A Deep Research in Classic Classification Network

Hanyue He *

International School, Beijing University of Posts and Telecommunications, Beijing 100876, China
* Corresponding author e-mail:2017213127@bupt.edu.cn

Abstract. This article is a further study and reflection of the classification network after the completion of the article "The Comparison and Analysis of Classic Convolutional Neural Network in the Field of Computer Vision". Different from previous work comparing network performance of VGGNet and ResNet, this paper compares and analyzes the importance of depth, breadth and residual learning of GoogLeNet and ResNet in convolutional neural networks. And considering that most of the time, the basic classification network nowadays appears in the form of backbone in the detection and segmentation network, this paper compares the advantages and disadvantages of several major classification networks as backbones.

1. Introduction

Convolutional neural network is a common deep learning network architecture\cite{17,18,19,20} inspired by a natural visual cognition mechanism in which the visual cortex is hierarchical in the biological vision system (Hubel & Wiesel proposed the concept of receptive field, 1959). Later, in the 1980s, Fukushima proposed the concept of Neocognitron based on the concept of receptive field, which can be regarded as the first realized network of convolutional neural network. The concept of CNN was proposed by Yann Lecun of New York University in 1998, which is essentially a multi-layer perceptron with local connections and weight sharing\cite{1}.

Many excellent network models were invented during the development of CNN, such as LeNet5\cite{2}, AlexNet\cite{3}, VGGNet\cite{4}, GoogLeNet\cite{5} and ResNet\cite{6}. This paper will make a more in-depth comparison and analysis of GoogLeNet and ResNet classical convolutional neural networks based on the article "The Comparison and Analysis of Classic Convolutional Neural Network in the Field of Computer Vision". Among them, GoogLeNet consists of multiple inception blocks, which enables efficient dense computation while using sparse connections; The main idea of ResNet is to add a direct connection channel to the network, which is the idea of the Highway Network. The two structural designs make the network layer deeper and the convolution layer wider, which makes the performance of the convolutional neural network reach a higher level. At the same time, this paper will also compare the advantages and disadvantages of the three most popular classification networks (VGGNet, GoogleNet and ResNet) as the backbone of subsequent complex tasks.

This paper will make a review of the development of the basic classification network in Sect. 2, analyze the performance of GoogleNet and ResNet in Sect. 3, compare the three most popular classification networks (as the backbone) in Sect. 4. Finally in Sect. 5, we will summarize the main features of the two algorithms and propose some views and understandings on future applications.
2. Review
The first typical CNN is LeNet5 [2], but the first network that draws attention is AlexNet [3], as shown in Figure 1. AlexNet is an article published after a model that won the championship in the 2012 ImageNet competition. This article was published in 2012. The model in the article was contested by ImageNet LSVRC-2010. The ImageNet dataset has 1.2 million high-resolution images with a total of 1000 categories. The test set was divided into top-1 and top-5, and AlexNet received error rates of 37.5% and 17%, respectively. This result exceeded the previous level of technology at the time.

Next, in 2014, VGGNet was developed by the University of Oxford Computer Vision Portfolio and Google DeepMind researchers. It explores the relationship between the depth of the convolutional neural network and its performance. By repeatedly stacking 3*3 small convolution kernels and 2*2 maximum pooling layers, 16–19 deep Convolutional Neural Network are successfully constructed. VGGNet won the runner-up for the ILSVRC 2014 competition and champion for the positioning project, with an error rate of 7.5% on top5. So far, VGGNet is still used to extract image features.

3. Comparison
3.1. GoogLeNet
GoogLeNet is a new convolutional neural network structure proposed by Christian Szegedy in 2014[5]. Prior to this, the conventional design idea of CNN is to increase the network depth by continuously stacking convolutional layers and then classify them by multiple layers of fully connected layers. However, this network design idea also has obvious drawbacks. The amount of parameters tends to cause over-fitting of the network and incurs high computational costs. As a result, GoogLeNet appears, which enables efficient dense computation while using sparse connections[7,13].
The entire network is built on the basis of the inception module. The module consists of four branches: 1*1 convolution, 3*3 convolution, 5*5 convolution, and max-pooling (Figure 3). The 1*1 convolution has the function of merging the information in small regions of the previous layer while reducing the dimension of the feature (to prevent the parameter from explosive growing). Using 3*3 convolution and 5*5 convolution to obtain highly relevant information in a large region. The max-pooling layer can play the role of extracting features and is an indispensable part of the convolutional neural network. The inception module designed in this way convolves the image using multi-branch and multi-size convolution kernels, and the convolution kernel (branch) of each size extracts the corresponding high correlation feature. The results obtained by each branch are spliced in the channel dimension, which is equivalent to the aggregation of the high correlation features extracted from each branch in the channel dimension, thus realizing the use of dense computation to approximate sparse connections.

![Figure 3. Inception module[5]](image1)

3.2. ResNet
As the number of network layer increases, people use Xaiver initialization[11] and batch normalization[12] to greatly alleviate the problem of vanishing gradient. However, a new problem in the deep network appears—the degradation problem. When the performance of the network gradually reaches the bottleneck, it will suddenly deteriorate, and the accuracy of this network is not as good as that of a shallower network. To address the degradation problem, a deep residual learning framework was introduced to allow the network to fit a residual mapping instead of directly fitting the underlying mapping, which makes network optimization easier[5,14].

![Figure 4. A residual building block[6]](image2)

![Figure 5. A “bottleneck” building block[6]](image3)

In forward propagation, the residual block has the function of identity mapping of the deep network. And in the process of backpropagation[2,17], the parameters in the network are more sensitive to the loss value because of the reduction of the value of the parameters in the block. By reducing the parameters, the effect of backpropagation loss is increased and a certain regularization effect is generated, which further increase the number of network layers, make the optimization easier and improve the performance. However, it does not solve the problem of vanishing gradients fundamentally and on the extremely deep model of more than 1000 layers, there is still a problem of slow training and over-fitting.
3.3. Evaluation
On the CIFAR-10 and CIFAR-100 datasets, we select GoogLeNet, ResNet-50 and ResNet-101 separately for training. The parameter setting is similar to that in the article "The Comparison and Analysis of Classic Convolutional Neural Network in the Field of Computer Vision"(we use Batch Normalization and ReLU after all convolutional layers, a weight decay of 0.0005, a momentum of 0.9 and a learning rate starting from 0.1). All networks are trained on two GeForce RTX 2080Tis with a batch size of 128. When the number of Epoch reaches 15 and its loss is no longer falling, we use the strategy of adjusting the learning rate dynamically to multiply the original learning rate by a scale factor of 0.2. We follow the simple data augmentation in [15] for training: 4 pixels are padded per side and a 32*32 crop is randomly sampled from the padded image or its horizontal flip. In testing, we only use the original 32*32 image as the input to the network. Besides, we use Nvidia's DALI library in the data loading and preprocessing stage. By loading data through Pipeline and using GPU to perform data augmentation, the loading time of all data is 6 times shorter than before, which greatly improved GPU utilization and therefore accelerate the model training.

![Figure 6. Experimental results on CIFAR-100. Top: train loss. The green curve presents GoogLeNet and the red presents ResNet-50. Middle: train accuracy. The orange curve presents ResNet-101. The purple curve presents GoogLeNet and the blue presents ResNet-50. Bottom: test accuracy. The dark blue curve presents VGG16. The red curve presents ResNet-101 and the light blue presents ResNet-50.](image-url)
The experimental results suggest that our reproducibility of the CNN models is better than the original, which has a lot to do with the change of weight decay and the adjustment of the learning rate. On CIFAR-10, ResNet-101 has the highest accuracy in Top-1 and Top-5 (Top1: 3.91%, Top5: 0%). On CIFAR-100, ResNet-101 has the highest accuracy in Top-1 (21.34%), and GoogLeNet has the highest accuracy in Top-5 (5.66%). In general, the top-performing model is ResNet-101. And from Figure 6, we can find that the convergence speed of ResNet-50 is evidently faster than VGG-16 and GoogLeNet, which proves that the residual learning is effective to alleviating the vanishing gradients and accelerating the training process.

4. Comparison of backbones
Many new networks are designed to achieve high scores on ImageNet data sets. AlexNet was the first network to increase the depth of CNN. In order to reduce the cost of network computing and increase the effective experience field, AlexNet downsampled 32 times, which became a standard for network design later. The VGGNet overlay 3x3 convolution layers to build a deeper network and still contains 32 times downsampling. The latter research is basically based on the VGG network to design each stage a better part. GoogleNet proposes an inception block to contain more features. However, these networks are designed to deal with classification tasks instead of directly dealing with detection and segmentation tasks. But now people have two ways to train and segment the network: a. full training, b. pretrained. Typically, a pretrained network has better performance. In order to perform pretrained operations, people often choose VGGNet, GoogleNet or ResNet (of different layers) as the backbone or encoder of the network. For example: Unet [8] often uses VGG16 as the Backbone of the network, Yolo [9] uses a modified GoogleNet as the backbone, and LinkNet [10] attempts to apply different depths of ResNet as the Backbone.

But the design principle of the classification task is not suitable for the positioning task, because the spatial resolution of the feature map in the traditional network such as VGG16 and Resnet is gradually reduced. Some techniques such as FPN, hole convolution are used for these networks to maintain high spatial resolution. However, when training with these backbone networks, there are several problems: different numbers of network stages, weak visibility of large objects, invisibility of small objects, and most critically——the translation invariance and translational variability.

5. Conclusion
This article analyzes the network structure changes between GoogleNet and ResNet, and compares the two neural network models GoogLeNet and ResNet. As can be seen from GoogleNet and ResNet, the direction of network structure changes is mainly in depth and width. The design strategy of split-transform-merge in GoogLeNet became the design paradigm of CNN. ResNet's residual learning ensures the normal work of the network after deepening. The two algorithms greatly alleviate the vanishing gradient problem, the large amount of model parameters and the degradation problem, whose recognition ability has exceeded the human level. However, problems such as vanishing gradient problem still remain to be solved. Finally, this article compares the performance of several popular classification networks as backbones.

References
[1] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
[2] Lecun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., & Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4), 541-551.

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. NIPS(Vol.25). Curran Associates Inc.

[4] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Computer Science.

[5] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., & Anguelov, D., et al. (2014). Going deeper with convolutions.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition.

[7] Lin, M., Chen, Q., & Yan, S. (2013). Network in network. Computer Science.

[8] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). Unet++: a nested u-net architecture for medical image segmentation.

[9] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015). You only look once: unified, real-time object detection.

[10] Chaurasia, A., & Culurciello, E. (2017). Linknet: exploiting encoder representations for efficient semantic segmentation.

[11] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9), 1-1.

[12] Ioffe, S., & Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. International Conference on International Conference on Machine Learning. JMLR.org.

[13] Arora, S., Bhaskara, A., Ge, R., & Ma, T. (2014, January). Provable bounds for learning some deep representations. In International Conference on Machine Learning (pp. 584-592).

[14] Szeliski, R. (1989). Fast surface interpolation using hierarchical basis functions. IEEE Computer Society Conference on Computer Vision & Pattern Recognition, Cvpr. IEEE.

[15] Lee, C. Y., Xie, S., Gallagher, P., Zhang, Z., & Tu, Z. (2015, February). Deeply-supervised nets. In Artificial intelligence and statistics (pp. 562-570).

[16] #Deep Learning Review#: LeNet, AlexNet, GoogLeNet, VGG, ResNet.(n.d.). Retrieved August 21, 2019, from https://zhuanlan.zhihu.com/p/24657060

[17] Lecun, Y., Boser, B. E., Denker, J. S., Henderson, D., & Jackel, L. D. (1989). Handwritten Digit Recognition with a Back-Propagation Network. Advances in Neural Information Processing Systems 2, [NIPS Conference, Denver, Colorado, USA, November 27-30, 1989]. Morgan Kaufmann.

[18] Jarrett, K., Kavukcuoglu, K., Ranzato, M., & Lecun, Y. (2009). What is the Best Multi-Stage Architecture for Object Recognition?. Proc. International Conference on Computer Vision (ICCV'09). IEEE.

[19] Lecun, Y., Huang, F. J., & Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. IEEE.

[20] (2010). Convolutional networks can learn to generate affinity graphs for image segmentation. Neural Computation, 22(2), 511-538.