Dense Multimodal Fusion for Hierarchically Joint Representation

Abstract

Multiple modalities can provide more valuable information than single one by describing the same contents in various ways. Hence, it is highly expected to learn effective joint representation by fusing the features of different modalities. However, 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 three typical multimodal learning tasks, including audiovisual speech recognition, cross-modal retrieval, and multimodal classification. The noticeable performance in the experiments demonstrates that our model can learn more effective joint representation.

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. For example, the seen flying plane is always accompanied by loud roar (vocal+visual), speech signal can be expressed by audio and lip movements (vocal+visual), image and caption can jointly describe the concrete contents (verbal+visual), and language can be recorded in both text and audio (verbal+vocal). As these modalities actually describe the same contents, they are closely related [32]. 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 [12]. 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) [25], image-text classification [33], and cross-modal retrieval [37].

Although these works benefit from the valuable multimodal data, different modalities take diverse representations and statistical properties. For example, audio data is usually represented with spectrogram or Mel-Frequency Cepstrum Coefficient in real-value, while textual data is represented using sparse one-hot or word-count vector in discrete value. These different representations make it difficult to capture the complex correlation across modalities [33]. 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 [29, 28]. Some works, such as *Multimodal Deep Autoencoder* (MDAE) [22] and *Multimodal Deep Boltzmann Machine* (MDBM) [33], consider that the extracted high-level representations of different modali-

Preprint. Work in progress.

ties are semantically correlated. 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 (e.g, SIFT descriptor, simple word-count vector). Second, if the shared patterns only exist in the semantic level or the representation in specific single layer? In other words, if it is suitable to perform the fusion across modalities for only once?

Actually, the fusion across the high-level representations works like the classical late fusion that fuses the semantic concepts from unimodal features [31]. 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 [30, 18], the hierarchical supervision throughout the whole network [9]. Meanwhile, cognitive researchers have also verified that the multisensory integration exists in not only the temporal lobe but also parietal and frontal lobe [34, 11], which indicates that the integration occurs in the different stages of information processing in our brain. 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, image-text classification, and cross-modal retrieval. Extensive experiments 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 Multimodal Deep Learning

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 [23], which is named as *Early Multimodal Fusion* (EMF), as shown in Fig. 1. When the deep network is viewed as a kind of discriminative model, such fusion scheme is truly based on the feature level. Early fusion is easy to capture the covariation between modalities, or other correlations exist at the feature level, meanwhile, it is the simplest to implement [18]. Unfortunately, although such fusion increases the dimensionality, it lacks the ability in capturing more complex correlation across modalities [22]. Another issue is that different features have different properties but need to be in the same space type, which requires to convert and scale them.

To tackle the problems of EMF, recent works propose to build a joint hidden layer based on the outputs of modality-specific networks. The general idea behind such fusion scheme is to reduce the influence of individual differences and improve the shared semantic [33]. The earliest representative work is MDAE [22]. It is a AVSR network that learns a new shared representation layer across separated audio and visual networks, and employs it for multimodal reconstruction. As the shared layer exists in the middle part of the whole multimodal network, such fusion is named as *Intermediate Multimodal Fusion* (IMF) [26], as shown in Fig. 1. In view of the explainable rationale in semantic fusion, IMF has derived multiple variants in different multimodal task, such as multimodal deep belief network (MDBN) in AVSR [14], MDBM in image-text classification [33], *Correspondence Autoencoder* (Corr-AE) in cross-modal retrieval [9, 38], deep fusion in affect prediction [23] and video description generation [16], etc. More related works can be found in the recent survey [2].

Although these IMF networks show promised results in different tasks, the fusion is just performed on one specific level, which therefore fails to capture the correlation in other layers, such as co-

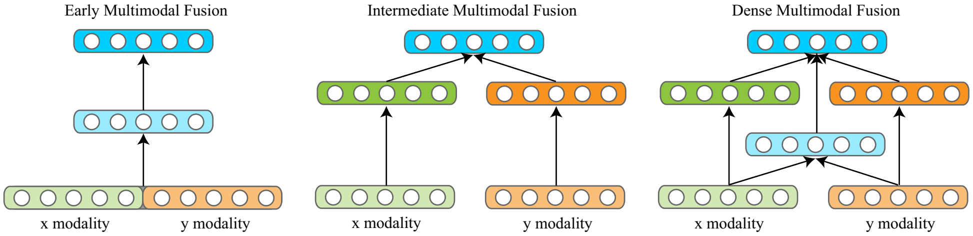


Figure 1: Different multimodal fusion strategies based on deep networks: early fusion (left), intermediate fusion (middle), and the proposed dense fusion (right).

variation [30]. Moreover, there exists only one path that connects two modalities, and such prolix inferring progress weakens the interaction and supervision across modalities. Hence, the effective fusion at different levels is highly expected as opposed to the single fusion layer [26]. In the unimodal scenario, Karpathy *et al.* [17] show that gradually fusing the video temporal information can provide the higher layers more global information in the video classification task. Du *et al.* [6] propose to hierarchically fuse the skeleton parts and gradually constitute the whole skeleton at the higher layers for action recognition. While, there is few work that models the intricate multimodal correlation via densely fusing the representations at different depths. Although Suk *et al.* [35] propose to perform the fusion based on patch-level and the image-level representation, the shared representation learning is only considered for patch feature learning via the standard MDBM [33], hence it is actually the same as conventional IMF scheme. The most related multimodal work [21] is to progressively fuse different modalities (e.g., fusing gray scale and depth firstly, then combing the joint representation with audio/pose descriptor), hence it has not learned the hierarchical correlation between each two modalities.

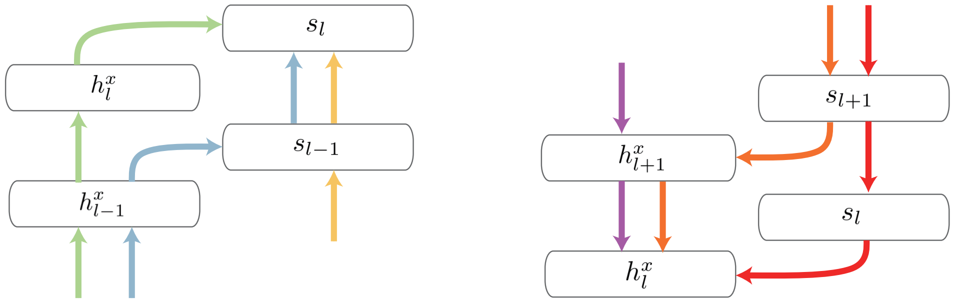
Apart from multimodal learning, other tasks also hope to capture effective correlations between different networks. For example, action recognition aims to capture the correspondence between temporal and spatial network by resorting to arithmetical operation [24, 8] and residual connection [7] on pre-defined levels. But they fail to simultaneously capture the hierarchical correlation. More importantly, they do not share the same problem with multimodal learning, as we aim to establish reliable joint representation (via multiple shared layers) instead of the direct spatiotemporal correspondence.

3 Dense Multimodal Fusion

Based on the analysis about the previous multimodal networks, we can easily find that they share the same fusion scheme that consists of one shared layer and two modality-specific layers, i.e., the bottom two layers in EMF and the top two in IMF, as shown in Fig. 1. Such multimodal units have the ability in capturing the correlation between different input layers [22, 32]. To simultaneously capture the correlations in each layer, multiple shared layers should be considered. In this paper, we employ dense multimodal fusion to learn the complex hierarchical correlations between the representations of different modalities, as shown in Fig. 1.

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.

Actually, when there are only two multimodal units in the DMF (i.e., the example in Fig. 1), it can be viewed as a kind of combination of EMF and IMF. Hence, DMF can simultaneously enjoy the merits of early fusion and intermediate fusion. Moreover, when there are two more stacked



(a) Feed forward

(b) Back propagation

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

multimodal units, DMF can model more correlations in the middle layers, which are neglected in EMF and IMF. In the following sections, we will provide a detailed analysis about this.

3.1 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. Different from EMF, the feed-forward and back-propagation path of IMF rely on the modality-specific networks. Concretely, for a IMF of L layers, the top shared layer is obtained by the following equation[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 $\partial \varepsilon \big/ \partial h_{l+1}^x$, 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:

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

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

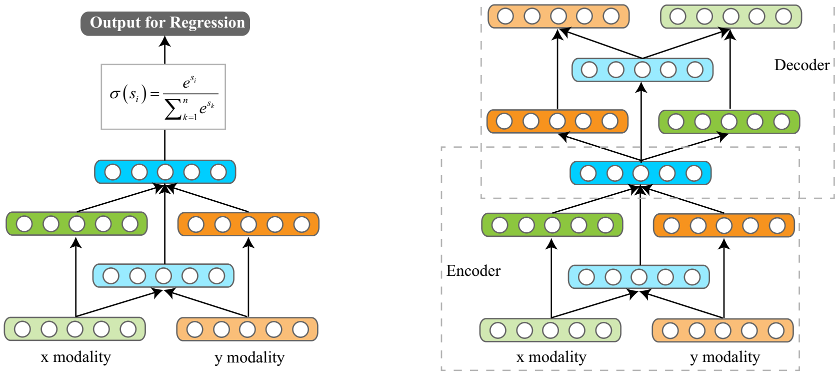


Figure 3: The DMF variants in the discriminative (left) and generative (right) task, respectively.

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

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 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

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

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

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 $\partial \varepsilon \big/ \partial h_{l+1}^x$ 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.

3.2 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

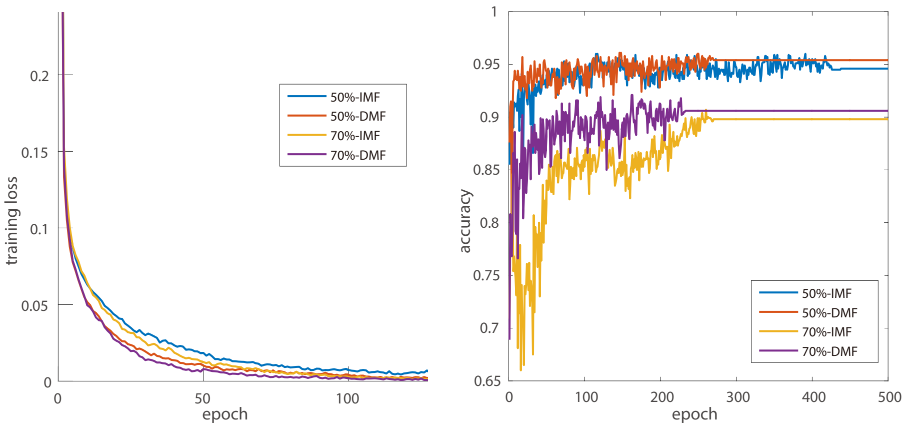


Figure 4: The training loss and testing accuracy of DMF and IMF on the mnist dataset. The right half of digital image is manually damaged at different levels. Best viewed in color.

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. Another common task is to infer the robust joint representation with different input modalities, which could be then used for cross-modal retrieval [37] or recognition [12]. The cross-modal retrieval task takes the representation of one modality as the query to retrieve relevant items of another modality [20]. 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. When there is only one modality available in the retrieval phase, DMF has to encode the joint representation based on single modality input and decode the representation into different modalities. Such "encoder-decoder" network can also be used for recognition. We can take the encoded joint representation as the inputs to classifier (e.g., SVM). Such operation has been widely used in the traditional multimodal network. As for the optimization of these DMF model variants, we initialize each multimodal unit with the Contrastive Divergence [10] in the layer-wise fashion, then fine-tune the whole network based on the respective supervision above.

4 Experiments

4.1 Toy Example

In this section, following [32, 13], we first evaluate different fusion strategies (i.e., EMF, IMF, and DMF) on the MNIST dataset [19]. The MNIST dataset consists of 60,000 training and 10,000 testing images. These images of 28×28 pixels are equally divided into right and left halves which are considered as two modalities describing the same digits [32]. 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

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**

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.

4.2 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) [3], MDAE [22], MDBN [14], *Recurrent Temporal Multimodal RBM* (RTMRBM) [12], *Conditional RBM* (CRBM) [1], and CorrRNN [39]. 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 [4]. It is about reading letters from A to Z, spoken by five people, seven times for each letter. Similar with [12], we use the letters spoken by four people for training and the rest for testing. Both unimodal and multimodal inputs are considered, where 4 audio frames and 1 video frame are used as the multimodal inputs to the networks.

Table 5 shows the results in accuracy. We can find that DMF shows significant improvement over the other ones. Note that RTMRBM, CRBM, and CorrRNN are temporal models that can capture the dependence among the multimodal sequence. However, DMF is not a temporal model but still outperforms these methods, which shows its ability in learning effective joint representation. 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. Even so, the visual modality still lowers the performance of multimodal inputs to some extent, but it is a common situation [12].

Figure 5: 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**

4.3 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 [27], CMFH (without binary constraint) [5], LCFS [36], and Corr-Full-AE [9]. The benchmark image-text dataset of Wiki is chosen for evaluation [27]. 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.

We show the ranking performance in Table 2. It is obvious that DMF enjoys the best results among these methods. Specifically, the non-linear methods outperform the linear ones, which illustrates that they are better at capturing the complex non-linear correlation across modalities. On the other hand, 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. Such framework makes it difficult to train and optimize. In contrast, DMF is easier to optimize and shows better performance.

Table 2: 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**

4.4 Multimodal Classification

For multimodal classification, we compare with different variants of IMF, including MDAE [22], MDBN [14], MDBM [33], and MinVI [32], where these models take the same number of layers and units as DMF except the dense shared layers. We employ DMF to reconstruct both modalities based on the multimodal inputs, and use the top joint representation as the input features to the same discriminative model (SVM) as other models. MIR-FLICKR dataset [15] is employed for evaluation, which consists of 1 million images and corresponding user tags collected from the photography website Flickr. We use the same visual and text features as [33], where the image feature is preprocessed into zero-mean and unit variance for each dimension and the text input is represented by a word count of the 2000 most frequency tags. We randomly select 15,000 for training and the rest 10,000 for testing.

Table 3 shows the comparison results. DMF outperforms other IMF models without additional fine-tuning. MDBM is an undirect multimodal model, where the responsibility of the multimodal modeling is spread over the entire network. Hence it has the ability to indirectly capture the correlation in the lower layers and is slightly better than MDBN and MDAE but worse than DMF. DMF can directly learn the correlation in each layer as well as the dependency between shared layers, hence it shows the best performance among these methods. On the other hand, MinVI aims to reduce the variant information across modalities, which is good at inferring mutual information but weak in learning the private properties of each modality. However, DMF can simultaneously preserve the mutual and private information via the dense shared layers and modality-specific layers.

Table 3: Classification results on Flickr dataset. MDRNN† is the same as MinVI but fine-tuned by a second-stage multimodal RNN.

Models	MDAE	MDBN	MDBM	MinVI	DMF	MDRNN†
mAP	0.600	0.599	0.609	0.574	**0.618**	0.686

5 Conclusion

Deep model can provide representations in different levels for single modality, such as raw features in the bottom and abstract semantic in the intermediate layer. Hence, there exist multiple kinds of hierarchical correlations between different modality-specific networks, which have not been effectively explored so far. In this paper, we propose to densely correlate these 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. Moreover, DMF can also provide hierarchical cross-modal supervision for the modality with high uncertainty.

References

[1] M. R. Amer, B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney. Multimodal fusion using dynamic hybrid models. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 556–563. IEEE, 2014.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[3] H. L. Bear, S. J. Cox, and R. W. Harvey. Speaker-independent machine lip-reading with speaker-dependent viseme classifiers. In *AVSP*, pages 190–195, 2015.

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

[5] G. Ding, Y. Guo, J. Zhou, and Y. Gao. Large-scale cross-modality search via collective matrix factorization hashing. *IEEE Transactions on Image Processing*, 25(11):5427–5440, 2016.

[6] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

[7] C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In *Advances in Neural Information Processing Systems*, pages 3468–3476, 2016.

[8] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.

[9] F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16. ACM, 2014.

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

[11] N. P. Holmes and C. Spence. Multisensory integration: space, time and superadditivity. *Current Biology*, 15(18):R762–R764, 2005.

[12] D. Hu, X. Li, et al. Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3574–3582, 2016.

[13] D. Hu, X. Lu, and X. Li. Multimodal learning via exploring deep semantic similarity. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 342–346. ACM, 2016.

[14] J. Huang and B. Kingsbury. Audio-visual deep learning for noise robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7596–7599. IEEE, 2013.

[15] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43. ACM, 2008.

[16] Q. Jin and J. Liang. Video description generation using audio and visual cues. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 239–242. ACM, 2016.

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[18] A. K. Katsaggelos, S. Bahaadini, and R. Molina. Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE*, 103(9):1635–1653, 2015.

[19] 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.

[20] X. Li, D. Hu, and F. Nie. Deep binary reconstruction for cross-modal hashing. *arXiv preprint arXiv:1708.05127*, 2017.

[21] N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.

[22] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[23] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288. ACM, 2016.

[24] E. Park, X. Han, T. L. Berg, and A. C. Berg. Combining multiple sources of knowledge in deep cnns for action recognition. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

[25] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22:23, 2004.

[26] D. Ramachandram and G. W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[27] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM, 2010.

[28] R. Salakhutdinov and G. Hinton. Deep boltzmann machines. In *Artificial Intelligence and Statistics*, pages 448–455, 2009.

[29] R. Salakhutdinov and G. Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, 2009.

[30] S. T. Shivappa, M. M. Trivedi, and B. D. Rao. Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE*, 98(10):1692–1715, 2010.

[31] C. G. Snoek, M. Worring, and A. W. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM, 2005.

[32] K. Sohn, W. Shang, and H. Lee. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149, 2014.

[33] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[34] B. E. Stein and M. A. Meredith. *The merging of the senses.* The MIT Press, 1993.

[35] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative, et al. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. *NeuroImage*, 101:569–582, 2014.

[36] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2088–2095. IEEE, 2013.

[37] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[38] W. Wang, Z. Cui, H. Chang, S. Shan, and X. Chen. Deeply coupled auto-encoder networks for cross-view classification. *arXiv preprint arXiv:1402.2031*, 2014.

[39] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo. Deep multimodal representation learning from temporal data. *CoRR abs/1704.03152*, 2017.