Multi-level Image Feature Fusion Algorithm for Image-Text Matching Tasks (Postprint)

Authors: Hao Zhifeng, Li Junfeng, Ruichu Cai, Wen Wen, Wang Lijuan, Li Yiting

Date: 2019-01-28T00:00:00+00:00

Abstract

Existing mainstream approaches for extracting image features using pre-trained convolutional neural networks suffer from the following limitations: they represent images using only single-layer pre-trained features, and there exists a discrepancy between the pre-training task and the actual research task. This prevents existing text-image matching methods from fully exploiting image features and makes them highly vulnerable to interference from noisy features. To address these issues, we utilize multi-layer features from pre-trained networks and propose a multi-level image feature fusion algorithm. Under the guidance of the text-image matching learning objective, we employ a Multi-Layer Perceptron (MLP) to supervisedly fuse and reduce the dimensionality of multi-level pre-trained image features, generating fused image features that thereby enable full utilization of pre-trained features while mitigating noise interference. Experimental results demonstrate that the proposed fusion algorithm achieves more effective utilization of pre-trained image features and yields superior text-image matching performance compared to methods employing single-level features.

Full Text

Fusion of Multi-level Image Features for Image-Text Matching

Hao Zhifeng[1],[2], Li Junfeng[1]†, Cai Ruichu[1], Wen Wen[1], Wang Lijuan[1], Li Yiting[1] [1]College of Computer Science, Guangdong University of Technology, Guangzhou 510006, China [2]College of Mathematics & Big Data, Foshan University, Foshan, Guangdong 528000, China

Abstract: Existing mainstream methods that utilize pre-trained convolutional neural networks for image feature extraction suffer from two key limitations: (a) they represent images using only a single layer of pre-trained features, and

(b) there is inconsistency between the pre-training task and the actual research task. These limitations prevent existing image-text matching approaches from fully leveraging image features and make them highly susceptible to noise interference. To address these issues, this paper employs multi-layer features from pre-trained networks and proposes a multi-level image feature fusion algorithm. Guided by the learning objective of image-text matching, the algorithm uses a Multi-Layer Perceptron (MLP) to supervise the fusion and dimensionality reduction of multi-level pre-trained image features, generating fused image features that make full use of pre-trained features while reducing noise. Experimental results demonstrate that the proposed fusion algorithm enables more effective utilization of pre-trained image features and achieves better image-text matching performance compared to methods using single-level features.

Keywords: image-text matching; multi-level image features; pre-trained features; fusion features; recommendation system

0 Introduction

In recent years, the image-text matching task has gained increasing attention in artificial intelligence and machine learning. Traditionally, selecting appropriate images for text required manual searching through massive image collections based on textual content, consuming substantial human time and effort. Building upon prior achievements, we can now leverage machine learning techniques to construct an image-text matching system that recommends suitable images based on textual content, eliminating tedious and repetitive manual searches. As an image-text matching system must simultaneously consider text and images —two modalities belonging to different domains—this task is inherently multimodal.

To accomplish this task, three fundamental problems must generally be addressed: how to represent text, how to represent images, and how to jointly analyze text and image features to accurately measure their similarity. The first two representation problems are particularly crucial as they form the foundation for solving the third. For decades, solving representation problems required careful engineering design and considerable domain expertise to craft feature extractors that transform raw data (such as unprocessed text or image pixel values) into appropriate internal representations or feature vectors, making the modeling process overly complex and often lacking in representational power.

Drawing upon remarkable achievements in deep learning, we can utilize general artificial neural networks for representation learning, such as Multi-Layer Perceptrons (MLP) [1], Recurrent Neural Networks (RNN) [2], Convolutional Neural Networks (CNN) [3], and Long Short-Term Memory networks (LSTM) [4]. These networks comprise multiple simple, non-linear feature layers that transform features at one level into more abstract, higher-level features. With sufficient transformation combinations, the network can learn highly complex

functions. Most importantly, the feature layers in these neural networks are not designed by human engineers but are learned from data under the guidance of learning objectives. Consequently, deep learning methods simplify the modeling process and enhance representational capacity for research objects.

Generally, there are two approaches to feature extraction in deep learning: (a) training an artificial neural network in a supervised manner under the learning objective of the research task, then using that network to extract task-relevant features for the research object; or (b) training an artificial neural network on a pre-training task with high-quality datasets, then using features from a specific layer of that network as general features for the research object. For research tasks with insufficient dataset quality, the mainstream approach adopts the second method to achieve richer and more efficient representation. For example, to better extract image features, one can pre-train a CNN on the ImageNet dataset under the guidance of an image recognition task, then use the output from a specific feature layer (typically the fully connected layer before classification output) as image features for further processing.

The hierarchical structure of artificial neural networks naturally determines that high-level features are abstractions and summaries of low-level features. In other words, different feature layers in the network represent features at different levels, with deeper layers expressing more abstract and higher-level features. During learning, the network must, under the guidance of the task' s learning objective, supervise the induction of features useful for the task. However, deep learning-based image-text matching typically uses only a single layer of features from pre-trained networks as image features, or performs fine-tuning on that single layer. Consequently, it can only utilize the single-level features induced by the pre-training task or further induce from that single level. Unfortunately, there is a certain discrepancy between the pre-training task and the actual image-text matching task (task inconsistency). Directly using a single-level pre-trained feature may result in failure to induce features necessary for image-text matching, while also introducing numerous ineffective noise features. Moreover, fine-tuning single-level pre-trained features fails to leverage useful features from other levels. Therefore, directly using or fine-tuning a single layer of pre-trained features does not adequately or reasonably utilize these features; it is necessary to extract multi-level pre-trained features and perform further induction and fusion based on them.

Specifically addressing the above problems, this paper innovatively employs multi-layer features from pre-trained networks and proposes a multi-level image feature fusion algorithm (referred to as the fusion algorithm). This algorithm uses a Multi-Layer Perceptron to supervise the fusion and dimensionality reduction of multi-level pre-trained image features under the guidance of the image-text matching task' s learning objective, ultimately generating fused image features. The use of multi-level pre-trained image features enables fuller utilization of features from different levels, while the fusion and dimensionality reduction process induces features useful for image-text matching and elimi-

nates useless features, thereby reducing noise interference. This paper adopts MLP for fusion because multi-level pre-trained image features cannot be simply stacked and fused—they exhibit complex non-linear relationships. As MLP is a generalization of the perceptron that can effectively process features with such non-linear relationships, using MLP to supervise the processing of multi-level features is a concise and effective approach.

By introducing fused image features into our implemented image recommendation algorithm based on textual content, better recommendation performance can be achieved. Finally, experimental results on two datasets demonstrate that the proposed method can indeed utilize pre-trained image features more effectively and generate fused image features with stronger expressive power for image-text matching tasks.

The main contributions of this paper include: (a) employing multi-level image features extracted by pre-trained convolutional neural networks; (b) constructing an MLP that fuses and reduces the dimensionality of multi-level pre-trained image features under the guidance of the image-text matching task' s learning objective, and using it to generate fused image features for images; and (c) building an image recommendation algorithm based on collaborative filtering [5] that can recommend images according to text content, and incorporating the proposed fused image features into this algorithm.

1 Related Work

Image-text matching holds an important position in recommendation systems and machine learning. Yan et al. [6] proposed using deep networks to represent images and text, then employing joint hidden space learning with deep canonical correlation analysis to solve image-text matching problems. Ma et al. [7] constructed an image feature extraction network for image-text matching tasks and proposed initializing this feature extraction network with pre-trained CNNs. Wang et al. [8] built a general framework for deep learning-based joint hidden space learning for image and text, proposing that both images and text have their own structure-preserving constraints and that image-text matching has bidirectional ranking constraints. Nam et al. [9] proposed using attention mechanisms to solve multimodal tasks including image-text matching and visual question answering, achieving state-of-the-art results on standard datasets.

In the field of image feature engineering, CNN architectures for image processing [10–16] have become increasingly deep and modular. These networks continuously 刷新 (set new records) in image recognition tasks, now achieving high performance that even surpasses human recognition capabilities. Therefore, it is reasonable to believe that these excellent network models can extract large amounts of advanced image semantic features, and utilizing these pre-trained networks to extract image feature information is logical.

Based on these excellent image recognition networks, Zeiler et al. [17] attempted to visualize and understand CNNs and observed which structures in the input

image influence given feature maps. Subsequently, Garcia-Gasulla et al. [18] attempted to unsupervisedly extract visual expression features about abstract semantics. They used convolutional layer features from pre-trained CNNs to represent images, then evaluated the semantics of this feature space using correlation with WordNet distances, finding that vector distances in this space are strongly correlated with linguistic semantics. Through clustering experiments, they discovered that elements close in WordNet could be clustered together, clear gaps existed between categories like "canine" and "wheeled vehicle," and higher-level semantic categories like "living thing" and "non-living thing" could also be clearly distinguished. This evidence demonstrates that this representation method can successfully capture high-level visual semantic information. Agrawal et al. [19] analyzed whether pre-trained features are effective for actual object recognition research tasks. Experiments proved that in most cases, both pre-trained features and fine-tuned features perform better than features trained from scratch (except when the dataset has substantial supplementation, where retrained features may outperform pre-trained features, but fine-tuned features still perform best).

In applied research, numerous experiments demonstrate that utilizing pre-trained image features can achieve excellent results. For example, Vinyals et al. [20] used pre-trained CNNs to represent images and built a model that generates descriptive text for images. Peng et al. [21] used fine-tuned pre-trained CNNs to obtain multi-scale image features and proposed the concept of label inheritance. Liu et al. [22] used convolutional layers from pre-trained CNNs to obtain local image features for image recognition tasks. Since convolutional layers preserve spatial information, there is no need to repeatedly use the network to obtain local image features, thus eliminating the impact of scale inconsistency between training images and image patches. This work proved that with proper usage, not only pre-trained fully connected layer features are useful, but pre-trained convolutional layer features can also contain highly useful information.

Doersch et al. [23] utilized image self-information to self-supervisedly pre-train networks without using any labels beyond the actual research dataset. Experimental results from this work showed that features pre-trained using this method can also play a role in computer vision tasks and improve performance. Although the above works all use pre-trained networks to represent images and achieve good performance, they only use a single-level feature from the pre-trained network as image features. This leads to problems such as insufficient utilization of pre-trained features and susceptibility to noise feature interference. Therefore, researchers began attempting to use and fuse multi-level features from pre-trained networks. For example, Gatys et al. [24] used features from a single convolutional layer of a pre-trained network as content features for images in image style transfer tasks, and used multiple Gram Matrices generated from multi-layer convolutional features to jointly represent image style features. They then used backpropagation to modify noise images, ultimately producing images with specified style and content. This work utilized multi-level pre-trained fea-

tures through loss function superposition. Liu et al. [25] used the BoVW (bag of visual words) algorithm to convert pre-trained convolutional layer features into word histograms for image representation. Finally, they obtained a total deep pyramid matching kernel through weighted summation of histogram distance kernels from each layer, which was used to optimize the SVM (support vector machine) algorithm. Ronneberger et al. [26] constructed a U-shaped network to solve biomedical image segmentation tasks. To enable better localization, this network combined high-resolution features from the network' s contraction path with outputs from the upsampling path, allowing subsequent continuous convolutional layers to learn to combine this information into more precise outputs.

2.1 Multi-level Image Feature Fusion Algorithm

The fusion algorithm fuses and reduces the dimensionality of multi-level pre-trained image features and constitutes the core algorithm of this paper. The overall algorithm framework is shown in Figure 1 [Figure 1: see original paper].

Figure 1 Framework of fusion algorithm

Given a convolutional neural network trained on a pre-training task (such as ImageNet image recognition), we can selectively extract and use features from either convolutional or fully connected layers. Assuming the pre-trained network has n feature layers, after inputting image k into the network, we concatenate features from all layers to generate a total multi-level pre-trained feature:

where represents the i-th layer feature of image k in the pre-trained network.

To enable concatenation of features from different layers, when the i-th layer feature of the pre-trained network is a convolutional layer feature , which contains spatial information, we need to perform a pooling operation on that layer feature to eliminate spatial information:

When the i-th layer feature is a fully connected layer feature , which does not contain spatial information, no pooling operation is needed:

Note that we do not always need to use all feature layers from the pre-trained network; instead, we can selectively use some layers based on specific circumstances. Therefore, in Equation (1) may contain features from only a subset of layers.

To induce features useful for image-text matching tasks from multi-level pre-trained features and discard useless noise features, this paper constructs a Multi-Layer Perceptron (MLP) that fuses and reduces the dimensionality of image k' s features under the guidance of the image-text matching task's learning objective, ultimately outputting fused image features :

This MLP is a standard fully connected artificial neural network with the following design characteristics: (a) both hidden and output layers have non-linear activation functions to enhance network expressiveness; (b) the dimensions of network layers decrease with depth to fuse and reduce the dimensionality of

high-dimensional multi-level pre-trained features containing substantial noise; and (c) the dimension of the output fused image features must match that of text features to enable similarity measurement. Since the activation functions, layer dimensions, and number of layers in the MLP are related to the actual research object and dataset, this section does not define the network' s detailed structure more specifically but provides only a general definition. More details about the actual MLP used will be presented in Chapter 3.

To train the MLP' s network parameters, we define a constraint where and represent the positive (matching) and negative (non-matching) image sets corresponding to training text , respectively; is the feature vector of text (extracted through unsupervised methods such as Latent Semantic Analysis (LSA) topic model [27] and doc2vec [28]); and represent the fused image features output when images and are input to the fusion algorithm; represents the cosine similarity between and ; and m is the enforced margin size.

The constraint in Equation (5) indicates that for a given training text , the feature similarity with each of its positive images must be greater than the margin size m plus its similarity with each negative image .

By using the standard form of Hinge Loss, the constraint in Equation (5) is converted into the MLP' s training loss function:

The loss function in Equation (6) includes all triplets composed of training texts, their corresponding positive images, and negative images from the training set. However, using all triplets for training is impractical due to the enormous number of possible combinations. Therefore, in each iteration of MLP training, for each training text, we randomly select only one negative image to construct a triplet with its corresponding positive images for matching training.

In practice, the margin size m in Equation (6) can vary for different training examples. However, to facilitate optimization, we set a fixed margin size m for all training examples in the dataset, with its specific value to be provided in Chapter 3.

2.2 Image Recommendation Algorithm

This section implements an image recommendation algorithm based on collaborative filtering that can recommend images according to text content. This algorithm was used in the Sohu 2017 Image-Text Matching Competition (https://www.biendata.com/competition/luckydata/), where it achieved third place. Let represent the text set in the training set, represent the image set in the test set, and result represent the set of recommended images. The specific steps of the algorithm are:

a) Given a text in the test set, use a text topic model to find the text in text set with the most similar content (Equation (7)). Correspondingly, we can obtain the matching image of text in the training set.

b) In image set , use image feature information to find the image most similar to image as a recommendation candidate (Equation (8)), and add this to image set result.

This recommendation process is shown in Figure 2 [Figure 2: see original paper]. Repeat step b) until result contains K recommendation candidate images.

Figure 2 Recommendation flowchart

Obviously, using image features with different representational capabilities in this recommendation algorithm will produce different recommendation performance. This paper attempts to use various image features (including fused image features generated by the fusion algorithm in Section 2.1) as image feature information in the recommendation algorithm, and then directly evaluate the representational capabilities of various image features based on recommendation performance.

3 Experiments and Results Analysis

This chapter conducts comparative experiments on the Sohu Image-Text Matching Competition dataset and the Flickr30K dataset [29] to evaluate the performance of various image features. Experimental recommendation performance is reported using the commonly used evaluation metric in image-text matching tasks, recall@K% (K=1, 5, 10), which represents the proportion of matching images retrieved within the top K of recommendation results.

3.1 Sohu Image-Text Matching Competition Dataset

This dataset originates from the Image-Text Matching Competition held by Sohu Company in 2017. The data used in this experiment includes: an initial training set of 100,000-level samples, a semifinal training set of million-level samples, a final 400 small test set (a subset divided from the final 20,000 complete test set, containing 400 news articles and corresponding 400 matching images from the complete test set), and the final 20,000 complete test set. Each news article in this dataset has a corresponding matching image.

In this experiment, to better represent text, we trained a 500-topic Latent Semantic Analysis topic model [27] using all news texts from the million-level semifinal training set, and used this model to generate feature vectors for all news texts. The pre-trained network in the fusion algorithm is the Inception v3 network [14] pre-trained on the ImageNet image recognition task. The network structure outline is provided in Table 1, with more detailed structure available in reference [14]. This experiment uses two feature layers from the pre-trained network as multi-level pre-trained image features: the final Pool layer feature (referred to as fc feature) and the convolutional layer feature output from the first Inception module 3 (Figure 6 [Figure 6: see original paper]) after max pooling (referred to as mixed9 feature).

Table 1 Network structure of Inception v3

To determine the structural parameters of the MLP in the fusion algorithm, we conducted the following investigation. Cybenko [30] has proven that an MLP needs at most one hidden layer to approximate functions. Based on this conclusion, our MLP uses only one hidden layer. We further investigated the impact of non-linear activation functions on fusion performance, with results presented in Figure 3 [Figure 3: see original paper] (using recommendation performance to indirectly reflect fusion performance). We observed that adding non-linear activation functions to the hidden layer decreased the network's fusion performance, but adding L2 regularization constraints on network parameters improved the network' s fitting capability and generalization, achieving optimal fusion performance. Therefore, we finally adopted a network structure with non-linear activation functions in the hidden layer and used L2 regularization to optimize network parameter training.

Specifically, the MLP used in the fusion algorithm (referred to as) that takes fc features and mixed9 features as input has the following structure: an input layer with 4096 dimensions; a hidden layer with 2048 dimensions and sigmoid activation function; and an output layer with 500 dimensions and tanh activation function. We used the 100,000-level initial training set to supervise the training of the network parameters by minimizing Equation (6), with L2 regularization constraints on parameters (weight 0.0005) added during training to optimize performance. The margin m in Equation (6) loss function was set to 0.5, and parameters were updated using the Adam algorithm (learning_rate=0.001, beta1=0.9, beta2=0.999, epsilon=1e-08).

Figure 3 Influence of MLP network structure on fusion performance

This experiment also designed an MLP using only fc features as input (referred to as feature), with its specific structure and training details consistent with except that the input layer dimension was 2048. Since MLP training constraints enable its output image feature vectors to directly match text feature vectors via cosine similarity, Table 3 presents the recommendation performance of directly matching image features output by different MLPs with text features, to verify whether using multi-level features is more advantageous than using single-level features during the fusion and dimensionality reduction process.

Table 3 Comparison of similarity matching between image features and text features on SOHU dataset

Table 3 shows that fused image features outperform features in most recall@K metrics (except for recall@1 in the small test set where fused image features performed poorly). This result indicates that under the guidance of the image-text matching task' s learning objective, generates features with stronger expressive power than . In other words, using multi-level features is indeed more advantageous than using single-level features during the fusion and dimensionality reduction process. Additionally, compared with the image recommendation algorithm, the method of directly using fused image features for similarity matching demonstrates superior performance in recall@5 and recall@10, especially with a

14.0 improvement in recall@10 on the small test set and a 1.3 improvement on the complete test set compared to the image recommendation algorithm using only fc features.

To directly compare the recommendation performance of fused image features generated by (referred to as fused image features), fc single-level features, and fc+mixed9 multi-level features, this paper conducted comparative experiments in the image recommendation algorithm from Section 2.2 using these three image features. Results are presented in Table 2 (using fc single-level features in the image recommendation algorithm was our team's approach in the 2017 Sohu Image-Text Matching Competition, and the top two teams' image representation methods were similar). Based on experimental results from the final 400 small test set, fc+mixed9 features show slight improvements in recall@K compared to fc features, with a 0.5 improvement in recall@10. This proves that multi-level features yield better recommendation performance than single-level features, but due to their excessively high dimensionality containing numerous noise features for image-text matching, the performance improvement is not substantial. However, our method can supervise the fusion and dimensionality reduction of multi-level image features under the image-text matching task' s learning objective. Consequently, fused image features show substantial improvements in recall@K compared to fc features, with a remarkable 9.5 improvement in recall@10. Clearly, the proposed method is effective.

To further verify the method' s effectiveness on large test sets, we also conducted this experiment on the final 20,000 complete test set, with results still presented in Table 2 . As the search space for recommendations expands, recall@K performance on the complete test set is significantly worse than on the small test set. Nevertheless, most recall@K metrics for fc+mixed9 features still show slight improvements over fc features. fused image features demonstrate even greater improvements over fc features, with a 0.6 improvement in recall@10. Thus, the proposed method is also effective on large test sets.

Table 2 Comparison of image recommendation algorithms using different image features on SOHU dataset

3.2 Flickr30K Dataset

To verify the universality of the proposed fusion algorithm, we also conducted comparative experiments on the Flickr30K dataset. This dataset contains 31,783 images, each with 5 descriptive short sentences. Following the publicly available dataset split scheme [31], we divided the dataset into 29,783 training images, 1,000 validation images, and 1,000 test images.

In this experiment, we trained a 500-topic Latent Semantic Analysis topic model using all descriptive short sentences to generate text feature vectors. Consistent with Section 3.1, this experiment also used fc features and mixed9 features from the pre-trained Inception v3 network as multi-level pre-trained image features. The structure and training details in the fusion algorithm remained consistent

with the Sohu dataset experiments, as described in Section 3.1. Ultimately, we could supervise the training of network parameters using 29,783 training images and their corresponding descriptive short sentences.

Table 4 Comparison of image recommendation algorithms using different image features on Flickr30K dataset

Table 4 presents comparative experimental results of various image features in the image recommendation algorithm on this dataset. Unlike the results in Section 3.1, most recall@K metrics for fc+mixed9 features are worse than fc features, proving that noise features have severe adverse effects in this dataset. fused image features, compared to fc features, show decreased performance in recall@1 but better performance in recall@5 and recall@10, with a 1.0 improvement in recall@10. Therefore, the proposed method is also effective on the Flickr30K dataset.

Similarly, we conducted comparative experiments on the Flickr30K dataset between MLP using single-level features and MLP using multi-level features, with results presented in Table 5 . The results show that fused image features comprehensively outperform features in recall@K metrics. Thus, on the Flickr30K dataset, we can also conclude that using multi-level features is more advantageous than using single-level features during the fusion and dimensionality reduction process. Moreover, compared with the image recommendation algorithm, the method of directly using fused image features for similarity matching also demonstrates comprehensive superiority in recall@K performance, with an impressive 20.6 improvement in recall@10 compared to the image recommendation algorithm using only fc features.

Table 5 Comparison of similarity matching between image features and text features on Flickr30K dataset

3.3 Summary

Combining experimental results from both datasets, the proposed fusion algorithm can indeed effectively fuse and reduce the dimensionality of multi-level pre-trained image features, fully utilizing pre-trained features to generate fused image features with stronger expressive power for image-text matching tasks. Particularly, since the Sohu competition dataset consists of real-world news articles and their matching images, solving the image-text matching task on this dataset is more challenging. However, our method remains effective on this dataset and achieves better recommendation performance.

4 Conclusion

Image-text matching has always been a challenging task that requires accurate extraction of text and image features. For images in particular, obtaining image features is especially difficult due to their richer expressions of the same concepts. Numerous previous studies have proposed various methods for obtaining image

features, with the current mainstream approach being the use of pre-trained deep learning networks for image feature extraction. However, this mainstream approach fails to fully utilize useful features and is susceptible to noise feature interference.

To address these issues, this paper proposes a fusion algorithm that leverages multi-level deep representations of pre-trained image features: by utilizing the learning objective of the image-text matching task, the algorithm supervises the fusion and dimensionality reduction of multi-level pre-trained image features to generate fused image features, thereby making full use of more useful features and reducing noise interference. In the experimental section, by introducing fused image features into our implemented image recommendation algorithm for comparative experiments, we demonstrate that fused image features indeed possess stronger representational capabilities and achieve better recommendation performance. Furthermore, we designed an additional experiment proving that using multi-level features is more advantageous than using single-level features during the fusion and dimensionality reduction process, yielding better results. Based on all experimental results, we conclude that the proposed method is effective. Notably, although this paper focuses on the image-text matching task, the method can be extended to different tasks by changing the learning objective that guides the fusion and dimensionality reduction.

Currently, image-text matching on short sentence texts has achieved excellent results. However, for long sentence texts, due to their complex content and difficulty in extracting key features, methods applicable to short sentences are not suitable for long sentences. Therefore, image-text matching on long sentence texts will be a challenge to overcome.

References

[1] Song Yibin. Quick training method for multi-layer perception and its application [J]. Control and Decision, 2000, 15 (1): 125-127.

[2] Lipton Z C, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning [EB/OL]. (2015-10-17). https://arxiv.org/abs/1506.00019.

[3] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86 (11): 2278-2324.

[4] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9 (8): 1735-1780.

[5] Leng Yajun, Lu Qing, Liang Changyong. Survey of recommendation based on collaborative filtering [J]. Pattern Recognition and Artificial Intelligence, 2014, 27 (8): 720-734.

[6] Yan Fei, Mikolajczyk K. Deep correlation for matching images and text [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2015: 3441-3450.

[7] Ma Lin, Lu Zhengdong, Shang Lifeng, et al. Multimodal convolutional neural networks for matching image and sentence [C]// Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2015: 2623-2631.

[8] Wang Liwei, Li Yin, Lazebnik S. Learning deep structure-preserving image-text embeddings [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016: 5005-5013.

[9] Nam H, Ha J W, Kim J. Dual attention networks for multimodal reasoning and matching [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017: 2156-2164.

[10] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012: 1097-1105.

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10). https://arxiv.org/abs/1409.1556.

[12] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2015: 1-9.

[13] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C]// Proc of ACM International Conference on Machine Learning. New York: ACM Press, 2015: 448-456.

[14] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016: 2818-2826.

[15] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning [C]// Proc of AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2017: 4278-4284.

[16] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016: 770-778.

[17] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2014: 818-833.

[18] Garcia-Gasulla D, Ayguadé E, Labarta J, et al. A visual embedding for the unsupervised extraction of abstract semantics [EB/OL]. (2016-12-16). https://arxiv.org/abs/1507.08818.

[19] Agrawal P, Girshick R, Malik J. Analyzing the performance of multilayer neural networks for object recognition [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2014: 329-344.

[20] Vinyals O, Toshev A, Bengio S, et al. Show and tell: a neural image caption generator [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2015: 3156-3164.

[21] Peng Kuanchuan, Chen T. A framework of extracting multi-scale features using multiple convolutional neural networks [C]// Proc of IEEE International Conference on Multimedia and Expo. Piscataway, NJ: IEEE Press, 2015: 1-6.

[22] Liu Lingqiao, Shen Chunhua, van den Hengel A. The treasure beneath convolutional layers: cross-convolutional-layer pooling for image classification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2015: 4749-4757.

[23] Doersch C, Gupta A, Efros A A. Unsupervised visual representation learning by context prediction [C]// Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2015: 1422-1430.

[24] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016: 2414-2423.

[25] Liu Qingshan, Hang Renlong, Song Huihui, et al. Adaptive deep pyramid matching for remote sensing scene classification [EB/OL]. (2016-11-11). https://arxiv.org/abs/1611.03589.

[26] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation [C]// Proc of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234-241.

[27] Evangelopoulos N E. Latent semantic analysis [J]. Wiley Interdisciplinary Reviews: Cognitive Science, 2013, 4 (6): 683-692.

[28] Le Q, Mikolov T. Distributed representations of sentences and documents [C]// Proc of ACM International Conference on Machine Learning. New York: ACM Press, 2014: 1188-1196.

[29] Young P, Lai A, Hodosh M, et al. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions [J]. Transactions of the Association for Computational Linguistics, 2014, 2: 67-78.

[30] Cybenko G. Approximation by superpositions of a sigmoidal function [J]. Mathematics of Control, Signals and Systems, 1989, 2 (4): 303-314.

[31] Mao Junhua, Xu Wei, Yang Yi, et al. Deep captioning with multimodal recurrent neural networks (M-RNN) [EB/OL]. (2015-06-11). https://arxiv.org/abs/1412.6632.

Note: Figure translations are in progress. See original paper for figures.