Fine-grained Image Classification Combined with Label Description

Xiruo Shi
School of Computer Science
Beijing University of
Posts and Telecommunications,
National Engineering Laboratory of
Big Data Application on
Improving Government Governance Capabilities
Beijing, China
shixiruo@bupt.edu.cn

Liutong Xu
School of Computer Science
Beijing University of
Posts and Telecommunications
Beijing, China
xliutong@bupt.edu.cn

Pengfei Wang
School of Computer Science
Beijing University of
Posts and Telecommunications
Beijing, China
wangpengfei@bupt.edu.cn

***Abstract*—Fine-grained image classification faces huge challenges because fine-grained images are similar overall, and the distinguishable regions are difficult to find. Generally, in this task, label descriptions contain valuable semantic information that is accurately compatible with discriminative features of images (i.e., the description of the "Rusty Black Bird" corresponding to the morphological characteristics of its image). Bringing these descriptions into consideration is benefit to discern these similar images. Previous works, however, usually ignore label descriptions and just mine informative features from images, thus the performance may be limited. In this paper, we try to take both label descriptions and images into consideration, and we formalize the classification task into a matching task to address this issue. Specifically, Our model is based on a combination of Convolutional Neural Networks (CNN) over images and Graph Convolutional Networks(GCN) over label descriptions. We map the resulting image representations and text representations to the same dimension for matching and achieve the purpose of classification through the matching operation. Experimental results demonstrate that our approach can achieve the best performance compared with the state-of-the-art methods on the datasets of Stanford dogs and CUB-200-2011.**

I. Introduction

Fine-grained image classification aims to recognize sub-categories under some basic-level categories (e.g., classifying different bird types [28], dog breeds [17], car models [14], flower species [27], etc.). Due to the development of generic image recognition on large scale datasets [3], models of fine-grained image classification have made great progress in recent years. However, the large intra-class variance and small inter-class variance make this work still faces a big challenge, as shown in Fig. 1. The fine-grained images are similar in general and can only be distinguished by local information (such as claw shape). Most of the fine-grained image classification methods now consist of the following two steps:(1) localizing the object or its discriminative parts depend on the bounding box of manual annotation [12][41][42], or analyzing convolutional responses from neural networks in an unsupervised fashion [32][36][4] (2) using Deep CNN to extract features from the obtained regions for classification. However, these methods have the following limitations:

Fig. 1. Example of large intra-class variance and small interclass variance from CUB-200-2011 and Stanford Dogs.

manual annotations are very expensive, and not all of the extracted discriminative parts are valid for the final classification. He and Peng [8] proposed a two-stream model combining vision and language (CVL) for learning latent semantic representations. They divided the experiment into two streams: image stream and language stream. The image stream is used to extract image features for classification. Language stream refers to a method in which images and fine-grained visual descriptions can be jointly embedded for

2375-0197/19/$31.00 ©2019 IEEE
DOI 10.1109/ICTAI.2019.00148

classification using DS-SJE [31] algorithm.

Label descriptions contain valuable semantic information that is accurately compatible with discriminative features of images. It is difficult to find the subtle, distinguishable details expressed in the image, but these details can be well expressed by label descriptions. Good label description expressions may complement image detail information, which have a positive impact on fine-grained image classification. Recently, a new research method called graph neural networks [18][6][2] or graph embeddings which can preserve global structure information of a graph in graph embeddings has received extensive attention. It is used to introduce information representations of nodes and edges, and may be effective for tasks that could have rich relationship information between different entities. Inspired by He [8] and graph neural networks, we propose a fine-grained image classification model combining images and image label descriptions. Our model converts image classification tasks into matching tasks for images and label descriptions. The matching score is designed to increase if the image has the same category as label description and decrease otherwise. Our approach uses the well-acknowledged VGG-19 model with batch normalization as an image feature extractor. The label description graph is built using the method in TextGCN [40], and the features of the label description on the graph are extracted. Different from He's [8] dual stream network, our model is implemented with merely a single stream and has achieved positive results. Our contributions can be summarized as follows:

(1) We combine image label information with image information to learn latent semantic representations which can induce the image feature extractor to extract distinguishable features represented in the label description, and convert the classification problem into a matching problem. Our model does not need any annotation information or any preprocessing (e.g., object localization) of the image.

(2) In order to extract the features of the label description better, we use the graph neural networks which can preserve the global structure information to get the text features. The words in the label description which are better (or more representative) to describe the image can be found through our model.

(3) We conduct comprehensive experiments on two challenging datasets (CUB Birds, Stanford Dogs), and the proposed approach outperforms the state-of-the-art approaches on both datasets.

II. Related Work

A. Fine-grained Image Classification

A variety of methods have been developed for fine-grained object recognition. Most methods of fine-grained image classification consist of first finding discriminative parts and then extracting features by CNN. To focus on the discriminative regions, some methods depend on the annotation information of the data [42][23][37][25][41]. Part R-CNN [42] learned detectors and part models under a geometric prior which extended R-CNN [5], then predicted a fine-grained category from a pose-normalized representation. Zhang et al. [41] proposed a CNN architecture that integrates semantic part detection and abstraction for fine-grained classification.

To reduce the use of manual annotation information, the methods of finding object parts using little or no supervision of parts have been widely proposed. Simon and Rodner [32] proposed a neural activation constellations part model (NAC) to localize parts with the constellation model. Tianjun Xiao et al. [36]aimed to select relevant proposals to the object and the discriminative parts by two-level attention. Zhang et al. [44] proposed a model named PDFR, picking deep filter responses proposes to find distinctive filters and learn part detectors. Fu et al. [4] recurrently predicted the location of one attention area and extracted the corresponding features through a novel recursive attention convolutional neural network(RA-CNN). Recently, the WS-DAN model [11] introduced the data augmentation method into the task of fine-grained image classification which has achieved good results on both the CUB-200-2011 dataset and the Stanford Dogs dataset.

B. Combination Analysis of Image and Text

A number of approaches have been developed for grounding text in the visual domain [20][47][19]. Popular approaches include learning joint image-word embeddings as well as embedding images and sentences into a common space. Hodosh et al. [10] applied canonical correlation analysis(CCA) to find embeddings that maximize the correlation between images and sentences, which is further improved by incorporating deep neural networks. Karpathy et al. [16] decomposed images and sentences into fragments and infer their inter-modal alignment using a ranking objective. In 2017, they [15] further introduced an alignment model based on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks (RNN) over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. Peng et al. [29] proposed to build multiple deep networks for cross-media shared representation learning. It integrated the intra-modal and inter-modal representations to learn the cross-modal correlation using hierarchical neural networks. CNN is widely used for image modeling, and LSTMs [9] and character-based convolutional networks [43] are widely used for text modeling. In this paper, we apply the extension of Graph Convolutional Network(GCN) to get a visual semantic embedding. We build a graph with all the label descriptions, and perform a convolution operation on the graph. Each label description node is connected to each word in the label description, and each word is connected to the word co-occurring with it. We combine image information and label descriptions to build a model to improve the accuracy of fine-grained image classification.

III. Approach

In this section, we introduce the model that combines image features and label descriptions features for fine-grained image classification. We convert a classification task into a matching task. Image label description is another expression of

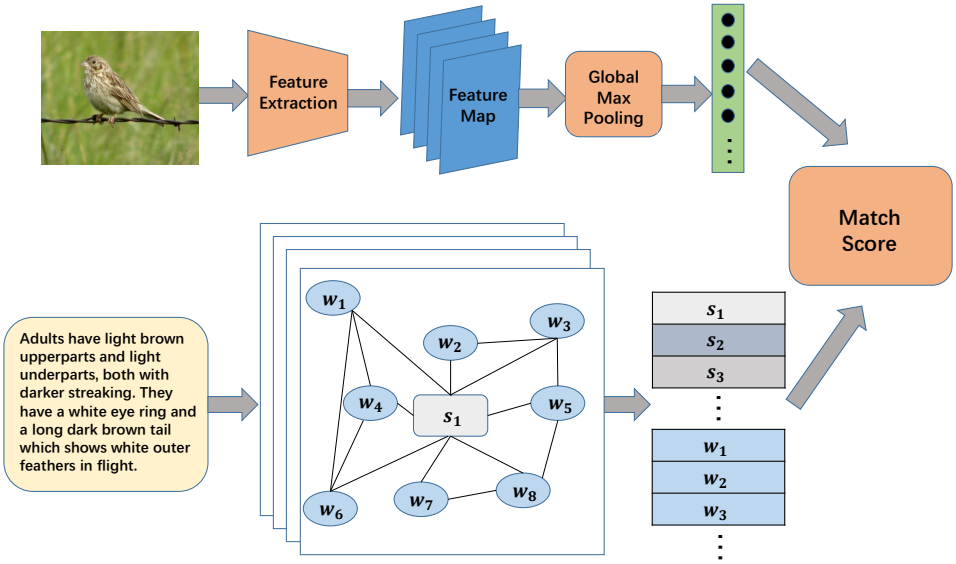


Fig. 2. Overview of our approach. The inputs are unprocessed images of size 224×224 and label descriptions for each image category. VGG-19 model is used to extract image features. The label description is used to build a graph. The ellipse w represents the word node of label description, and the square s represents the label description node. Each label description node is connected to each word in the label description, and each word is connected to the word co-occurring with it. Performing the convolution operation on the graph to extract the label description features. Image features and label description features are matched to get matching scores. The model combines image features and label description features to achieve better classification.

image information, natural language description features can complement the detailed features of the image. Therefore, we propose a model that combines image and label description to classify fine-grained images, which combines the advantages of image and text, as shown in Fig. 2.

A. Vision Encoder Model

In this paper, we do not need to do any pre-processing of the image, nor do we need to locate the object in the image during training and testing, and the object's location annotation(e.g. bounding boxes or keypoints) is not available. We use the Oxford VGGNet-19 [33] pre-trained on ImageNet without fine-tuning as our image encoder to get image features. Given an input image I, We extract image features vectors v using VGGNet19 without fully connected layers, we get the $7\times7\times512$ feature map of the sixteen convolution layer, shown as:

$$v = CNN_\theta(I) \qquad (1)$$

We perform global maximum pooling on the resulting $7\times7\times512$ feature map to get the $1\times1\times512$ feature map, then the $1\times1\times512$ feature map is compressed into 512 vectors v.

B. Text Encoder Model

To get a better text expression, we introduce the TextGCN [43] to build a graph with all the label descriptions, and perform a convolution operation on the graph. Each document node in the TextGCN [43] is connected to each word in the document, and each word node is connected to the word co-occurring with it. Since GCN can preserve the global structure information of a graph in graph embeddings, it can be better compared to LSTM [35] [26], TextCNN [43] [22] and other methods to get the expression of text features.

1) Graph Convolutional Networks(GCN): The GCN [18] model is a multi layer neural network that operates directly on a graph and induces embedding vectors of nodes based on properties of their neighborhoods. The primary thought of GCN is to realize the convolution operation on the topology map utilizing the theory of the spectrum. The GCN model learns node representations based on the node features and their connections. A multi-layer Graph Convolutional Network (GCN) with the following layer-wise propagation rule [40]:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \qquad (2)$$

where $\tilde{A} = A + I$ indicates adjacency matrix added self-connection, $\tilde{D_{ii}} = \sum_j \tilde{A_{ij}}$ is a diagonal matrix, $H^{(l)}$ and $W^{(l)}$ are the node representation matrix and the trainable parameter matrix for the lth layer, $H^{(0)}$ can be regarded as the original feature matrix, $\sigma(\cdot)$ is the activation function.

2) Build Graph Convolutional Networks(GCN): Our task is to use label descriptions to construct graph representations and obtain text features through graph convolution. We treat the documents (label descriptions) and all the words in the label descriptions as nodes to construct a large heterogeneous text graph proposed in TextGCN [40]. We construct the edge between the document node and the word node in the document, and initialize the edge weight with the term frequency-inverse

Category	Image	Label Description
Rusty Black Bird		They have a pointed bill and a pale yellow eye. They have black plumage with faint green and purple gloss; the female is greyer. "Rusty" refers to the brownish winter plumage. They resemble the western member of the same genus, the Brewer's blackbird; however, Brewer's has a longer bill and the male's head is iridescent green
Purple Finch		Adults have a short forked brown tail and brown wings and are about 15cm in length and weigh 34g (1.2oz). Adult males are raspberry red on the head, breast, back and rump; their back is streaked. Adult females have light brown upperparts and white underparts with dark brown streaks throughout; they have a white line on the face above the eye.
Brown Creeper		Adults are brown on the upper parts with light spotting, resembling a piece of tree bark, with white underparts. They have a long thin bill with a slight downward curve and a long stiff tail used for support as the bird creeps upwards. The male creeper has a slightly larger bill than the female. The is 11.7–13.5cm (4.6–5.3in) long.
Cape Glossy Starling		The Cape starling, red-shouldered glossy-starling or Cape glossy he Cape starling has an adult length of about 25 cm (10 in) and weight of about 100 grams (3.5oz). The plumage is a fairly uniform bright, glossy color. The head is blue with darker ear coverts and the upper parts of the body are greenish-blue.

Fig. 3. Sample label description of CUB-200-2011

document frequency (TF-IDF) method:

$$\mathrm{A}(i,j) = \begin{cases} \mathrm{PMI}(i,j), & i,j \text{ are words}, \mathrm{PMI}(i,j) > 0 \\ \mathrm{TF-IDF}_{ij}, & i \text{ is document}, j \text{ is word} \\ 1, & i = j \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

We construct edges between the word and the word co-occurring with it and initialize the edge weight with the point-wise mutual information(PMI):

$$\mathrm{PMI}(i,j) = log(\frac{p(i,j)}{p(i)p(j)}) \quad (4)$$

$$p(i,j) = \frac{\#W(i,j)}{\#W} \quad (5)$$

$$p(i) = \frac{\#W(i)}{\#W} \quad (6)$$

where $\#W(i)$ is the number of sliding windows in label descriptions that contain word i, $\#W(i,j)$ is the number of sliding windows that contain both word i and j, and $\#W$ is the total number of sliding windows in the label descriptions. After building the graph, we feed the graph into a simple one layer GCN according to Eqn. (2) to get a $nodesize \times 512$ dimensions text feature map $S = \{s_1, s_2, s_3 \dots s_n, w_1, \dots w_m\}$ where s_i represents document (label description) features, w_i represents word features, n represents the number of label descriptions, m represents the number of words in all label descriptions and $n + m = nodesize$. We use the label description features to match the image, and use the word features to find the words in the label description which are better (or more representative) to describe the image.

C. Final Prediction

We have described the transformations that map every image and label description into a set of vectors in a common h-dimensional space. Inspired by Karpathy et al [16], the model interprets the dot product $v_i^T s_j$ between the ith image and jth label description as a measure of similarity and use it to define the score between image v_i and label description s_j.

$$Z_{ij} = v_i^T s_j \quad (7)$$

Given images $I = \{I_1, I_2, I_3, I_4 \dots I_n\}$, label descriptions $T = \{T_1, T_2, T_3, T_4 \dots T_n\}$, n indicates number of image categories, we can get the image feature vectors $V = \{v_1, v_2, v_3, v_4 \dots v_n\}$ and text feature vectors $S = \{s_1, s_2, s_3 \dots s_n, w_1, \dots w_m\}$, assuming that $i = j$ denotes a corresponding image and label description pair, the final structured loss remains:

$$L(\Theta) = \sum_{i}^{b} max(0, max(Z_{ij}) - Z_{ii} + margin) + \alpha \left\| \Theta \right\|_2^2 \quad (8)$$

where $margin$ is a hyperparameter, in the experiments, we set $margin$ as 2, and b represents the batch size. The matching score is designed to increase if the image has same category as label description. Conversely, the matching score is designed to be reduced, if the image has a different category than the label description. This way we can guarantee that the final positive sample score is higher than the negative sample.

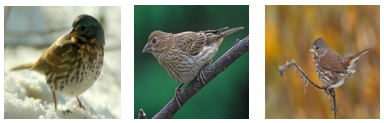
Category	Image	The Word of Match Score Top 3	Label Description
Olive sided Flycatcher		tail: 2.8098495, flanks: 0.9393288, short: 0.20994028	It is a medium-sized tyrant flycatcher. Adults are dark olive on the face, upperparts and flanks. They have light underparts, a large dark bill and a short tail. The song is a whistled quick-three beers.
White necked Raven		nape: 10.256281 purple: 4.109243 neck: 3.6913671	They have a much shorter tail than the common raven...Though predominantly black, the throat, breast and neck show a faint purple gloss. There is a large patch of white feathers on the nape of the neck.
Fox Sparrow		spot: 5.806372, streaked: 5.40088, breast: 4.902195	Adults are among the largest sparrows, heavily spotted and streaked underneath. All feature a messy central breast spot though it is less noticeable on the thick billed and slate-colored varieties.

Fig. 4. Shows the matching score between the image and the word of image label description. The higher the score, the more important the word is to the image, and the red font represents the top 3 words in the matching score. By comparing the label description with the image of the label, it can be found that the top three words obtained correspond to the key distinguishable areas of the image.

IV. Experiments

A. Datasets and Baselines

In this section, we describe our experiments on the Caltech-UCSD Birds dataset (CUB) and Stanford Dogs dataset which are widely used to evaluate fine-grained image recognition. Caltech-UCSD Birds dataset contains 11,788 images of 200 types of birds, 5,994 for training and 5,794 for testing. Every image has detailed annotations: 15 part locations, 312 binary attributes, and 1 bounding box. Stanford Dogs dataset contains 20,580 images of 120 types of dogs, 12,000 for training and 8,580 for testing. We expand the CUB-200-2011 dataset and Stanford Dogs dataset by collecting fine-grained image label descriptions for every category from Wikipedia, as shown in Fig. 3. We compare with the following baselines, due to their state-of-the-art results. All the baselines are listed as follows:

- **DeepLAC** [23]: **d**eep **l**ocalization, **a**lignment and **c**lassification proposes to use a pose-aligned part image for classification.
- **Part-RCNN** [42]: extends **R-CNN** [7] based framework by part annotations.
- **PA-CNN** [21]: **p**art **a**lignment-based method generates parts by using co-segmentation and alignment.
- **MG-CNN** [37]: **m**ultiple **g**ranularity descriptors learn multi-region of interests for all the grain levels.
- **FCAN** [25]: **f**ully **c**onvolutional **a**ttention **n**etwork adaptively selects multiple task-driven visual attention by reinforcement learning.
- **B-CNN** [24]: uses two separate feature extractors to capture pairwise feature interactions for classification.
- **SPDA-CNN** [41]: **s**emantic **p**art **d**etection and **a**bstraction proposes to generate part candidates and extract features by detection/classification networks.
- **Mask-CNN** [39]: localizing parts and selecting descriptors by learning masks.
- **PN-CNN** [1]: **p**ose **n**ormalized CNN proposes to compute local features by estimating the object's pose.
- **TLAN** [36]: **t**wo-**l**evel **a**ttention **n**etwork proposes domain-nets on both objects and parts to classification.
- **DVAN** [45]: **di**verse **a**ttention **n**etwork attends object from coarse to fine by multiple region proposals.
- **NAC** [32]: **n**eural **a**ctivation **c**onstellations find parts by computing neural activation patterns.
- **PDFR** [44]: **p**icking **d**eep **f**ilter **r**esponses proposes to find distinctive filters and learn part detectors.
- **RA-CNN** [4]: **r**ecursively learns discriminative region **a**ttention and region-based feature representation at multiple scales in a mutually reinforced way.
- **CVL** [8]: two-stream model **c**ombining **v**ision and **l**anguage for learning latent semantic representations.
- **RAN** [38]: a convolutional neural network using attention mechanism which can incorporate with state-of-art feed forward network architecture in an end-to-end training fashion.
- **MAMC** [34]: applies the **m**ulti-**a**ttention **m**ulti-**c**lass constraint in a metric learning framework, a novel attention-based convolutional neural network which regulates mul-

tiple object parts among different input images.

- **WS-DAN** [11]: proposes **w**eakly **s**upervised **d**ata **a**ugmentation **n**etwork to explore the potential of data augmentation to improve the performance of fine-grained image classification.

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON CUB-200-2011 TESTING DATASET

Methods	Train Anno.	Accuracy
DeepLAC [23]	✓	80.3
Part-RCNN [42]	✓	81.6
PA-CNN [21]	✓	82.8
MG-CNN [37]	✓	83.0
FCAN [25]	✓	84.3
B-CNN [24]	✓	85.1
SPDA-CNN [41]	✓	85.1
PN-CNN [1]	✓	85.4
Mask-CNN [39]	✓	85.4
VGG-19 [33]	×	77.8
TLAN [36]	×	77.9
NAC [32]	×	81.0
MG-CNN [37]	×	81.7
FCAN [25]	×	82.0
B-CNN(250k-dims) [24]	×	84.1
PDFR [44]	×	84.5
RA-CNN [4]	×	81.0
CVL [8]	×	85.5
MACNN [46]	×	86.5
WS-DAN [11]	×	89.4
ours	×	**90.2**

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON STANFORD DOGS TESTING DATASET

Methods	Train Anno.	Accuracy
PDFR [44]	×	72.0
VGG-19 [33]	×	76.7
ResNet-50 [7]	×	81.1
DVAN [45]	×	81.5
RAN [38]	×	83.1
FCAN [25]	×	84.2
ResNet-101 [7]	×	84.9
MAMC [34]	×	85.2
RA-CNN [4]	×	87.3
WS-DAN [11]	×	92.2
ours	×	**92.3**

TABLE III
THE EFFECT OF IMAGE SIZE ON CLASSIFICATION ACCURACY ON CUB-200-2011 DATASET

Image Size	Accuracy
224×224	90.2
448×448	93.2

B. Implement Details

In our experiments, we apply the widely used VGG-19 model with batch normalization [13] pre-trained on ImageNet as the vision encoder as the same settings with baselines for image encoder, and the model can be replaced with any CNN model. We empirically set $margin$ as 2 in Eqn. (8). We train the models using Stochastic Gradient Descent (SGD) with the momentum of 0.9, epoch number of 70, weight decay of 0.000005, and a mini-batch size of 16. The initial learning rate is set to 0.0008. For text encoder, we train the Text GCN model using Adaptive Moment Estimation (Adma) with the learning rate is 0.0008, the dropout rate is 0.5, L2 loss is 0. and we use 300 dimensional GloVe [30] word embeddings. The code will be released in the near feature.

C. Comparison with Stage-of-the-Art Methods

Table I compares our main result with prior work on CUB-200-2011 dataset. Without using the annotation information of image, the accuracy of our model is 12.4% higher than the baseline VGG-19 [33]. The latest model WS-DAN [11] in 2019 has achieved an excellent accuracy of 89.4% on the CUB-200-2011 dataset, but our model is superior to it, and the accuracy of our model is 0.8% higher than WS-DAN. Compared to CVL [8] that also incorporates textual information, the accuracy of our model is 90.2%, 4.7% higher than that of its dual stream model (image stream and text stream). Similarly, our model is more outstanding than some models that use annotation information. Our model is 4.6% more accurate than Mask-CNN [39] using annotation information. We experiment on different sizes of input images, as shown in Table III, increase the size of the input image, the classification accuracy will be improved.

Table II shows the comparison of results and baselines on Stanford Dogs dataset. On Stanford Dogs dataset, our model can still achieve good results. Compared with the WS-DAN [11] model, our model accuracy has increased by 0.1%. Compared with the basic model VGG-19 [33], the accuracy of our model has increased by 15.6%. These results and comparisons show that our model is effective.

We match the image feature with the word features to get which words in the label description are more important to the image. The higher the match score, the more important the word is. We list the words of processed label descriptions (removing stop words, removing unrelated symbols, etc.) and image matching scores, as shown in Fig. 4. The red fonts are the top 3 words in the ranking, indicating that these words are positive for fine-grained image classification. By comparing the image and label description of the same label, it can be found that the top three words obtained correspond to the key distinguishable areas of the image. For example, for the "White necked Raven" bird, we get the words "neck", "purple" and "nape", which is corresponding to the key parts or features of the image. The picture shows that the neck is covered by a white area and the breast appears a faint purple gloss, which is its key distinguishable part. This shows that we can indeed find the key parts or properties in the image by matching the features of the words in the label description with the image features. The label description feature contains the features of all the words in the label description which can induce the image feature extractor to extract distinguishable features represented in the label description.

Our model introduces more information (label description information), and this information is well expressed by GCN.

There is a relationship between text information and image information. We extract key features from the matching operation. This is the reason for our model to achieve good results.

V. Conclusion

In this paper, we present a method for fine-grained image classification combining image and label information. This model does not require any annotation information and can achieve better classification accuracy. We use the VGG-19 model pre-trained on ImageNet without fully connected layer as the feature extractor for the image to obtain image features. We build a graph with all the label descriptions, and then perform a convolution operation on this graph to get the text feature vector. We convert the classification problem into a matching problem and increase the matching scores of image features and label description features under the same label, and decrease them under different ones to match and thus classify the images. The performance of our model on the CUB-200-2011 dataset and the Stanford Dog dataset is superior to the current state-of-the-art method, confirming the validity of the model.

VI. Acknowledgments

This research work was supported by the National Natural Science Foundation of China under Grant No.61802029, the fundamental Research for the Central Universities No.500418800.

References

[1] Steve Branson, Grant Van Horn, Serge J. Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *CoRR*, abs/1406.2952, 2014.

[2] Jie Chen, Tengfei Ma, and Cao Xiao. Fastgcn: Fast learning with graph convolutional networks via importance sampling. *CoRR*, abs/1801.10247, 2018.

[3] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848.

[4] J. Fu, H. Zheng, and T. Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4476–4484, July 2017. doi: 10.1109/CVPR.2017.476.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, June 2014. doi: 10.1109/CVPR.2014.81.

[6] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *CoRR*, abs/1706.02216, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[8] Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7332–7340, 2017. doi: 10.1109/CVPR.2017.775.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

[10] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 05 2013. doi: 10.1613/jair.3994.

[11] Tao Hu and Honggang Qi. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *CoRR*, abs/1901.09891, 2019.

[12] S. Huang, Z. Xu, D. Tao, and Y. Zhang. Part-stacked cnn for fine-grained visual categorization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1173–1182, June 2016. doi: 10.1109/CVPR.2016.132.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015.

[14] J.Krause, H. Jin, J. Yang, and L. Fei-Fei. Fine-grained recognition without part annotations. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5546–5555, Los Alamitos, CA, USA, Jun 2015. IEEE Computer Society. doi: 10.1109/CVPR.2015.7299194.

[15] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014.

[16] Andrej Karpathy, Armand Joulin, and Fei Fei Li. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in Neural Information Processing Systems*, 3, 06 2014.

[17] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[18] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016.

[19] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[20] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3558–3565, June 2014. doi: 10.1109/CVPR.2014.455.

[21] Jonathan Krause, Hailin Jin, Jianchao Yang, and Fei-Fei Li. Fine-grained recognition without part annotations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5546–5555, 2015. doi: 10.1109/CVPR.2015.7299194.

[22] Hoa T. Le, Christophe Cerisara, and Alexandre Denis. Do convolutional networks need to be deep for text classification. *CoRR*, abs/1707.04108, 2017.

[23] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep LAC: deep localization, alignment and classification for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1666–1674, 2015. doi: 10.1109/CVPR.2015.7298775.

[24] Tsung-Yu Lin, Aruni Roy Chowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1449–1457, 2015. doi: 10.1109/ICCV.2015.170.

[25] Xiao Liu, Tian Xia, Jiang Wang, and Yuanqing Lin. Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition. *CoRR*, abs/1603.06765,

2016.

[26] Yuan Luo. Recurrent neural networks for classifying relations in clinical notes. *Journal of Biomedical Informatics*, 72, 07 2017. doi: 10.1016/j.jbi.2017.07.006.

[27] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[28] Welinder P.and Branson S.and Mita T.and Wah C.and Schroff F.and Belongie S.and Perona P. Caltech-ucsd birds 200. In *California Institute of Technology*, 2010.

[29] Yuxin Peng, Xin Huang, and Jinwei Qi. Cross-media shared representation by hierarchical learning with multiple deep networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 3846–3853, 2016.

[30] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[31] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 49–58, June 2016. doi: 10.1109/CVPR.2016.13.

[32] M. Simon and E. Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1143–1151, Dec 2015. doi: 10.1109/ICCV.2015.136.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[34] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, pages 834–850, 2018. doi: 10.1007/978-3-030-01270-0_49.

[35] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1556–1566, 2015.

[36] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 842–850, June 2015. doi: 10.1109/CVPR.2015.7298685.

[37] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang. Multiple granularity descriptors for fine-grained categorization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2399–2406, Dec 2015. doi: 10.1109/ICCV.2015.276.

[38] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6450–6458, 2017. doi: 10.1109/CVPR.2017.683.

[39] Xiu-Shen Wei, Chen-Wei Xie, and Jianxin Wu. Mask-cnn: Localizing parts and selecting descriptors for fine-grained image recognition. *CoRR*, abs/1605.06878, 2016.

[40] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. *CoRR*, abs/1809.05679, 2018.

[41] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1143–1152, June 2016. doi: 10.1109/CVPR.2016.129.

[42] Ning Zhang, Jeff Donahue, Ross B. Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 834–849, 2014. doi: 10.1007/978-3-319-10590-1_54.

[43] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657, 2015.

[44] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1134–1142, 2016. doi: 10.1109/CVPR.2016.128.

[45] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, and Shuicheng Yan. Diversified visual attention networks for fine-grained object classification. *CoRR*, abs/1606.08572, 2016.

[46] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5219–5227, 2017. doi: 10.1109/ICCV.2017.557.

[47] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, Los Alamitos, CA, USA, dec 2015. IEEE Computer Society. doi: 10.1109/ICCV.2015.11.