Deep Networks for Image Classification

Zhongjie Xu, 190622114, ec19500@qmul.ac.uk

1. Introduction

CNN has a long history in computer vision. In recent years, with the emergence of large-scale category-level training data (such as ImageNet [1]) and the improvement of GPU performance, CNN has shown excellent performance in large-scale visual recognition. In 2012, Krizhevsky et al. [2] used CNN in LSVRC competition for the first time. They expanded the depth of CNN model and obtained the best classification result at that time. Compared with traditional CNN, ReLU is used to replace the saturated nonlinear tanh function in AlexNet, which reduces the computational complexity of the model and increases the training speed of the model by several times. In the training process, some neurons in the middle layer are randomly set to zero by dropout technology, which makes the model more robust and reduces the over-fitting of the full connection layer. At the same time, the training samples are increased through image translation, image horizontal mirror transformation, image grayscale and other data enhancement methods, so as to reduce model overfitting. Since then, CNN has attracted more and more attention from researchers. In this paper, three representative networks of convolutional neural networks (Googlenet, Resnet and SEnet) in the deep learning system will be discussed and their performance on MNIST and CIFR10 [3] data sets will be evaluated. In addition, some new technologies will be used to test and evaluate their improvement on CNN network.

2. Related Work

2.1 GoogleNet

Compared with AlexNet, Szegedy et al [4] greatly increased the depth of CNN, and proposed a CNN structure (called GoogLeNet) with more than 20 layers. In GoogleNet, an Inception block is proposed and the network [5] (NIN) structure is applied. in this block, three types of convolution operations (1*1, 3*3, 5*5) and a pooling operation (3x3) are adopted. Using convolution kernels with different sizes means that receptive fields with different sizes, and finally splicing means the fusion of features with different scales. At the same time, pooling operations in blocks can reduce the size of space and over-fitting. At the same time, the 1x1 convolution kernel is introduced to reduce the dimension and is also used to modify the activation function (ReLU). In order to avoid the gradient disappearing, two auxiliary classifiers are added to the whole network structure to help training, and the two classifiers are used to forward to the gradient. The auxiliary classifier tries to use the data of the middle layer as classification and adds it to the classification results with a smaller weight (0.3), which is equivalent to model fusion. At the same time, it adds the gradient signal of back propagation to the network and also provides external regularization, thus improving the training of the whole network. At that time, during the actual test, these two additional classifiers will be removed. At the end of the network, the full connection layer is abandoned and the average pooling layer is used instead, which greatly reduces the parameters of the model. Its

parameters are 12 times less than AlexNet, and GoogleNet has a higher accuracy rate. in LSVRC-14, it won the first place in the "specified data" group of image classification.

2.2 ResNet

KaiMing et al [6] used Residual Networks (ResNet) to solve the degradation problem. ResNet's main feature is cross-layer connection, which transfers the input across layers and adds it to the convolution result by introducing Shortcut Connections. There is only one sampling layer in ResNet, which is connected behind the last convolution layer. ResNet enables the underlying network to be fully trained, and the accuracy rate is significantly improved with the deepening of depth. ResNet with a depth of 152 layers was used in the image classification competition of LSVRC-15, and it won the first place. In this document, we also try to set ResNet's depth to 1000, and validate the model in CIFAR-10 image processing data set.

2.3 SEnet

Convolution kernel, as the core of convolutional neural network, is usually regarded as an information aggregate that aggregates spatial information and feature dimension information on local receptive fields. Convolution neural network is composed of a series of convolution layers, non-linear layers and down sampling layers, so they can capture image features from the global receptive field to describe the image. Recently, many works have been proposed to improve the performance of the network from the spatial dimension. For example, the Inception structure embeds multi-scale information and aggregates features on different receptive fields to obtain performance gains. Contextual information in space is considered in Inside-Outside network [7]. SEnet [8] is motivated by the desire to explicitly model the interdependence between feature channels. Squeeze operation represents the global distribution of responses on the feature channel, and also enables the layer close to the input to obtain the global receptive field. The Excitation operation is a mechanism similar to gates in a cyclic neural network. Weights are generated for each feature channel by a parameter W, wherein the parameter W is learned to explicitly model correlation between feature channels. In addition, the network adopts a brand-new feature recalibration strategy. Specifically, the importance of each feature channel is automatically obtained through learning, and then useful features are promoted and features that are not useful for the current task are suppressed according to the importance.

3. Method / Model Description

3.1 Dataset

The MNIST database of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.

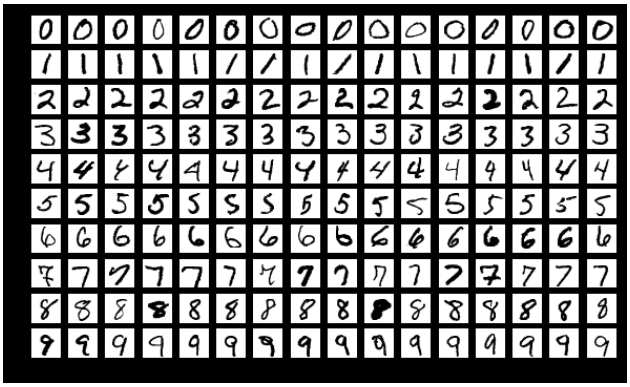

Figure 1. Mnist Example

The CIFAR-10 dataset (Canadian Institute of Advanced Research) is a machine learning and computer vision algorithm commonly used for training images. It is one of the most widely used data sets in machine learning research. The CIFAR-10 dataset contains 60,000 32x32 colour images in 10 different categories. 10 different categories represent aircraft, cars, birds, cats, deer, dogs, frogs, horses, ships and trucks. There are 6,000 images in each category.

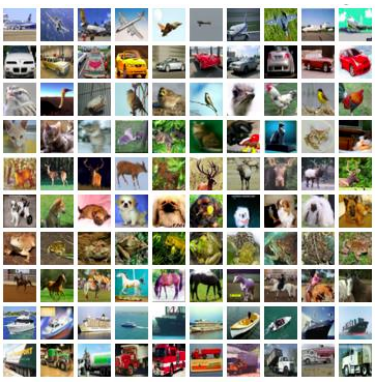

Figure 2. Cifr10 Example

3.2 Data Augmentation

For the two types of data sets described earlier, data augmentation is performed before model training. There are several ways to enhance image data: flip, crop, move, and rotate. Generally, for all data, the size is adjusted to 244x244 and normalized. For all training set data, random adjustment and cropping and random horizontal flipping will be used. Only the size of the mnist data test set will be adjusted to 256 for center cropping, while the Cifr10 test set does not need to perform other processing. In the training of the SEnet network, the training set of Cifr10 will keep the original size input, and do random cropping and horizontal flip between years. The test set does nothing. The Mnist dataset does not do anything.

3.3 Training Process

This article trained a total of three networks: Googlenet, Resnet34, and SEnet. Among them, Googlenet used Adam as the optimizer, and the learning rate setting was 0.0003. It used cross-entropy loss as the loss function. Resnet uses Adamw as the optimizer, the learning rate is set to 0.01, and a fixed step size is used to decay the learning rate. Its loss function is the same as that of Googlenet. The loss function of SEnet is consistent with the above two networks, but

it uses SGD as an optimizer. The learning rate is 1e-1, the weight decay is 1e-4, the impulse is 0.9, and nesterov is set to True.

4. Experiments

For these three networks, we will conduct the following experiments. All three networks will use mnist and cifr10 for training, and resnet34 will try to use PReLU instead of the ReLU function used in the original paper. We will analyse by comparing the accuracy and loss of these networks on these two datasets. At the same time, we give the corresponding confusion matrix and some misclassified results visualization.

4.1 Experiments Results

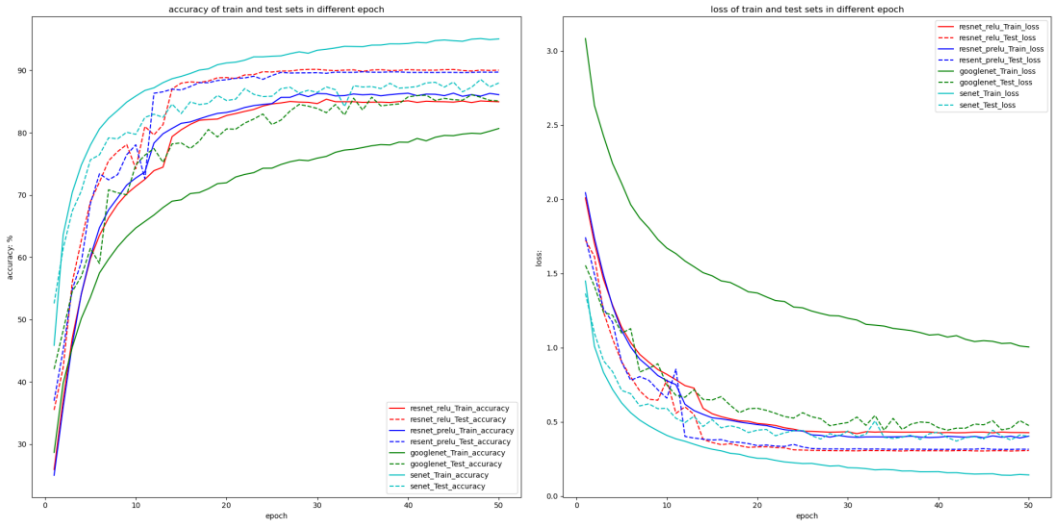

Figure 3. Accuracy & Loss for Three Network (Cifr10)

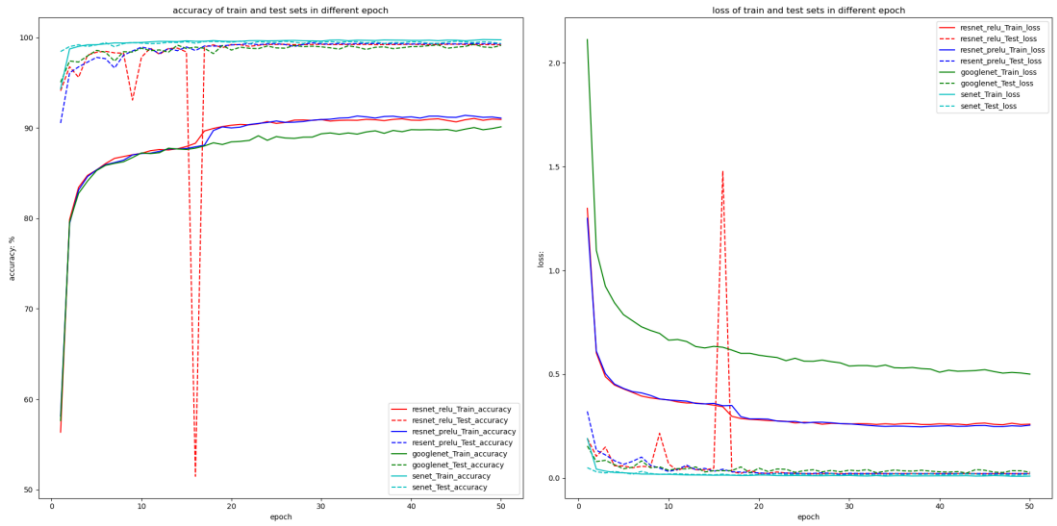

Figure 4. Accuracy & Loss for Three Network (Mnist)

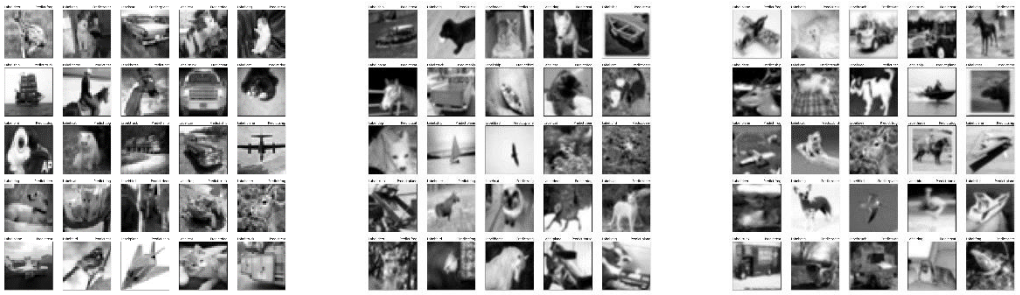

Figure 5. Misclassification for Three Network (Cifr10.From left to right, SEnet, Resnet and Googlenet.)

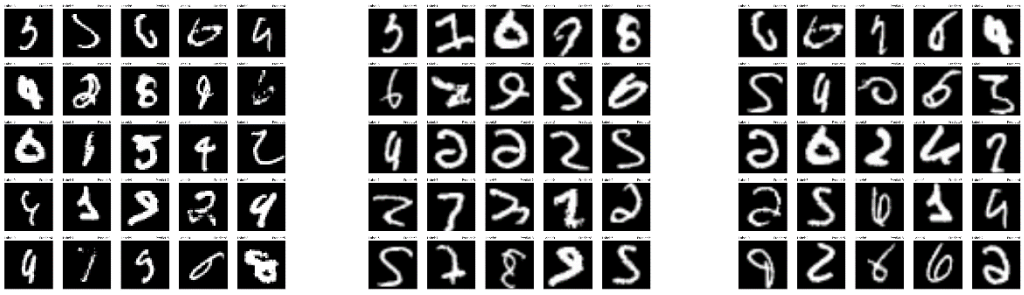

Figure 6. Misclassification for Three Network (Mnist. From left to right, SEnet, Resnet and Googlenet.)

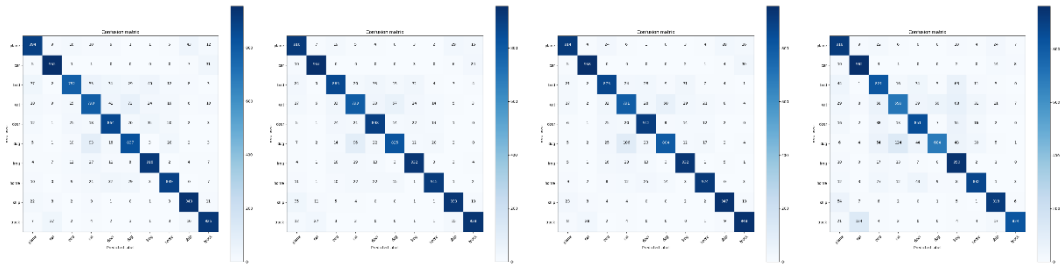

Figure 7. Confusion matrix for Three Network (Cifr10. From left to right, SEnet, Resnet (PReLU), Resnet (ReLU) and Googlenet.)

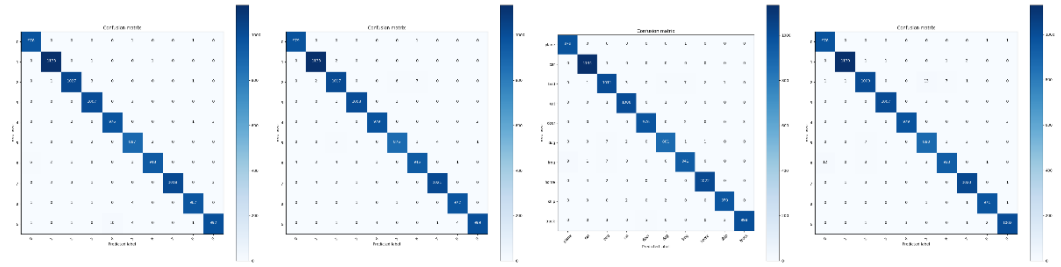

Figure 8. Confusion matrix for Three Network (Mnist. From left to right, SEnet, Resnet (PReLU), Resnet (ReLU) and Googlenet.)

Table 1. Training Time & Best Accuracy (Cifr10)

	Senet	Resnet (PReLU)	Resnet (ReLU)	Googlenet
Cost Time	96m 33s	99m 2s	115m 33s	132m 51s
Best Accuracy	0.885400	0.897700	0.901600	0.860900

Table 2. Training Time & Best Accuracy (Mnist)

	Senet	Resnet (PReLU)	Resnet (ReLU)	Googlenet
Cost Time	73m 2s	141m 38s	141m 42s	131m 59s
Best Accuracy	0.996000	0.994100	0.992600	0.992500

4.2 Analysis

As can be seen from Table 1, Resnet34 can obtain the best test accuracy on Cifr10 data set, and using ReLU can obtain better accuracy than PReLU, but PReLU takes less time than ReLU. Senet is the fastest of the three networks, and its accuracy is second only to Resnet34.Googlenet is the worst of the three networks. However, from fig. 3, it can be found that Googlenet may not have completely converged, and may need to add more algebra, and may get better results. However, we can conclude that Googlenet does not converge as fast as the other two networks. Resnet using PReLU converges faster than ReLU.

As can be seen from table 2, all networks on the Mnist data set can achieve a test accuracy of more than 99%. The three networks have little difference in accuracy. In terms of time expenditure, Senet can still spend the least time. However, Resnet spent the most time, which is in sharp contrast to the results on Cifr10 dataset. Since the Mnist data set is too simple, all networks can converge very quickly from fig. 4. An interesting phenomenon can be found in fig. 4, the curve of Resnet (ReLU) has experienced two oscillations, one smaller but the other larger. This phenomenon may be caused by the use of ReduceLROnPlateau function to dynamically adjust the learning rate.

From the misclassification examples of figs. 5 and 6, a large part of the results leading to misclassification of the model are due to too much acquaintance between the composition of the image and the tag content of the image miscalculation. Most misclassified images of Mnist data sets are difficult to distinguish by human eyes.

5. Conclusion & Future Work

In general, deeper networks can indeed achieve better results. As to how to make the network change deeply and avoid the gradient disappearance (explosion), Googlenet proposes an Inception structure to enable the network to obtain more scale features, and also introduces 1x1 convolution kernel to prevent overfitting and adds an auxiliary classifier to solve the gradient disappearance problem. Resnet proposes a residual network to solve the gradient degradation problem caused by deeper networks and points out the possibility of deeper networks. Senet focuses on extracting more useful information from feature channels to enhance the network. Senet can achieve good results under the condition that the acceptable calculation amount is increased and the calculation speed is not affected. At the same time, this SE module can be easily embedded into mainstream neural networks.

Verify and solve the strange oscillation of Resnet on the Mnist data set in this experiment. At the same time, verify again the overhead time of Resent using different activation functions. Is it because of the GPU server's error that the calculation speed drops, or is it because Resnet's calculation speed on the Mnist data set is very slow?

Reference

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009.

[2] Alex Krizhevsky, IIya Sutskever, Geoffrey Hinton. Imagenet classification with deep convolutional neural networks//Proceedings of Advances in Neural Information Processing Systems, Lake Tahoe, USA, 2012:1097-1105.

[3] A. Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto, 2009.

[4] Christian Szegedy, Liu Wei, Jia Yang-Qing, et al. Going deeper with convolutions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015:1-9.

[5] Lin Min, Chen Qiang, Yan Shui-Cheng. Network in network. arXiv:1312.4400v3, 2013.

[6] He Kai-Ming, Zhang Xiang-Yu, Ren Shao-Qing, et al. Deep residual learning for image recognition// Proceedings of the IEE E Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016:770-778.

[7] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in CVPR, 2016.

[8] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507.