Fine-grained Image Classification via Combining Vision and Language

Xiangteng He and Yuxin Peng*
Institute of Computer Science and Technology, Peking University
hexiangteng@pku.edu.cn, pengyuxin@pku.edu.cn

Abstract

*Fine-grained image classification is a challenging task due to the **large** intra-class variance and **small** inter-class variance, aiming at recognizing hundreds of sub-categories belonging to the same basic-level category. Most existing fine-grained image classification methods generally learn part detection models to obtain the semantic parts for better classification accuracy. Despite achieving promising results, these methods mainly have two limitations: (1) not all the parts which obtained through the part detection models are beneficial and indispensable for classification, and (2) fine-grained image classification requires more detailed visual descriptions which could not be provided by the part locations or attribute annotations. For addressing the above two limitations, this paper proposes the two-stream model **combining vision and language (CVL)** for learning latent semantic representations. **The vision stream** learns deep representations from the original visual information via deep convolutional neural network. **The language stream** utilizes the natural language descriptions which could point out the discriminative parts or characteristics for each image, and provides a flexible and compact way of encoding the salient visual aspects for distinguishing sub-categories. Since the two streams are complementary, combining the two streams can further achieves better classification accuracy. Comparing with 12 state-of-the-art methods on the widely used CUB-200-2011 dataset for fine-grained image classification, the experimental results demonstrate our CVL approach achieves the best performance.*

1. Introduction

Fine-grained image classification aims to recognize sub-categories under some basic-level categories. Models of fine-grained image classification have made great progress in recent years[1, 2, 3, 4, 5], due to the progress of deep neural networks. And on the data side, more fine-grained
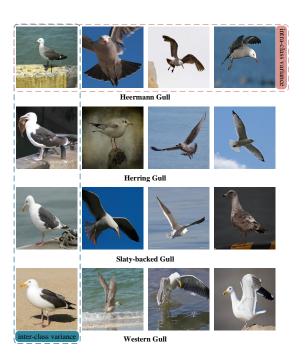
*Corresponding author.

Figure 1. Examples from CUB-200-2011. Note that it is a technically challenging task even for humans to categorize them due to large intra-class variance and small inter-class variance.

domains have been covered, such as bird types [6, 7], dog species [8], plant breeds [9] , car types [10] and aircraft models [11].It is easy for an inexperienced person to recognize basic-level categories such as bird, flower and car, but highly hard to recognize 200 or even more sub-categories. Consequently, fine-grained image classification is a technically challenging task, due to the large intra-class variance and small inter-class variance, as shown in Figure 1.

The sub-categories are generally same in global appearance, and distinguished by the subtle and local differences, such as the color of abdomen, the shape of toe and the texture of feather for bird. These subtle differences are located at the regions of object or its parts, so the localization of object and its parts is crucial for fine-grained image classification. A two-stage learning framework is adopted by most of

the existing methods: the first stage is to localize the object or its discriminative parts, and the second stage is to extract the deep features of the object or its parts through Convolutional Neural Network (CNN) and train a classifier for the final prediction. However, these methods mainly have two limitations. First, parts are crucial for achieving better accuracy, but not all of them are beneficial and indispensable. Huang et al. [1] show that the classification accuracy declines when the number of parts increases from 8 to 15 in the experiments of their Part-stacked CNN method. Zhang et al. [5] pick only 6 parts in the experiment for achieving the best classification accuracy. And Zhang et al. [3] only use the head and body parts for classification in Part-based R-CNN. Generally speaking, the number of the parts used in the experiments is highly empirical. This is highly limited in flexibility, and difficult for generalizing to the other datasets or domains. Second, fine-grained image classification requires more detailed visual descriptions which could not be provided by the part locations or attribute annotations. The part locations could not point out which part is the discriminative parts for sub-categories recognition and tell the discriminative features, such as the color of bill and the shape of wing. The attribute annotations may tell us the color of bill, but they do not tell us whether the bill is significantly important for distinguishing sub-categories than other parts. Hence, we need a kind of promising information to tell us the attributes as well as the importances of the parts, and further facilitate the classification accuracy. Fortunately, text descriptions from natural language satisfies the above promising requirements.

How to exactly relate the natural language descriptions to the visual content of images is the key of image classification. Inspired by the progress of the cross-modal analysis which reduces the multi-modal representation gap between visual information and natural language descriptions, this paper proposes a two-stream model combining vision and language (CVL) for learning latent semantic representations. The vision stream first localizes the object of image via saliency extraction and co-segmentation, and then learns deep representations of the original image and its discriminative object via deep convolutional neural network. The language stream utilizes the cross-modal analysis to learn the correlation between the natural language descriptions and the discriminative parts, and provides a flexible and compact way of encoding the salient visual aspects for distinguishing sub-categories. Vision stream focuses on the locations of the discriminative regions, while language stream focuses on the attributes of the discriminative regions. They are complementary, combining the two streams further exploit the correlation between visual feature and nature language descriptions, and enhances their mutual promotion to achieve better classification accuracy. Comparing with 12 state-of-the-art methods on the widely used CUB-200-2011

dataset for fine-grained image classification, the experimental results demonstrate our CVL approach achieves the best performance.

The rest of this paper is organized as follows: Section II briefly reviews related works on fine-grained image classification and cross-modal analysis. Section III presents our proposed CVL approach, and Section IV introduces the experiments as well as the results analyses. Finally Section V concludes this paper.

2. Related Work

2.1. Fine-grained Image Classification

Most existing works follow the pipeline: first localizing the object or its parts, and then extracting discriminative features for fine-grained image classification. Some fine-grained image classification datasets, e.g. CUB-200-2011 [7], have the detailed annotations: object annotation (i.e. bounding box of object) and parts annotations (i.e. parts locations), an intuitive idea is that using these annotations for the localizations of object and its parts. Object annotation is used in the works of [12, 13] to learn part detectors in an unsupervised or latent manner. And even part annotations are used in these methods [14, 15]. Since the annotations of the testing image are not available in practical applications, some researchers use the object or part annotations only at training stage and no knowledges of any annotations at testing stage. Object and Part annotations are directly used in training phase to learn a strongly supervised deformable part-based model [16] or directly used to fine-tune the pre-trained CNN model [17]. Further more, Krause et al. [18] only use object annotation at training stage to learn the part detectors, then localize the parts automatically in the testing stage. Recently, there are some promising works attempting to learn the part detectors under the weakly supervised condition, which means that neither object nor part annotations are used in both training and testing phase. These works make it possible to put the fine-grained image classification into practical applications. Simon et al. [19] propose a neural activation constellations part model (NAC) to localize parts with constellation model. Xiao et al. [2] propose a two-level attention model, which combines two level attentions to select relevant proposals to the object and the discriminative parts. And Zhang et al. [5] incorporate deep convolutional filters for both part detection and description. The problem of fine-grained image classification is still far from solved.

2.2. Cross-modal analysis

With the rapid growth of multimedia information, the cross-modal data, e.g. image, text, video and audio, has been the main form of the big data. Cross-modal data carries different kinds of information, which needs to be in-
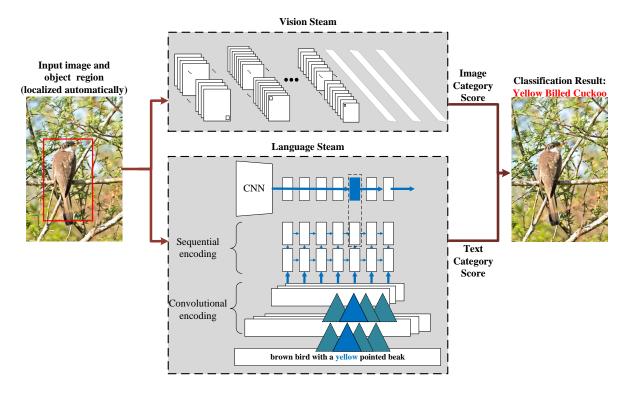
Figure 2. Overview of our CVL approach. The two-stream model conducts on the original images and their object localizations. One learns the deep representations directly from the vision information. The other learns the salient visual aspects for distinguishing sub-categories via jointly modeling vision and language. The classification results of the two streams are merged in later phase to combine the advantages of vision and language.

tegrated to get comprehensive results in many real-world applications. How to learn multi-modal representation for cross-modal data is a fundamental research problem. A traditional representation method is the canonical correlation analysis (CCA) [20], which learns a subspace to maximize the correlation among data of different media types, and is widely used for modeling multi-modal data [21, 22, 23]. Zhai et al. [24] propose to learn projection functions by the metric learning, and this method is further improved as Joint Representation Learning (JRL) [25] by adding other information such as semantic categories and semi-supervised information. Inspired by the progress of deep neural networks, some works have been focused on deep multi-modal representation learning. Ngiam et al. [26] propose a multimodal deep learning (MDL) method to combine the audio and video into an autoencoder, which improves the speech signal classification for noisy inputs as well as learns a shared representation across modalities. Recently, a surge of progress has been made in image and video captioning. LSTMs [27] are widely used in modeling captions at word level. Besides LSTMs, character-based convolutional networks [28] have been used for language modeling. In this paper, we apply the extension of Convolutional and Recurrent Networks (CNN-RNN) to learn a visual semantic em-

bedding. In this paper, we bring the multi-modal representation learning into fine-grained image classification to boost the performance, and jointly modeling vision and language.

3. Our CVL Approach

Our method is based on a very simple intuition: natural language descriptions could point out the discriminative parts or characteristics from other sub-categories, and are complementary with vision information. Therefore, we propose a two-stream model combining vision and language for learning latent semantic representations, which takes the advantages of vision and language jointly, as shown in Figure. 2. Since the object is crucial for fine-grained image classification, we take the original images and their object localizations as the inputs of the two-stream model.

3.1. Object localization

In this paper, we apply an automatic object localization method based on saliency extraction and co-segmentation proposed in TSC [4], which allows to localize the object in a weakly-supervised manner that means neither object nor part annotations are used. Saliency extraction is to localize the object preliminarily with the saliency map generated by
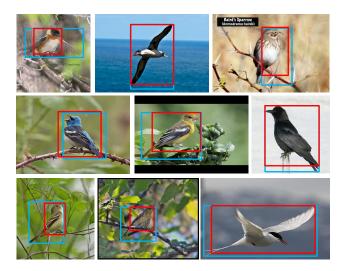
Figure 3. Sample object localization results in this paper. The blue rectangles indicate the ground truth object annotations, i.e. bounding boxes of objects, and the red rectangles indicate the object regions generated by jointly applying saliency extraction and co-segmentation.

the CNN model. However, only through saliency extraction, the object region is not accurate enough so that co-segmentation is conducted to make the object region more accurate for fine-grained image classification. The sample object localization results are shown in Figure 3.

3.2. Jointly Model Vision and Language

Considering that the two different descriptions of an image are complementary, i.e. visual information and natural language descriptions, we jointly model the two different forms of descriptions to learn deep representations for better classification accuracy.

3.2.1 Vision Stream

A natural candidate for the visual classification function f is a CNN model, which is consist of a hierarchy of convolutional and fully connected layers. We can benefit from model pre-training due to the additional training data. This has been proved by a large amount of recognition tasks, such as object detection, texture recognition and fine-grained image classification [29, 30, 31, 32] etc. Therefore, we use a CNN model pre-trained on the ImageNet dataset [33] as the base model in our experiments. And then, we fine-tune the pre-trained CNN model on the fine-grained dataset.

Given an image I, its object region b is generated at object localization stage, then the object region is clipped from the original image and saved as image I'. We take the original image I and its object image I' as the inputs of the CNN model to obtain the prediction, which is the result of the vision stream.

3.2.2 Language Stream

Deep Structured Joint Embedding We apply the deep structured joint embedding method [34], because it can jointly embedding images and fine-grained visual descriptions (i.e. natural language descriptions for images). This method learns a compatibility function of image and text, which can be seen as an extension of the multimodal structured jointed embedding [35]. Instead of using a bilinear compatibility function, we use the inner product of features generated by deep neural encoders, and maximize the compatibility between a description and its matching image as well as minimize compatibility with images from other classes.

Given data $D = (v_n, t_n, y_n), n = 1, ..., N$, in which $v \in V$ indicates visual information, $t \in T$ indicates text description and $y \in Y$ indicates the class label, then the image and text classifier functions $f_v : V \to Y$ and $f_t : T \to Y$ are learned by minimizing the empirical risk:

$$\frac{1}{N} \sum_{n=1}^{N} \Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n)) \tag{1}$$

where $\Delta : y \times y \to \mathbb{R}$ is the 0-1 loss and

$$f_v(v) = arg \max_{y \in Y} \mathbb{E}_{t \sim T(y)}[F(v, t)] \tag{2}$$

$$f_t(t) = arg \max_{y \in Y} \mathbb{E}_{v \sim V(y)}[F(v, t)] \tag{3}$$

We then define the compatibility function $F : V \times Y \to \mathbb{R}$ that uses features from the learnable encoder functions $\theta(v)$ for images and $\phi(t)$ for texts:

$$F(v, t) = \theta(v)^T \phi(t) \tag{4}$$

We apply the GoogleNet [36] as the image encoder model, and Convolutional Recurrent Net (CNN-RNN) [34] as the text encoder model which will be discussed in the next paragraph.

Text encoder model We apply the CNN-RNN [34] for learning the fine-grained visual descriptions. A mid-level temporal CNN hidden layer is at the bottom of CNN-RNN model, and a recurrent network is stacked on it. We extract the average hidden unit activation over the sequence as the text feature, as shown in equation 5. The resulting scoring function is defined as a linear accumulation of evidence for compatibility with the image which needs to be recognized.

$$\phi(t) = \frac{1}{L} \sum_{i=1}^{L} h_i \tag{5}$$

where h_i indicates the hidden activation vector for the i-th frame and L indicates the sequence length.

Category	Vision	Language	Vision	Language
Heermann Gull		(1)A large bird with different shades of grey all over its body, white and black tail feathers, and a long sharp orange beak. (2)This bird is gray and black in color, with a orange beak. (3)This bird has black outer retices and white inner retires and an orange beak. ...		(1)This bird has a grey body, a white head with an orange bill and black wings, tarsus and feet. (2)A white bird with black wings and orange beak and eyes. (3)This white bird has a bright orange bill with a black tip. ...
Red Legged Kittiwake		(1)This bird has a white head, breast and belly with gray wings, red feet and thighs, and a red beak. (2)This is a white bird with gray wings, red webbed feet and a red beak. (3)Long bird with an orange beak and white feathers with grey colored wings. ...		(1)This white bird has grey wings and tail, with orange feet and tarsus and a pointed yellow bill. (2)This bird has a mellow yellow bill coloration and deep red feet (3)The bird has a yellow beak with a white head and orange web feet ...
Bohemian Waxwing		(1)This bird is light gray with a light orange patch on its under-tail covets, neck and crown, and a black malar stripe and nape. (2)This is a grey bird with a red and yellow tail and a red face. (3)This bird has wings that are gray and black and has a red crown ...		(1)This colorful bird has an orange crown, black eyebrows and secondaries with the primaries being rimmed in yellow. (2)This bird is grey with red and has a long, pointy beak. (3)The bird has a tan spiked crown and short bill. ...

Figure 4. Sample natural language descriptions of CUB-200-2011.

3.3. Final Prediction

Given an Image I, its object bounding region is obtained automatically through localization method. The two-stream model conducts on the original images and their object localizations. The vision stream gives the prediction from the view of the image only, while the language stream gives the prediction via measuring the image and text description with the shared compatibility function. Finally, we fuse the prediction results of the two streams to utilize the advantages of the two via the follow equation:

$$f(I) = f_v(v) + \beta * f_t(t) \tag{6}$$

where $f_v(v)$ and $f_t(t)$ are the image and text classifier functions as mentioned above, and β is selected by the cross-validation method. In the experiments, we set β as 3.

4. Experiments

This section presents the evaluations and analyses of our CVL approach on the challenging fine-grained image classification benchmark CUB-200-2011 [7]. It 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. Scott Reed et al. [34] expand the CUB-200-2011 dataset by collecting fine-grained visual descriptions. Ten single-sentence visual descriptions are collected for each image, as shown in Figure 4. The fine-grained visual descriptions are collected through the Amazon Mechanical Turk (AMT) platform, and are at least 10 words, without any information of species, background and actions.

4.1. Implementation Details

Vision Stream In our experiments, we apply the widely used model of VGGNet [42] as the vision stream model. The reason of choosing VGGNet is for fair comparison with state-of-the-art methods. It is important to note that the model used in our proposed method can be replaced with any CNN model. The model is pre-trained on ImageNet dataset, and then fine-tuned on the CUB-200-2011 dataset. In the fine-tuning step, we follow the strategy of TL Atten [2]. First, we apply the selective search [43] to generate patches for each image. Then the pre-trained CNN model on ImageNet dataset is used as a filter net for selecting the patches relevant to the object. With the selected patches, we fine-tune the pre-trained model.

Language Stream In our experiments, we apply the GoogleNet [36] with batch normalization [44] as image encoder and CNN-RNN [34] as text encoder. For image encoder, we take the strategy used in vision stream for better accuracy. And for text encoder, the CNN input size (sequence length) is set to 201 for character-level model. We keep the image encoder fixed, and used RMSprop with base learning rate 0.0007 and minibatch size 40. All the configurations and source code [1] used for training and testing follow the work of Scott Reed et al. [34].

4.2. Comparisons with state-of-the-art methods

For comparison purpose, we adopt 12 state-of-the-art fine-grained image classification methods. Table 1 shows the comparison results on CUB-200-2011. Bounding box

---

[1] https://github.com/reedscot/cvpr2016

Method	Train Annotation		Test Annotation		Accuracy (%)
	Bbox	Parts	Bbox	Parts	
Our CVL approach					**85.55**
PD [5]					84.54
Spatial Transformer [37]					84.10
Bilinear-CNN [38]					84.10
NAC [19]					81.01
TL Atten [2]					77.90
VGG-BGLm [39]					75.90
PG Alignment [18]	√		√		82.80
Triplet-A (64) [40]	√		√		80.70
VGG-BGLm [39]	√		√		80.40
Part-based R-CNN [3]	√	√			73.50
SPDA-CNN [41]	√	√	√		85.14
Part-based R-CNN [3]	√	√	√	√	76.37
POOF [14]	√	√	√	√	73.30
GPP [15]	√	√	√	√	66.35

Table 1. Comparisons with state-of-the-art methods on CUB-200-2011, sorted by amount of annotation used. "Our CVL" indicates our full method combining vision and language. "Bbox" indicates the object annotation (i.e. bounding box of object) provided by the dataset, and "Parts" indicates the parts annotations (i.e. parts locations). "√" indicates that one of bounding box and part locations is used in training or testing stage. Since the exact amount of annotation used varies from method to method, we defer to the original sources for details.

and part annotations used in the methods are listed for fair comparison. Early works [14, 15] choose SIFT [45] as features, and the performance is limited. When applying CNN model, our CVL approach is the best. In our experiments, both object and part annotations are not used, due to labeling is heavily labor consuming. Compared with the methods [37, 38, 19, 2, 5] which do not use object and part annotations, our CVL approach obtains a 1.01% higher accuracy than the best performing result of PD [5]. Moreover, our CVL approach outperforms methods which use object annotation [18] (82.50%) or even part annotations [41, 46] (85.14%, 76.37%). It proves the effectiveness of our CVL approach, which jointly integrates the vision and language streams to exploit the correlation between visual feature and nature language descriptions and enhance their complementarity.

4.3. Performances of components in our CVL approach

4.3.1 Effectivenesses of vision stream and language stream

We perform detailed analyses by comparing different variants of our CVL approach. "Language-stream" refers to the classification result of the language stream, "Vision-stream" refers to the classification result of the vision stream, "CVL" refers to our CVL approach combining vision and language, and "Original" refers to the classification result that only use the original image for prediction. From Table 2, we can observe that:

- Two-stream model combining vision and language

Method	Accuracy (%)
our CVL approach (Language-stream+Vision-stream)	**85.55**
Language-stream	81.81
Vision-stream	82.98
Original	76.17

Table 2. Effects of different variants of our method on CUB-200-2011. "Language-stream" refers to the classification result of the language stream, "Vision-stream" refers to the classification result of the vision stream, "CVL" refers to our CVL approach combining vision and language, and "Original" refers to the classification result that only use the original image for prediction.

boosts the performance significantly. CVL brings about a nearly 10% (76.17% → 85.55%) improvement compared with the "Original".

- The classification result of the language stream is promising. From the first line of each row in Figure 5, we can find that the text description with the highest score always points out the discriminative parts or characteristics. As shown in Figure 5 , the red words are the important visual descriptions for distinguishing sub-categories, and the blue ones are the visual descriptions of the easily confused sub-categories.

- Combining vision and language can achieve more accurate result than only one stream (85.55% vs. 81.81% and 82.98%), which demonstrates that visual information and text descriptions are complementary in fine-

Category	Image	Text Rank List(Top3)
Sooty Albatross		(1)This bird has **wings** that are **grey** and has a **black bill**. (2)This bird is **gray** in color, with a large curved beak. (3)This bird is white and brown in color, and has a **black beak**.
California Gull		(1)This bird has large feet, a short **yellow bill**, and a **black and white body**. (2)This bird has wings that are grey and has a **white belly** and **yellow bill**. (3)This bird has a **yellow beak** as well as a **white belly.**
Cerulean Warbler		(1)A little bird with a short, **grey bill**, **blue crown**, **nape**, **white breast.** (2)The bird has a **white** abdomen, black breast and **white** throat, **blue** specks. (3)This bird is **blue** and **white** in color with a black beak, and black eye rings.

Figure 5. Some results of the language stream. The red words are the important visual descriptions for distinguishing sub-categories, and the blue ones are the visual descriptions of the easily confused sub-categories.

grained image classification. The two streams have the different but complementary focuses. (1) The vision stream localizes the object region of image and extracts the visual features from the original pixels through the CNN model, which focuses on the location of the discriminative region and the texture, color or even the semantic parts we called. However, we do not know or learn correctly which parts or features are the most discriminative representations from other sub-categories. (2) The language stream learns correlation between the nature language descriptions and the visual features to exploit the attributes of the discriminative regions for distinguishing sub-categories. The natural language descriptions directly point out key parts or features distinguished from other sub-categories, e.g. Cerulean Warbler has blue crown, back and white breast, while Sooty Albatross has grey wings and black bill.

4.3.2 Effect of fine-tuning and object localization

There are two differences from the work [34]: (1) instead of directly using the GoogleNet, we first fine-tune it on the CUB-200-2011, and (2) extract the features of the original image and its object region for each image. We find that both (1) and (2) are important for not only fine-grained image classification but also zero-shot recognition, as shown in Table 3 and Table 4 respectively. It also proves the effectiveness of object localization in vision stream of our CVL approach, which focuses on the discriminative region of the image and eliminate the side effect of the background noise.

5. Conclusions

In this paper, the CVL approach has been proposed, which jointly models vision and language for learning latent

Method	Accuracy (%)
Language+ft+box	**81.81**
Language+ft	77.80
Language	50.54

Table 3. Effect of fine-tuning and object localization for fine-grained image classification. "ft" indicates fine-tuning is applied, and "box" indicates object localization is applied.

Method	Top-1 Accuracy (%)
DS-SJE+ft+box	**65.1**
DS-SJE+ft	60.0
DS-SJE [34]	54.0

Table 4. Effect of fine-tuning and object localization for zero-shot recognition. "ft" indicates fine-tuning is applied, and "box" indicates object localization is applied.

semantic representations. The vision stream learns deep representations from the original visual information via deep convolutional neural network. The language stream utilizes the natural language descriptions which could point out the discriminative parts or characteristics for each image, and provides a flexible and compact way of encoding the salient visual aspects for distinguishing sub-categories. Since the two streams are complementary, combining the two streams can further achieves better classification accuracy. Experimental results on CUB-200-2011 dataset demonstrate the superiority of our method compared with state-of-the-art methods. The results are promising, and point out a few future directions. First, combining vision and language can boosts the classification accuracy, but the two streams are trained respectively, we will focus on the work of training the two streams end-to-end. Second, from

Table 3 we can find that small improvement on the original language stream boosts the performance a lot. And nowadays there are a lot of works focusing on how to relate images to natural language descriptions. So improving the performance of the language stream will be significantly helpful to the fine-grained image classification.

6. Acknowledgments

References

[1] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. *arXiv preprint arXiv:1512.08086*, 2015. 1, 2

[2] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015. 1, 2, 5, 6

[3] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014. 1, 2, 6

[4] Xiangteng He and Yuxin Peng. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In *AAAI Conference on Artificial Intelligence*, pages 4075–4081, 2017. 1, 3

[5] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1134–1142, 2016. 1, 2, 6

[6] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2019–2026. IEEE, 2014. 1

[7] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 2, 5

[8] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, 2011. 1

[9] Anelia Angelova and Shenghuo Zhu. Efficient object detection and segmentation for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 811–818, 2013. 1

[10] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 1

[11] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[12] Yuning Chai, Victor Lempitsky, and Andrew Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 321–328, 2013. 2

[13] Shulin Yang, Liefeng Bo, Jue Wang, and Linda G Shapiro. Unsupervised template learning for fine-grained object recognition. In *Advances in Neural Information Processing Systems*, pages 3122–3130, 2012. 2

[14] Thomas Berg and Peter Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 955–962, 2013. 2, 6

[15] Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan, and Bo Zhang. Hierarchical part matching for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1641–1648, 2013. 2, 6

[16] Ning Zhang, Ryan Farrell, Forrest Iandola, and Trevor Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *2013 IEEE International Conference on Computer Vision*, pages 729–736. IEEE, 2013. 2

[17] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014. 2

[18] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5546–5555, 2015. 2, 6

[19] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1143–1151, 2015. 2, 6

[20] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. 3

[21] Hervé Bredin and Gérard Chollet. Audio-visual speech synchrony measure for talking-face identity verification. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pages II–233. IEEE, 2007. 3

[22] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004. 3

[23] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015. 3

[24] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In *AAAI*, 2013. 3

[25] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6):965–978, 2014. 3

[26] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011. 3

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[28] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657, 2015. 3

[29] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 4

[30] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, pages 647–655, 2014. 4

[31] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 4

[32] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014. 4

[33] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 4

[34] Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee. Learning deep representations of fine-grained visual descriptions. *arXiv preprint arXiv:1605.05395*, 2016. 4, 5, 7

[35] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015. 4

[36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 4, 5

[37] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 6

[38] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015. 6

[39] Feng Zhou and Yuanqing Lin. Fine-grained image classification by exploring bipartite-graph labels. *arXiv preprint arXiv:1512.02665*, 2015. 6

[40] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. *arXiv preprint arXiv:1512.05227*, 2015. 6

[41] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. pages 1143–1152, 2016. 6

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[43] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 5

[44] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 5

[45] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 6

[46] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014. 6