

Pre-training CNN model Several Methods of Comparative Analysis

Jun Hu*

School of Information Engineering and Artificial Intelligence, Lanzhou University of Finance and Economics, Lanzhou, Gansu, 730000, China

*Corresponding author: Bryanchiquita@email.phoenix.edu

Abstract. In recent years, the rapid evolution of neural networks, exemplified by models such as VGG and GoogLeNet, has reshaped the landscape of artificial intelligence. This comparative exploration, focusing on transfer learning and attention mechanisms, not only reveals subtle performance distinctions but also provides crucial insights for refining these models in practical applications. This paper delves into the comparative analysis of two prominent neural network models, VGG and GoogLeNet, within the context of image classification. VGG, known for its deep structure, and GoogLeNet, incorporating innovative inception modules, are evaluated through experiments involving flower image datasets. The study includes transfer learning and explores the integration of attention mechanisms, specifically CBAM attention, into both models. Results indicate GoogLeNet's superior performance in terms of parameter efficiency, convergence speed, and overall accuracy. Furthermore, the addition of attention mechanisms enhances classification accuracy for both models. The paper concludes with insights into potential areas for further research, emphasizing optimal attention mechanism placement and a comparison with traditional methods in future studies.

Keywords: VGG, GoogLeNet, transfer learning, attention mechanism.

1. Introduction

Today, with the development of artificial intelligence in the world today, the development of neural networks is getting faster and faster. The emergence of various neural network models has made artificial intelligence a considerable future. For example, the ChatGPT released last year uses the Tranofromer's neural network architecture. The function makes it popular quickly. As one of the mainstream of neural networks now, convolutional neural networks have become a major area in computer vision. The introduction of convolutional layers and pooling layers allows the network to automatically learn the local characteristics and sharing weights in the image The robustness of position and scale change. Therefore, it is widely used in the fields of image classification, and target recognition, and has achieved remarkable achievements.

The rapid development of convolutional neural network models has made many scholars invest in their research and learning of them, and gradually emerge from various capabilities of powerful models. From LENET in the early 1990s to 2012, they won The ImageNet champion's Alexnet, The VGG proposed by the 2014 runner-up visual geometry group, the champion Googlenet, and the RESNET proposed by He Kaiming's team in 2015 has defeated all models in the image classification and other aspects. Get excellent performance. Therefore, through continuous improvement and optimization of convolutional neural network models, and comparative analysis, we can better understand and solve problems in the fields of computer vision, natural language processing, audiovisual processing, and intelligent systems to promote the development and application of artificial intelligence technology.

The article primarily conducts a comparative analysis of the VGG model and the GoogLeNet model. It analyzes their architectures, and parameters, and performs transfer learning on flower-class data using both VGG16 and GoogLeNet models. The article also compares the training results of models with added attention mechanisms to both VGG16 and GoogLeNet. The aim is to provide relevant suggestions and prospects for the future development of these models.

2. Theoretical Introduction

2.1. VGG Model

It is proposed by the visual geometry group of Oxford, which is mainly divided into Vgg16 and Vgg19, Vgg16 contains 13 hidden layers and 3 fully connected layers, and Vgg19 contains 16 hidden layers and 3 fully connected layers. The VGG model all uses a 3×3 convolution kernel and a 2×2 pooling. The VGG model mainly extracts more small features through the combination and stacking of 3×3 convolutional kernels, the receptive field generated by two 3×3 convolutional kernels is the same as the effect of a 5×5 convolutional kernel, three 3×3 convolutional kernels are equivalent to a 7×7 convolution kernel, each convolutional kernel has a step size of 1, a total of 5 pooling layers, the size is 2×2 , and the step size is 2, but the deep network structure model of Vgg is prone to overfitting problems, and because there are too many network layers, it is quite time-consuming to train. However, the number of parameters is effectively reduced by stacking multiple small convolution kernels, with three 3×3 convolution kernels having 27c parameters and one 7 by 7 convolution kernel having 49c parameters. The VGG model also shows that increasing the depth of the network can affect the final performance of the network to a certain extent.

VGG has been widely used in the field of image recognition, and Zhang Jianhua, Kong Fantao et al. accurately identified and distinguished cotton disease species through transfer learning under the improved model based on VGG[1].

2.2. Googlenet Model

Proposed by Google in 2014, the biggest innovation in the design of GoogleNet is the use of multi-scale convolution, local densification of sparse matrix operations, parallel processing of features, and then feature stitching[2]. Googlenet mainly introduces the inception module, which mainly performs multiple convolutions in the same layer, and adds a 1 by 1 convolution kernel in front of each convolution kernel to reduce the dimensionality, effectively reducing the calculation of parameters. The Googlenet model replaces the fully connected layer with average pooling, but the dropout still exists.

Wan Junjie et al. effectively identified orchard pests and diseases through transfer learning under the improvement of the GoogleNet model and found that the performance of the GoogleNet transfer learning was better than that of VGG16 and other models under the same parameter conditions[3]. Chen Bin et al. used the InceptionV3 network structure based on GoogleNet, constructed a remote sensing classification model through transfer learning, and realized the classification of several typical features in the main urban area of Wuhan[4].

2.3. Transfer Learning

When we train a new task, we generally don't start the training again because of the training time constraints and the fact that the training samples are not allowed to take up too much storage space[5]. ImageNet is an initiative aimed at creating an extensive image database, complete with annotations such as pictures and their corresponding labels. Models such as InceptionV1, InceptionV2, VGG-16, and VGG-19, which have been pre-trained, utilize ImageNet's diverse range of image categories.[6]. Transfer learning is to use the weights of the trained model, and then only to retrain the classification part of the fully connected layer, generally if you want to train the whole model even if it is trained on the GPU, it will consume a lot of time, if you use transfer learning, it can effectively shorten the training time, Transfer learning involves setting up a source domain (DS) and a source task (TS), as well as a target domain (DT) and a target task (TT), to enhance the performance of the target function (Ft) [7].

2.4. Attention Mechanism

The attention mechanism is widely present in human daily life. We are constantly receiving various types of information, but ultimately we will pay attention to the key information, which is the attention

mechanism. The attention mechanism mainly weights different parts of the input data, assigning different weights, which can make the model pay more attention to key information. Here we mainly mention the attention mechanisms in convolutional neural networks, which are usually divided into SeNet Attention and CBAM Attention. SeNet Attention mainly adds attention to the channel dimension, while CBAM Attention mainly adds attention in the channel and spatial dimensions. Therefore, CBAM generally has a more significant effect. The spatial attention mechanism primarily operates within each feature map. In traditional convolutional neural networks, equal attention is given to all regions of a feature map. By using the spatial attention mechanism, regions within the feature map are weighted based on their contribution to the classification task, resulting in varying weights for different points within the feature map[8]. The channel attention mechanism, on the other hand, mainly operates between channels. Traditional convolutional neural networks apply the same processing to all channels. Leveraging the inter-channel attention mechanism, we can allocate weights to different feature maps with precision, effectively filtering the information contained within each feature map. Based on the contribution of each feature map, we assign higher weights to those with stronger feature-extraction capabilities, while allocating relatively lower weights to less-contributing feature maps. This process helps emphasize critical features while reducing interference from less important ones.

3. Analysis and Discussion

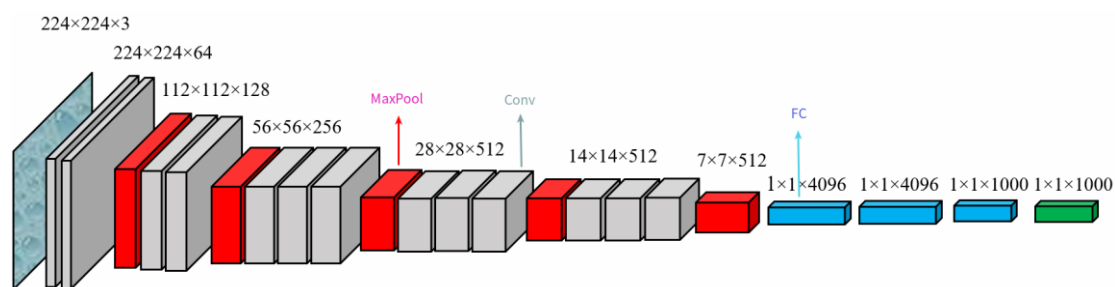


Fig. 1 Vgg network architecture [9]

As shown in the Fig 1, we can find that it consists of 5 VGG blocks, 5 maxpools, 3 FCs, and finally a softmax layer. However, there are some problems with this approach. First, the training process is time-consuming because the network has multiple layers. Secondly, there are many network parameters, and if there is little training data, it is easy to lead to the occurrence of overfitting.

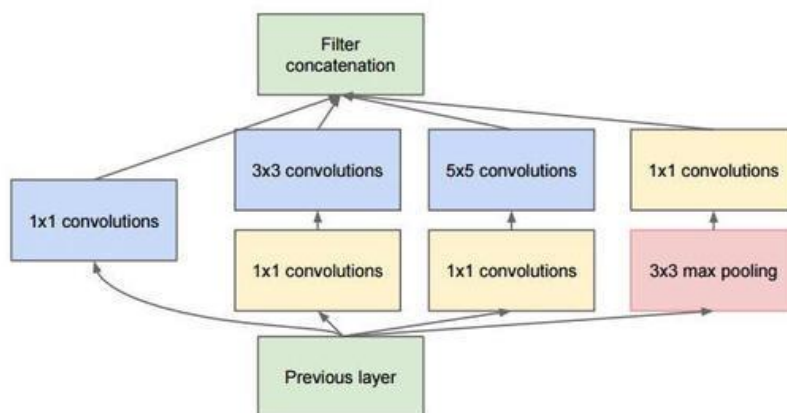


Fig. 2 Inception architecture diagram [10]

As shown in the Fig 2, the Inception module of GoogleNet mainly uses multiple convolutions in one layer, using the convolution kernels of 1×1 , 3×3 , 5×5 , and finally the feature splicing, and then derives the InceptionV2, InceptionV3, and InceptionV4 versions, which make the function more complete through some improvements.

This topic describes the data runtime environment

Pytorch is an open-source deep learning framework developed by Facebook's artificial intelligence research team, which is widely used in business, research, education and other fields, PyTorch supports the use of GPUs for computing, which can greatly improve the speed of training and inference, this article is based on the pytorch deep learning framework, configured as win10, GPU is NVIDIA GeForce MX150, Calculations performed on a platform with 8.00GB of RAM.

The dataset taken in this paper is 5 common flower species, namely daisy, dandelion, rose, sunflowers, and tulips, a total of 3670 photos, which are divided into 3306 training sets and 364 verification sets.

Figures 3 and 4 show the results trained by vgg-16 and googlenet when epoch is 30, batchsize is 32, lr is 0-0001 and the optimizer is adam. Figure 3-5 shows vgg16 and googlenet through the freezing convolution layer and pooling layer under transfer learning. Change the 1000 classifier to 5, and the results are trained when the epoch is 20, batchsize is 32, lr is 0.0001, and the optimizer is adam.

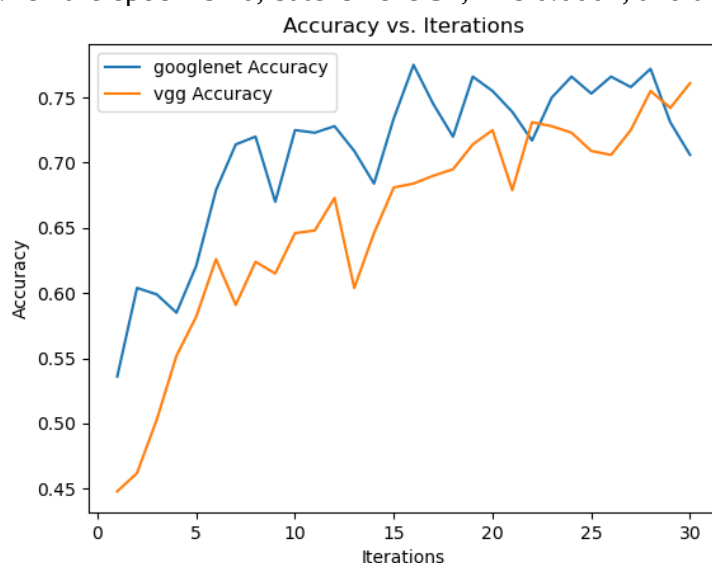


Fig. 3 Comparison chart of accuracy between vgg16 and Google Net (Photo/Picture credit: Original).

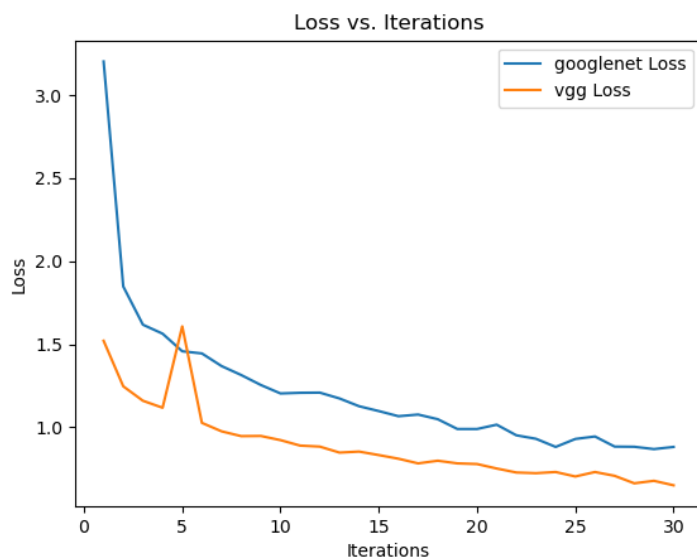


Fig. 4 Comparison of losses between Vgg16 and Google Net (Photo/Picture credit: Original).

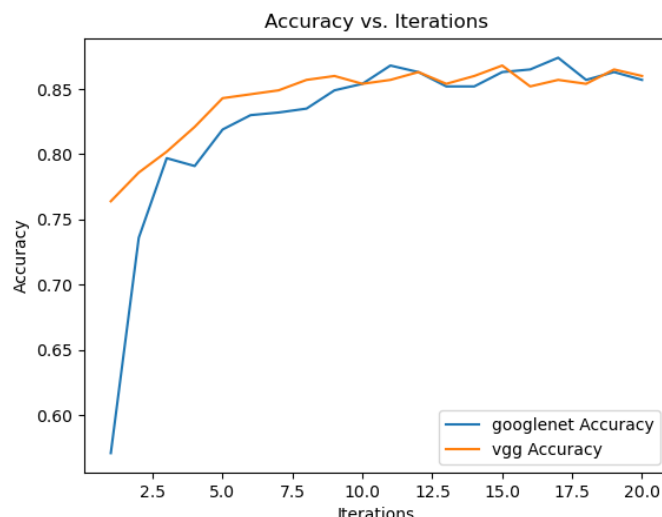


Fig. 5 Comparison chart of transfer learning accuracy between vgg16 and Google Net (Photo/Picture credit: Original).

Fig 3, Fig 4 and Fig 5 show the variation curves of the number of iterations, accuracy, and loss of vgg-16 and googlenet, as well as the variation curves of the number of iterations and accuracy after transfer learning. As the number of iterations increases, the final loss of vgg-16 decreases to 0.615, and the accuracy reaches 0.761. The final loss of googlenet decreases to 0.882, with the highest accuracy reaching 0.772. The highest accuracy obtained after transfer learning of vgg-16 is 0.868. Although the accuracy obtained after transfer learning of googlenet is initially lower than that of vgg-16, as the number of iterations increases, googlenet gradually exceeds vgg-16. It can be seen from the above figure that in comparison, the accuracy obtained by googlet is higher than that of vgg16. Higher, as the number of training sessions increases, the Google T loss decreases faster and converges faster.

The following Fig 6 represents the change in loss and accuracy with the number of iterations when channel attention and spatial attention are added to VGG16 and GoogLeNet based on transfer learning. In this experiment, CBAM modules were added after the feature extraction layers of VGG16, and only the convolutional layers were frozen. The model output was modified to a 5-class model through fine-tuning, and then the newly added attention modules were trained. In GoogLeNet, CBAM modules were added in front of each auxiliary classifier and in front of the final output layer, and the output of each auxiliary classifier and the final main classifier was changed to a 5-class model. All parameters were then frozen, and the corresponding attention modules, as well as the output classifier module, were unfrozen, followed by model training.

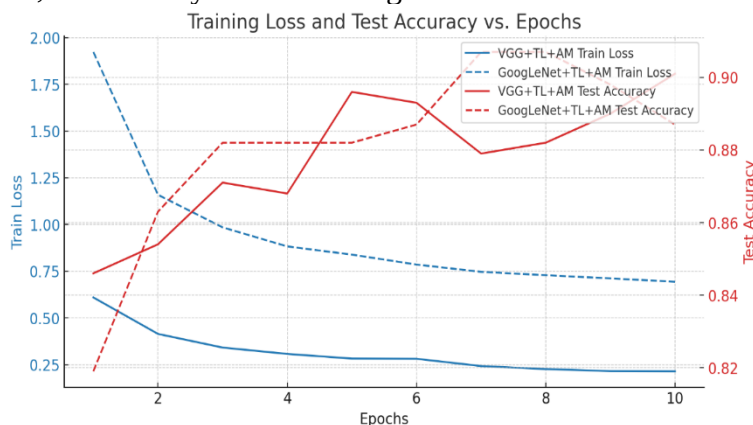


Fig. 6 The comparison between CBAM-enhanced VGG16 and GoogLeNet based on transfer learning (Photo/Picture credit: Original).

Below are the class activation maps for the two models, visualizing the attention mechanism. The deeper the color region, the greater the contribution to image recognition.



Fig. 7 Origin picture (Photo/Picture credit: Original).

The Fig7 image is from our validation set, and it is a picture of a tulip. In order to reduce the impact of the background on the recognition of its flowers, a spatial attention mechanism has been introduced into it.

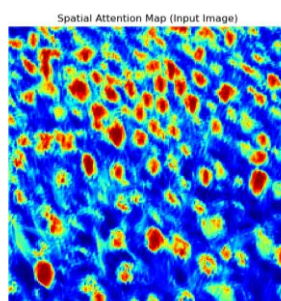


Fig. 8 Class activation mapping with VGG16 (Photo/Picture credit: Original).

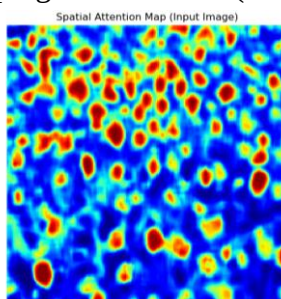


Fig. 9 Class activation mapping with Googlenet (Photo/Picture credit: Original).

There is a certain difference in the attention distribution between the two images. In the attention map of Fig8, the focused regions appear to be more scattered, and warm-colored areas are relatively evenly distributed, indicating that the model pays attention to multiple areas when recognizing the image. In contrast, in the attention map of Fig9, the focused regions seem to be more concentrated, especially on specific tulip flowers, suggesting that the model may be more focused on specific features or patterns.

As shown in the Fig8, although there are multiple focused regions, the intensity of attention appears to be less concentrated compared to the image from Fig9. As shown in the Fig9, the most focused areas have more vivid colors, indicating a higher level of attention intensity. This may suggest that GoogLeNet is more effective at recognizing key features when processing images of this kind.

4. Conclusion

Through comparison, it was found that GoogLeNet outperforms VGG in terms of performance. Whether it's the original models or models based on transfer learning, GoogLeNet has fewer parameters than VGG, higher accuracy, faster convergence, and effectively improves model performance through transfer learning. This study also trained VGG16 and GoogLeNet by adding

attention mechanisms and found that attention mechanisms can significantly enhance accuracy. In this flower recognition model experiment, both models enhanced with attention mechanisms improved accuracy by exactly 3.3 percentage points compared to models based solely on transfer learning. In comparison, the GoogLeNet model with added attention mechanisms based on transfer learning achieved a higher accuracy than the corresponding VGG16, reaching up to 90.7%, while VGG achieved 90.1%. Although attention mechanisms were added in this study, the optimal positions for adding attention mechanisms require further experimental research. In this experiment, attention mechanisms were only added in front of the max-pooling layer before the model's classifier. Perhaps adding them at other positions could yield more optimized results. Furthermore, this study did not compare with traditional methods, so the next research steps will focus on finding the optimal positions for adding attention mechanisms in comparison to traditional methods.

References

- [1] Zhang Jianhua, Kong Fantao, Wu Jianzhai et al. Cotton disease recognition model based on improved VGG convolutional neural network. *Journal of China Agricultural University*, 2018, 23(11): 161-171.
- [2] Fu Ruonan. Research on object Detection based on Deep learning. Beijing Jiaotong University, 2017.
- [3] Wan Junjie, Qi Lijun, Lu Zhongao et al. Identification and classification of GoogLeNet orchard pests and diseases based on transfer learning. *Journal of China Agricultural University*, 2019, 26(11): 209-221.
- [4] Chen Bin, Wang Hongzhi, Xu Xinliang et al. Automatic classification of medium resolution remote sensing images based on GoogleNet model of deep learning. *Surveying and mapping bulletin*, 2019, (6): 29-33 + 40.
- [5] Qin Min. The flower image classification recognition model based on deep learning research. Guangxi Normal University, 2020.
- [6] Tammina, Srikanth. Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. *International Journal of Scientific and Research Publications (IJSRP)*, 2019.
- [7] Feng Guohui. Small-scale image classification based on Convolutional Neural network VGG model. Lanzhou University, 2018.
- [8] Zhang Mengyu. Flower Recognition Based on ResNet and Attention Mechanism. *Computer and Modernization*, 2021, (04): 61-67.
- [9] Zhou Feiyan, Jin Linpeng, Dong Jun. Review of convolutional neural networks. *Chinese Journal of Computers*, 2017, 40(06):1229-1251.
- [10] Zhang Chi, Guo Yuan, Liming. Review on development and application of Artificial neural network model. *Computer Engineering and Applications*, 2021, 57(11): 57-69.