# Deep Multi-Modal Learning with Cascade Consensus⋆

Yang Yang, Yi-Feng Wu, De-Chuan Zhan, Yuan Jiang

National Key Laboratory for Novel Software Technology, Nanjing University,
Nanjing 210023, China
{yangy, wuyf, zhandc, jiangy}@lamda.nju.edu.cn

**Abstract.** Multi-modal deep learning has achieved great success in many applications. Previous works are mostly based on auto-encoder networks or paired networks, however, these methods generally consider the consensus principle on the output layers and always need deep structures. In this paper, we propose a novel Cascade Deep Multi-Modal network structure (CDMM), which generates deep multi-modal networks with a cascade structure by maximizing the correlations between each hidden homogeneous layers. In CDMM, we simultaneously train two nonlinear mappings layer by layer, and the consistency between different modal output features is considered in each homogeneous layer, besides, the representation learning ability can be forward enhanced by considering the raw feature representation simultaneously for each layer. Finally, experiments on 5 real-world datasets validate the effectiveness of our method.

**Keywords:** Multi-modal Learning, Deep Learning, Cascade Structure

## Introduction

In most real-world data analysis problems as image processing, medical detection and social computing, complicated objects can always be described from diverse domains and are naturally with multi-modal feature presentations. However, the representations of various modalities are quite different from each other and it is a challenge to fuse the multiple modalities directly with large discrepancy. Recently, substantial efforts have been dedicated to consider the modal consensus problem, which generally maximizes the correlation between different modalities in the projected subspace. Modern multi-modal subspace learning methods mainly derived from the CCA method [7]. However, these methods are most linear ones, though they can be extended to non-linear models with kernel tricks as KCCA [1], it is difficult to design a suitable kernel and also inefficient to deal with the large datasets.

Recently, multi-modal methods based on deep networks have attracted more attention, which more easily to process large amounts of data[12, 9]. Different from KCCA, these methods generally maximize the correlation between the output features of multiple distinct modal networks for learning more discriminative feature representations. Though deep networks are powerful, it is notable that the structures of deep CCA are

very complicated and always require deeper structure for better representation learning, while leaving the consensus principle of the homogeneous hidden layers among different modal networks without considering during the training phase. Thus, in recent, [19] proposed the gcForest, which generates a deep forest ensemble method with a cascade structure, it is notable that the number of cascade levels can be adaptively determined such that the model complexity can be automatically set.

Inspired by this fact, we therefore propose the CDMM (Cascade Deep Multi-Modal networks) approach to learn multiple maximal correlated deep networks simultaneously, which trains the multiple deep networks with a cascade structure by maximizing the correlation between each homogeneous layers of different modal networks. Specifically, we train multiple deep nonlinear networks layer by layer, and consider the consistency between each homogeneous layer of different modal networks carefully, and then output the processing result to the next level without retraining anymore. As a consequence, the number of layers can be adaptively determined. On the other hand, we forward enhance the network representational learning ability by concatenating the raw input with the output of each hidden layer.

## Related Work

The exploitation of multiple modal subspace learning has attracted many attentions recently. Most proposed methods are mainly derived from the CCA methods, which are devoted to fully utilize the relationships between multiple modalities, and leveraging the consistency among different modalities is one of the significant principles. CCA style subspace learning approaches have been well developed in decades [15, 3, 14]. However, these methods are most linear ones. Thus, Kernel canonical correlation analysis (KCCA) [1] extended the CCA to nonlinear projections. Nevertheless, these methods are limited by the fixed kernel and are difficult to handle a large amount of data.

Therefore, considering deep networks can learn nonlinear feature representations without suffering from the drawbacks of nonparametric models, and have achieved great success in many scenarios [20, 10, 16]. Recently [2] used the DCCA to learn complex nonlinear transformation for two modalities; [17] proposed the DCCAE, which combined the DCCA and deep auto-encoder in one unified framework for more discriminative feature representation. All these methods employ the deep neural network to maximize the correlation on the output feature representation of multiple distinct modalities. Nevertheless, they only expect the output feature representations of different distinct modal networks to be maximally correlated, which need deeper networks for learning better discriminative features, while ignoring the correlations among the homogeneous hidden layers.

To the best of our knowledge, previous linear or kernelized multi-modal methods, which improved the performance by considering the consistency among different modalities on the projected subspace, are difficult to handle a large amount of data and are restricted to the reproducing kernel Hilbert space. Though deep CCA based methods solved these problem, yet they only consider the consensus principle of the output feature representation. In this paper, we propose the CDMM (Cascade Deep Multi-Modal networks), which trains multiple separate deep network with the cascade structure by

considering the consistency between homogeneous hidden layers of different modalities layer by layer, moreover, the representation learning ability can be forward enhanced gradually by considering the raw input for each layer. Consequently, we can obtain a competitive performance with a controlled number of hidden layers.

## Proposed Method

Suppose we have $N$ instances, denoted by $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_N\}$, where each instance $\mathbf{x}_i = [x_{i_1}, x_{i_2}, \cdots, x_{i_d}] \in \mathcal{R}^d$. Meanwhile, in multi-modal learning, instance space can be denoted as $M$ parts without overlap, $v = \{v_1, v_2, \cdots, v_M\}$, where $\mathbf{x}^{v_i} \in \mathbb{R}^{d_i}$ is raw features from the $i$-th modality, $d = d_1 + d_2 + \cdots + d_M$. Without any loss of generalities, each instance $\mathbf{x}_i$ can be denoted as $(\mathbf{x}_i^{v_1}, \mathbf{x}_i^{v_2}, \cdots, \mathbf{x}_i^{v_M})$.

### Deep canonical correlation analysis (DCCA)

Recently, several works are proposed to combine the deep neural network and CCA for better feature representation learning, [2] proposed the deep canonical correlation analysis (DCCA) approach. In DCCA, two deep neural networks $f_{v_1}$ and $f_{v_2}$ are used to extract nonlinear features for different modalities, and then maximize the canonical correlation between the extracted features $f_{v_1}(X^{v_1})$ and $f_{v_2}(X^{v_2})$, which can be represented as:

$$
\begin{aligned}
\max_{\theta_{v_1}, \theta_{v_2}, U, V} \quad & \frac{1}{N} \operatorname{tr}(U^\top f_{v_1}(X^{v_1}) f_{v_2}(X^{v_2})^\top V) \\
s.t. \quad & U^\top (\frac{1}{N} f_{v_1}(X^{v_1}) f_{v_1}(X^{v_1})^\top + r_{v_1} I) U = I, \\
& V^\top (\frac{1}{N} f_{v_2}(X^{v_2}) f_{v_2}(X^{v_2})^\top + r_{v_2} I) V = I,
\end{aligned}
\tag{1}
$$

where $\theta_{v_1}$ and $\theta_{v_2}$ are the weight parameters of networks $f_{v_1}$ and $f_{v_2}$, $U$ and $V$ are the CCA directions which project the output features to the same subspace. $(r_{v_1}, r_{v_2}) > 0$ are regularization parameters for same covariance estimation [4], the $U^\top f_{v_1}(X^{v_1})$ and $f_{v_2}(X^{v_2})^\top V$ are the final projection mapping for testing. Nevertheless, DCCA method and the extensions most concentrate on the correlation between the output feature representation, while ignoring the correlation between homogeneous hidden layers.

### Cascade Deep Multi-Modal networks (CDMM)

In this section, we mainly introduce the concrete steps on learning the discriminative deep multi-modal feature representations with a novel cascade structure, which takes the consensus principle into consideration for each homogeneous hidden layer. We simultaneous train paired deep networks layer by layer, and maximize the consistency between the homogeneous output feature representation of the hidden layers, consequently, we can learn more discriminative feature representations for different modalities, meanwhile, the layers of different modal networks can be adaptively induced by the performance measure, rather than designed in advance manually. On the other hand,

the estimated output of hidden layer forms a feature representation vector, which is then concatenated with the raw feature vector to be the input of the next cascade layer for more robust feature representation.
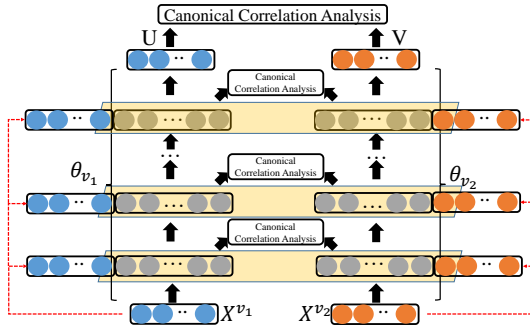


Fig. 1: The overall flowchart. CDMM consists two homogeneous deep networks, which trains with cascade structure different from previous DNN-based method. During the training phase, CDMM maximally correlates each homogeneous hidden layer, besides, the raw features are concatenated with each hidden layer output as next input for more robust representations.

Representation learning in deep neural networks mostly relies on the layer by layer processing of the raw features. Inspired by this recognition, [19] proposed the gc-Forest, which employs a cascade structure, where each layer of the cascade structure receives feature information processed by its preceding level, and output its processing result to the next level. Thus, we propose a novel deep multi-modal networks with the cascade structure as shown in Fig. 1. Specifically, CDMM can be with different deep structure, and for simplicity, we use fully connection matrix for each layer as the DNN-based multi-modal representation learning models as [2], which can be further implemented to convolution structure as CNN-based model. Then, in order to maximize the correlation of each homogeneous layer of different modalities, we consider the hidden layer output (shown in yellow shadows) of each modal networks as the $f_{v_1}(X^{v_1})$ and $f_{v_2}(X^{v_2})$ in Eq. 1, and optimize the parameters of current hidden layers as the DCCA.

It is notable that the objective couples all training samples through the whitening constraints, so stochastic gradient descent (SGD) cannot be applied in a standard way, yet it has been observed by [2] that DCCA can still be optimized efficiently as long as the gradient is estimated using a sufficiently large minibatch. Intuitively, this approach works due to a large minibatch contains enough information for estimating the covariances. Then, the outputs of the estimated hidden layers form a feature representation, considering the representations of the shallow layers of the deep network structure are usually weak features, the hidden layer output is then concatenated with the raw feature vector to be input to the next level of cascade as shown in Fig. 1, i.e., the dimension of the hidden layer output is 1024, and the raw feature is 798 dimensionality, thus, the next level of cascade will receive 1822 (= 1024+798) augmented features. It is notable that the transformed feature vectors, augmented with the raw feature representations, will then be used to train the next grade of cascade multi-modal networks respectively, and the parameters of preceding hidden layers remain unchanged.

Table 1: The correlation of CDMM with compared methods. The significant best classification performance on each dataset is bolded.

| | MNIST | AVLETTER | XRBM | WIKI | FLICKR8K |
|---|---|---|---|---|---|
| CCA | 3.59 | 6.55 | 15.97 | 5.23 | 4.87 |
| KCCA | 1.29 | 2.37 | 35.51 | 7.25 | 6.80 |
| DCCA-10-2 | 7.49 | 7.21 | 42.14 | 10.21 | 7.04 |
| DCCA-10-3 | 7.84 | 7.24 | 43.00 | 10.86 | 7.17 |
| DCCAE-10-2 | 7.81 | 7.37 | 42.23 | 10.09 | 7.08 |
| DCCAE-10-3 | 7.94 | 7.41 | 42.50 | 10.80 | 7.24 |
| CDMM-10-2 | 8.03 | 14.11 | 42.14 | 11.18 | 8.04 |
| CDMM-10-3 | **8.07** | **14.21** | **43.20** | **11.40** | **8.06** |

## Experiments

### Datasets and Configurations

CDMM can learn more discriminative multi-modal feature representation with self-adaption networks. In this section, we will provide the empirical investigations and performance comparison of CDMM. In particular, we demonstrate these phenomenon on 5 real datasets, i.e., MNIST generates two modal data using the original MNIST dataset [11]. As in [17], we randomly rotate the images and the resulting images are used as modal $v_1$ inputs. For each $v_1$, we randomly select an image of the same identity from the original dataset, add independent random noise to obtain the corresponding modal $v_2$ sample; AVLETTER contains 10 speakers speaking the letters A to Z at 3 times for each one. This dataset provides pre-extracted lip regions of $60 \times 80$ pixels as modal $v_1$ and audio features (raw audio is not provided) Mel-Frequency Cepstrum Coefficient (MFCC) as modal $v_2$; XRBM follows the setup of [17]. Inputs to multi-modal feature learning are acoustic features as modal $v_1$, and articulatory features concatenated over a 7-frame window around each frame as modal $v_2$; WIKI [13] is a rich-text web document dataset with images, which has 2,866 documents extracted from Wikipedia as modal $v_1$. Each document is accompanied by an image as modal $v_2$. Text is represented by TF-IDF feature with 7343-dimensional; FLICKR8K [6] consists of 8,000 images that are each paired with five different captions, similarly, we denote the image as model $v_1$ and text information as modal $v_2$.

For WIKI and FLICKR8K datasets, 70% instances are chosen as training set, 20% are chosen as validation set and the remains are test set as [18]. In other three datasets, training and test splits are provided by [2, 8]. For DNN-based models, feature mappings $(f_{v_1}, f_{v_2})$ are implemented by networks of 2 or 3 hidden layers, each of 1,024 sigmoid units, and a linear output layer of L units, we refer to a DNN-based model with an output size of o and d layers (including the output) as *-o-d, i.e., CDMM-o-d, DCCA-o-d, DCCAE-o-d. The two networks $(f_{v_1}, f_{v_2})$ are pre-trained in a layerwise manner using restricted Boltzmann machines [5], and SGD is used for optimization with minibatch size as 800, learning rate and momentum tuned on the tuning set, a small weight decay parameter of $10^{-4}$ is used for all layers.

(*a*) **t-SNE embedding of the first hidden layers of CDMM**



(*b*) **t-SNE embedding of the third hidden layers of CDMM**
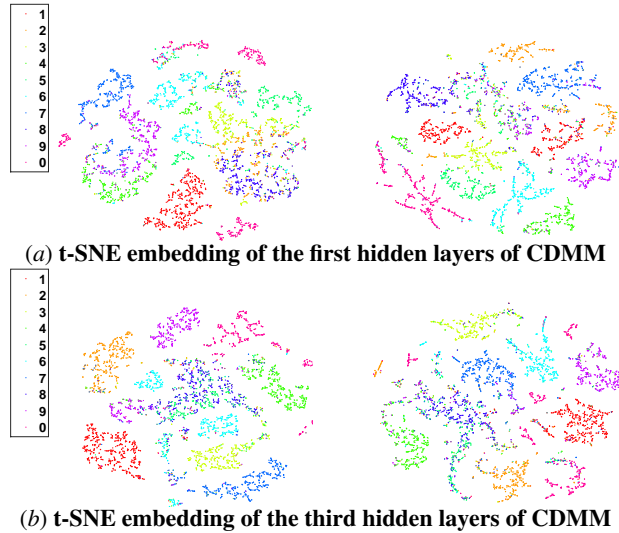
Fig. 2: t-SNE embedding of the projected MNIST and noisy MNIST digits. Left represents the projected original MNIST modality, and right denotes the noisy MNIST modality. Each sample is denoted by a maker located at its coordinates of embedding and color coded by its label.

**Comparing with CCA-based Multi-Modal Methods**

CDMM is firstly compared to linear and kernerlized multi-modal CCA-based methods. Since there are deep networks in CDMM, DNN-based multi-modal methods are also compared in the experiments. In detail, the compared methods are listed as: Linear CCA (CCA), Kernel CCA, DCCA, DCCAE.

Table 1 compares the total correlation on the test sets obtained for the 10 most correlated dimensions with compared methods. It clearly reveals that on all datasets, with the same number of layers, the CDMM total correlation is the highest. Besides, note that CDMM also has exceeded other compared methods only with 2 layers on most datasets except XRBM. Thus, CDMM can acquire more discriminative feature representation with shallow deep network structures.

**Investigation on Embedding of Different Layers**

In order to explore the influence of the cascade structure, more experiments are conducted. We qualitatively investigate the features by embedding the projected features in 2D using t-SNE of each pair homogeneous hidden layers, the resulting visualizations are given in Fig. 2. Each sample is denoted by a marker located at its coordinates of embedding and color coded by its label. Due to the page limits, we only list the noisy MNIST digits dataset for verification. From the Fig.2, we can find that CDMM gives more accurate embedding with the cascade structure from the initial layers, i.e., CDMM pushed different digits far apart from the initial layers.
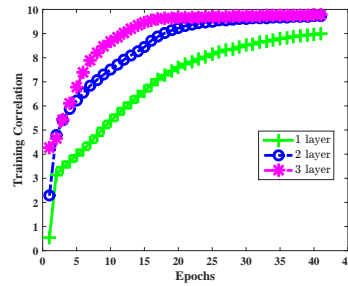
Fig. 3: The correlation of different layers in training phase of Noisy MNIST digit dataset.

**Empirical Investigation on Convergence**

To investigate the convergence of CDMM iterations empirically. The objective function value, i.e., the value of Eq. 1 of CDMM in each iteration of each homogeneous layers are recorded. Due to the page limits, only results on noisy MNIST digits dataset are plotted in Fig. 3. It clearly reveals that the correlation value between different modalities increases as the iterations increase, and the performance is stable after several layers in Fig. 3, i.e., the variations between the correlation values of second hidden layers and third hidden layers less than the predefined threshold, which can be used to control the layers self-adaptively.

## Conclusion

Previous DNN-based multi-modal networks have been used for learning more discriminative feature representations. However, these methods only consider the consensus principle on output layers and always need predefined the network structures, i.e., number of layers, which lead complex deep network structures and high computation expense, while neglect considering the correlation between the homogeneous hidden layers of different deep modal structures. In this paper, we propose a novel Cascade Deep Multi-Modal networks (CDMM). This method generates a deep multi-modal networks with a cascade structure which fully maximizes the correlations between homogeneous hidden layers of different modal networks, and can acquire representative networks with shallow layers. Besides, the representational learning ability can be further enhanced by concatenating the raw features with each hidden layer output. And empirical studies show that we can learn more discriminative features with shallow layers. How to extend the scalability with improved performance is an interesting future work.

## References

1. Akaho, S.: A kernel method for canonical correlation analysis pp. 263–269 (2007)
2. Andrew, G., Arora, R., Bilmes, J.A., Livescu, K.: Deep canonical correlation analysis. In: Proceedings of the 30th International Conference on Machine Learning. pp. 1247–1255. Atlanta, GA (2013)

3. Arora, R., Mianjy, P., Marinov, T.V.: Stochastic optimization for multiview representation learning using partial least squares. In: Proceedings of the 33rd International Conference on Machine Learning. pp. 4847–4855. New York, NY (2016)

4. Hardoon, D.R., Szedmak, S.R., Shawe-Taylor, J.R.: Canonical Correlation Analysis: An Overview with Application to Learning Methods. MIT Press (2004)

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

6. Hodosh, M., Young, P., Hockenmaier, J.: Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. JAIR 47, 853–899 (2013)

7. Hotelling, H.: Relations between two sets of variates. Biometrika 28(3-4), 321–377 (1936)

8. Hu, D., Li, X., Lu, X.: Temporal multimodal learning in audiovisual speech recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3574–3582. Las Vegas, NV (2016)

9. Kan, M., Shan, S., Chen, X.: Multi-view deep network for cross-view classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4847–4855. Las Vegas, NV (2016)

10. Kang, G., Li, J., Tao, D.: Shakeout: A New Regularized Deep Neural Network Training Scheme. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. pp. 1751–1757. Phoenix, Arizona (2016)

11. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition 86(11), 2278C2324 (1998)

12. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning. pp. 689–696. Bellevue, Washington (2011)

13. Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A New Approach to Cross-modal Multimedia Retrieval. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 251–260. Firenze, Italy (2010)

14. Rupnik, J., Shawe-Taylor, J.: Multi-view canonical correlation analysis. In: Slovenian KDD Conference on Data Mining and Data Warehouses. pp. 1–4. Ljubljana, Yugoslavia (2010)

15. Shrivastava, A., Rastegari, M., Shekhar, S., Chellappa, R., Davis, L.S.: Class consistent multi-modal fusion with binary features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2282–2291. Boston, MA (2015)

16. Tian, F., Gao, B., Cui, Q., Chen, E., Liu, T.Y.: Learning Deep Representations for Graph Clustering. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence. pp. 1293–1299. Qu—bec, Canada (2014)

17. Wang, W., Arora, R., Livescu, K., Bilmes, J.: On deep multi-view representation learning. In: Proceedings of the 32nd International Conference on Machine Learning. pp. 1083–1092. Lille, France (2015)

18. Yang, Y., Zhan, D.C., Jiang, Y.: Deep learning for fixed model reuse. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. pp. 1033–1039. New York, NY (2017)

19. Zhou, Z.H., Feng, J.: Deep forest: Towards an alternative to deep neural networks. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia (2017)

20. Zhu, X., Huang, Z., Wu, X.: Multi-view visual classification via a mixed-norm regularizer. In: Proceedings of the 17th Pacific-Asia Knowledge Discovery and Data Mining. pp. 520–531. Gold Coast, Australia (2013)