Image-text Dual Neural Network with Decision Strategy for Small-sample Image Classification

Fangyi Zhu[a], Zhanyu Ma[a,*], Xiaoxu Li[a,b,*], Guang Chen[a], Jen-Tzung Chien[c], Jing-Hao Xue[d], Jun Guo[a]

[a]*Pattern Recognition and Intelligent System Lab., Beijing University of Posts and Telecommunications, Beijing, China*
[b]*School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China*
[c]*Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan, China*
[d]*Department of Statistical Science, University College London, London, UK*

Abstract

Small-sample classification is a challenging problem in computer vision. In this work, we show how to efficiently and effectively utilize semantic information of the annotations to improve the performance of small-sample classification. First, we propose an image-text dual neural network to improve the classification performance on small-sample datasets. The proposed model consists of two sub-models, an image classification model and a text classification model. After training the sub-models separately, we design a novel method to fuse the two sub-models rather than simply combine their results. Our image-text dual neural network aims to utilize the text information to overcome the training problem of deep models on small-sample datasets. Then, we propose to incorporate a decision strategy into the image-text dual neural network to further improve the performance of our original model on few-shot datasets. To demonstrate the effectiveness of the proposed models, we conduct experiments on the LabelMe and UIUC-Sports datasets. Experimental results show that our method is superior to other models.

☆This paper was presented in part at the CCF Chinese Conference on Computer Vision, Tianjin, 2017. This papaer was recommended by the program committee.

*Corresponding author
Email addresses: mazhanyu@bupt.edu.cn (Zhanyu Ma), lixiaoxu@lut.cn (Xiaoxu Li)

Keywords: Small-sample image classification, few-shot, ensemble learning, deep convolutional neural network

1. Introduction

With the wide use of the Internet, the amount of image data on the network is increasing dramatically. How to retrieve and understand the image data correctly is a hot but difficult problem in computer vision. Recently, with the development of deep learning, learning and extracting semantic information from massive images by using convolutional neural network provides an effective approach to image understanding. However, such an approach requires a large amount of trainging data, while in many applications, there is only a small amount of labeled data, e.g., in the LabelMe dataset [1] and the UIUC-Sports dataset [2]. These two datasets consist of 8 classes. Each class has less than 326 images. The total numbers of images in the two datasets are much smaller than those of other image datasets, such as Flickr [3] or MS COCO [4], which are frequently used in many image classification tasks. However, adding annotations to data will cost a lot of manual work. In addition, manual image tagging is very unstable, easy to get subjective and individual tagging errors. Therefore, how to achieve good classification performance on small-sample datasets is an important problem computer vision. Moreover, how to learn semantic information from a small amount of labeled samples, among others, is one striking challenge. Motivated by this observation, we aim to study image semantic learning on small-sample datasets.

Topic model-based methods have been a focus for learning semantic features[5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. A topic model refers to a statistical model that discovers or learns abstract topics of documents, which originates from natural language processing (NLP) [8, 9]. In recent years, with the fast developing of neural network research, the research about neural topic model, which is the topic model based on neural network [6], and image classification based on the neural topic model has been proposed [5, 6, 10]

Larochelle et al. proposed a model named Document Neural Autoregressive Distribution Estimator (DocNADE) [10], which can obtain good topic features. The model assumes that the generation of each word is only associated with the words that generated before it, and the document is the product of a set of conditional probabilities, which is generated by a feedforward neural network. This method models the relationship between words and the calculation of latent variables does not require complex approximate reasoning as other probabilities generation models. Zheng et al. presented the SupDocNADE, a shallow model based on the DocNADE [5, 6]. The model got 83.43% accuracy on the LabelMe dataset and 77.29% accuracy on the UIUC-Sports dataset [5, 6]. Recently, there has been remarkable progress in the direction under big data with the development of deep learning. AlexNet trained on ImageNet obtained perfect performance for image classificaton [15]. Zhou et al. [16] proposed a method of extracting features trained on the Places dataset. The model is similar to AlexNet, and got perfect performance on several datasets. Zisserman et al. proposed VGG-Net [17], which achieved the first place on the localization task of ILSVR and the second place on the classification task of ILSVR. However, the models mentioned above are all deep models. Since deep models have a large number of parameters needed to be trained, the models cannot be trained adequately on small-sample datasets, and thus, the performance will be constrained. To tackle this problem, we utilize the annotations to learn deep semantic features of the images and improve the performance of image classification on small-sample datasets.

In recent methods, e.g. SupDocNADE, Fu-L, Mv-sLDA[5, 6, 10, 18, 19, 20], there are two ways to learn the connections between images and the corresponding annotations. One was to input the joint features of images and annotations into one classification model [19, 11, 12, 13, 21, 22, 23, 24, 25], the other was to input the features of images and annotations to different classification models separately[20, 18]. In the existing methods, the classification models on small-sample datasets mostly utilize traditional non-deep models. This is because the deep models cannot be trained reliably on small-sample datasets. Hence,

we propose an image-text dual neural network in order to utilize the semantic information of the annotations to overcome the insufficient training problem. By compared with some recently proposed methods, we find that our semantic information can achieve significant improvement in terms of image classification accuracy.

To tackle the challenges, we decompose the problem of image classification into two manageable sub-problems in the prior CCCV paper: an image classification problem and a text classification problem, where the text is from the image annotations. An image classification model is trained on the given images, and a text classification model is built on the annotations of given images. On top of the two classification models, we introduce a fusion process to learn the connections between the two sub-models. To get better performance on few-shot datasets, we futher propose a novel decision strategy and build the image-text dual neural network with the decision strategy based on the conference version.

The main contribution of our paper is to propose a new image-text dual neural network with decision strategy. The model utilizes two models to learn image and text features separately and fuse results of the two models in the end. Our model can utilize the semantic information to improve the performance of image classification model. Compared with some existing models for image classification on small-sample datasets, our model achieves the best performance in terms of classification accuracy on the LabelMe dataset and the UIUC-Sports dataset [5, 6, 1, 2]. Our model achieves 97.75% classification accuracy on the LabelMe dataset [1] and 99.51% classification accuracy on the UIUC-Sports dataset [10]. Moreover, our model can also save computational resources significantly. Then, we propose to incorporate a decision strategy into the original image-text dual neural network to further improve the model performance on few-shot datasets [26].

2. Image-text Dual Neural Network with Decision Strategy

In this section, we describe our new image-text dual neural network, which is inspired by DocNADE [5, 6] to learn jointly from multimodal data.

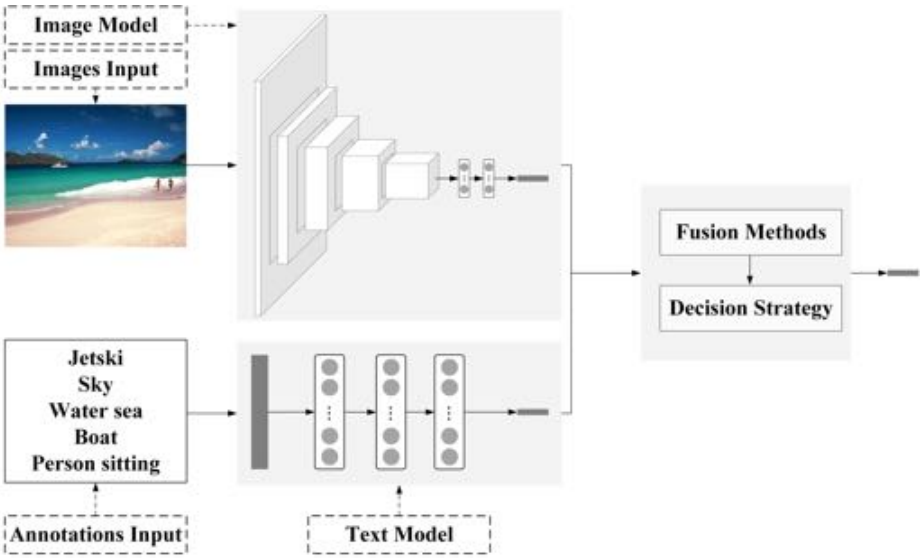

Figure 1: Illustration of the image-text dual neural network. Images and annotations are trained separately. On top of the two classification models, we add another fusion processing to merge results of the two models. To learn the connections between the two models, we propose a method in the fusion process.

2.1. Image-text Dual Neural Network

In order to utilize the semantic information of the annotations to improve the performance of image classification model, we first propose a simple yet effective image-text dual neural network. It decomposes image classification into two manageable sub-models.

Image Model. It is an end-to-end neural network fine-tuned by VGG16 [17], which gives the classification results of images.

Text Model. It is an end-to-end neural network as well, which gives the classification results of annotations.

Then, we propose a method to fuse the classification results of images and annotations of images to predict the final class of the input images. The architecture of the proposed image-text dual neural network is shown in Fig.1.

2.1.1. Image Model

Instead of simply joining the image features and the annotations features as input of one model, we train two models separately. We use transfer learning [27] to build the models. Transfer learning allows to utilize known relevant tasks to slove new unknown tasks. We use VGG16 [17] as the pre-trained model and fine-tune the model on our pre-processed datasets [1, 2]. The structure of our image model is shown in Fig.2.

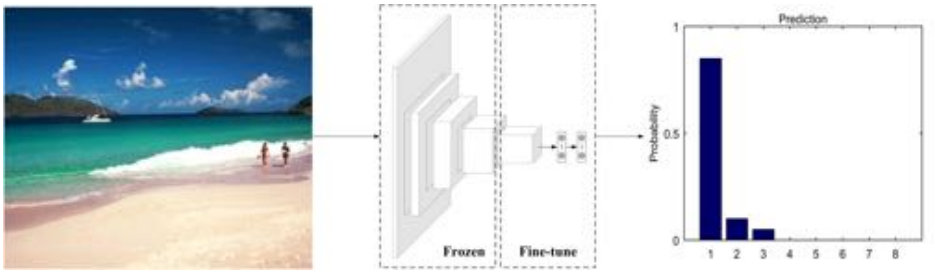

Figure 2: Illustration of *image model*. We take only the convolutional layers of VGG16 and drop the FC layers. On top of the convolutional layers, we add two FC layers. When fine-tune the model, we freeze the first four convolution blocks and fine-tune the fifth convolution block and new FC layers.

Implementation details. The model consists of five convolutional blocks and two full-connected (FC) layers. The convolutional blocks consist of 3×3 stride 1 convolutions and ReLU. Between two convolutional blocks, there is 2×2 stride 2 maxpooling. The stack of convolutional layers is followed by two FC layers: the first contains 512 channels, and the second is a soft-max layer with 8 channels (one for each class).

2.1.2. Text Model

To learn the semantic information of the annotations, we build a text classification model consisting of three FC layers as shown in Fig.3. There are

many ways to build word embedding. We applied the word2vec [28, 29] and bag of words model to build the text vectors. The two methods work well in text classification tasks. Considering redundant information of text features for classification, we further use PCA [30] to reduce the word vectors to lower dimensions.

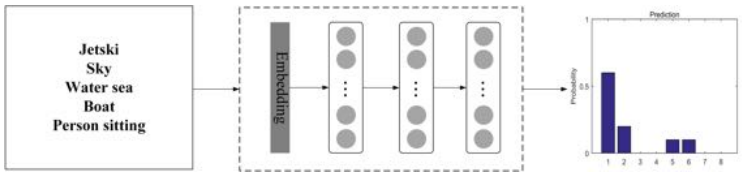

Figure 3: Illustration of *text model*.

Implementation details. The model consists of three FC layers: the first contains 64 channels, the second contains 512 channels, and the third is a soft-max layer with 8 channels (one for each class).

2.1.3. Fusion Models

Given data (v, t) containing images v, and text annotations t. We get the output $P_{img}(v)$ of soft-max layer of the image model and the output $P_{text}(t)$ of soft-max layer of the text model. $P_{img}(v)$ and $P_{text}(t)$ are both of eight dimensions. Each of the dimensions represents the probability of the sample belongs to a specific class. The final result c_{final} is predicted by merging results of the two models. A simple way of fusion is to add results of the two sub-models together. However this fusion strategy cannot get good results. We further propose a method to combine the results of the two models as shown in (1), where λ is a regularization parameter that controls the balance between the two sub-models. The range of values of λ is between 0 to 1.

$$c_{final} = max_c[\lambda P_{img}(v) + (1 - \lambda)P_{text}(t)]. \quad (1)$$

2.2. Image-text Dual Neural Network with Decision Strategy

Although our original image-text dual neural network has achieved better performance than the other models as reported in our pervious work [31], here

we incorporate it with a decision strategy to further improve the performance of the original model.

By analyzing the annotations, we find that many annotations of images are not discriminative, e.g. sky, trees, people, athlete and so on. These words could appear in every class and have no help for classification task. If an image only has the indiscriminative words, the text classifier cannot get the correct classification result. Therefore, we add a decision strategy after the fusion. If the image annotations are all indiscriminative, we will directly use the result of image model and neglect the result of text model. With this decision strategy, the semantic information of the annotations can be used selectively to ensure high performance.

Given data (v, t), we define a compatibility function $f : v \times t \to [0, 1]$ that uses features $\theta(v)$ for images and $\varphi(t)$ for annotations after removing the indiscriminative words [32]:

$$f(v,t) = f(\theta(v)^T \varphi(t)) = \begin{cases} 0, & \text{if } \theta(v)^T \varphi(t) = 0, \\ 1, & \text{if } \theta(v)^T \varphi(t) \neq 0. \end{cases} \tag{2}$$

We then formulate the predicted class as follows:

$$c^*_{final} = max_c[\lambda P_{img}(c) + f(v,t)(1-\lambda)P_{text}(c)]. \tag{3}$$

3. Experiment Results and Discussions

In this section, we compare the performance of our model with the other models for image classification on small-sample datasets. Specifically, we first test our original image-text dual neural network [31] to learn from multimodal data on two real-world datasets: the LabelMe dataset and the UIUC-Sports dataset [1, 2] which are widely used in the image classification tasks. Then we futher test our original image-text dual neural network and our image-text dual neural network with decision strategy on few-shot datasets.

3.1. Experiments for our Original Image-text Dual Neural Network

To evaluate the proposed image-text dual neural network, we conduct extensive quantitative and qualitative evaluations on the LabelMe and the UIUC-Sports datasets [1, 2]. The two datasets contain image annotations and are popular classification benchmarks. We provide a quantitative comparison between SupDocNADE, Fu-L, Mv-sLDA [5, 6, 18] and our original image-text dual neural network. We use the average classification accuracy of the five subsets to measure the performance of image classification.

3.1.1. Datasets

The LabelMe dataset [1] has eight classes: coast, forest, highway, inside city, open country, street, and tall building. For each class in one of our subsets, 200 images are randomly selected and split evenly into the training and test sets, yielding 1600 images.

The UIUC-Sports dataset [2] contains 1792 images, classified into eight classes. Each subset we constructed consists of 1720 images: badminton (300 images), bocce (130 images), croquet (300 images), polo (190 images), rock-climbing (190 images), rowing (250 images), sailing (180 images), and snow-boarding (190 images). The images are randomly selected and split evenly into the training and test sets.

Following Li et al. [18], we randomly extract five subsets of the LabelMe dataset [1] and five subsets of the UIUC-Sports dataset [2].

3.1.2. Performance Analysis

In this section, we analyze our proposed dual model on the LabelMe dataset [1] and the UIUC-Sports dataset [2] with baseline models, as shown in Fig.4.

The average image classification accuracy of our single image model on the LabelMe dataset [1] is 80.9% and on the UIUC-Sports dataset[2] is 70.7%. Our single image model also can get better performance than single the image classification model in Fu-L and Mv-sLDA [18] for the reason that we adopt deep learning to slove the task.

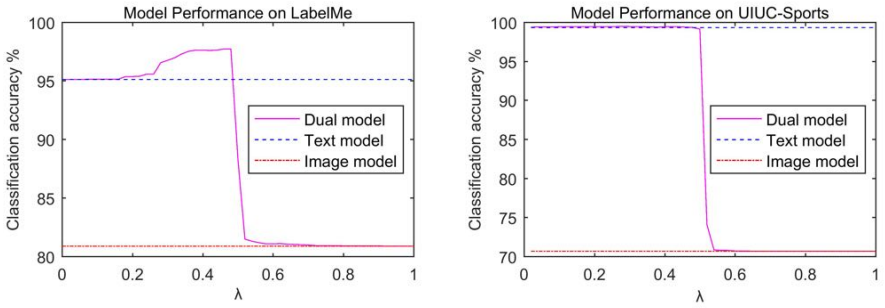

Figure 4: The vertical axis presents the classification accuracy and the horizontal axis represents the value of the weight λ in (1).

We use word2vec [28, 29] and bag of words to build text vectors. However, word2vec [28, 29] in our task does not have good performance. The reason is that word2vec fits for sentences but not for separate words. Furthermore, with only less than 1000 words in total, neural network cannot get a good word2vec embedding. In our work, we choose the bag of words model to build the text vectors. We further use PCA [30] to reduce word vectors to lower dimensions for removing useless information for classification task. The average text classification accuracy on the LabelMe dataset without PCA [30] is 95.13%. The accuracy is 94.73% after reducing to 480 dimensions with PCA. The accuracy is 93.78% after reducing to 240 dimensions. The model without PCA is always higher than the model with PCA [30]. This is due to the fact that the dimension of the original vectors is not high for the neural model to train and without much redundant information for classification. Therefore, better result can be obtained without dimension reduction. The average text classification accuracy on the UIUC-Sports dataset [2] without PCA [30] is 99.35%. The aforementioned text classification model also can get better performance than the text classification models in Fu-L and Mv-sLDA [18].

Figure 4 also shows the trend of image classification accuracy with different values of λ. When λ is 0, the model is the single text classification model. When λ is 1, the model is the single image classification model. When λ is 0.5, the model has the same result with simply combining the two sub-models. The

better performance of the model can be obtained when λ falls in [0.4, 0.5] on the LabelMe dataset and can be obtained when λ falls in [0.2, 0.3] on the UIUC-Sports dataset. As illustrated in Fig. 4, the dual model achieves significant improvement than the single models. It exhibits that our original image-text dual neural network is reliably effective. These experimental results demonstrate that the proposed dual model can overcome the insufficient training of deep models on small-sample datasets. The idea that utilizing semantic information of the image annotations to improve the capability of the image classification model is proved to be feasible. In addition, the accuracy of the dual model drops rapidly when λ is equal or greater than 0.5 verifies the effectiveness of our proposed fusing method.

Table 1: Performance comparison of our original deep image-texet dual neural network between different setting on the LabelMe dataset.

Model	**Accuracy(average)%**	**Accuracy(max)%**
Our model ($\lambda = 0.40$)	97.65	98.25
Our model ($\lambda = 0.42$)	97.63	98.25
Our model ($\lambda = 0.44$)	97.65	98.25
Our model ($\lambda = 0.46$)	**97.75**	**98.25**
Our model ($\lambda = 0.48$)	97.75	98.13
Our model ($\lambda = 0.50$)	88.35	90.63

From Table 1 and Table 2, we can conclude that the proposed original image-text dual neural network can get the highest image classification accuracy on the LabelMe dataset when λ is 0.46 and on the UIUC-Sports dataset when λ is 0.26.

3.1.3. Experiment Results

In this section, we describe quantitative comparison between our original image-text dual neural network and other methods. The classification results

Table 2: Performance comparison of our original image-text dual neural network between different settings on the UIUC-Sports.

Model	**Accuracy(average)%**	**Accuracy(max)%**
Our model ($\lambda = 0.20$)	99.48	99.76
Our model ($\lambda = 0.22$)	99.48	99.76
Our model ($\lambda = 0.24$)	99.48	99.76
Our model ($\lambda = 0.26$)	**99.51**	**99.76**
Our model ($\lambda = 0.28$)	99.48	99.76
Our model ($\lambda = 0.30$)	99.46	99.76

are illustrated in Table 3. Our original model obtains an accuracy of 97.75% on the LabelMe [1] dataset and 99.51% on the UIUC-Sports dataset. This is significantly superior to performance of other models on the two datasets, as shown in Table 3.

We compare the proposed image-text dual neural network with the SupDocNADE [5, 6], which uses the image-text joint embedding during training and inputs images only during the test stage. Since using the code published by the authors cannot get the accuracy reported in the papers, we compare directly with the results reported in the corresponding papers [5, 6].The reported accuracy is 83.43% on the LabelMe dataset and 77.29% on the UIUC-Sports dataset for SupDocNADE [5, 6], which is lower than our image-text dual neural network as shown in Table 3. Therefore, we can conclude that training the image and text separately can yield better results. In addition, the max iteration number for training SupDocNADE [5, 6] is 3000. Our image model only needs 350 epochs to train at most, 100 epochs to fine-tune, and the text model also only needs 100 epochs at most, as the number of parameters of our model is less than that of SupDocNADE. Moreover, the two sub-models can be trained parallelly at the same time. Our single image model's accuracy is lower than SupDocNADE because deep models have the insufficient training problem

and the single image model do not utilize the annotations. In conlusion, the final results show that our model can learn the semantic information of annotations better and overcome the insufficient training problem of deep models on small-sample datasets.

Table 3: Performance comparison of different models.

Datasets	SupDocNADE	Fu-L	Mv-sLDA	Our model
LabelMe	83.43%	82.3%	92.2%	**97.75%**
UIUC-Sports	77.29%	84.1%	99.0%	**99.51%**

We also compare the proposed model with existing methods requiring both images and annotations during the test stage, i.e., Fu-L and Mv-sLDA [18]. Fu-L and Mv-sLDA all utilize traditional non-deep learning [18]. The structure of Fu-L is similar to our model. Fu-L builds two separate traditional models for image and text. Then Fu-L utilizes the third model to fuse the two models. Mv-sLDA also trains the image model and the text model separately, and then fuses the results of the two models. As shown in Table 3, our model achieves the best performance in terms of image classification, due to the deep models we adopted.

3.2. Experiments for our Image-text Dual Neural Network with Decision Strategy

We now test the performance of our image-text dual neural network with decision strategy. In this section, we will show that our image-text dual neural network with decision strategy achieves better performance than our original image-text dual neural network on small-sample datasets, and it also can achieve good perfoemance on few-shot datasets.

3.2.1. Datasets

To test the ability of our original model and the proposed decision strategy, we measure their performance on few-shot data. We treat our original model

as the baseline method and test on reduced datasets with only 10, 20, and 40 images in the training sets and keep the test sets unchanged. For each number of images in the training set, we randomly extract twenty subsets from the whole datasets and use the average classification accuracy over the twenty subsets.

3.2.2. Experimental Results

The classification results are illustrated in Table 4. We observe the image-text dual neural network with decision strategy outperforms the original model. From Fig.5, we can also observe the improvement obviously. Furthermore, we can conclude that the classification accuracy of 10-shot, 20-shot, 40-shot improves with the number of images in the training set. We can conclude that our model not only can get good results on small-sample datasets, but also can apply to large amount of data. In the following paper, we call the original image-text dual neural network without the decision strategy as *our model*, and call the image-text dual neural network with the decision strategy as *our model**.

Table 4: Performance comparison on LabelMe and UIUC-Sports datasets. *Our model* presents the original image-text dual neural network, and *our model** presents the image-text dual neural network with decision strategy.

		10-shot	**20-shot**	**40-shot**	**All**
LabelMe	Our model	86.93%	88.73%	91.28%	92.33%
	Our model*	89.80%	92.65%	94.39%	96.36%
UIUC	Our model	82.88%	83.10%	84.14%	99.35%
	Our model*	85.20%	85.48%	85.98%	99.97%

To further demonstrate these improved results are not obtained by chance, we do the Paired t-Test . We use p-values of the Paired t-Test of image classification accuracies obtained by models to evluate if the two given methods have significant difference. The p-values between our original model and our improved model are listed in Table 5. In the table of hypothesis testing results,

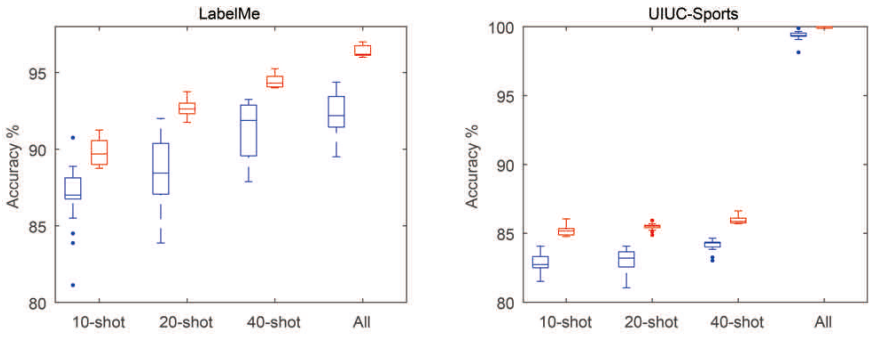

Figure 5: Comparison of accuracies obtained by *our model* and *our model** via box plot on the LabelMe and UIUC-Sports datasets. The blue boxes are obtained by *our model*, and the red boxes are obtained by *our model**. The central mark is the median, and the edges of the box are the 25th and 75th percentiles. The outliers are marked individually.

Table 5: P-values of the Paired t-Test between *our model* and *our model** on different datasets.

	10-shot	**20-shot**	**40-shot**	**All**
LabelMe	***	***	***	***
UIUC-Sports	***	***	***	***

instead of showing the exact p-values, we use "The * methods: * means p-value ≤ 0.05; ** means p-value ≤ 0.01; *** means p-value ≤ 0.001." All the p-values bewteen the original image-text dual neural network and the image-text dual neural network with decision strategy are smaller than 0.05, where 0.05 is the significant level. Therefore, our original model has statistically significant different performance from the image-text dual neural network with decision strategy. Morever, as shown in Table 6 and Table 7, the results of the Paired t-Test of the same model between different reduced sets and the whole set also show statistically significant difference. These results further prove the model performance.

Table 6: P-values of the Paired t-Test of accuracies obtained by *our model* and *our model** between different reduced sets and the whole set on the LabelMe dataset.

Our model	**10-shot**	**20-shot**	**40-shot**	**All**
10-shot	-	*	***	***
20-shot	*	-	**	***
40-shot	***	**	-	*
All	***	***	*	-
Our model*	**10-shot**	**20-shot**	**40-shot**	**All**
10-shot	-	***	***	***
20-shot	***	-	***	***
40-shot	***	***	-	***
All	***	***	***	-

Table 7: P-values of the Paired t-Test of accuracies obtained by *our model* and *our model** between different reduced sets and the whole set on the UIUC-Sports dataset.

Our model	**10-shot**	**20-shot**	**40-shot**	**All**
10-shot	-	*	***	***
20-shot	*	-	***	***
40-shot	***	***	-	***
All	***	***	***	-
Our model*	**10-shot**	**20-shot**	**40-shot**	**All**
10-shot	-	*	***	***
20-shot	*	-	***	***
40-shot	***	***	-	***
All	***	***	***	-

4. Conclusions

In this paper, we propose an image-text dual neural network for small-sample image classification. The proposed method decomposes the image classification model into two manageable sub-models, i.e., an image classification model and a text classification model. Furthermore, we propose a method to fuse the two sub-models. Finally, we propose a decision strategy to improve the model performance. Extensive quantitative and qualitative results demonstrate the effectiveness of our proposed model and the decision strategy. Compared with some existing models, our method can better incorporate the semantic information of the annotations, and thus can get higher image classification accuracy. Moreover, our image-text dual neural network needs fewer epochs to train because it contains fewer parameters. In our work, the two sub-models can be trained at the same time, which also contributes to saving the computational efficiency. In addition, the structure of our dual model can be extended to other modalities, e.g. image-sketch, image-video, text-video.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61773071, Grant 61628301 and Grant 61563030, in part by the Beijing Nova Program under Grant Z171100001117049, in part by the Beijing Natural Science Foundation (BNSF) under Grant 4162044, in part by the Gansu Natural Science Foundation of China under Grant 2017GS10830.

References

[1] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, International Journal of Computer Vision 42 (3) (2001) 145–175.

[2] L. Li, F. Li, What, where and who? classifying events by scene and object recognition, in: IEEE International Conference on Computer Vision, IEEE, 2007, pp. 1–8.

[3] J. S. Hare, P. H. Lewis, Automatically annotating the mir flickr dataset: Experimental protocols, openly available data and semantic spaces, in: ACM Sigmm International Conference on Multimedia Information Retrieval, ACM, 2010, pp. 547–556.

[4] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girhick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, , P. Dollr, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[5] Y. Zheng, Y. Zhang, H. Larochelle, Topic modeling of multimodal data: an autoregressive approach, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1370–1377.

[6] Y. Zheng, Y. Zhang, H. Larochelle, A deep and autoregressive approach for topic modeling of multimodal data, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (6) (2016) 1056–1069.

[7] D. Putthividhya, H. T. Attias, S. S. Nagarajan, Topic regression multimodal latent dirichlet allocation for image annotation, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 3408–3415.

[8] T. Hofmann, Probabilistic latent semantic indexing, in: ACM Special Interest Group on Information Retrieval, ACM, 1999, pp. 50–57.

[9] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (1) (2003) 993–1022.

[10] H. Larochelle, S. Lauly, A neural autoregressive topic model, in: Annual Conference on Neural Information Processing Systems, 2012, pp. 2708–2716.

[11] Z. Niu, G. Hua, X. Gao, Q. Tian, Context aware topic model for scene recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2743–2750.

[12] N. Rasiwasia, N. Vasconcelos, Latent dirichlet allocation models for image classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (11) (2013) 2665–2679.

[13] J. D. Mcauliffe, D. M. Blei, Supervised topic models, Annual Conference on Neural Information Processing Systems 3 (2010) 327,332.

[14] D. M. Blei, M. I. Jordan, Modeling annotated data, in: ACM Special Interest Group on Information Retrieval, 2003, pp. 127–134.

[15] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Annual Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.

[16] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Annual Conference on Neural Information Processing Systems, 2014, pp. 487–495.

[17] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[18] X. Li, R. Li, F. Feng, J. Cao, X. Wang, Multi-view supervised latent dirichlet allocation, Acta Electronica Sinica 42 (10) (2014) 2040–2044.

[19] L. Wu, S. Oviatt, P. Cohen, Multimodal integration - a statistical view, IEEE Transactions on Multimedia 1 (4) (1999) 334–341.

[20] G. Wang, D. Hoiem, D. Forsyth, Building text features for object image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1367–1374.

[21] W. J. Frawley, G. P. Shapiro, S. J. Matheus, Knowledge discovery in databases: An overview, AI Magazine 13 (3) (1992) 57–70.

[22] R. Memisevic, On multi-view feature learning, in: International Conference on Machine Learning, 2012.

[23] N. Chen, J. Zhu, E. P. Xing, Predictive subspace learning for multi-view data: a large margin approach, in: Annual Conference on Neural Information Processing Systems, 2010, pp. 361–369.

[24] C. Chang, C. Jin, Libsvm: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (3).

[25] G. C. Cawley, N. L. C. Talbot, M. Girolami, Sparse multinomial logistic regression via bayesian l1 regularisation, in: Annual Conference on Neural Information Processing Systems, 2007, pp. 209–216.

[26] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning, in: International Conference on Learning Represenntations, 2017.

[27] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (10) (2010) 1345–1359.

[28] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[29] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International Conference on Machine Learning, 2014, pp. 1188–1196.

[30] C. M. Bishop, Pattern recognition and machine learning, springer, 2006.

[31] F. Zhu, X. Li, Z. Ma, G. Chen, P. Peng, X. Guo, J.-T. Chien, J. Guo, Image-text dual model for small-sample image classification, in: The Chinese Conference on Computer Vision, 2017.

[32] S. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.