EasyChair Preprint

№ 4401

Domain Adaptive Transfer Learning on Visual Attention Aware Data Augmentation for Fine-grained Visual Categorization

Ashiq Imran and Vassilis Athitsos

October 14, 2020

Domain Adaptive Transfer Learning on Visual Attention Aware Data Augmentation for Fine-grained Visual Categorization

No Author Given

No Institute Given

Abstract. Fine-Grained Visual Categorization (FGVC) is a challenging topic in computer vision. It is a problem characterized by large intra-class differences and subtle inter-class differences. In this paper, we tackle this problem in a weakly supervised manner, where neural network models are getting fed with additional data using a data augmentation technique through a visual attention mechanism. We perform domain adaptive knowledge transfer via fine-tuning on our base network model. We perform our experiment on six challenging and commonly used FGVC datasets, and we show competitive improvement on accuracies by using attention-aware data augmentation techniques with features derived from deep learning model InceptionV3, pre-trained on large scale datasets. Our method outperforms competitor methods on multiple FGVC datasets and showed competitive results on other datasets. Experimental studies show that transfer learning from large scale datasets can be utilized effectively with visual attention based data augmentation, which can obtain state-of-the-art results on several FGVC datasets. We present a comprehensive analysis of our experiments. Our method achieves state-of-the-art results in multiple fine-grained classification datasets including challenging CUB200-2011 bird, Flowers-102, and FGVC-Aircrafts datasets.

Keywords: Domain Adaptation · Transfer Learning · Fine-Grained Visual Categorization · Visual Attention.

1 Introduction

Deep neural networks have provided state-of-the-art results in many domains in computer vision. However, having a big training set is very important for the performance of deep neural networks [7, 15]. Data augmentation techniques have been gaining popularity in deep learning and are extensively used to address the scarcity of training data. Data augmentation has led to promising results in various computer vision tasks [15]. There are different data augmentation methods for deep models, like image flipping, cropping, scaling, rotation, translation, color distortion, adding Gaussian noise, and many more. Previous works mostly choose random images from the dataset and apply the above operations to enlarge the amount of training data. However, applying random cropping to generate new

training examples can have undesirable consequences. For example, if the size of the cropped region is not large enough, it may consist entirely of background, and not contain any part of the labeled object. Moreover, this generated data might reduce accuracy and negatively affect the quality of the extracted features. Consequently, the disadvantages of random cropping might cancel out its advantages. More specific features need to be provided to the model to make data augmentation more productive.

In Fine-Grained Visual Categorization (FGVC), same-class items may have variation in the pose, scale, or rotation. FGVC contains subtle differences among classes in a sub-category of an object, which includes the model of the cars, type of the foods or the flowers, species of the birds or dogs, and type of the aircrafts. These differences are what make FGVC a challenging problem, as there are significant intra-class differences among the sub-categories, and at the same time, items from different classes may look similar. In contrast with regular object classification techniques, FGVC aims to solve the identification of particular subcategories from a given category [10, 12].

Convolutional Neural Networks (CNNs) have been extensively used for various applications in computer vision. To achieve good performance with CNNs, typically, we need large amounts of labeled data. However, it is a tedious process to collect labeled fine-grained datasets. That is why there are not many FGVC datasets, and existing datasets are not as large compared to standard image recognition datasets like ImageNet [7]. Normally, a model pre-trained on large scale datasets such as ImageNet is used, then that is fine-tuned using data from an FGVC dataset. Typically, FGVC datasets are not too big. Without being big in size and with a limited number of large scale FGVC datasets, it becomes critical to design methods that can compensate for the limited amount of data. In this paper, we investigate some techniques that allow the model to learn features more effectively, and that perform well on large scale datasets with fine-grained categories. Generally, there are two kinds of the domain involved in fine-tuning a network. One is the source domain, which typically includes large scale image datasets like ImageNet [7], where initial models are pre-trained. Another is the target domain, where data is used to fine-tune the pre-trained models. In this paper, the target domain is FGVC datasets, and we are interested in developing techniques that can boost accuracy on these type of datasets. Modern FGVC methods use pre-trained networks with ImageNet dataset to a large extent. We explore the possibility of achieving better accuracy than what has been achieved so far using ImageNet. A model first learns useful features from a large amount of training data, and is then fine-tuned on a more evenly-distributed subset to balance the efforts of the network among different categories and transfer the already learned features. In short, our research tries to find out the answer of two questions: 1) What approaches beyond transfer learning do we need to take to boost the performance on FGVC datasets? 2) How can we determine which large scale dataset for transfer learning we choose, given that the target domain is FGVC?

We calculate the domain similarity score between the source and target domains. This score gives us a clear picture of selecting the source domain for transfer learning to achieve better accuracy in the target domain. Then, we focus on a visual attention guided network for data augmentation. Typically, FGVC datasets are relatively smaller in size, we leverage the feature learning from fine-tuning as well as data augmentation to achieve better accuracy. The performance of the combination of these two strategies outperforms the baseline approach.

In summary, the main contributions of this work are:

1. We propose a simple yet effective improvement over recently proposed network WS-DAN [12], which is used for generating attention maps to extract sequential local features to tackle the FGVC challenge. A domain similarity score can play a vital role before applying transfer learning. Based on the score, we decide which source domain is necessary to use for transfer learning. Then, we can employ WS-DAN [12] to achieve better results among FGVC datasets.
2. We demonstrate a domain adaptive transfer learning approach combining with visual attention based data augmentation can achieve state-of-the-art results on CUB200-2011 [28], and Flowers-102 [20], and FGVC-Aircrafts [19] datasets. Additionally, we match the current state-of-the-art accuracy on Stanford Cars [14], Stanford Dogs [13] datasets.
3. We present the relationship of top-1 accuracy and domain score on six commonly used FGVC datasets. We illustrate the effect of image resolution in transfer learning in detail.

2 Related Work

In this section, we present a brief overview of data augmentation, fine-grained visual categorization, visual attention mechanism and transfer learning.

2.1 Data Augmentation

Machine learning theory suggests that a model can be more generalized and robust if it has been trained on a dataset with higher diversity. However, it is a very difficult and time-consuming task to collect and label all the images which involve these variations [33]. Data augmentation methods are proposed to address this issue by adding the amount and diversity of training samples. Various methods have been proposed focusing on random spatial image augmentation, specifically involving in rotation variation, scale variation, translation, and deformation, etc. [12]. Classical augmentation methods are widely adopted in deep learning techniques. The main drawback of random data augmentation is low model accuracy. Additionally, it suffers from generating a lot of unavoidable noisy data. Various methods have been proposed to consider data distribution rather than random data augmentation. A search space based data augmentation method has been proposed [5]. It can automatically search for improving

data augmentation policies in order to obtain better validation accuracy. In contrast, we leverage WS-DAN [12], which generates augmented data from visual attention features of the image.

Peng *et al.* proposed a method for human pose estimation, by introducing a complicated augmentation network whose task is to generate hard data online, and thus improving the robustness of models [21]. Nevertheless, their augmentation system is complicated and less accurate compared to the network that we experimented with. Additionally, attention-aware data segmentation is much simpler and proven effective in terms of accuracy.

2.2 Fine-Grained Visual Categorization

Fine-grained Visual Categorization (FGVC) is a challenging problem in the field of computer vision. Normally, object classification is used for categorize different objects in the image. In contrast with typical object classification which concentrates to find correct labels such as a humans, objects or animals, fine-grained image classification concentrates more on detecting sub-categories of a given category like various types of bird, dogs or cars. The purpose of FGVC is to find subtle differences among various categories of a dataset. It presents significant challenges for building a model that generalizes patterns. This problem provides insights to a wide range of applications such as image captioning [2], image generation [4], image search engines, and so on. Various methods have been developed to differentiate fine-grained categories. Due to the remarkable success of deep learning, most of the recognition works depend on the powerful convolutional deep features. Several methods were proposed to solve large scale real problems [24, 11, 26]. However, it is relatively hard for the basic models to focus on very precise differences of object's parts without adding special modules [12]. A weakly supervised learning based approach was adapted to generate class-specific location maps by using pooling methods [18]. Adversarial Complementary Learning (ACoL) [34] is a weakly supervised approach to identify entire objects by training two adversarial complementary classifiers, which aims at locating several parts of objects and detects complementary regions of the same object. However, their method fails to accurately locate the parts of the objects due to having only two complementary regions. On the contrary, our proposed approach depends on attention-guided data augmentation and domain adaptive transfer learning. Our method extracts fine-grained discriminative features and provides a generalization of domain features to achieve state-of-the-art performance in terms of accuracy.

2.3 Attention

Attention mechanisms have been getting a lot of popularity in the deep learning area. Visual attention has been already used for FGVC. Xiao *et al.* proposed two-way attention (one is for finding the object-level attention and another is for finding the part-level attention) based method to train domain-specific deep networks [30]. Fu *et al.* proposed an approach that can predict the location of

one attention area and extract corresponding features [9]. However, this method can only focus on local object's part at the same time. Zheng *et al.* addressed this issue and introduced Multi-Attention CNN (MA-CNN) [35], which can simultaneously focus on multiple body parts. However, selected parts of the object are limited and number of selected parts is fixed (2 or 4) which might hamper accuracy because of limited number of object's parts. The works mentioned above mostly focus on object localization. In contrast, our research concentrates more on data augmentation with visual attention, which has not been much explored. We use the attention mechanism for data augmentation purposes. Moreover, the benefit of guided attention based data augmentation [12] helps the network to locate object precisely which makes our trained model learn about closer object details and hence, improve the predictions.

2.4 Transfer Learning

Transfer learning is a machine learning technique where a model trained on one task is re-trained on a second related task. The purpose of transfer learning is to improve the performance of a learning algorithm by utilizing knowledge that is acquired from previously solved similar problems. CNNs have been widely used for transfer learning. They are mostly used in the form of pre-trained networks that serve as feature extractors [23, 8]. Considerable amounts of effort have been made to understand transfer learning [31, 25, 3]. Initial weights for a certain network can be obtained from an already-trained network even if the network is used for different tasks [31]. Some prior work has shown some results on transfer learning and domain similarity [6]. Their contribution mostly addresses the effect of image resolution on large scale datasets and choosing different subsets of datasets to boost accuracy. In our work, we show domain adaptive transfer learning can be very useful if we incorporate visual attention based data augmentation with it.

Unlike previous works, our proposed technique takes account of domain adaptive transfer learning between the source and target domains. Then, it incorporates the attention-driven approach for data augmentation. Our main goal is to guide the training model to learn relevant features from the source domain and augment data with the visual attention of the target domain. The combination of two processes can be useful to achieve better performance.

3 Domain Adaptive Transfer Learning (DATL)

In our research, we consider different types of large scale datasets to find out the similarity with FGVC datasets. We compute domain similarity score initially. Based on the domain similarity score we choose large scale datasets for transfer learning and then we perform WS-DAN [12] to train and evaluate the accuracy.

3.1 Domain Similarity

Generally, transfer learning performs better if it has been trained on bigger datasets. Chen *et al.* showed that transfer learning performance increases logarithmically with the number of data [25]. In our work, we observe that using a bigger dataset does not always provide a more accurate result. Yosinski *et al.* [31] mentions that there is some correlation between the transferability of a network from the source task to the target task and the distance between the source and target tasks. Furthermore, they show fine-tuning on a pre-trained network towards a target task can boost performance. For measuring domain similarity, we use the approach of Cui *et al.* [6] who introduce a method which can calculate domain similarity by the Earth Mover's Distance (EMD) [22]. Furthermore, they show transfer learning can be treated as moving image sets from the source domain S to the target domain T. The domain similarity [6] can be defined as

$$d(S,T) = EMD(S,T) = \frac{\sum_{i=1,j=1}^{m,n} f_{i,j} d_{i,j}}{\sum_{i=1,j=1}^{m,n} f_{i,j}} \tag{1}$$

where s_i is i-th category in S and t_j is j-th in T, $d_{i,j} = ||g(s_i) - g(t_j)||$, feature extractor $g(.)$ of an image and the optimal flow $f_{i,j}$ computes total work as a EMD minimization problem. Finally, the similarity is calculated as:

$$sim(S,T) = e^{-\gamma d(S,T)} \tag{2}$$

where γ is a regularization constant of value 0.01.

Domain similarly score can be calculated between the source and target domains. In our approach, we use large scale datasets as source domains, and target domains are selected from six commonly used FGVC datasets.

3.2 Attention Aware Data Augmentation

In our method, we consider using the Weakly Supervised Data Augmentation Network (WS-DAN) [12]. Firstly, we extract features of the image I and feature maps $F \in R^{H \times W \times C}$, where H, W, and C correspond to height, width, and number of channels of a feature layer. Then, we generate attention maps $A \in R^{H \times W \times M}$ from feature maps, where M is the number of attention maps. One more critical component is bi-linear attention pooling, which is used to extract features from part objects. Element-wise multiplication between feature maps and attention maps is computed to get part-feature maps, and then, pooling operation is applied on part-feature maps afterward. Randomly generated data from augmentation is not much efficient. However, attention maps can be handy for data augmentation. This way model can be guided to focus on essential parts of the data and augment those data to the network. With an augmentation map, part's region can be zoomed, and detailed features can be extracted. This process is called attention cropping. attention maps can represent similar object's part. Attention dropping can be applied to the network to distinguish multiple object's

part. Both attention cropping and attention dropping are controlled through a threshold value.

During the training process, no bounding box or keypoints based annotation are available. For each particular training image, attention maps are generated to represent the distinguishable part of object. Attention guided data augmentation component is responsible to select attention maps efficiently utilizing attention cropping and attention dropping. Bilinear Attention Pooling (BAP) is used to extract feature from object's parts. Element-wise multiplication between the feature maps and attention maps is used to generate part feature matrix. Then, part features are extracted by convolutional or pooling operation. In the last step, the original data along with attention generated augmented data are trained as input data.

During the testing process, at first, the probability of the object's categories and attention maps are produced from input images. Then, the selected part of the object can be enlarged to refine the category's probability. The final prediction is evaluated as the average of those two probabilities.

4 Experiments

In this section, we show comprehensive experiments to verify the effectiveness of our approach. Firstly, we calculate the domain similarity score using EMD [22] to demonstrate the relationship between the source and target domains. Then, we compare our model with the state-of-the-art methods on six publicly available fine-grained visual categorization datasets. Furthermore, we perform additional experiments to demonstrate the effect of image resolution on transfer learning. We compare input images in the iNaturalist (iNat) dataset from 299×299 to 448×448 to observe the effect of image resolution in terms of accuracy. We train the baseline inceptionV3 model with iNat datasets for this experiment. Additionally, we combine both iNat and imageNet dataset to make a bigger dataset. We perform detailed experimental studies with different types of large scale datasets and apply the WS-DAN [12] method to observe the impact.

4.1 Datasets

We present a detailed overview of the datasets that we use for our experiments.

ImageNet: The ImageNet [7] contains 1.28 million training images and 50 thousand validation images along with 1,000 categories.

iNaturalist(iNat) : The iNat dataset introduces in 2017 [27]. It contains more than 665,000 training and around 10000 test images from more than 5000 natural fine-grained categories. Those categories include different types of species, including mammals, birds, insects, plants, and more. This dataset is quite imbalanced and varies a lot in terms of the number of images per category.

Fine-grained object classification dataset: Table 1 summarizes the information of each dataset in detail.

Table 1. Six commonly used FGVC datasets.

Datasets	Objects	Classes	Training	Test
CUB200-2011	Bird	200	5,994	5,794
FGVC-Aircraft	Aircraft	100	6,667	3,333
Stanford Cars	Car	196	8,144	8,041
Stanford Dogs	Dog	120	12,000	8,580
Flowers-102	Flowers	102	2,040	6,149
Food-101	Food	101	75,750	25,250

4.2 Implementation Details

In our experiment, we used open-source implementation of popular deep learning framework, Tensorflow [1] to train all the models on multiple Nvidia Geforce GTX 1080Ti GPUs. The machine has Intel Core-i7-5930k CPU@ 3.50GHz x 12 processors with 64GB of memory. During training, we adopted Inception v3 [26] as the backbone network. We employed WS-DAN [12] technique to perform experiments to demonstrate the effectiveness of transfer learning. For all the datasets, we used Stochastic Gradient Descent (SGD) with a momentum of 0.9, the number of epoch 80, mini-batch size 12. The initial learning rate was set 0.001, with exponential decay of 0.8 after every 2 epochs.

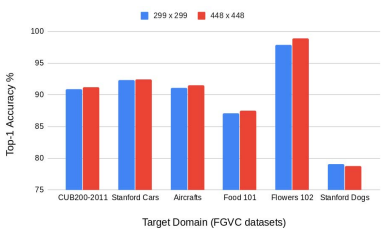

Fig. 1. Effect of transfer learning with different sizes of image resolution on iNat dataset.

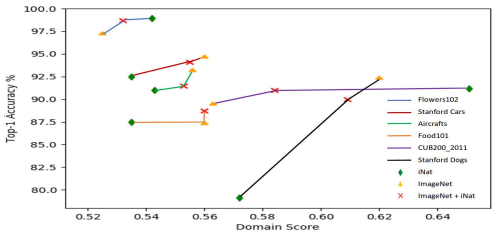

Fig. 2. Co-relation between transfer learning accuracy and domain similarity score between the source and target domain.

5 Results

When training a CNN, some pre-processing is usually done to the input image to match with a specific size. Higher resolutions images usually contain essential information and precise details that are important to visual recognition. We compare results on six FGVC datasets with different sizes of image resolution of the iNat dataset. In summary, images with higher resolution yields better accuracy except for stanford dogs dataset. Figure 1 represents the effect of transfer learning with various sizes of image resolution on iNat dataset.

Table 2. Comparison to different types of FGVC datasets. Each row represents a network pre-trained on source domain for transfer learning and each column represents top-1 image classification accuracy by fine-tuning different networks on the target domains.

Pre-trained InceptionV3	CUB200 2011	Stanford Cars	Aircrafts	Food 101	Flowers 102	Stanford Dogs
ImageNet	82.8	91.3	85.5	88.6	96.2	84.2
ImageNet on WS-DAN	89.3	**94.5**	**93.0**	87.2	97.1	**92.2**
iNat on WS-DAN	**91.2**	92.5	91.0	87.5	**98.9**	79.1
ImageNet + iNat on WS-DAN	91.0	94.1	91.5	**88.7**	98.7	90.0

Table 3. Comparison in terms of accuracy with existing state-of-the-art FGVC methods.

Method	CUB200 2011	Stanford Cars	Aircrafts	Food 101	Flowers 102	Stanford Dogs
Bilinear-CNN [18]	84.1	91.3	84.1	82.4	-	-
DLA [32]	85.1	94.1	92.6	89.7	-	-
RA-CNN [9]	85.4	92.5	-	-	-	87.3
Improved Bilinear-CNN [17]	85.8	92.0	88.5	-	-	-
GP-256 [35]	85.8	92.8	89.8	-	-	-
MA-CNN [9]	86.5	92.8	89.9	-	-	-
DFL-CNN [29]	87.4	93.8	92.0	-	-	-
MPN-COV [16]	88.7	93.3	91.4	-	-	-
Subset B [6]	89.6	93.5	90.7	90.4	-	88.0
WS-DAN [12]	89.4	94.5	93.0	87.2	97.1	92.2
DATL + WS-DAN	**91.2**	94.5	**93.1**	88.7	**98.9**	**92.2**

In Table 2, we present the top-1 accuracy of the target domains on various source domains. These results show the impact of transfer learning from pre-trained model. Large scale datasets are essential for getting improved accuracy when transfer learning is conducted. ImageNet dataset is much larger than iNat dataset; still, it shows worse accuracy in the CUB200-2011 dataset. So, we cannot conclude that using a bigger dataset while transfer learning can always yield better results. Moreover, the domain similarity score also supports this hypothesis. Hence, transfer learning can be effective if the target domain can be trained with similar source domain. We compare our method with state-of-the-art baselines on six commonly used fine-grained categorization datasets. The summary of the comparison is presented in Table 3. We visually represent the relationship between the top-1 accuracy and the domain similarity score. We can observe from Figure 2 that the domain similarity score positively correlated with transfer learning accuracy between large scale datasets and FGVC datasets. Each marker represents a source domain. With the right selection of source do-

main, better transfer learning performance can be achieved. For example, the domain similarity score between iNat and CUB200-2011 is around **0.65**, which is the reason it shows higher accuracy **(91.2)** when iNat is used as pre-training the source domain compared to others. For Flowers-102 dataset, the accuracy is **98.9** with iNat as the source domain which has the highest domain simiarity score **0.54**, among other source domains. Similarly, Stanford Cars, Stanford Dogs and Aircrafts dataset show higher domain similarity score supports better accuracy. Only for the Food101 dataset, the accuracy from transfer learning remains similar while domain similarity changes. We believe this is due to having vast number of training images in Food101. Consequently, the target domain contains enough data; thus, transfer learning has a little contribution. We can observe that both ImageNet and iNat are highly biased, achieving dramatically different transfer learning accuracy on target datasets. Intriguingly, when we transfer networks trained on the combined ImageNet + iNat dataset and perform WS-DAN [12] method over it, we got better results in Food-101 dataset. Intriguingly, the resulted accuracy of the combination of ImageNet and iNat, fell in-between ImageNet and iNat pre-trained model. It means that we can not attain good accuracy on target domains by just using a larger (combined) source domain.

Our work demonstrates utilizing proper domain similarity score can be used initially to identify which large scale dataset to employ. That way, target datasets can learn essential features from large training data. Furthermore, we can employ attention aware data augmentation techniques to achieve state-of-the-art accuracies on FGVC datasets.

6 Conclusion

In this paper, we describe a simple technique that takes attention mechanism as a data augmentation technique. attention maps are guided to focus on the object's parts and encourage multiple attention. We demonstrate the effect of domain adaptive transfer learning to play a vital role in boosting performance. Depending on the domain similarity score on the source datasets, we can consider which target datasets we can train to get better accuracy. We show that the effect of adequately selected datasets in the source domain with attention-based augmentation technique can achieve the state-of-the-art result in multiple fine-grained visual classification datasets. We also analyze the effect of image resolution on transfer learning between the source and target domains. In future work, we are planning to explore the various types of domain similarity metrics on transfer learning to boost performance.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). pp. 265–283 (2016)

2. Anne Hendricks, L., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., Darrell, T.: Deep compositional captioning: Describing novel object categories without paired training data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–10 (2016)
3. Azizpour, H., Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: Factors of transferability for a generic convnet representation. IEEE transactions on pattern analysis and machine intelligence **38**(9), 1790–1802 (2015)
4. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: Cvae-gan: fine-grained image generation through asymmetric training. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2745–2754 (2017)
5. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018)
6. Cui, Y., Song, Y., Sun, C., Howard, A., Belongie, S.: Large scale fine-grained categorization and domain-specific transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4109–4118 (2018)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
8. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning. pp. 647–655 (2014)
9. Fu, J., Zheng, H., Mei, T.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4438–4446 (2017)
10. Ge, Z., Bewley, A., McCool, C., Corke, P., Upcroft, B., Sanderson, C.: Fine-grained classification via mixture of deep convolutional neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–6. IEEE (2016)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hu, T., Qi, H.: See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv preprint arXiv:1901.09891 (2019)
13. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fgvc: Stanford dogs. In: San Diego: CVPR Workshop on FGVC. vol. 1 (2011)
14. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 554–561 (2013)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
16. Li, P., Xie, J., Wang, Q., Gao, Z.: Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 947–955 (2018)
17. Lin, T.Y., Maji, S.: Improved bilinear pooling with cnns. arXiv preprint arXiv:1707.06772 (2017)
18. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 1449–1457 (2015)

19. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
20. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008)
21. Peng, X., Tang, Z., Yang, F., Feris, R.S., Metaxas, D.: Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2226–2234 (2018)
22. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International journal of computer vision **40**(2), 99–121 (2000)
23. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 806–813 (2014)
24. Simon, M., Rodner, E.: Neural activation constellations: Unsupervised part model discovery with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1143–1151 (2015)
25. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision. pp. 843–852 (2017)
26. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
27. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
28. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
29. Wang, Y., Morariu, V.I., Davis, L.S.: Learning a discriminative filter bank within a cnn for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4148–4157 (2018)
30. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: 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. pp. 842–850 (2015)
31. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in neural information processing systems. pp. 3320–3328 (2014)
32. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2403–2412 (2018)
33. Zhang, P., Zhong, Y., Deng, Y., Tang, X., Li, X.: A survey on deep learning of small sample in biomedical image analysis. arXiv preprint arXiv:1908.00473 (2019)
34. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1325–1334 (2018)
35. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 5209–5217 (2017)