Fine-grained Visual Classification

Jinna Cui
College of Information Science and Engineering
Ocean University of China
Qingdao 266100, China
Email: Jinna_ouc@163.com

***Abstract*—Fine-grained visual classification aims to distinguish subordinate-level categories, such as bird species or dog breeds. This task is extremely challenging due to high intra-class and low inter-class variance. Deep learning technology has shown impressive performance in various vision tasks such as image classification, object detection and semantic segmentation. In particular, recent advance of deep learning techniques bring encouraging performance to fine-grained image classification. In this report, we did some image classification and fine-grained image classification experiments with deep neural network.**

1. Introduction

Fine-grained recognition refers to the task of distinguishing sub-ordinate categories, such as bird species [2], dog breeds [3], car models [4], flower categories [5], food dishes [6], etc. With the great potential in rivaling human experts, it has shown tremendous applications in real world ranging from e-commerce [7] to education [8]. Although great success has been achieved for basic-level recognition in the last few years, fine-grained recognition still faces several challenges. First, it is more difficult and time-consuming to gather a large amount of labeled fine-grained data because it calls for experts with specialized domain knowledge. In addition, the difference between fine-grained classes is very subtle. The most discriminative features are often not based on the global shape or appearance variation but contained in the mis-alignment of local parts or patterns while the precise part annotations are usually expensive to acquire. A common approach is to first localize carious parts of the object and model the appearance conditioned on their detected locations. The parts are often defined manually and the part detectors are trained in a supervised manner. A drawback of these approaches is that annotating parts is significantly more challenging than collecting image labels. Another approach is to use a robust image representation. Traditionally these included descriptors such as VLAD [9] or Fisher vector [10] with SIFT features [11]. By replacing SIFT by features extracted from convolutional layers of a deep network pre-trained on ImageNet [12], these models achieve state-of-the-art results on a number of recognition tasks. Although these models are easily applicable as they don't rely on part annotations, their performance is below the best part-based models, especially when objects are small and appear in clutter. More over, the effect of end-to-end training of such architectures has not been fully studied.

The requirement of our experiment is shown in the following table:

Networks	Datasets	Details
MLP	MNIST	At least one experiment
LeNet-5	MNIST	One experiment
LeNet-5	CIFAR-10	One experiment
CIFAR-10 net	CIFAR-10	ReLU + Max pooling
CIFAR-10 net	CIFAR-10	Sigmoid + Max pooling
CIFAR-10 net	CIFAR-10	tanh + Max pooling
CIFAR-10 net	CIFAR-10	ReLU + Average pooling
AlexNet	CUB-200-2011	One experiment

1.1. Neural Networks

1.1.1. MLP. An MLP can be viewed as a logistic regression classifier where the input is first transformed using a learnt non-linear transformation. This transformation probects the input data into a space where it becomes linearly separable. This intermediate layer is referred to as a hidden layer. A single hidden layer is sufficient to make MLPs a universal approximator. An MLP with two hidden layers can be represented graphically as follows:

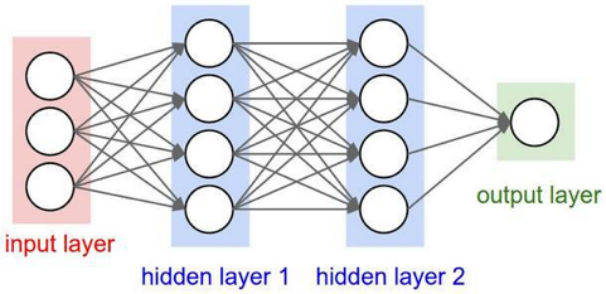

Figure 1. MLP with two hidden layers

1.1.2. LeNet-5. Convolutional Neural Networks are a special kind of multi-layer neural networks. Like almost every other neural networks they are trained with a version of the back-propagation algorithm. Where they differ is in the architecture. Convolutional Neural Networks are designed to recognize visual patterns directly from pixel images with minimal preprocessing. They can recognize patterns with extreme variability, and with robustness to distortions and simple geometric transformations.

LeNet-5 [13] is the very earliest convolutional network designed for handwritten and machine-printed character recognition. LeNet combine three architectural ideas to ensure some degree of shift, scale, and distortion invariance: 1) local receptive fields; 2) shared weights (or weight replication); and 3) spatial or temporal subsampling.

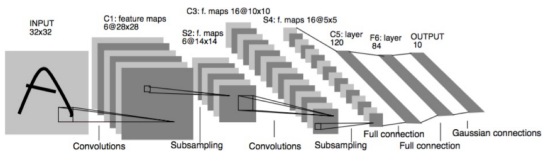

Figure 2. LeNet-5

The input plane receives images of characters that are approximately size normalized and centered. Each unit in a layer receives inputs from a set of units located in a small neighborhood in the previous layer. The idea of connecting units to local receptive fields on the input goes back to the perceptron in the early 1960s, and it was almost simultaneous with Hubel and Wiesels discovery of locally sensitive, orientation selective neurons in the cats visual system [14]. With local receptive fields neurons can extract elementary visual features such as oriented edges, endpoints, corners (or similar features in other signals such as speech spectrograms). These features are then combined by the subsequent layers in order to detect higher order features. Units in a layer are organized in planes within which all the units share the same set of weights.

The input of LeNet-5 is a 32×32 pixel image. The reason is that it is desirable that potential distinctive features such as stroke endpoints or corner can appear in the center of the receptive field of the highest level feature detectors.

In the following, convolutional layers are labeled Cx, subsampling layers are labeled Sx, and fully connected layers are labeled Fx, where x is the layer index.

Layer C1 is a convolutional layer with six feature maps. Each unit in each feature map is connected to a 5 5 neighborhood in the input. Layer S2 is a subsampling layer with six feature maps ofsize 14×14. Each unit in each feature map is connected to a 2×2 neighborhood in the corresponding feature map in C1. Layer C3 is a convolutional layer with 16 feature maps. Each unit in each feature map is connected to several 5×5 neighborhoods at identical locations in a subset of S2s feature maps. Layer S4 is a subsampling layer with 16 feature maps of size 5 5. Each unit in each feature map is connected to a 2×2 neighborhood in the corresponding feature map in C3, in a similar way as C1 and S2. Layer C5 is a convolutional layer with 120 feature maps. Each unit is connected to a 5×5 neighborhood on all 16 of S4s feature maps.

1.1.3. CIFAR-10 net. CIFAR-10 net is a convolutional neural network with a simple structure. The activation function of CIFAR-10 net is ReLU function and the pooling method is a combination of max pooling and average pooling. In this report, we use CIFAR-10 net to compare the classification effects of different activation functions, pooling methods.

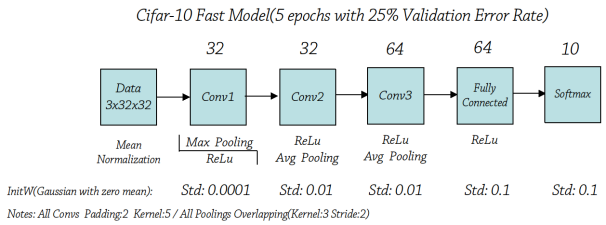

Figure 3. CIFAR-10 net

1.1.4. AlexNet. AlexNet [15] contains eight learned layers – five convolutional and three fully-connected. In AlexNet,

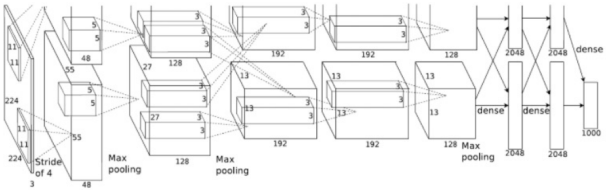

Figure 4. AlexNet

they refer to neurons with nonlinearity as Rectified Linear Units (ReLUs). Deep convolutional neural networks with ReLUs train several times faster than other activation functions. AlexNet used overlap pooling method based on max pooling. The dropout layer consists of setting to zero the output of each hidden neuron with probability of 0.5 which is a reasonable approximation to taking the geometric mean of the predictive distributions produced by the exponentially-many dropout networks. AlexNet uses dropout in the first two fully-connected layers. Without dropout, AlexNet exhibits substantial overfitting. Dropout roughly doubles the number of iterations required to converge.

1.2. Datasets

1.2.1. MNIST. MNIST is a database of handwritten digits. MNIST has a training set of 60,000 examples, and a test set of 10,000 examples. It is a good database for people who want to try learning techniques and pattern recognition methods.

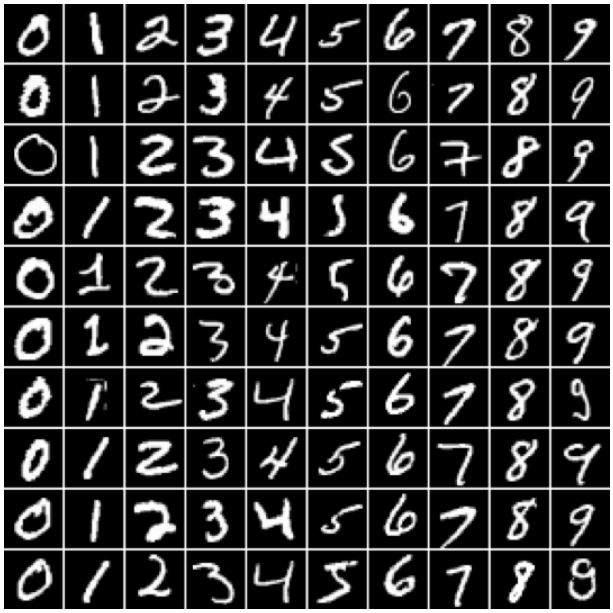

Figure 5. Example images in MNIST dataset

1.2.2. CIFAR-10. The CIFAR-10 dataset consists of 60000 32×32 color images in classes, with 6000 images per class. There are 50000 training images and 10000 test images.

The dataset is divided into five training batches and one test batch, each with 10000 images. The test batch contains exactly 1000 randomly-selected images from each class. The training batches contain the remaining images in random order, but some training batches may contain more images from one class than another. Between them, the training batches contain exactly 5000 images from each class.

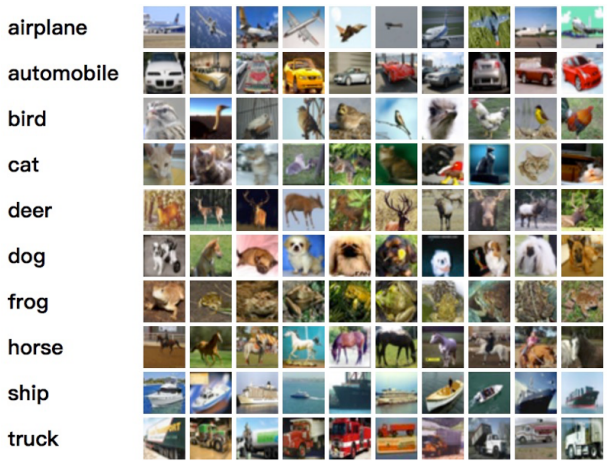

Figure 6. Example images in CIFAR-10 dataset

1.2.3. CUB-200-2011. Caltech-UCSD Birds-200-2011 (CUB-200-2011) is an extended version of the CUB-200 dataset, a challenging dataset of 200 bird species with roughly double the number of images per class and new part location annotations. The number of categories is 200 and the number of images is 11788. All images are annotated with bounding boxes, part locations, and attribute labels.

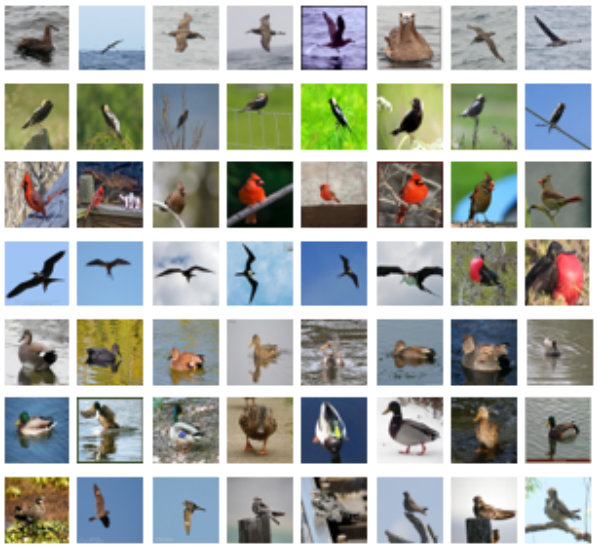

Figure 7. Example images in CUB-200-2011 dataset

2. Experimental Results

The experimental requirements are shown in the following table:

Networks	Datasets	Details
MLP	MNIST	At least one experiment
LeNet-5	MNIST	One experiment
LeNet-5	CIFAR-10	One experiment
CIFAR-10 net	CIFAR-10	ReLU + Max pooling
CIFAR-10 net	CIFAR-10	Sigmoid + Max pooling
CIFAR-10 net	CIFAR-10	tanh + Max pooling
CIFAR-10 net	CIFAR-10	ReLU + Average pooling
AlexNet	CUB-200-2011	One experiment

2.1. MLP on MNIST dataset

I did several experiments on MLP. Firstly, I compare the classification accuracy of ReLU activation function and the classification accuracy of Sigmoid activation function. The MLP net we use in this experiment is shown in the following image. The MLP contains one hidden layer. The input layer (fc6) contains 784 nodes, the hidden layer (fc7) contains 800 nodes and the output layer (fc8) contains 10 nodes.

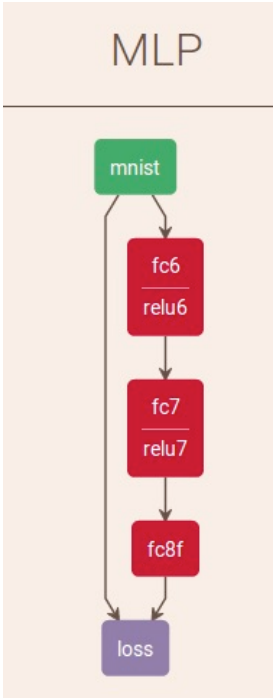

Figure 8. The MLP net

In the first experiment, the activation function is Sigmoid function. In the second experiment, the activation function is ReLU function. From the result we can see that the

Networks	Datasets	Details	Iteration	Accuracy
MLP	MNIST	Sigmoid	500000	96.46%
MLP	MNIST	ReLU	10000	98.03%

ReLU activation function can get higher classification accuracy. And the iteration of ReLU function is 10000 and the iteration of Sigmoid is 500000. The training speed of ReLU activation function is faster than Sigmoid activation function.

2.2. LeNet-5 on MNIST dataset

The training result of LeNet-5 on MNIST dataset is shown in the following table: From the experimental result

Networks	Datasets	Iteration	Accuracy
LeNet-5	MNIST	10000	99.09%

we can see that the classification accuracy of LeNet-5 on MNIST dataset is 99.09 while the iteration is 6000. The classification effectiveness of LeNet-5 on MNIST is pretty good.

Networks	Datasets	Details	Iteration	Accuracy
CIFAR-10 net	CIFAR-10	ReLU&AVE	50000	72.66%
CIFAR-10 net	CIFAR-10	ReLU&MAX	50000	78.31%
CIFAR-10 net	CIFAR-10	Sig&MAX	60000	54.63%

2.3. CIFAR-10 net on CIFAR-10 dataset

From the classification result we can see that the classification effectiveness of max pooling is better than average pooling in CIFAR-10 net. This is why most convolutional neural network choose max pooling method. And the ReLU activation function can get much higher classification accuracy than sigmoid activation function.

2.4. AlexNet on CUB-200-2011 dataset

Networks	Datasets	Iteration	Accuracy
AlexNet	CUB	60000	26.28%

From the experimental result we can see that the AlexNet can not get high classification accuracy on CUB-200-2011 dataset. As a result, the fine-grained classification problem can not just use the traditional classification method. And more fine-grained classification problem remain to be proposed.

3. Conclusion

Nowadays, most image classification tasks use deep neural network method and deep learning method helps improving the classification accuracy effectively. From the experiments we can see that the general convolutional neural networks including the AlexNet, VGG net and GoogLeNet can not be directly used on fine-grained image classification task.

In the feature, I will mainly do some researches on the fine-grained image classification methods to get higher fine-grained classification accuracy.

References

[1] Turner Jefferson T: The importance of small planktonic copepods and their roles in pelagic marine food webs. In: Zoological Studies, IEEE, pp. 255–266 (2004)

[2] Wah, Catherine, et al. "The caltech-ucsd birds-200-2011 dataset." (2011)

[3] Khosla, Aditya, et al. "Novel dataset for fine-grained image categorization: Stanford dogs." Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC). Vol. 2. 2011.

[4] Krause, Jonathan, et al. "3d object representations for fine-grained categorization." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2013.

[5] Nilsback, Maria-Elena, and Andrew Zisserman. "Automated flower classification over a large number of classes." Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on. IEEE, 2008.

[6] Bossard, Lukas, Matthieu Guillaumin, and Luc Van Gool. "Food-101mining discriminative components with random forests." European Conference on Computer Vision. Springer International Publishing, 2014. APA

[7] Bell, Sean, and Kavita Bala. "Learning visual similarity for product design with convolutional neural networks." ACM Transactions on Graphics (TOG) 34.4 (2015): 98.

[8] Kumar, Neeraj, et al. "Leafsnap: A computer vision system for automatic plant species identification." Computer VisionECCV 2012 (2012): 502-516. APA

[9] Jgou, Herv, et al. "Aggregating local descriptors into a compact image representation." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010. APA

[10] Perronnin, Florent, Jorge Snchez, and Thomas Mensink. "Improving the fisher kernel for large-scale image classification." Computer VisionECCV 2010 (2010): 143-156.

[11] Lowe, David G. "Object recognition from local scale-invariant features." Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Vol. 2. Ieee, 1999.

[12] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[13] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

[14] Hubel, David H., and Torsten N. Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." The Journal of physiology 160.1 (1962): 106-154.

[15] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.