

Comprehensive Study of Multiple CNNs Fusion for Fine-Grained Dog Breed Categorization

Minori Uno

Xian-Hua Han

Yen-Wei Chen

Faculty of Science, Yamaguchi University, 1677-1 Yoshida, Yamaguchi City, Yamaguchi, 753-8511, Japan.

Graduate School of Science and Technology for Innovation, Yamaguchi University, 1677-1 Yoshida, Yamaguchi City, Yamaguchi, 753-8511, Japan.

Collegen of Information Science and Engineering, Ritsumeikan University, 1-1-1, Noji Higashi, Kusatsu, Shiga, 525-8577, Japan

Abstract—Fine-grained visual categorization aims to distinguish objects in subordinate classes instead of basic class, and is a challenge visual task due to the high correlation between subordinated classes and large intra-class variation (e.g. different object poses). Although, deep convolutional neural network (DCNN) has brought dramatic success on generic object classification, detection and segmentation with the availability of the large-scale training samples, direct application of DCNN on fine-grained visual categorization, where only decades or at most hundreds of training samples for each subordinate class are available in most public fine-grained image datasets, cannot lead to satisfactory classification results due to small number of training samples. This study explores the transfer learning strategy for fine-grained dog breed categorization based on the learned CNN models with the large-scale image dataset: ImageNet, and prove promising performance with two DCNN models: AlexNet and VGG-16. Furthermore, we argue that different DCNN architecture may extract the representation of different image aspects due to the previously defined CNN kernel sizes, number and various operations in the model learning procedure, and thus result in different performance for visual categorization. This study proposes to fusion multiple CNN architectures for combining different aspect representations to give more accurate performance. We compressively study the fusion of different layers such as Fc6 and Fc7 in AlexNet and VGG-16, and manifest 2.88% improvement of the fusion architecture over the best performance of the only one DCNN model: VGG-16 from 81.2% to 84.08%.

Keywords—Dog breed categorization, transfer learning, CNN fusion, multiple CNN architecture

I. INTRODUCTION

Fine-grained visual categorization (FGVC) is to categorize objects into subordinate classes instead of basic classes [1-3]. Dog breed categorization is representative one of a set of FGVC problems. Dogs are the most popular domestic animal as a pet, and thus After human, dogs are possibly the most photographed species. Extracting information from images of pets and recognizing the pet breeds also has a practical side beyond the technique interest of FGVC. As shown in the large amount of social network for sharing the pet animals: Dogster [4], Pet Finder [5], My dog space [6] and several others [7-10], people dedicate to a lot of attention to their domestic pets and share the cute picture in the social networks. The posted domestic animal images are

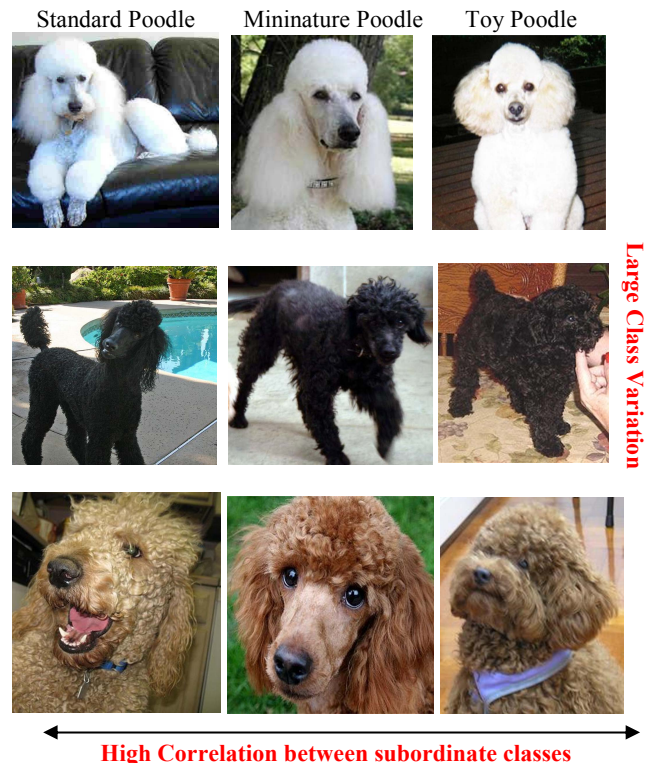


Fig. 1 the contradictory issues in FGVC

not always provided with breed description or maybe with the incorrect breed the pet owner believed. How to automatically recognize the pet breed or correct the incorrect breed of the posted animal images are needed for pet image management and effective search.

As the representative one of the FGVC problem, the challenge of the dog breed categorization mainly lies on two aspects: 1) simultaneously handling the co-occurrence of two somewhat contradictory issues: the high correlation between subordinated classes and large intra-class variation (e.g. different object poses); 2) only small number of images are available for training compared with more than thousands of available training images for each basic classes. The examples of the high correlation between subordinated classes and large intra-class variations for dog breed categorization shown in Fig. 1. A common approach for robustness against the contradictory issues is to first localize various parts or keypoints of the object and model the

appearance conditional on the detected location or keypoints. For accurate location detection, the parts are needed to previously defined and manually labeled to train the parts detector in a supervised manner. Recently deep convolutional neural network (DCNN) has brought dramatic success on generic object classification, detection and segmentation with the availability of the large-scale training samples. DCNN model have also been applied the object part detection in FGVC problem, and shown the significantly improve over earlier work based on hand-crafted features. However, these approaches need to annotate object parts manually which is significantly more difficult than collecting image labels. In addition, the previously defined parts may not optimal for the final categorization task.

Another research direction is to firstly extract robust image representation and conduct recognition with a classifier such as support vector machine (SVM [11]). The conventional feature extraction approaches mainly include the bag-of-feature (BOF) model [12] and its extensions such as ScSPM [13], LLCSPM [14], VLAD [15] or Fisher vector [16] with SIFT Feature, and manifested the performance in some extend for different classification problems. Deep convolutional neural network (DCNN) construct hierarchical network architecture for combining image representation extraction and classification into a unified end-to-end learning procedure in most application tasks. However, there need a lot of training samples for learning the acceptable CNN model, while collection for large-scale labeled images in the subordinate class of FGVC problem is much more challenge than the basic classes in generic object, and only decades or at most hundreds of training samples for each subordinate class are available in most public fine-grained image datasets. Therefore, directly learning the DCNN model on fine-grained visual categorization dataset usually cannot obtain satisfactory classification results due to small number of training samples.

This study explores the transfer learning strategy for fine-grained dog breed categorization based on the learned CNN models with the large-scale image dataset: ImageNet, and prove promising performance with two DCNN models: AlexNet [17] and VGG-16 [18]. Furthermore, we argue that different DCNN architecture may extract the representation of different image aspects due to the previously defined CNN kernel sizes, number and various operations in the model learning procedure, and thus result in different performance for visual categorization. We propose to fusion multiple CNN architectures for combining different aspect representations to give more accurate performance. In detail, we compressively study the fusion of different layers such as Fc6 and Fc7 in AlexNet and VGG-16, and manifest 2.88% improvement of the fusion architecture over the best performance of the only one DCNN model: VGG-16 from 81.2% to 84.08%.

II. FUSION OF MULTIPLE CNN ARCHITECTURES

Since the amazing performance with the deep CNN model: AlexNet [17] in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) has been achieved in 2012, deep CNNs have been rapidly applied different visual problems, and more complex and deeper CNN architectures such as VGG [18], GoogLeNet [19], ResNet [20] have been developed for pursuing performance. In spite of the success

of deep CNNs on different applications, training high-generalized CNN model requires large-scale labeled dataset especially for complicated and very-deep CNN architectures. However, the fine-grained dog breed usually can only collect small number of labelled images for model learning, and thus cannot obtain reasonable categorization model. Therefore, this study explores the transfer learning strategy for dog-breed categorization with the pre-trained models using the large-scale ImageNet dataset. The common solution for using deep CNNs on small dataset is parameter transfer learning, which drop the classifier layer (or several later layers near the final classifier layer) of a pre-trained CNN model and fine-tune it on target dataset. We replace the final classifier layer (1000-neurons) in the pre-trained model for ImageNet dataset with a layer of the dog breed number of neurons, and retain the same conditions such as kernel size numbers of other layer for reusing the learned parameters of the pre-trained CNN model as initial values.

As we mentioned that the recently proposed CNN architectures such as ResNet with hundreds or thousands of layers can improve the performance about several percent in several image classification problems [20]. However, it is more difficult to train due to the complicated connection and much deeper structure. Instead of the increased deeper structure, we advocate to parallel several not-so-deep CNN architectures with different kernel sizes and numbers, etc., which may learn the image representation of different aspects such as multi-scale properties, and concatenate the feature maps in some later layers for dog breed categorization learning. Due to the infeasibility of the CNN model training with small number of images, we investigate the comprehensive fusion of the popularly used AlexNet and VGG-net, where the pre-trained models using large-scale ImageNet dataset have been released and state-of-the-art performances in different computer vision problems was provided regardless to their no-so-deep architectures. Next, we will describe the used basic CNN architecture and the detail fusion method.

A) The used Basic CNN Models for Transfer Learning

This study explores the popularly used CNN architectures: AlexNet and VGG. AlexNet is proposed by Krizhevsky and Hinton in 2012 [17] and achieved state of the art performance on the ImageNet 2012 classification benchmark [17]. The original model contains 8 layers and the last three layers are two fully connected (FC) layers (same as hidden layers in BP Neural Network) and output layer. We replace the last FC layer with dog breed number of neurons and retain the same hyper-parameters (kernel size, numbers, neuron numbers in FC layers etc.) of the previous 7 layers with the learned parameters of the pre-trained AlexNet model using ImageNet dataset and set the initial value of the last layer parameter randomly conditioned on Gaussian distribution. The basic learning rate for of the previous 7 layers is set as 0.001, and the one of the last FC layer is as 0.02. We initialize the parameters (weights, bias) to extract features from images. The fine-tune AlexNet model for dog breed categorization is trained using 10 epochs with the training dog samples. VGG-net proposed by Simonyan applies very small (3x3) convolutional filters in all layers and steadily increase the depth of the network by adding more convolution layers. As the original VGG-net

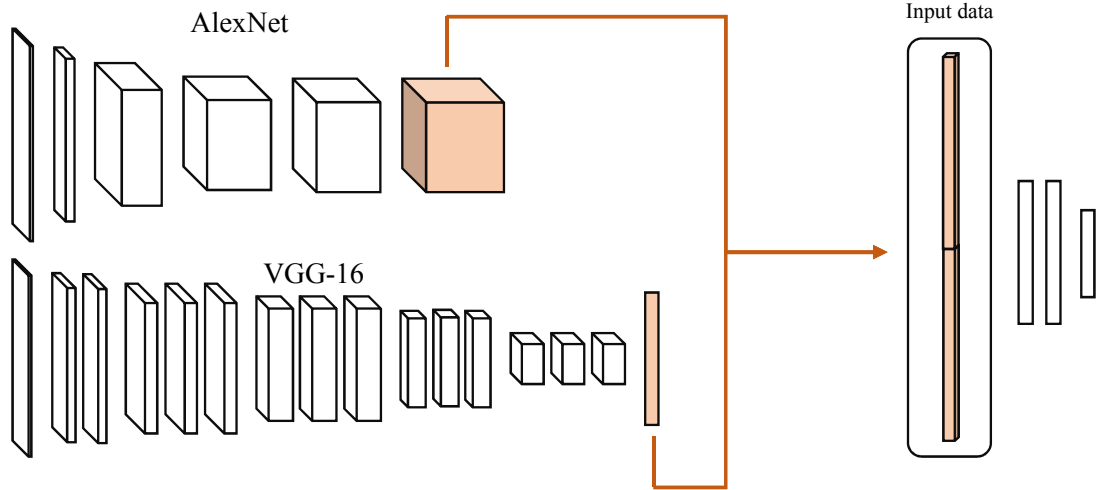


Fig. 2 Fusion architecture of two CNN streams with AlexNet and VGG-16

work claimed that combination of several convolutional layers with small filter can account for one convolutional layer with large effective receptive field, and the VGG-net can be considered as an equivalent version of AlexNet with small size filters in more convolutional layers. Related work has shown that the VGG-net can achieve much better performance than AlexNet for image classification problems, and has been popularly used in different applications such as object detection, image segmentation and so on. We also conduct transfer learning with the pre-trained VGG CNN model using ImageNet dataset, where we replace the final FC layer of VGG-net with dog breed number of neurons and retain others in the original VGG. The experimental conditions are set as same with AlexNet transfer learning.

B) Fusion of Multiple CNN Architectures

Although the existed CNN models provided promising performance in different classification problems, and proved some other improvement with more complicated and deeper architectures such as ResNet. The depth-increased CNNs have more complicated structures, and are more hard to understand. In addition, the model is much difficult for training. On the other hand, we argue that paralleling several basic CNN models (increasing width) with different kernel sizes and numbers, etc. may provide research insight for the learned representation, where each stream of CNN has not so many layers and is much easier to be maintained. Different streams of CNN may learn different representations of image aspects such as multi-scale properties, and the concatenation of the later layers' feature maps can take consideration of all the representation in different CNN streams for the final task learning. In addition, each stream of CNN can firstly be independent-learned for providing a good initialization of the fusion CNN architecture, which is much easier to be trained and is more stable. In our experiments, we use AlexNet and VGG-16 as our basic CNN streams, and concatenate the feature maps of different layers from the basic CNNs for comprehensive study of the dog breed categorization task.

The schematic structure of the fusion CNN architectures is shown in Fig. 2.

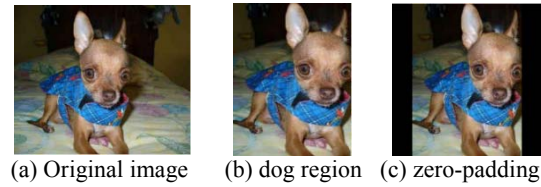


Fig.3 Pre-processing of dog image

III. EXPERIMENTAL RESULTS

We evaluate the dog breed categorization performance on Stanford Dog database using the single AlexNet, VGG-16, the fusion architectures with concatenation of different layers in the basic CNN streams. Stanford Dog database was constructed by Khosla [21,22], and has 120 class dog breeds, which is a challenging dataset aimed at fine-grained image categorization. The dataset includes 20,580 annotated images of dogs belonging to 120 species. Each image is annotated with a bounding box and object class label. The dataset are divided into training subset and test subset, where training subset consists of 12,000 images with 100 images for each dog breed and test subset has 8,580 images with varied 50~120 images for each dog breed. For using the CNN models, we firstly unify the image into the same size with a pre-processing procedure. The pre-processing procedure includes three-steps as shown in Fig. 3: 1) Extracting the dog region using the annotated bounding box from the original image; 2) zero-padding the direction with small pixel number to the same size of the other direction which avoids the deformation of the object in the image; 3) resizing all images into the same size such as 224x224 for AlexNet, 227x227 for VGG-16.

We train the Alexnet, VGG-16 with the Stanford Dog dataset using scratch learning (randomly initialize the network parameters) and fine-tuning from the pre-trained model with ImageNet dataset. The top-one accuracy is calculated for evaluation, and the compared results are shown in Fig. 4. The baseline result [21,22] is taken from the released site of Stanford Dog dataset with bag-of-features representation and SVM classification [11]. Since the scratch learning of the AlexNet and VGG-16 model for dog breed categorization cannot provide reasonable results even in training procedure (maintaining about 10% recognition rates) in our conducted experiments, which shows the un-convergent character for dog breed categorization, we do not give the accuracies with scratch learning in Fig. 4. As we analyzed the scratch learning of the AlexNet and VGG-16 model cannot perform dog breed categorization correctly due to small number of training samples while the fine-tuning from the pre-trained CNN model on ImageNet dataset manifests promising performance as shown in Fig. 4. Furthermore, we conduct dog breed categorization experiments using the fusion architecture via concatenating the feature maps of different layers in the basic CNN streams: AlexNet and VGG-16. The concatenation patterns in the fusion CNN architectures are shown in Table 1. The compared results of different fusion architectures are shown in Table 2, which manifests most fusion architectures can improve the accuracy from the one of VGG-16, and the best improvement with concatenation pattern of Relu7 feature maps in both AlexNet and VGG-16 is 2.88%.

As we known that the correlation among different dog breeds are very high as shown in Fig. 1, where for example

Table.1 Concatenation patterns of the feature maps in different layers

AlexNet	VGG-16	Concatenation Pattern
Fc6	Fc6	Pt1
Relu6	Relu6	Pt2
Fc7	Fc7	Pt3
Relu7	Relu7	Pt4
Fc8	Fc8	Pt5

Table.2 Top-one Accuracy of different concatenation patterns

Concatenation Pattern	Top-one Accuracy
Pt.1	78.97%
Pt.2	83.87%
Pt.3	82.80%
Pt.4	84.08%
Pt.5	82.45%

‘Standard Poodle’, ‘Mininature Poodle’ and ‘Toy Poodle’ are undistinguishable even for human, and thus the categorized class of the misclassified image may have very similar properties with ground-truth breed. Some easily misclassified dog breeds are shown in Fig. 5, where the first column manifests that the percentage of the input images was classified into the ground-truth class, and the second and third column show the easily mis-classified dog breeds with categorization error rates. From Fig. 5 it obvious that some

dog breeds are difficult to distinguished, such as "Siberian husky" and "Border collie" even by human. Therefore, we explore Top-K accuracy for considering the possible mis-classification even by human. The results of the Top-K accuracy evaluations are given in Fig. 6. Finally, we manifest the confusion matrix of the dog breed categorization using the fusion architecture of AlexNet and VGG-16 Relu7 in Fig. 6.

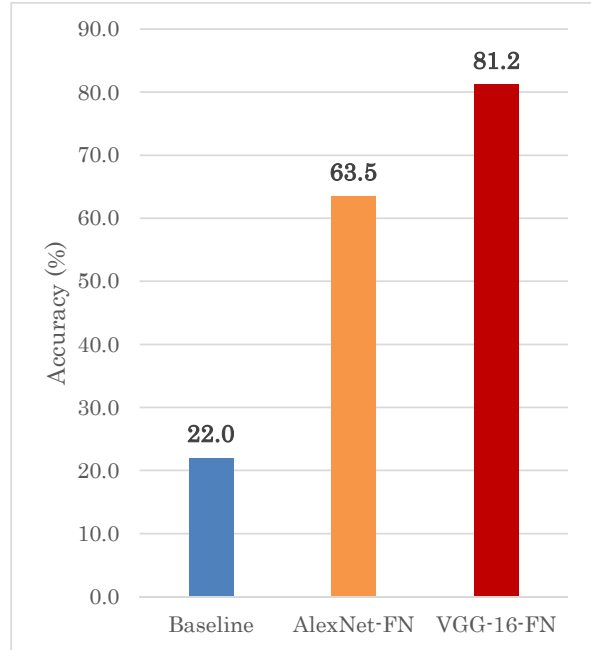


Fig. 4 The compared accuracies of the baseline result with Bag-of-Features and the results with fine-tuning AlexNet and VGG-16.

IV. CONCLUSIONS

This paper proposed to fusion multiple basic CNN models for dog breed categorization and proved promising performance improvement. Due to the small number of training samples in the fine-grained dog breed categorization problem, we exploited transfer learning strategy based on the pre-trained CNN models with the large-scale ImageNet dataset, which set the learned parameters in the pre-trained CNN model as initial values for fine-tuning the CNN model again with dog breed dataset. Experimental results showed that outstanding performance can be achieved. In addition, we argued that different DCNN architecture may extract the representation of different image aspects due to the previously defined CNN kernel sizes, number and various operations in the model learning procedure, and thus result in different performance for visual categorization. Thus we proposed to fusion multiple CNN architectures for combining different aspect representations to give more accurate performance. We compressively study the fusion of different layers such as FC6, Relu6, FC7, Relu7 and FC8 in AlexNet and VGG-16, and manifested 2.88% improvement of the fusion architecture over the best performance of the only one DCNN model: VGG-16 from 81.2% to 84.08%.
















Input dog breeds and accuracy	The misclassified dog breeds with Top-1 and Top-2 errors	
Siberian husky  31.5%	Eskimo dog  43.5%	Malamute  15.2%
Border collie  42.0%	Collie  46.0%	Borzoi  2.0%
Norwich terrier  43.5%	Cairn  25.9%	Norfolk terrier  12.9%
Miniature poodle  45.5%	Toy Poodle  21.8%	Standard poodle  9.1%
American Staffordshire terrier  50.0%	Staffordshire bullterrier  37.5%	Boxer  1.6%

Fig. 5 Easily mis-classified dog breeds and their accuracy and Top-1/Top-2 error rates

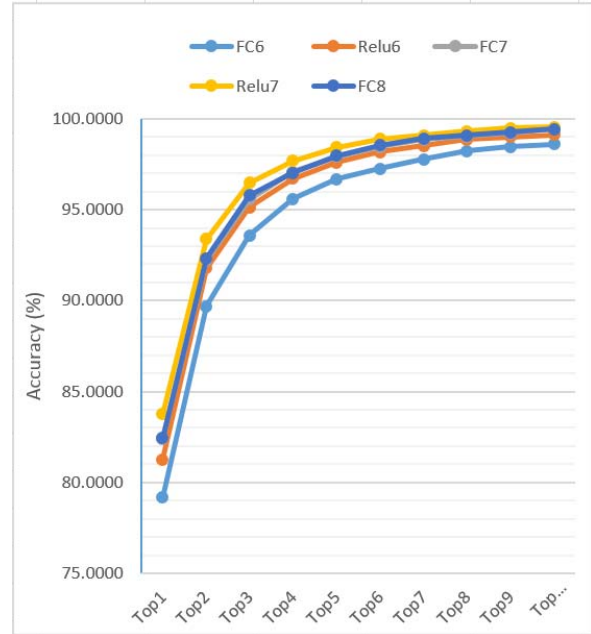


Fig. 6 The computed accuracies considering Top-K classified dog breeds.

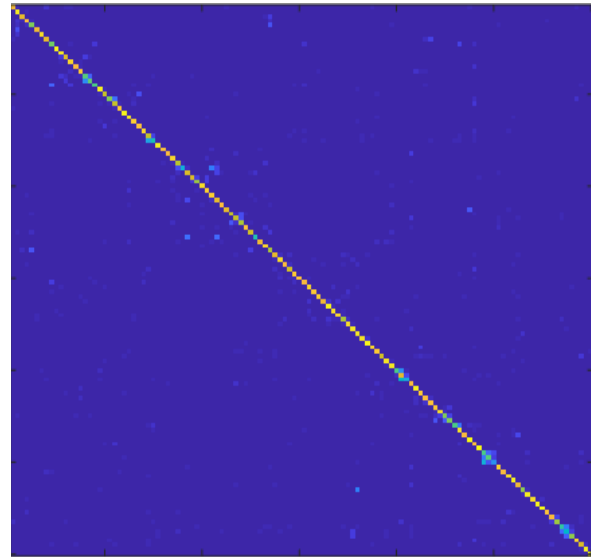


Fig. 7 Confusion matrix of dog breed categorization with the fusion architecture of Relu7 concatenation pattern

V. REFERENCE

- [1] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In Proc. CVPR, 2006.
- [2] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In Proc. ICVGIP, 2008.
- [3] A. Khosla, N. Jayadevaprakash, B. Yao, and F. F. Li. Novel dataset for fine-grained image categorization. In First Workshop on Fine-Grained Visual Categorization, CVPR, 2011.
- [4] Dogster. <http://www.dogster.com/>.

- [5] My dog space. <http://www.mydogspace.com/>.
- [6] Petfinder. <http://www.petfinder.com/index.html>
- [7] The international cat association. <http://www.tica.org/>.
- [8] My cat space. <http://www.mycatspace.com/>.
- [9] The cat fanciers association inc. <http://www.cfa.org/Client/home.asp>.
- [10] Cats in sinks. <http://catsinsinks.com/>.
- [11] C. Cortes and V. Vapnik, Support-Vector Networks, Machine Learning. Journal Machine Learning, Vol. 20, No. 3, pp. 273-297, 1995.
- [12] G. Csuka, C.R. Dance, Li Fan, J. Willamowski and C. Bray. Visual Categorization with Bags of Kerpoints. ECCV International Workshop on Statistical Learning in Computer Vision (2004).
- [13] Jianchao Yang and Kai Yu. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In Proc. CVPR, 2009.
- [14] Kai Yu, Tong Zhang and Yihong Gong. Nonlinear Learning Using Local Coordinate Coding. In Proc. NIPS, 2009.
- [15] Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., & Schmid, C.. Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 34, No. 9, pp. 1704-1716, 2012.
- [16] Jorge Sanchez, Florent Perronnin, Thomas Mensink, Jakob Verbeek, Image Classification with the Fisher Vector: Theory and Practice, International Journal of Computer Vision, Vol. 105, No. 3, pp. 222-245, 2013.
- [17] K. Alex, S. Ilya and H. E. Geoffrey. ImageNet Classification with Deep Convolutional Neural Networks. In Proc. BIPS, Vol. 1, pp. 1097-1105, 2012.
- [18] K. Simonyan and Z. Zisserman. Very Deep Convolutional Networks for Large-scale Image Recognition. In Proc. ICLR, 2015.
- [19] S. Christian, L. Wei and J. Yangqing, S. Pierre, R. Scott. Going Deeper with Convolutions. In Proc. CVPR, 2015.
- [20] H. Kaiming, Z. Xiangyu and R. Shaoqing. Deep Residual Learning for Image Recognition. In Proc. CVPR, 2016.
- [21] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao and Li Fei-Fei. Novel dataset for Fine-Grained Image Categorization. First Workshop on Fine-Grained Visual Categorization (FGVC). IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
- [22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.