Journal of Electronic Imaging

JElectronicImaging.org

Discriminative feature representation for image classification via multimodal multitask deep neural networks

Shuang Mei
Hua Yang
Zhouping Yin

SPIE. IS&T imaging.org

Shuang Mei, Hua Yang, Zhouping Yin, "Discriminative feature representation for image classification via multimodal multitask deep neural networks," *J. Electron. Imaging* **26**(1), 013023 (2017), doi: 10.1117/1.JEI.26.1.013023.

Discriminative feature representation for image classification via multimodal multitask deep neural networks

Shuang Mei, Hua Yang,* and Zhouping Yin
Huazhong University of Science and Technology, State Key Laboratory of Digital Manufacturing Equipment and Technology, Wuhan, China

Abstract. A good image feature representation is crucial for image classification tasks. Many traditional applications have attempted to design single-modal features for image classification; however, these may have difficulty extracting sufficient information, resulting in misjudgments for various categories. Recently, researchers have focused on designing multimodal features, which have been successfully employed in many situations. However, there are still some problems in this research area, including selecting efficient features for each modality, transforming them to the subspace feature domain, and removing the heterogeneities among different modalities. We propose an end-to-end multimodal deep neural network (MDNN) framework to automate the feature selection and transformation procedures for image classification. Furthermore, inspired by Fisher's theory of linear discriminant analysis, we improve the proposed MDNN by further proposing a multimodal multitask deep neural network (M2DNN) model. The motivation behind M2DNN is to improve the classification performance by incorporating an auxiliary discriminative constraint to the subspace representation. Experimental results on five representative datasets (NUS-WIDE, Scene-15, Texture-25, Indoor-67, and Caltech-101) demonstrate the effectiveness of the proposed MDNN and M2DNN models. In addition, experimental comparisons of the Fisher score criterion exhibit that M2DNN is more robust and has better discriminative power than other approaches. [DOI: 10.1117/1.JEI.26.1.013023]

Keywords: feature representation; feature selection; feature transformation; deep neural networks.

Paper 16779 received Sep. 10, 2016; accepted for publication Jan. 24, 2017; published online Feb. 24, 2017.

1 Introduction

Learning effective and discriminative representations is a long-standing goal in the field of image classification. Many researchers have attempted to design descriptors that suffice to represent image characteristics.[1,2] Although great successes have been achieved using many well-known descriptors, the task still has some shortcomings that strongly affect performance. For example, the scale-invariant feature transform (SIFT)[3] is a widely used feature descriptor that has made huge contributions to the fields of activity recognition,[4] image retrieval,[5] motion estimation,[6] and others. However, this descriptor fails to consider the global context, which can resolve ambiguities when an image has multiple similar subregions.[7] Deep neural network (DNN)-based models are efficient at encoding the semantic information of targets; however, they are too coarse to extract the more detailed information required for visual tracking,[8] object verification,[9] and so on. Consequently, the existing descriptors are good at extracting characteristics only along specific aspects.

To gain more robust and complete representations for common tasks, multimodal feature learning has been widely used in recent years. Traditional multimodal features involve relating information from multiple sources in signal processing, while the broad concepts also refer to features extracted by multiple descriptors in the image-processing field. Therefore, rather than focusing only on single-modal feature representation, multimodal features have been shown to be more sufficient and effective on many occasions.[10,11] Exploring the potentially complementary characteristics of different modalities, reducing the heterogeneities among them, and then transforming them to a common latent space are the key issues in this research. These issues can be summarized into two subproblems in feature engineering: feature selection and feature transformation. The former selects efficient features from each modality. The latter transforms the multimodal features to a common latent space and removes the heterogeneities among them. Optimizing these two subproblems is crucial in learning multimodal features.

Numerous researchers have devoted substantial efforts to these issues. Xu et al.[12] integrated information from multiple views to generate a latent intact representation and found that the complementarities among these views were beneficial for the stability and generalization of the latent feature representation. Their multiview intact space learning algorithm exhibited an effective and promising performance in 3-D model reconstruction. However, it directly transformed information from different views without removing the ineffective features, which may decrease the overall performance. Furthermore, procedures for the intact feature transformation and feature classification were separately optimized, which may result in a suboptimal performance. These two drawbacks could potentially weaken the ability for discriminative image classification. Luo et al.[13] proposed a large margin multimodal multitask feature extraction algorithm (LM3FE) to explore the nature of different modalities. This approach exhibits good performance for multimodal feature extraction. However, the partitive optimization of each subproblem in

*Address all correspondence to: Hua Yang, E-mail: huayang@mail.hust.edu.cn

1017-9909/2017/$25.00 © 2017 SPIE and IS&T

feature extraction may limit the performance for image classification. Zhou et al.[14] proposed a fully conjugate multiple kernel learning (MKL) algorithm to train a classifier for head detection. MKL is a classical intermediate-combination-based method and is completely different from the two algorithms described above. It utilizes mixed kernels to classify features from different modalities and shows good performance for image recognition. Deep-learning-based methods have recently become popular in this research area and have been utilized in applications such as activity recognition,[15,16] emotion recognition,[17] and image classification.[18] Ngiam et al.[19] proposed an unsupervised multimodal deep-learning (MDL)-based method for learning the shared representation between modalities of audio-only data and video-only data. Experiments demonstrated that this model is capable of audio-visual speech classification. Ren et al.[20] proposed a maximum margin multimodal deep neural network (3mDNN) model for learning joint features from multiple descriptors. The famous maximum margin strategy was utilized to constrain features to be more discriminative. This model exhibits good performance on the classical Scene-15, Indoor-60, and Caltech-101 datasets. However, these two multimodal feature representation models (Ngiam's and Ren's) share the same shortcomings: the feature representation and recognition procedures are separately optimized, which may result in suboptimal results and damage the overall performance. Motivated by the properties of good generative models of multimodal data, Sohn et al.[21] proposed a multimodal representation learning framework. Rather than learning with the maximum likelihood, the model was trained to minimize the variation of information. This objective enables learning of a good shared representation of multiple heterogeneous data modalities that better predicts missing input modalities. It showed state-of-the-art performance on many datasets. However, the features from different modalities were directly transformed—but without removing the heterogeneities, which may introduce unwanted feature components. Kavi et al.[16] coupled convolutional neural networks (ConvNets) with long short-term memory (LSTM) networks for multiview driver activity recognition. This ConvNets LSTM architecture combined the automatic feature-extraction capability of neural networks with additional memory in the temporal domain and significantly improved the accuracy of the driver activity recognition system. Song et al. also proposed a ConvNets LSTM model to tackle the egocentric activity recognition problem in which the ConvNets was utilized to learn the spatial and temporal features from egocentric videos and the LSTM model was used to learn the features from multiple sensor streams. Experimental results demonstrated the powerful performance of this multistream multimodal architecture. However, these two ConvNets LSTM models are both unsuitable for use in traditional image classification tasks, which do not include temporal information.

To overcome the limitations of the existing methods, we proposed a multimodal deep neural network (MDNN) model for image classification. This model is an end-to-end network framework that merges feature selection, transformation, and classification procedures into a single unified architecture. It combines the feature selection attribute of the rectified linear units (ReLu) activation function[22] with the representational power of multilayer neural networks.

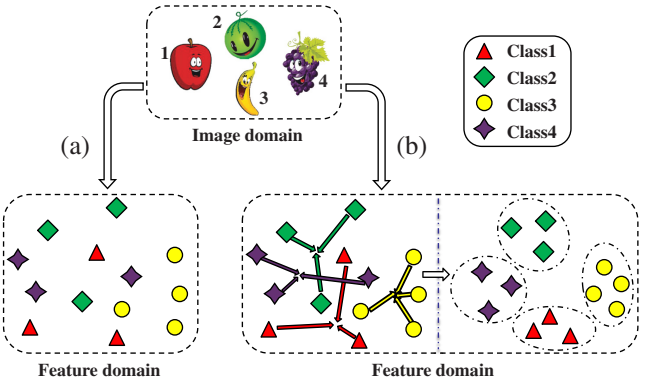

Fig. 1 Image classification task: (a) traditional classification task that maps features only to the output domain and (b) discriminative classification task that not only maps features to the output domain but also makes them discriminative.

In this way, features from different modalities can be automatically selected and transformed based on their contributions. However, the MDNN model is still not good enough. As shown in Fig. 1(a), the main objective of the MDNN is to map features from the feature space to the output domain with minimum classification error—just as the traditional methods do. Thus, it fails to constrain the generated features to be more discriminative. Inspired by Fisher's theory of linear discriminant analysis, we have improved the MDNN model in the multitask field and propose a more powerful multimodal multitask deep neural network (M2DNN) model. The auxiliary task in M2DNN is able to promote the original classification task by maximizing the between-class scatter and minimizing the within-class scatter of the learned features. Therefore, the shared feature representations tend to be more discriminative, as shown in Fig. 1(b).

We evaluated the proposed M2DNN model using four criteria: the average classification accuracy (Accu),[23] the mean ranking performance computed under each label (mAp),[24] the average agreement of the data class labels (mPre),[23] and the average F-score criterion (F-score).[23] Then, we conducted real-world experiments on the NUS-WIDE, Scene-15, Texture-25, Indoor-67, and Caltech-101 datasets. The remainder of this paper is organized as follows. Related works are summarized in Sec. 2. The procedures involved in the MDNN and M2DNN methods are introduced in detail in Sec. 3; specifically, we present strategies for recognizing a candidate image with M2DNN. The results of experiments and analyses are provided in Sec. 4. Finally, Sec. 5 concludes the paper.

2 Related Works

As described above, the M2DNN model is a multimodal multitask deep-learning-based framework for image classification. The multimodal feature learning technique aims to gain an intact feature representation from multiple modalities, while the multitask learning technique attempts to improve the performance of the classification task by assisting a related auxiliary task. Many researchers have focused on these fields recently. We will briefly summarize some relevant methods.

2.1 *Multimodal Feature Representation Methods*

The various multimodal feature representation methods have different appearances based on the combination strategy they use, as shown in Fig. 2. In Figs. 2(a)–2(c), we name these

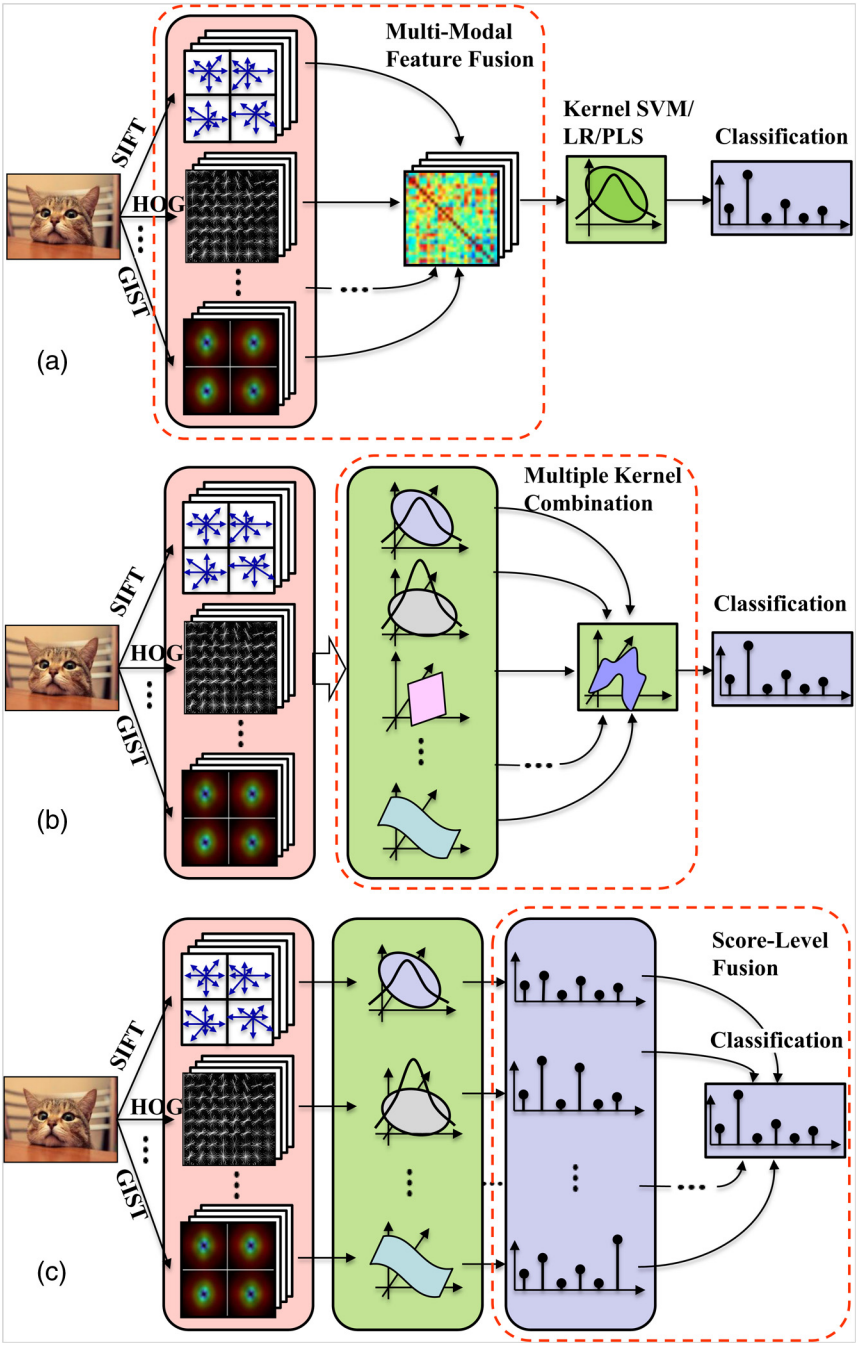

Fig. 2 The multimodal feature representation methods: (a) prior combination—the subspace feature fusion method, (b) intermediate combination—the multiple kernel learning method, and (c) late combination—the score-level fusion method.

models the prior, intermediate, and late combination approaches, respectively, based mainly on the corresponding position of the combination strategy in the entire processing chain. Detailed descriptions and comparisons are discussed as follows.

2.1.1 *Prior combination: the subspace feature fusion method*

The prior combination approach in multimodal feature representation also refers to the subspace feature fusion method. It aims to find a common latent space and maps features from different modalities to this space. As shown in Fig. 2(a), the fused subspace features are regarded as the final feature representations for image classification. Directly concatenating features from multiple modalities is a simple prior combination method that is efficient and saves computational time; however, this approach generally includes poor features because it is without feature selection. Moreover, it fails to reduce the heterogeneities among different modalities. Therefore, extracting discriminative feature representations from the shared modality is of crucial importance for the prior combination methods. The majority of researchers in the multimodal feature representation field have focused

on this prior combination approach. The MDL,[19] 3mDNN,[20] and LM3FE[13] are all superior subspace feature fusion methods. The proposed MDNN and M2DNN models also fall into this prior combination category.

2.1.2 *Intermediate combination: the multiple kernel learning method*

The intermediate combination approach refers to its association with multiple classifiers via different kernels. These kernels may correspond to the use of different similarity concepts or to the different information from multiple modalities.[25] The architecture of the intermediate combined multimodal feature representation methods is shown in Fig. 2(b). Many practical applications have demonstrated that using multiple kernels instead of a single kernel is a more useful and promising approach.[26] Guillaumin et al.[18] proposed a semisupervised-learning-based approach with MKL to deal with the classification scenario in which keywords are associated with the training images. The model exhibited good performance in classifying images even when some image labels were missing. Many current algorithms, such as the SimpleMKL[27] and sparse MKL,[28] have been derived from the MKL framework, and they have been utilized in numerous studies.[25,27]

2.1.3 *Late combination: the score-level fusion method*

As shown in Fig. 2(c), the late combined approach mainly refers to the score-level fusion method.[11] The cotraining, co-EM, and coregression style algorithms are classical late combined methods. This approach forces them to be consistent across multiple modalities. Late combined methods are typically used in semisupervised learning. In this paper, we will mainly concentrate on the prior and intermediate combined approaches for supervised image classification.

2.2 *Multitask Learning-Based Methods*

Multitask learning is a learning paradigm that simultaneously learns a series of task-specific models by capturing the inner correlations between different tasks. Multitask learning has been widely employed in applications, such as face recognition[29] and visual tracking.[30] If all the tasks in a multitask learning-based model are well designed, the joint learning procedure of these tasks can result in better performance than learning each task independently.[29,30] Therefore, for an image classification task (the main task in this paper), the problem of choosing the auxiliary task that best promotes the performance of the main task is crucial. Yan et al.[31] proposed a multitask linear discriminant analysis framework for multiview action image classification. In this work, the auxiliary task is used to optimize the generated features, making them more discriminative. Ding et al.[29] jointly learned features from different poses to profit from the latent interpose correlations. Each pose in this model is viewed as a task. For general image classification, we assigned a feature similarity constraint task to the proposed MDNN model to achieve the more discriminative and robust multimodal framework called M2DNN. Section 3 describes the procedures in detail.

3 Proposed Approach

In this section, we first propose the MDNN model. Then, by extending MDNN to the multitask learning field, we propose the M2DNN model. Finally, we compare and discuss the discriminative attributes of M2DNN and MDNN.

3.1 *Multimodal Deep Neural Network*

The MDNN model is an end-to-end MDNN framework, as shown in Fig. 3. It combines the feature selection attribute of the ReLu activation function[22] with the representational power of multilayer neural networks.

Assume that the training set is $\xi = \{(x_1^{(i)}, x_2^{(i)}, \cdots, x_P^{(i)}; y^{(i)}) | i = 1, \cdots, N\}$. Here, P and N are the numbers of modalities and training samples, respectively, and $x_k^{(i)} \in \mathbb{R}^{V_k}$ refers to the feature of the k'th modality for the i'th sample. The $y^{(i)} \in \mathbb{R}^{N_C}$ is the corresponding label. It is a one-hot code in which all the elements are 0 except for the category to which this sample belongs. N_C is the number of total categories. Consequently, the object function of the model MDNN can be expressed as follows:

$$\arg\min_{W} \sum_{i=1}^{N} L(W; y^{(i)}, x_1^{(i)}, x_2^{(i)}, \cdots, x_P^{(i)}) + \lambda \cdot \Omega(W),$$
$$\text{s.t. } \lambda > 0, \tag{1}$$

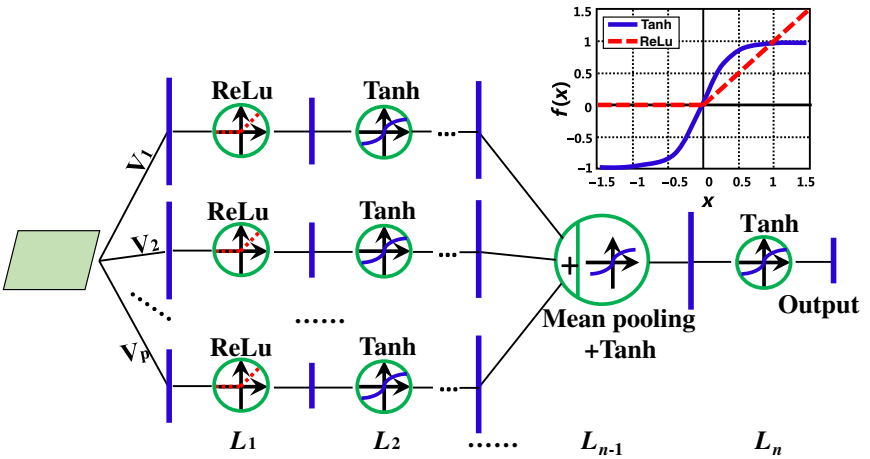

Fig. 3 Architecture of the MDNN model, the top-right image shows the activation function.

where λ is the trade-off parameter, L refers to the loss function, and Ω is the regularity term. Here, the Cauchy estimator is used to evaluate the loss between ground truths and fitted values. The Cauchy estimator is robust and has been used in discriminant learning for object recognition,[32] data reconstruction,[12] and many other applications.[33] Figure 4(a) illustrates the Cauchy estimator functions (in green) by $\rho(x) = \log[1 + (x/c)^2]$ and the L_2 estimator (square error, in blue) by $\rho(x) = x^2/2$. The former is a symmetric, positive-definite function with a unique minimum at zero and increases less than the least squares function, as shown in Fig. 4(a). The corresponding influence functions, with $\psi(x) = 2x/(x^2 + c^2)$ and $\psi(x) = x$, are shown in Fig. 4(b). As the figure shows, the influence function of the L_2 estimator is linear, that is, the influence of a sample increases linearly with the size of its error. Thus, the model may be easily destroyed by outliers. This analysis confirms the nonrobustness of the L_2 estimator.[12] However, for the Cauchy estimator, the influence of any single observation is insufficient for yielding a significant offset of the estimator because its influence function has upper and lower bounds, as shown in Fig. 4(b). With those bounds, the influence of any outlier will be constrained. Furthermore, the Cauchy estimator theoretically has a breakdown point of nearly 50%,[34] which means that the results of the Cauchy estimator are more reliable than those of traditional estimators. Therefore, we employed the Cauchy estimator in MDNN. The regularity term, which refers to the accumulated constraints for each modality, is used to avoid degradation and invalidation of the model. Consequently, the MDNN model can be optimized iteratively by solving the following problem:

$$\begin{aligned}&\arg\min_{W} \sum_{i=1}^{N} \log\left\{1 + \frac{\|y^{(i)} - s[f(W, x_1^{(i)}, x_2^{(i)}, \cdots, x_P^{(i)})]\|_F^2}{c^2}\right\} \\ &\quad + \lambda \cdot \sum_{l} \sum_{p} \|w_{(l,p)}\|_F^2, \\ &\text{s.t. } \lambda > 0, \quad c > 0, \quad \forall\, l = L_1, \cdots, L_n, \quad \forall\, p = 1, \cdots, P,\end{aligned} \tag{2}$$

where $w_{(l,p)}$ refers to the coefficients of the p'th modality in the l'th hidden layer, $s(\cdot)$ is a softmax regression classifier in which each component of its output denotes the probability of the class label to which the sample belongs, and $f(W, x_1^{(i)}, x_2^{(i)}, \cdots, x_P^{(i)})$ is the final feature representation of the i'th sample and is the comprehensive result of the feature selection and transformation processes. That is

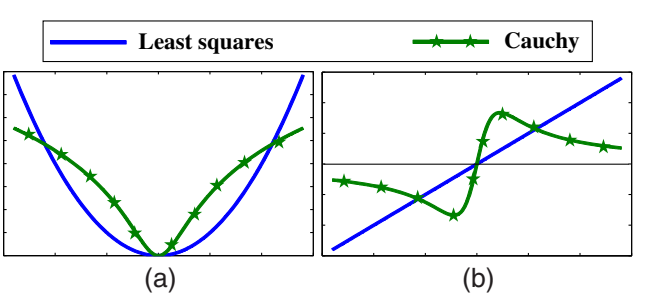

Fig. 4 Examples of the robust estimators: (a) the estimator function and (b) the influence function.

$$f(W, x_1^{(i)}, x_2^{(i)}, \cdots, x_P^{(i)}) = T[W_T, S(W_S, x_1^{(i)}, x_2^{(i)}, \cdots, x_P^{(i)})], \tag{3}$$

where $T(\cdot)$ and $S(\cdot)$ refer to the feature transformation and selection procedures, respectively, and W_T and W_S are the corresponding coefficients. According to maximum likelihood estimation theory and the object function [Eq. (2)], we can extend the branch and bound algorithm[35] and derive the multimodal feature selection criteria as follows:

Criteria1: the multimodal feature selection criteria

Given the input dataset ξ, tabled as N examples and P modalities with space $\mathbb{R}^{V_1}, \mathbb{R}^{V_2}, \cdots, \mathbb{R}^{V_P}$, $\xi = \left\{\left(x_{(1,1)}^{(i)}, \cdots, x_{(1,V_1)}^{(i)}\right), \left(x_{(2,1)}^{(i)}, \cdots, x_{(2,V_2)}^{(i)}\right), \cdots, \left(x_{(P,1)}^{(i)}, \cdots, x_{(P,V_P)}^{(i)}\right); y^{(i)} | i = 1, \cdots, N\right\}$, V_k is the dimension of the k'th modality. The objective of the feature selection process is to find a subspace of $m_i (i = 1, 2, \cdots, P)$ features for each of the P modalities that achieves the best possible classification performance. That is, for any set $B_k^{(i)} \subseteq \left\{x_{(k,1)}^{(i)}, \cdots, x_{(k,V_k)}^{(i)}\right\}$ in the k'th modality, $B_k^{(i)} \neq \emptyset$, if the equations

$$\begin{aligned}&\prod_{i=1}^{N}\left|1 + \frac{\|y^{(i)} - f(W, A_1^{(i)}, A_2^{(i)}, \cdots, A_k^{(i)}, \cdots, A_P^{(i)})\|_F^2}{c^2}\right| < \\ &\prod_{i=1}^{N}\left|1 + \frac{\|y^{(i)} - f(W, A_1^{(i)}, A_2^{(i)}, \cdots, A_k^{(i)} \bigcup B_k^{(i)}, \cdots, A_P^{(i)})\|_F^2}{c^2}\right|,\end{aligned} \tag{4}$$

and

$$\begin{aligned}&\prod_{i=1}^{N}\left|1 + \frac{\|y^{(i)} - f(W, A_1^{(i)}, A_2^{(i)}, \cdots, A_k^{(i)}, \cdots, A_P^{(i)})\|_F^2}{c^2}\right| < \\ &\prod_{i=1}^{N}\left|1 + \frac{\|y^{(i)} - f(W, A_1^{(i)}, A_2^{(i)}, \cdots, A_k^{(i)} - B_k^{(i)}, \cdots, A_P^{(i)})\|_F^2}{c^2}\right|,\end{aligned} \tag{5}$$

can always be met. With $A_k^{(i)} \subseteq \{x_{(k,1)}^{(i)}, \cdots, x_{(k,V_k)}^{(i)}\}$, the feature selection set $A_k^{(i)} (k = 1, 2, \cdots, P)$ is the optimal feature selection set for the classification task.

Equations (4) and (5) show that adding or removing any components in the optimal feature selection set will increase the classification error and reduce the recognition performance. However, just as with the branch and bound method,[35] searching for the optimal set is still impractical for problems with very large feature sets because of the huge computational complexity. Here, we adopt the DNN with a mini-batch technique in MDNN to reduce the computational cost. The ReLu activation function is applied to conduct the feature selection process automatically. Using the ReLu mechanism

$$x_{(k,q)} = \max(0, x_{(k,q)}), \quad k = 1, 2, \cdots, P; \quad q = 1, 2, \cdots, V_k, \tag{6}$$

where $x_{(k,q)}$ refers to the q'th component of the features in the k'th modality and V_k is the dimension of the features in

this modality. Thus, every component in each modality will be regularized to 0 if it has no effect on the model; otherwise, it remains unchanged. The ReLu activation function is shown in the top-right window in Fig. 3.

In addition, multilayer perceptron neural networks with the Tanh activation function are used for the feature transformation process in MDNN. Assuming $\mathrm{Out}_{(l,p,i)}$ is the output of the i'th neuron in the l'th layer and p'th modality, the transformed value of the j'th neuron in the next layer can be expressed as

$$\mathrm{Out}_{(l+1,p,j)} = \mathrm{Tanh}\left(\sum_i w_{(l,p,i,j)}\mathrm{Out}_{(l,p,i)} + b_{(l+1,p,j)}\right), \quad (7)$$

where $\mathrm{Tanh}(x) = (e^x - e^{-x})/(e^x + e^{-x})$, $b_{(l+1,p,j)}$ refers to the bias, and $w_{(l,p,i,j)}$ is the corresponding coefficient (the bias b can be concatenated with the coefficient w and will not be expressed explicitly for concise illustration in later statements). The multilayer perceptron neural network is an effective nonlinear feature transformation model that has exhibited good performance in many situations.[36]

The MDNN model jointly optimizes the feature selection, feature transformation, and softmax regression procedures in the same framework to achieve smaller classification errors. The reduction and transformation of the irrelevant or redundant features will be automatically conducted by iteratively solving Eq. (2). Thereby, an effective feature representation for multimodal image classification can be achieved.

3.2 *Multimodal Multitask Deep Neural Network*

The MDNN model examines the relationship between the feature domain and the output domain just as traditional classification models do; however, it cannot ensure that the attributes of the learned feature representations are discriminative, as shown in Fig. 1. That is, the features of samples from the same categories may be far apart, while those of samples from different categories may be close together. Therefore, any data perturbation can affect the feature representation results and cause the model to become unstable. Therefore, it is crucial to increase the discriminative attributes of the MDNN model while maintaining its original advantages.

However, how increasing the discriminative power of the feature representations learned by MDNN is a problem. Inspired by Fisher's linear discriminant analysis theory (Appendix), which attempts to maximize the "interclass distance" while minimizing the "intraclass scatter," we propose the M2DNN model for discriminative image classification. That is, we want to optimize the reconstruction loss as the MDNN model does but also constrain the learned feature representations as the Fisher criterion does. As shown in Fig. 5, we first pair two MDNN models to generate the final feature representations. Then, we add an auxiliary task to measure the similarity constraint (also the metric distance) between these two features (one for each MDNN model). Finally, the whole network is trained end-to-end by optimizing the reconstruction losses of the two MDNN models and the similarity constraint of the similarity term. Consequently, the global optimization property of the MDNN and the discrimination power brought by the auxiliary task are all considered, and the original multimodal feature representation model (MDNN) is extended to the multitask field. For the auxiliary task, the similarity constraint criteria are elaborated in Criteria2 as follows.

Criteria2: the similarity constraint criteria for M2DNN

Let ϑ_i^1 and ϑ_i^2 denote the feature representations for the i'th pair images; y_i^1 and y_i^2 are the corresponding labels, and $M_i^{1,2}$ is the penalty for their similarity constraint, which can be expressed as follows:

$$M_i^{1,2} = \begin{cases} \max[D(\vartheta_i^1, \vartheta_i^2) - d_0, 0] & y_i^1 = y_i^2 \\ \max[-D(\vartheta_i^1, \vartheta_i^2) + d_0, 0] & y_i^1 \neq y_i^2 \end{cases},$$
$$\text{s.t. } d_0 > 0, \quad (8)$$

where $D(\cdot)$ denotes the metric distance function. The M2DNN model uses the Euclidean distance function. Here, d_0 is the margin (cutoff value) that separates samples in different categories. It can be obtained by cross-validation.[37]

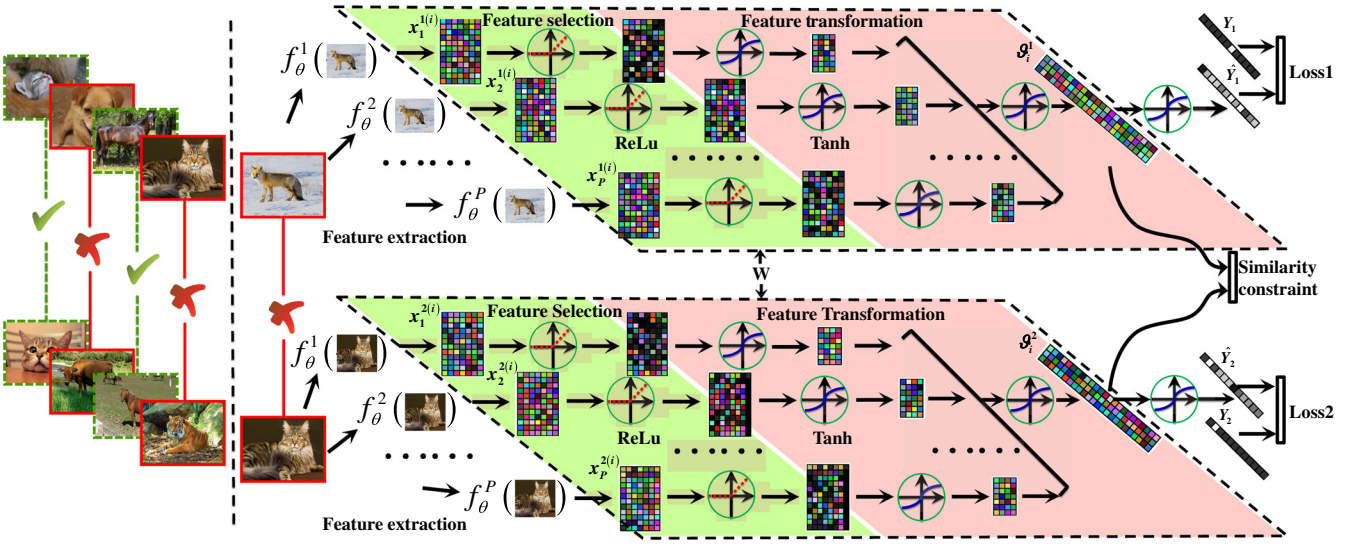

Fig. 5 Architecture of the M2DNN model.

Criteria2 aims to constrain the metric distance of sample pairs based on their similarities and categories. In other words, for a given image pair, if the two sample images belong to the same category and their Euclidean distance is larger than the margin or if they belong to different categories and their Euclidean distance is smaller than the margin, they will be punished by the residual $|D(\vartheta_i^1, \vartheta_i^2) - d_0|$.

Therefore, the model M2DNN in Fig. 5 can be expressed as follows:

$$\arg\min_W \Phi(W) = \arg\min_W \sum_{i=1}^{N_p} \underbrace{L^1 + L^2}_{①} + \lambda \cdot \underbrace{M^{1,2}}_{②} + \gamma \cdot \underbrace{\Omega}_{③},$$
$$\text{s.t. } \lambda > 0, \qquad \gamma > 0, \tag{9}$$

where Φ refers to the total losses of the model and N_p refers to the number of sample pairs. Here, ①, ②, and ③ refer to the reconstruction loss term, the similarity constraint term, and the weight regularity term, respectively. The parameters λ and γ determine the trade-off among them and are obtained by cross-validation. Taking the specific representations of these terms into consideration, the object function Eq. (9) can be expanded as follows:

$$\begin{aligned}
&\arg\min_W \sum_{i=1}^{N_p} \log\left\{1 + \frac{\|y_i^1 - s[f(W, x_1^{1(i)}, x_2^{1(i)}, \cdots, x_P^{1(i)})]\|_F^2}{c^2}\right\} \\
&+ \log\left\{1 + \frac{\|y_i^2 - s[f(W, x_1^{2(i)}, x_2^{2(i)}, \cdots, x_P^{2(i)})]\|_F^2}{c^2}\right\} \\
&+ \lambda \cdot M_i^{1,2} + \gamma \cdot \sum_l \sum_p \|w_{(l,p)}\|_F^2, \\
&\text{s.t. } \lambda > 0, \quad \gamma > 0, \quad c > 0, \quad \forall\, l = L_1, \cdots, L_n, \\
&\quad \forall\, p = 1, \cdots, P,
\end{aligned} \tag{10}$$

where $(y_i^1; x_1^{1(i)}, x_2^{1(i)}, \cdots, x_P^{1(i)})$ and $(y_i^2; x_1^{2(i)}, x_2^{2(i)}, \cdots, x_P^{2(i)})$ refer to the two samples of the i'th pair, respectively. The following equation

$$\begin{aligned}
M_i^{1,2} = &\mathbf{1}(y_i^1 = y_i^2) * \max[D(\vartheta_i^1, \vartheta_i^2) - d_0, 0] + [1 - \mathbf{1}(y_i^1 = y_i^2)] \\
&* \max[-D(\vartheta_i^1, \vartheta_i^2) + d_0, 0],
\end{aligned} \tag{11}$$

is a general and simplified expression for the similarity constraint term in Criteria2 in which $\mathbf{1}(y_i^1 = y_i^2)$ is an indicator function that equals 1 only when $y_i^1 = y_i^2$; otherwise, it is 0. As shown in Fig. 5, the two MDNN models in M2DNN share the same coefficients. The architecture and similar coefficient-shared models can be optimized according to the theory of Siamese networks.[38] To train the M2DNN, samples are combined stochastically and then fed to the model in pairwise fashion (this training phase can be done offline and it does not impact the effectiveness of M2DNN for recognizing candidate image samples). The green checkmark "$\surd$" and the red "x" in Fig. 5 indicate whether the two images in this pair belong to the same category. The procedures for training the M2DNN model with the mini-batch technique are summarized in Algorithm 1. After training, a candidate image can be recognized with either the upper or the lower MDNN model (they share the same coefficients) in M2DNN.

Algorithm 1 The optimal gradient method for solving W in M2DNN.

Input: The sample pair set $\{y_i^1, y_i^2; (x_1^{1(i)}, x_2^{1(i)}, \cdots, x_P^{1(i)}), (x_1^{2(i)}, x_1^{2(i)}, \cdots, x_P^{2(i)}) | i = 1, 2, \cdots, N_p\}$ and the initialized coefficients $W = \{w_{(l,p)} | l = L_1, L_2, \cdots, L_n; p = 1,2, \cdots, P\}$.

Parameters: The weights λ, γ in the loss function, the margin d_0 in the similarity constraint, the momentum coefficient μ and the learning rate lr to update W, the maximum epoch number n_{iter} and the iterative termination error ε, and the batch size N_{batch} in mini-batch training.

Output: The renewed coefficients W.

Initialize w in the feature selection procedure (activation function: ReLu) with method "MSRA"[39] and other procedures (activation function: Tanh) with method "Xavier."[40] Here, $\mu = 0.98$ and $lr = 0.1$, and λ and γ are obtained by cross-validation. The number of iterations is $t_{\text{epoch}} = 0$ and $t_{\text{batch}} = 0$.

while $(t_{\text{epoch}} < n_{\text{iter}}$ & $\Phi > \varepsilon)$ **do**

  $t_{\text{batch}} = 0$;

  for $[t_{\text{batch}} < \text{floor}(N_p / N_{\text{batch}})]$ **do**

    for all sample pairs in this batch, **do**

      if $(y_i^1 = y_i^2)$ **then**

        $1(\cdot) = 1$,

      else

        $1(\cdot) = 0$;

      end if

      Calculate the similarity constraint loss $M_i^{1,2}$ with $1(\cdot)$ [Eq. (11)];

      Calculate the reconstruction losses L^1, L^2 and Ω [Eq. (10)];

    end for

    Accumulate the total losses in this batch as $\Phi(W)$ [Eq. (9)];

    $\forall\, (l, p)$, $\Delta w_{(l,p)} \leftarrow \mu \cdot \Delta w_{(l,p)} + lr \cdot \frac{\partial \Phi(W)}{\partial w_{(l,p)}}$;

    $w_{(l,p)} \leftarrow w_{(l,p)} - \Delta w_{(l,p)}$;

    $t_{\text{batch}} \leftarrow t_{\text{batch}} + 1$;

  end for

  $t_{\text{epoch}} \leftarrow t_{\text{epoch}} + 1$;

end while

Save the trained coefficients W.

With the aid of the auxiliary task, the features learned by the M2DNN model are more discriminative than those in MDNN. The auxiliary term $\sum_{i=1}^{N_p} M_i^{1,2}$ can be divided into two parts as follows:

$$\sum_{i=1}^{N_p} M_i^{1,2} = \sum_{\substack{i=1 \\ j=i}}^{N_C} \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} \max[D(\vartheta_{i,p}, \vartheta_{j,q}) - d_0, 0] + \sum_{i=1}^{N_C} \sum_{\substack{j=1 \\ j\neq i}}^{N_C} \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} \max[-D(\vartheta_{i,p}, \vartheta_{j,q}) + d_0, 0], \tag{12}$$

where N_C refers to the number of the categories and n_i and n_j are the numbers of samples in the i'th and j'th categories, respectively. The first part of Eq. (12) refers to the similarity constraint within each class, while the second part refers to the similarity constraint between classes. Using this iterative optimization, the features of sample pairs whose metric distances are larger than the margin d_0 in the same class will be pushed closer to one another; therefore, the features of samples within the same class will be more compact. At the same time, the similarity constraint makes features of sample pairs whose metric distances are less than d_0 in different classes move further away from one another; therefore, the interclass distance will be larger. This tendency is analogous to a type of mechanical spring system[41] in which the constraints within or between classes can be thought of as masses attracting or repelling each other with springs. As the spring system oscillates, the model ultimately achieves a state of equilibrium. That is, by iteratively optimizing the similarity constraint term in Eq. (11), the intraclass scatter will decrease while the interclass distance will increase. This is exactly the essence of Fisher's discriminant analysis, which has been proven to be effective for discriminative classification on many occasions.[42,43] Thus, features learned with the M2DNN model are more discriminative than those of the MDNN model, as experimentally verified and discussed in Sec. 4.

4 Experimental Results and Discussion

In this section, we will first introduce the datasets, features, and evaluation criteria used in the subsequent experiments. Then, we evaluate the effectiveness of the proposed M2DNN model for image classification by comparing it with some state-of-the-art algorithms. Finally, we conduct a quantitative comparison of the robustness and discrimination power between the MDNN and M2DNN models. The results of these experiments demonstrate the good performance of the proposed M2DNN model for multimodal image classification.

4.1 *Datasets, Features, and Evaluation Criteria*

Five representative datasets were tested in this study: NUS-WIDE, Scene-15, Texture-25, Indoor-67, and Caltech-101 (Please refer to Ref. 44 for links to these datasets.). These datasets have an incremental number of categories, and, as shown in Fig. 6, they cover many aspects such as diverse animal categories, natural scene categories, texture-different wood categories, indoor scene categories, and various other

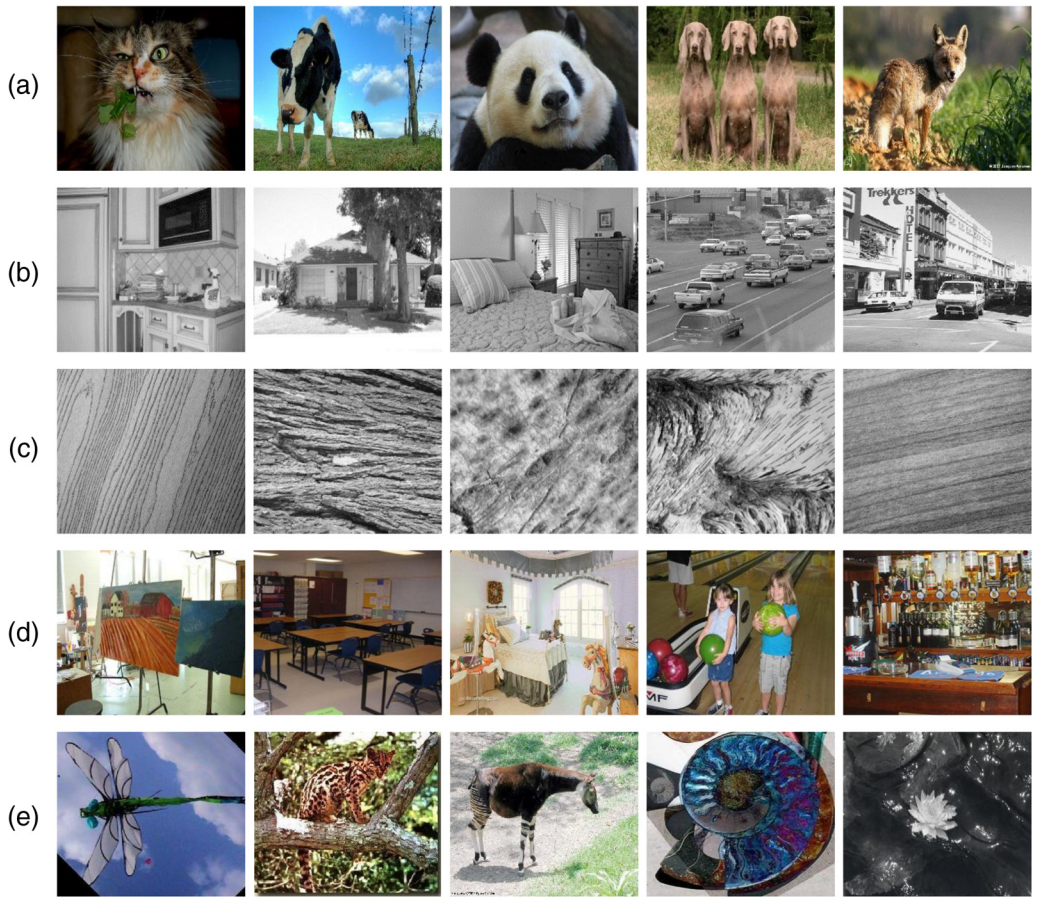

Fig. 6 The datasets used in this paper: (a) NUS-WIDE (12 categories), (b) Scene-15 (15 categories), (c) Texture-25 (25 categories), (d) Indoor-67 (67 categories), and (e) Caltech-101 (101 categories).

object categories. Varying the categories and objects is beneficial for evaluating the stability and robustness of the algorithms. The details are illustrated in Fig. 6.

The five datasets are publicly available image sets for classification and recognition tasks. (1) The NUS-WIDE dataset is a real-world web image database[45] that includes a total of 269,648 images with 5018 tags. In our experiments, we selected 12 animal image categories from this dataset: bird, fox, cow, horse, dog, elk, fish, bear, cat, tiger, zebra, and whale. In total, these categories include 16,519 images. The selected subset includes six off-the-shelf features for multimodal image retrieval: the color histogram (CH), the color correlogram (CC), the edge direction histogram (EDH), the wavelet texture (WT), the block-wise color moments (CM), and the SIFT descriptions. The dimensions of these features are shown in Table 1. (2) The Scene-15 dataset is composed of 15 scene categories. Containing 4485 images, it is one of the most complete scene category datasets used in studies thus far.[46] There are ~200 to 400 images in each category. (3) The Texture-25 database includes 25 texture classes, with 40 samples of each class. All the images have the same resolution: 640×480 pixels.[47] This dataset is a typical test bench for texture image recognition. (4) The Indoor-67 dataset contains 15,620 images in 67 indoor categories. Each category has a variable number of images and represents a challenging problem for indoor scene recognition.[48] (5) The Caltech-101 dataset contains 101 object categories with 40 to 800 images in each category.[49] It is a challenging database for object recognition because the objects in it vary widely. For the NUS-WIDE dataset, to evaluate the performance of the compared algorithms based on the number of labeled samples, we randomly chose subsets with 16, 64 and 128 samples in each category to form the training sets. For the Scene-15, Texture-25, Indoor-67, and Caltech-101 datasets, the number of images in each category varies widely; therefore, we used 80% of the images in each category for training and the rest for testing. None of the four datasets contain off-the-shelf features. Therefore, we extracted several effective features. As shown in Table 1, these features included the histogram of oriented gradient (HOG), the local binary pattern (LBP), the SIFT, and the GIST ("GIST" is an abstract representation of the scene[50]) descriptions. We also utilized a deep-learning-based descriptor (transferred from the ResNet50[51] model with the top softmax-classification layer removed) to make the algorithm performance comparisons more equitable. The dimensions for these features are listed in Table 1.

Table 1 Different features and their dimensions for the five datasets.

NUS-WIDE						
Feature type	CH	CC	EDH	WT	CM	SIFT
Dimension	64	144	73	128	255	500
Scene-15, Texture-15, Indoor-67, and Caltech-101						
Feature type	HOG	LBP	SIFT	GIST	ResNet50	
Dimension	576	928	1000	512	2048	

For the evaluation criteria, we adopted four different indicators to achieve an accurate and fair performance evaluation, including the general classification accuracy (Accu), the mean average precision (mAp),[24] the mean agreement for class labels (mPre),[23] and the average F-score indicator (F-score).[23] The Accu indicator is the most used criterion for evaluating the ability of a classifier to correctly distinguish candidate samples. The mAp is an effective criterion for image classification and retrieval. It refers to the mean ranking performance for each class and can be used to supplement Accu. The mPre is usually calculated from sums of per-class decisions to indicate label agreement. The F-score is another typical criterion often used in information retrieval and image classification.[52] It is the comprehensive result of the precision and recall indicators. These two indicators can be applied to reflect the repeatability and stability of algorithms.

In these experiments, several of the compared algorithms require parameters that must be determined by cross-validation. Therefore, we randomly selected 20% of the training images to form the validation set when needed. The parameters with the best performance on the validation set were selected to serve as the optimal parameter values in the subsequent experiments.

4.2 *Classification Performance Evaluation*

This section includes a discussion of the classification performance of several high-performing multimodal feature representation methods, including the following:

Best single feature (BSF)[13] uses the best single-modal features as the feature representations of images and then classifies them with the 1-nearest neighbor/recursive least squares (1-NN/RLS) classifier.

Concatenate (CAT)[13] concatenates the features of all modalities into a single vector and then applies the 1-NN/RLS classifier. BSF and CAT are used as the baselines for the other algorithms.

General canonical correlation analysis (GCCA)[53] is a multimodal generalized canonical correlation analysis algorithm that simultaneously transforms multiple data into one joint space. It optimizes the transformation matrix by comprehensively considering all the data from different modalities and categories.

MKL[54] learns general kernel combinations using gradient decent optimization and applies a support vector machine (SVM) solver to these nonlinear-combined kernels.

RFS[55] utilizes the joint $l_{2,1}$-norm minimization on both the reconstruction loss function and weight regularization. This $l_{2,1}$-norm is applied to select features across all modalities with joint sparsity. The trade-off parameter γ is set within the range $\{10^i | i = -4, -3, \cdots, 1\}$.

LM3FE[13] is a multimodal multitask feature extraction model that simultaneously optimizes the feature selection matrix and the combination coefficients. It exploits both the task relationships and the complementary nature of different modalities for feature transformation. The parameter γ_B is set within the range $\{10^i | i = -5, -4, \cdots, 1\}$, and γ_A, γ_C are tuned on the grid $\{10^i | i = -5, -4, \cdots, 5\}$.

MDL[19] is a multimodal multitask feature representation model. By jointly modeling the distribution of data from different modalities, the shared features in the representation layer are able to capture correlations across the modalities.

This model is extended from two modalities to multiple modalities and pretrained using sparse restricted Boltzmann machines[19] in this paper. The nonlinear correlations between the modalities in this model are modeled with a sigmoid transfer function.

3mDNN[20] is a deep autoencoder-neuron-network-based multimodal feature representation algorithm. Features from different modalities are encoded into the representation layer. The famous maximum margin strategy is used to constrain the learned features to be more discriminative. After encoding, samples are classified with a one-versus-one SVM classifier. The hyper parameters η and λ are set within the range $\{10^i | i = -5, -4, \cdots, 1\}$ using cross-validation.

MDNN is a deep-learning-based multimodal feature representation model. For the experiments described in this paper, we use the following model structure to conduct the image classification task with multiple modalities: 1 feature selection layer, 2 feature transformation layers, 1 pooling merge layer, 1 fully connected layer, and 1 softmax-classification layer. The dimension of the feature selection layer in each modality varies according to the feature selection ratio. The transformation layers and the merge layer share the same dimensions—unified to 512 in later experiments. A dimension of 256 was used for the fully connected layer. The softmax-classification layer is the last layer and has the same output dimension as the number of categories. The trade-off parameter λ in Eq. (2) is optimized within the range $\{10^i | i = -5, -4, \cdots, 1\}$.

M2DNN: the proposed M2DNN is a discriminative multimodal multitask feature representation model. The auxiliary task in this model is conducted by calculating the similarity term in Eq. (11), while the two reconstruction tasks are conducted by iteratively optimizing the reconstruction loss. These two tasks have the same structures and dimensions as those in the MDNN model. The trade-off parameter λ and γ in Eq. (10) are tuned on the grid $\{10^i | i = -5, -4, \cdots, 1\}$.

BSF and CAT use raw features with no selection or transformation strategies and perform as the baselines for comparison. For the other methods, the features are selected according to the dimension of each modality and the feature selection ratio, which varies within $\{0.1, 0.2, \cdots, 1.0\}$. Figure 7 shows the classification performance for these methods as the feature selection ratio varies from 0.1 to 1.0 and the number of labeled samples varies within {16, 64, 128}. The Accu, mAp, F-score, and mPre measures are calculated to evaluate the performances.

From the overall trends shown in Fig. 7, we conclude that (1) the classification performance of all the algorithms is improved as the number of labeled training samples increases (from the left column, to the center column to the right column of Fig. 7). This trend manifests similarly on all the Accu, mAp, F-score, and mPre indicators. For example, the Accu performance of all algorithms lies mostly within the interval [0.15, 0.25] when only 16 samples are used in each category, while with 128 samples in each category, the performance of this indicator improves, ranging within [0.25, 0.35]. The F-score performance also improves, increasing from [0.1, 0.2] to [0.25, 0.35]. (2) The deep-learning-based algorithms (MDL, 3mDNN, MDNN, and M2DNN) generally exhibit better performance than do the traditional numerical-optimization-based methods (GCCA, MKL, RFS, and LM3FS). The center column of Fig. 7 shows that the deep-learning-based algorithms—especially the 3mDNN, MDNN, and M2DNN algorithms—have obvious advantages. These models essentially have a stronger ability to fit the data than do the traditional methods because of the high nonlinearity of the models themselves. Moreover, the discriminative strategies used in the 3mDNN and M2DNN algorithms also contribute to the distinguishability of learned features, resulting in further classification performance improvements. (3) From an overall point of view, the results show that reserving more features does not necessarily result in a better classification performance. As shown in Fig. 7, when the feature selection radio ranges from 0.1 to 1.0, the classification performance first tends to rise, but subsequently there is a definite downward trend.

Regarding the performance of any single algorithm, we can draw the following conclusions from Fig. 7. (1) The CAT performs better than BSF on all four indicators. This occurs mainly because the CAT utilizes information from multiple modalities; therefore, it tends to generate an intact feature representation. In contrast, the BSF uses only features from a single modality, which may result in the loss of some important information. (2) The performance of the GCCA is largely affected by the correlations and distributions of data from different modalities. It exhibits relatively poor stability on the NUS-WIDE dataset because the complementary information from multiple modalities cannot be mined to the maximum extent. The RFS is a solid method with impressive stability and repeatability, as shown in Fig. 7 (F-score and mPre). However, because it is limited by a low degree of nonlinear transformation ability, this method exhibits relatively weak performance compared with LM3FE. MKL is a classical intermediate-combination multimodal feature representation method, and the performance of this method is related to the number of kernels. With more kernels, its ability to reconstruct data with different distributions is stronger than with fewer kernels; however, the computational complexity also becomes large. MKL shows a typical performance on this dataset. (3) MDL, 3mDNN, MDNN, and M2DNN are all deep-learning-based methods with relatively good performance. MDL generates multimodal feature representations using an autoencoder-based model; however, it has no feature selection process, which results in a relatively weaker classification performance (Accu and mAp) compared with MDNN. However, when this model is initialized with well pretrained models,[19] it exhibits better stability and repeatability (F-score and mPre) than MDNN. (4) 3mDNN is also an autoencoder-based model in which the feature representation from multiple modalities is constrained by the famous large margin strategy. Therefore, the generated features tend to be more robust and discriminative. As shown in Fig. 7, classification performance of 3mDNN is approximately (Accu and mAp), the same as that of MDNN; however, it is more robust and stable (see the F-score and mPre measures). (5) Both the 3mDNN and M2DNN algorithms are constrained with discriminative terms, and the features generated by these two models tend to be more robust. Nevertheless, 3mDNN does not include a feature selection process. Further, its transformation and classification processes are optimized separately. In contrast, M2DNN is globally optimized; consequently, 3mDNN does not peform as as well as M2DNN regardless of the recognition property or stability, as shown

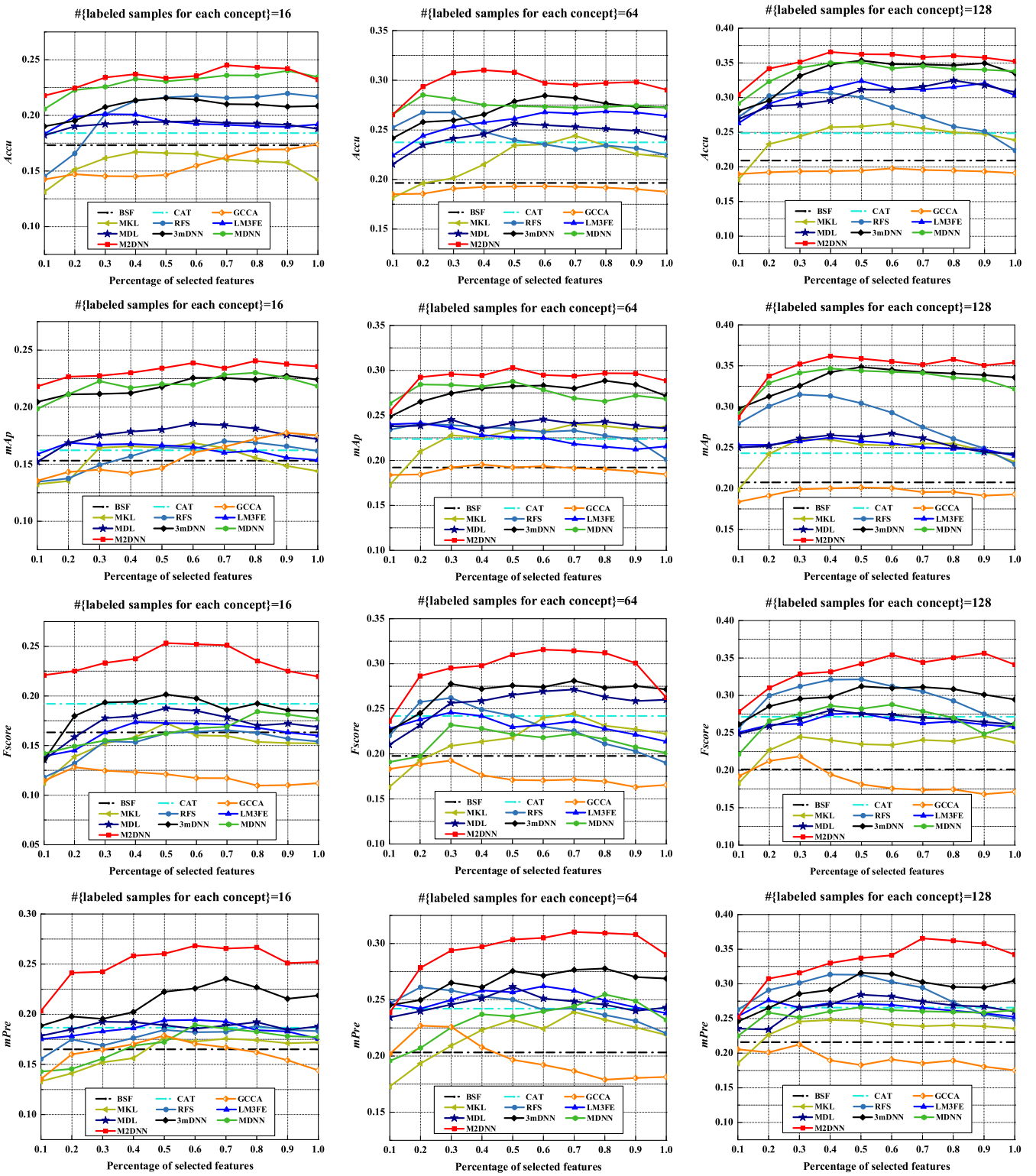

Fig. 7 Classification performance comparison w.r.t. the percentage of the features selected on the NUS-WIDE subset. The Accu, mAp, *F*-score, and mPre performances shown match the number of labeled-training samples (16, 64, and 128) in each column. The BSF and CAT methods serve as the baselines without any feature selection.

in Fig. 7. (6) LM3FE is a state-of-the-art method for extracting multimodal feature representations from the NUS-WIDE dataset. It also includes the same feature selection, transformation, and classification processes as M2DNN. However, these processes are optimized separately; therefore, the features it generates are also less discriminative than those of M2DNN, as shown in Fig. 7 (see the Accu and mAp measures). (7) M2DNN is based on MDNN and inherits that algorithm's good classification performance. However, in addition, M2DNN has been expanded to the

Table 2 Classification performance comparisons on the five datasets. The NUS-WIDE dataset was tested with 128 labeled-training samples in each category. The MDNN and M2DNN models were evaluated with a 0.5 feature selection ratio in the feature selection layer.

		Dataset									
		Baselines		Numerical-optimization-based methods				Deep-learning-based methods			
Method		BSF	CAT	GCCA	MKL	RFS	LM3FE	MDL	3mDNN	MDNN	M2DNN
NUS-WIDE	Accu	0.212 ± 0.022	0.246 ± 0.019	0.192 ± 0.019	0.258 ± 0.017	0.305 ± 0.027	0.319 ± 0.024	0.322 ± 0.035	0.351 ± 0.031	0.357 ± 0.027	**0.368 ± 0.035**
	mAp	0.203 ± 0.023	0.243 ± 0.014	0.208 ± 0.017	0.255 ± 0.018	0.307 ± 0.031	0.252 ± 0.034	0.267 ± 0.037	0.344 ± 0.032	0.348 ± 0.027	**0.361 ± 0.046**
	F-score	0.204 ± 0.021	0.273 ± 0.023	0.187 ± 0.027	0.233 ± 0.025	0.317 ± 0.025	0.282 ± 0.041	0.277 ± 0.026	0.312 ± 0.027	0.285 ± 0.021	**0.341 ± 0.027**
	mPre	0.213 ± 0.017	0.269 ± 0.016	0.186 ± 0.015	0.241 ± 0.013	0.309 ± 0.014	0.268 ± 0.039	0.286 ± 0.022	0.321 ± 0.022	0.261 ± 0.025	**0.332 ± 0.025**
Scene-15	Accu	0.631 ± 0.025	0.807 ± 0.025	0.502 ± 0.023	0.557 ± 0.021	0.657 ± 0.016	0.751 ± 0.031	0.779 ± 0.026	0.861 ± 0.037	0.849 ± 0.029	**0.869 ± 0.047**
	mAp	0.635 ± 0.015	0.812 ± 0.022	0.577 ± 0.013	0.556 ± 0.016	0.633 ± 0.021	0.764 ± 0.025	0.821 ± 0.031	0.853 ± 0.021	**0.885 ± 0.028**	0.878 ± 0.029
	F-score	0.627 ± 0.027	0.791 ± 0.025	0.611 ± 0.027	0.533 ± 0.019	0.664 ± 0.017	0.766 ± 0.031	0.826 ± 0.027	0.841 ± 0.032	0.851 ± 0.029	**0.865 ± 0.031**
	mPre	0.644 ± 0.015	0.812 ± 0.019	0.642 ± 0.016	0.557 ± 0.013	0.661 ± 0.017	0.761 ± 0.025	0.826 ± 0.033	0.844 ± 0.031	0.850 ± 0.025	**0.861 ± 0.034**
Texture-25	Accu	0.737 ± 0.022	0.756 ± 0.017	0.582 ± 0.013	0.637 ± 0.017	0.622 ± 0.013	0.781 ± 0.035	0.791 ± 0.029	0.817 ± 0.033	0.797 ± 0.022	**0.834 ± 0.033**
	mAp	0.747 ± 0.023	0.781 ± 0.021	0.561 ± 0.024	0.622 ± 0.017	0.628 ± 0.015	0.792 ± 0.027	0.783 ± 0.027	0.822 ± 0.031	0.807 ± 0.022	**0.838 ± 0.032**
	F-score	0.725 ± 0.027	0.783 ± 0.024	0.554 ± 0.015	0.644 ± 0.012	0.611 ± 0.022	0.787 ± 0.020	0.788 ± 0.035	0.781 ± 0.023	0.772 ± 0.026	**0.821 ± 0.027**
	mPre	0.754 ± 0.015	0.788 ± 0.017	0.686 ± 0.017	0.751 ± 0.017	0.655 ± 0.019	0.792 ± 0.027	0.787 ± 0.026	0.810 ± 0.027	0.813 ± 0.024	**0.824 ± 0.026**
Indoor-67	Accu	0.244 ± 0.015	0.332 ± 0.015	0.344 ± 0.016	0.257 ± 0.021	0.251 ± 0.028	0.379 ± 0.022	0.391 ± 0.021	0.444 ± 0.025	0.437 ± 0.021	**0.451 ± 0.033**
	mAp	0.223 ± 0.016	0.351 ± 0.015	0.357 ± 0.017	0.245 ± 0.016	0.230 ± 0.022	0.387 ± 0.025	0.391 ± 0.022	0.426 ± 0.025	0.437 ± 0.026	**0.462 ± 0.029**
	F-score	0.211 ± 0.016	0.292 ± 0.016	0.308 ± 0.016	0.233 ± 0.016	0.224 ± 0.017	0.342 ± 0.020	0.355 ± 0.016	0.377 ± 0.017	0.297 ± 0.022	**0.397 ± 0.028**
	mPre	0.225 ± 0.015	0.366 ± 0.021	0.328 ± 0.018	0.226 ± 0.017	0.213 ± 0.022	0.336 ± 0.026	0.347 ± 0.025	0.355 ± 0.017	0.311 ± 0.017	**0.382 ± 0.024**
Caltech-101	Accu	0.587 ± 0.019	0.682 ± 0.016	0.311 ± 0.016	0.331 ± 0.014	0.511 ± 0.012	0.677 ± 0.017	0.693 ± 0.021	0.722 ± 0.022	0.711 ± 0.016	**0.744 ± 0.025**
	mAp	0.585 ± 0.016	0.673 ± 0.013	0.274 ± 0.017	0.343 ± 0.021	0.511 ± 0.019	0.691 ± 0.022	0.701 ± 0.017	0.739 ± 0.020	0.730 ± 0.016	**0.766 ± 0.028**
	F-score	0.424 ± 0.017	0.522 ± 0.027	0.222 ± 0.026	0.273 ± 0.019	0.453 ± 0.021	0.651 ± 0.025	0.623 ± 0.027	0.655 ± 0.024	0.557 ± 0.017	**0.688 ± 0.025**
	mPre	0.469 ± 0.019	0.557 ± 0.017	0.389 ± 0.016	0.288 ± 0.015	0.466 ± 0.022	0.631 ± 0.027	0.627 ± 0.017	0.655 ± 0.025	0.574 ± 0.022	**0.681 ± 0.035**

Note: Values shown in bold are better.

multitask field in which an auxiliary task (a similarity constraint term) is able to constrain the learned features, making them more discriminative. M2DNN shows the best classification performance and stability compared with the other methods, as shown in Fig. 7.

We also evaluated these algorithms on the Scene-15, Texture-25, Indoor-67, and Caltech-101 datasets with four indicators, as shown in Table 2. Conclusions similar to those made previously can be made based on the algorithms' performances on these datasets. (1) The deep-learning-based methods tend to have better performances than do the traditional numerical-optimization-based methods. (2) The methods with discriminative constraints (i.e., M2DNN and 3mDNN) tend to perform better than the general learning-based methods (i.e., MDL and MDNN). (3) The methods with global optimization strategies (i.e., M2DNN and MDNN) tend to have better performances than do the methods that are optimized separately (i.e., RFS, LM3FE, MDL, and 3mDNN). (4) The M2DNN model outperforms the other methods even on the Indoor-67 dataset, which is a challenging problem in high-level vision[48] recognition. In Sec. 4.3, we conduct a quantitative evaluation of the robustness and discriminative capability of the M2DNN model.

4.3 *Discrimination Performance Comparison*

Fisher's linear discriminative criterion shows that for a proper metric function, features with larger interclass metric distances and smaller intraclass metric distances tend to be more discriminative and robust. Here, by using the Euclidean distance function, we evaluate the discriminative property of the features generated by adopting the MDNN and M2DNN models as follows:

$$S_{\text{Fisher}} = \frac{\text{Trace}(S_B)}{\text{Trace}(S_W)}, \tag{13}$$

where S_B and S_W refer to the inter- and intraclass metric distances, respectively. Features with a larger S_{Fisher} indicator tend to be more discriminative and are less likely to be confused even when some noise exists and, thus, are more robust.

We randomly selected 1500 samples from the NUS-WIDE test set and visualized their features as shown in Fig. 8. Figures 8(a)–8(c) refer to the performances with different numbers of training samples. The left column in this figure shows the features generated by the MDNN model, the right column shows the features generated by the

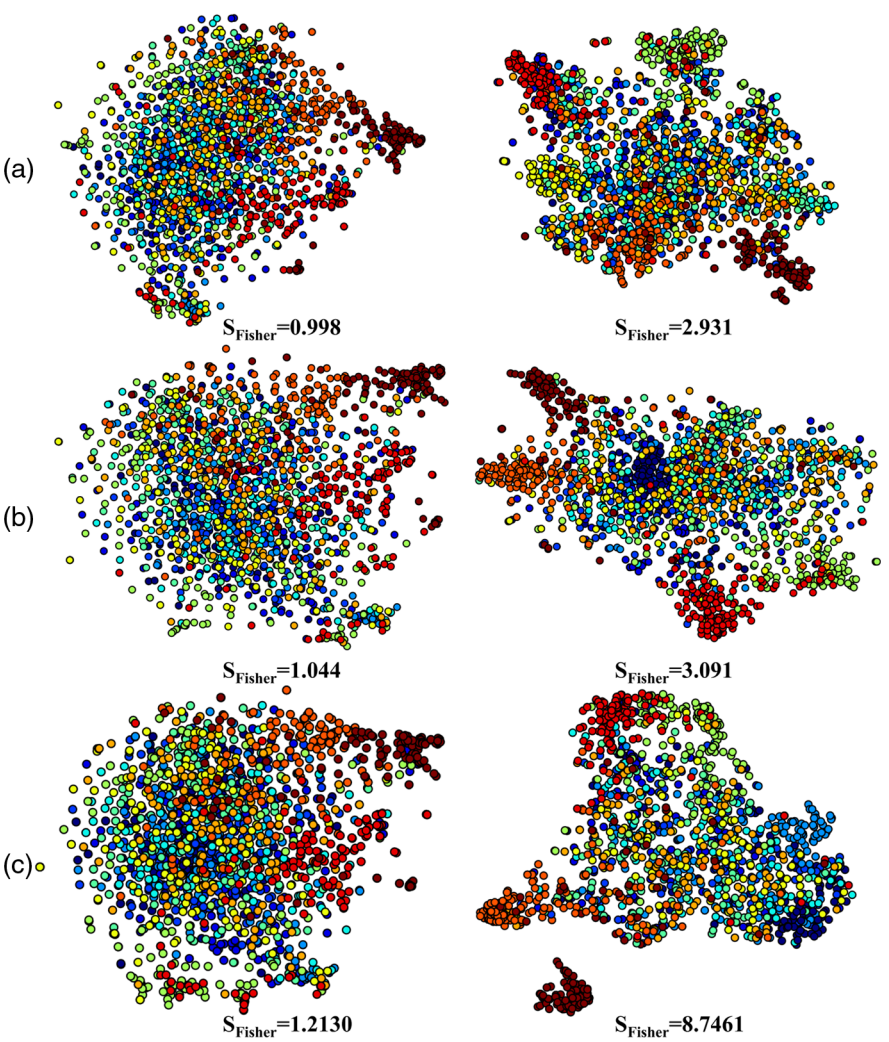

Fig. 8 Visualization of features generated by the MDNN model (the left column) and the M2DNN model (the right column). In (a), (b), and (c), the models were trained with 16, 64, and 128 samples from each category, respectively.

Table 3 Average Fisher score comparisons on the four datasets.

Dataset	Method	
	MDNN	M2DNN
Scene-15	1.158	**3.458**
Texture-25	1.501	**4.622**
Indoor-67	0.245	**0.741**
Caltech-101	1.123	**3.422**

M2DNN model. Each point in this figure refers to the generated feature of one image sample after feature selection, transformation, and dimension reduction (Please refer to Ref. 56 for the high-dimensional data visualization algorithm.). As the figure shows, the features generated by the M2DNN model are more discriminative and robust than those generated by the MDNN model. The indicator S_{Fisher} in Fig. 5 also proves this conclusion. Further, we evaluated the performance of this indicator on the other four datasets, as shown in Table 3. Similar conclusions can be made from these results.

Therefore, in summary, the multitask technique with a similarity constraint used in the M2DNN algorithm contributes to its classification robustness and discriminative ability when compared with the MDNN algorithm. The M2DNN model exhibits superior performance not only for classification accuracy but also for robustness and repeatability.

5 Conclusions

In this paper, we proposed the M2DNN model in the multimodal feature representation field for image classification. The M2DNN model is an end-to-end framework based on MDNN that is able to jointly optimize the feature selection, transformation, and classification procedures. This capability causes the learned feature representations to be more effective and efficient. Furthermore, the auxiliary task in M2DNN makes the generated features more discriminative and robust, contributing to its good classification performance. Experiments on five datasets illustrate the image classification ability of M2DNN, and the quantitative results using the S_{Fisher} evaluation indicator further demonstrate its capabilities. In the future, we plan to apply this model to additional application scenarios and further improve the classification performance.

Appendix: The Fisher Discriminant Analysis Theory

Fisher's linear discriminant theory[42] is briefly described as follows. Let $\chi_1 = \{x_1^1, \cdots, x_{l_1}^1\}$ and $\chi_2 = \{x_1^2, \cdots, x_{l_2}^2\}$ be samples from two classes. The discriminant is obtained by maximizing the Fisher criterion with the projection matrix w

$$w^* = \arg\max_w \frac{w^T S_B w}{w^T S_W w}, \tag{14}$$

where

$$S_B := (m_1 - m_2)(m_1 - m_2)^T \quad \text{and}$$
$$S_W := \sum_{i=1,2} \sum_{x \in \chi_i} (x - m_i)(x - m_i)^T, \tag{15}$$

are the inter- and intraclass scatter matrices, respectively, and m_i is defined by $m_i := \frac{1}{l_i} \sum_{j=1}^{l_i} x_j^i$. The goal of the Fisher criterion is to find the projection direction that maximizes the projected class while minimizing the class variance in that direction.

Acknowledgments

This work was supported by the National Science Foundation of China under Grant Nos. 51327801 and 51475193 and the Major Project Foundation of Hubei Province under Grant No. 2016AAA009. The authors fully appreciate the financial support.

References

1. P. Yan et al., "Local feature descriptor invariant to monotonic illumination changes," *J. Electron. Imaging* **25**(1), 013023 (2016).
2. C. Anant et al., "Comparative study of global invariant descriptors for object recognition," *J. Electron. Imaging* **17**(2), 023015 (2008).
3. D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. of the Seventh IEEE Int. Conf. on Computer Vision*, Vol. 2, pp. 1150–1157 (1999).
4. J. Wan et al., "3D SMoSIFT: three-dimensional sparse motion scale invariant feature transform for activity recognition from RGB-D videos," *J. Electron. Imaging* **23**(2), 023017 (2014).
5. X. Tian et al., "Feature integration of EODH and color-SIFT: application to image retrieval based on codebook," *Signal Process. Image Commun.* **29**(4), 530–545 (2014).
6. S. Battiato et al., "SIFT features tracking for video stabilization," in *14th Int. Conf. on Image Analysis and Processing*, pp. 825–830, IEEE (2007).
7. E. N. Mortensen, H. Deng, and L. Shapiro, "A SIFT descriptor with global context," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Vol. 1, pp. 184–190 (2005).
8. C. Ma et al., "Hierarchical convolutional features for visual tracking," in *IEEE Int. Conf. on Computer Vision*, pp. 3074–3082 (2015).
9. G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2518–2525 (2012).
10. S. W. Teng, M. T. Hossain, and G. Lu, "Multimodal image registration technique based on improved local feature descriptors," *J. Electron. Imaging* **24**(1), 013013 (2015).
11. F. He et al., "Score level fusion scheme based on adaptive local Gabor features for face-iris-fingerprint multimodal biometric," *J. Electron. Imaging* **23**(3), 033019 (2014).
12. C. Xu, D. Tao, and C. Xu, "Multi-view intact space learning," *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(12), 2531–2544 (2015).
13. Y. Luo, Y. Wen, and D. Tao, "Large margin multi-modal multi-task feature extraction for image classification," *IEEE Trans. Image Process.* **25**(1), 414–427 (2016).
14. T. Zhou, J. Yang, and A. Loza, "Crowd modeling framework using fast head detection and shape-aware matching," *J. Electron. Imaging* **24**(2), 023019 (2015).
15. S. Song, V. Chandrasekhar, and B. Mandal, "Multimodal multi-stream deep learning for egocentric activity recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 24–31 (2016).
16. R. Kavi et al., "Multiview fusion for activity recognition using deep neural networks," *J. Electron. Imaging* **25**(4), 043010 (2016).
17. S. E. Kahou et al., "Emonets: multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, **10**(2), 99–111 (2016).
18. M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 902–909 (2010).
19. J. Ngiam et al., "Multimodal deep learning," in *28th Int. Conf. on Machine Learning*, pp. 689–696 (2011).
20. Z. Ren, Y. Deng, and Q. Dai, "Local visual feature fusion via maximum margin multimodal deep neural network," *Neurocomputing* **175**, 427–432 (2016).
21. K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," in *Advances in Neural Information Processing Systems*, pp. 2141–2149 (2014).

22. X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. of the 14th Int. Conf. on Artificial Intelligence and Statistics (AISTATS'11)*, Fort Lauderdale, Florida, Vol. 15, No. 106, p. 275 (2011).
23. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.* **45**(4), 427–437 (2009).
24. M. Zhu, *Recall, Precision and Average Precision*, Vol. **2**, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo (2004).
25. M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.* **12**, 2211–2268 (2011).
26. X. Zhang and M. H. Mahoor, "Task-dependent multi-task multiple kernel learning for facial action unit detection," *Pattern Recognit.* **51**, 187–196 (2016).
27. A. Rakotomamonjy et al., "Simplemkl," *J. Mach. Learn. Res.* **9**, 2491–2521 (2008).
28. N. A. Subrahmanya and Y. C. Shin, "Sparse multiple kernel learning for signal processing applications," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(5), 788–798 (2010).
29. C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Trans. Image Process.* **24**(3), 980–993 (2015).
30. B. Jin et al., "Robust visual multitask tracking via composite sparse model," *J. Electron. Imaging* **23**(6), 063022 (2014).
31. Y. Yan et al., "Multitask linear discriminant analysis for view invariant action recognition," *IEEE Trans. Image Process.* **23**(12), 5599–5611 (2014).
32. X. Yang, J. Cheng, and W. Feng, "Cauchy estimator discriminant analysis for face recognition," *Neurocomputing* **199**, 144–153 (2016).
33. Y. Guo et al., "Multiview Cauchy estimator feature embedding for depth and inertial sensor-based human action recognition," *IEEE Trans. Syst. Man Cybern.: Syst.* **PP**(99), 1–11 (2016).
34. I. Mizera and C. H. Müller, "Breakdown points of Cauchy regression-scale estimators," *Stat. Probab. Lett.* **57**(1), 79–89 (2002).
35. P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Trans. Comput.* **26**(9), 917–922 (1977).
36. D. Soudry and D. Castro, "Memristor-based multilayer neural networks with online gradient descent training," *IEEE Trans. Neural Networks Learn. Syst.* **26**(10), 2408–2421 (2015).
37. P. S. Leo Breiman, "Submodel selection and evaluation in regression. the x-random case," *Int. Stat. Rev.* **60**(3), 291–319 (1992).
38. S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 539–546 (2005).
39. K. He et al., "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *Int. Conf. on Computer Vision* (2015).
40. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Int. Conf. on Artificial Intelligence and Statistics* (2010).
41. R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1735–1742 (2006).
42. S. Mika et al., "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX, Proc. of the 1999 IEEE Signal Processing Society Workshop* (1999).
43. Z. Wang, Q. Ruan, and G. An, "Uncorrelated regularized local Fisher discriminant analysis for face recognition," *J. Electron. Imaging* **23**(4), 043017 (2014).
44. S. Mei, "Multi-modal image classification," 2016, https://shuang-mei.github.io/blog/2016/08/10/multi-modal.html.
45. T.-S. Chua et al., "NUS-wide: a real-world web image database from National University of Singapore," in *Proc. of ACM Conf. on Image and Video Retrieval*, Santorini, Greece (2009).
46. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 2169–2178 (2006).
47. C. S. Svetlana Lazebnik and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1265–1278 (2005).
48. A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'09)*, pp. 413–420, IEEE (2009).
49. F.-F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," *Comput. Vision Image Understanding* **106**(1), 59–70 (2007).
50. A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *Int. J. Comput. Vision* **42**(3), 145–175 (2001).
51. K. He et al., "Deep residual learning for image recognition,"in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).
52. S. M. Beitzel, "On understanding and classifying web queries," PhD Thesis, Illinois Institute of Technology (2006).
53. Rupy, "Generalized canonical correlation analysis," 2015, https://github.com/rupy/GCCA (10 August 2016).
54. M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *The 26th Annual Int. Conf. on Machine Learning*, pp. 1065–1072 (2009).
55. F. Nie et al., "Efficient and robust feature selection via joint 2, 1-norms minimization," in *Advances in Neural Information Processing Systems*, pp. 1813–1821 (2010).
56. L. Maaten and G. Hinton, "Visualizing data using t-SNE[J]," *J. Mach. Learn. Res.* **9**(Nov), 2579–2605 (2008).

Shuang Mei is a PhD candidate at Huazhong University of Science and Technology (HUST), China. His research interests include pattern recognition and image recognition.

Hua Yang is an associate professor at HUST, China. His current research interests include high-speed vision and its applications.

Zhouping Yin is a professor at HUST, China. He was the recipient of the China National Funds for Distinguished Young Scientists in 2006. He leads a research group and conducts research into intelligent manufacturing and artificial intelligence.