *Conditional RBM* (CRBM) [16], and CorrRNN [17]. To be fair, we use the same discriminative model as MDAE and MDBN that employ the "encoder-decoder" framework and use SVM as the classifier. The same network architecture is also adopted except the shared pathway, which is [1024, 512, 256]. The experiments are conducted on the AVLetters2 dataset [18]. Similar with [2], we use the letters spoken by four people for training and the rest for testing.

Table 2 shows the results in accuracy. We can find that DMF shows significant improvement over the other ones. Moreover, when only the visual modality is available, DMF has a noticeable improvement. This is because DMF can establish efficient correlation between modalities in each layer, which helps to make one modality learn from the other one.

Table 2. The mean accuracy of speech recognition on AVLetters2. All the models are evaluated with different input modalities. For the unimodal input, one modality is preserved while the other one is set to zero.

Modality	A	V	A+V
AAM	15.2	-	-
MDAE	-	-	67.89
MDBN	-	-	54.1
CRBM	-	-	74.08
RTMRBM	75.85	31.21	74.77
CorrRNN	81.36	60.17	76.32
DMF	**86.43**	**75.87**	**80.43**

Even so, the visual modality still lowers the performance of multimodal inputs to some extent, but it is a common situation [2].

3.3. Cross-modal Retrieval

In this experiment, we focus on two cross-modal retrieval tasks, i.e., image2text (I2T) and text2image (T2I). We compare our model with four unsupervised methods, including CCA [19], CMFH (without binary constraint) [20], LCFS [21], and Corr-Full-AE [10]. The benchmark image-text dataset of Wiki is chosen for evaluation [19]. For each pair, the image modality is represented as 128-D SIFT descriptor histograms, and text is expressed as 10-D semantic vector. These pairs are annotated with one of 10 topic labels. In this paper, we choose 25% of the dataset as the query set and the rest for retrieval set. And we still use the "encoder-decoder" framework and reconstruct both modalities based on the query modality.

As shown in Table 3, it is obvious that DMF enjoys the best results among these methods. Specifically, Corr-Full-AE is similar as IMF, which attempts to capture the correlation between the middle layers of modality-specific networks. However, it aims to minimize the differences between the representations instead of learning the joint representation which makes it difficult to train and optimize. In contrast, DMF is easier to optimize and shows better performance.

Table 3. The ranking performance of cross-modal retrieval on Wiki dataset.

mAP	CCA	CMFH	LCFS	Corr-Full-AE	DMF
I2T	0.2490	0.2551	0.2798	0.2634	**0.2921**
T2I	0.1960	0.5407	0.2141	0.5418	**0.5612**

4. CONCLUSION

In this paper, we propose to densely correlate representations of different modalities layer-by-layer, where the shared layer not only models the correlation in the current level but also depends on the lower one. Such dense fusion not only rewards it the advantages of early and intermediate fusion multimodal network but also the multiple learning paths that help to capture more complex correlation and accelerate convergence.

5. REFERENCES

[1] Kihyuk Sohn, Wenling Shang, and Honglak Lee, "Improved multimodal deep learning with variation of information," in *Advances in Neural Information Processing Systems*, 2014, pp. 2141–2149.

[2] Di Hu, Xuelong Li, et al., "Temporal multimodal learning in audiovisual speech recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3574–3582.

[3] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in visual and audio-visual speech processing*, vol. 22, pp. 23, 2004.

[4] Nitish Srivastava and Ruslan R Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.

[5] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang, "A comprehensive survey on cross-modal retrieval," *arXiv preprint arXiv:1607.06215*, 2016.

[6] Ruslan Salakhutdinov and Geoffrey Hinton, "Semantic hashing," *International Journal of Approximate Reasoning*, vol. 50, no. 7, pp. 969–978, 2009.

[7] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[8] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 399–402.

[9] Shankar T Shivappa, Mohan Manubhai Trivedi, and Bhaskar D Rao, "Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010.

[10] Fangxiang Feng, Xiaojie Wang, and Ruifan Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 7–16.

[11] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 284–288.

[12] Dhanesh Ramachandram and Graham W Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[14] Helen L Bear, Stephen J Cox, and Richard W Harvey, "Speaker-independent machine lip-reading with speaker-dependent viseme classifiers.," in *AVSP*, 2015, pp. 190–195.

[15] Jing Huang and Brian Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7596–7599.

[16] Mohamed R Amer, Behjat Siddiquie, Saad Khan, Ajay Divakaran, and Harpreet Sawhney, "Multimodal fusion using dynamic hybrid models," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*. IEEE, 2014, pp. 556–563.

[17] Xitong Yang, Palghat Ramesh, Radha Chitta, Sriganesh Madhvanath, Edgar A Bernal, and Jiebo Luo, "Deep multimodal representation learning from temporal data," *CoRR abs/1704.03152*, 2017.

[18] Stephen J Cox, Richard W Harvey, Yuxuan Lan, Jacob L Newman, and Barry-John Theobald, "The challenge of multispeaker lip-reading.," in *AVSP*, 2008, pp. 179–184.

[19] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 251–260.

[20] Guiguang Ding, Yuchen Guo, Jile Zhou, and Yue Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5427–5440, 2016.

[21] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan, "Learning coupled feature spaces for cross-modal matching," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2088–2095.