Journal of Image Processing Theory and Applications (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jipta.2024.070110
ISSN 2560-6239 Vol. 7 Num. 1

Research on Image Recognition Based on Different Depths of VGGNet

Handong Song[1,*], Huimin Lu[2]

[1]*School of Mechanical Engineering, University of Jinan, Jinan, 250024, China*
[2]*School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, 430074, China*
[*]*Corresponding author: 2543520357@qq.com*

Keywords: VGGNet, Convolutional neural network, Image recognition, Deep learning

Abstract: With the advancement of image recognition technology, the significance of convolutional neural networks has continually increased. The VGGNet model, developed by the Visual Geometry Group at the University of Oxford, has proven successful, demonstrating that the depth of the network is crucial for model performance. This study aims to explore the impact of various depths of VGGNet models on image recognition tasks. Three classic VGG network models were selected: VGG-13, VGG-16, and VGG-19, along with two widely used image datasets, MNIST and CIFAR-10, for a comprehensive experimental analysis. The experimental results on the CIFAR-10 dataset indicated that as network depth increased, there was a significant enhancement in model accuracy, with VGG-19 performing the best. This outcome confirms the superiority of deep networks in processing complex image data. Conversely, in the simpler MNIST dataset, the deeper VGG-19 did not exhibit better performance compared to VGG-13, suggesting that excessively deep networks might not be necessary for simple datasets and could lead to overfitting and gradient vanishing.

1. Introduction

Image classification is an emerging technology that has developed rapidly in recent years. Its primary research tasks involve image categorization and feature description. In the history of deep learning, the rise of Convolutional Neural Networks (CNN) undoubtedly marks a significant milestone. The application of CNN in the field of image recognition has greatly advanced related technologies. Before the advent of VGGNet in 2014, several CNN architectures had already made notable contributions to image recognition tasks. As early as 1998, LeNet-5, one of the earliest convolutional neural networks, was used for handwritten digit recognition. Subsequently, AlexNet made a spectacular impact at the 2012 ImageNet challenge. Its deep network structure and the use of the ReLU activation function significantly enhanced the accuracy of image classification and garnered widespread attention from both the academic and industrial communities.

Despite the performance breakthroughs achieved by early CNN models such as AlexNet and ZFNet, these models still exhibit several significant shortcomings. For instance, they have relatively low architectural complexity, which limits their ability to fully utilize deeper abstract features, and

they are prone to overfitting when trained on large-scale image datasets[1]. Additionally, the configuration of convolutional and pooling layers in these early models is not optimally designed, impacting the learning efficiency and expressiveness of the models.

In response to these issues, the introduction of VGGNet marked a new phase in network depth and architectural standardization. VGGNet effectively increased network depth by repetitively using small 3x3 convolutional kernels and 2x2 max pooling layers, which not only enhanced the model's learning capabilities but also improved the local receptive fields of the features. This allowed the model to capture more detailed image nuances. Additionally, VGGNet simplified the training process of deep networks by constructing deeper architectures using standardized building blocks. By using multiple smaller filters in place of a single larger filter, VGGNet achieved more refined processing of input data, thereby overcoming some limitations of earlier models[1].

This study aims to assess the specific impact of network depth on the accuracy of image recognition by comparing the performance of different VGG models, and to explore how to further optimize deep convolutional networks to handle more complex visual tasks in the current context of deep learning technology.

In this paper, the second section presents related work on the development of deep learning. The third section details the network architecture of the VGGNet model. The fourth section reports the experimental results, explaining how the experiments were conducted and what data were obtained. The fifth section concludes the paper by summarizing and synthesizing the results of the experiments.

2. Related work

The early origins of deep learning can be traced back to the neural networks and perceptrons of the late 1950s. The perceptron, introduced by Rosenblatt in 1957, was one of the earliest artificial neural networks and was designed for simple binary classification problems. Although its structure was basic, the perceptron represented a preliminary attempt to mimic the way neurons work in the human brain. In the 1980s, the development of the multilayer perceptron (MLP) ushered in a new phase of neural network research. The MLP is a type of feedforward neural network that includes one or more hidden layers capable of capturing complex patterns and relationships in input data[2], as shown in Figure 1. In 1986, Rumelhart, Hinton, and Williams introduced the backpropagation algorithm, an effective method for training MLPs. By calculating error gradients and updating network weights to minimize errors[3], the backpropagation algorithm significantly advanced the research and application of multilayer neural networks.

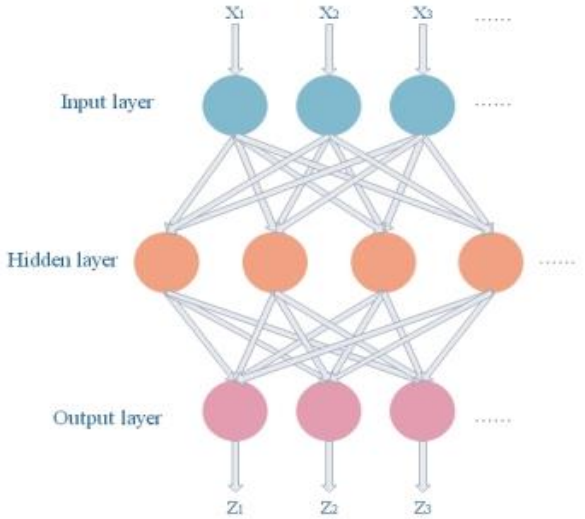

Figure 1: Basic structure of MLP

In 1989, the convolutional neural network marked a major breakthrough in the field of image processing within deep learning. LeCun and colleagues designed LeNet-5[4], the first convolutional neural network to be successfully applied to digit recognition. Convolutional neural networks utilize convolutional layers to automatically extract features from images, eliminating the need for manually set filters or feature detectors.

Convolutional layers operate on input images using filters to capture local features. Each filter corresponds to a specific feature type and produces a strong activation response when it passes over the corresponding feature. These local features are gradually combined into more sophisticated semantic information, such as parts of objects and overall structures, as the network deepens.

Convolutional neural networks also include pooling layers, which are responsible for reducing the spatial dimensions of feature maps, thereby decreasing the amount of computation and enhancing the robustness of the features. Pooling operations typically use either max pooling or average pooling to help the network maintain invariance to minor variations in the input while retaining critical information. The success of LeNet-5 not only demonstrated the efficiency and superiority of convolutional neural networks in processing image data but also laid the foundation for subsequent deeper and more complex convolutional network models such as AlexNet, VGGNet, and others.

In 2012, AlexNet, developed by Krizhevsky, Sutskever, and Hinton, achieved groundbreaking results in the ImageNet challenge[5], marking the advent of the deep learning era. AlexNet significantly improved the accuracy of image classification through the use of ReLU activation functions, multi-layer convolutional structures, Dropout, and data augmentation techniques[6]. ReLU stands for Rectified Linear Unit, as indicated in equation (1). It is a commonly used neural activation function. The ReLU function is essentially a piecewise linear function that turns all negative values to zero while keeping positive values unchanged, a process known as unilateral inhibition[7].

$$ReLU(x) = \begin{cases} x & if\ x > 0 \\ 0 & else \end{cases} \tag{1}$$

The development of the VGG network represents a significant milestone in the history of deep learning. The most famous models from the VGG network include VGG-16 and VGG-19, which performed exceptionally well in the 2014 ImageNet challenge. The main contribution of the VGG models lies in their use of repetitive, standardized convolutional layer structures to construct very deep network architectures. This design philosophy greatly simplified the configuration and optimization of deep networks and achieved better training results on large-scale datasets.

3. The VGGNet model

The VGG networks were developed by the Visual Geometry Group at the University of Oxford [8] and are renowned for their concise yet comprehensive network structures. The primary design concept behind the VGG network is to enhance model performance by increasing its depth, while employing a unified network architecture to simplify both design and optimization processes. Figure 2 illustrates the structural diagram of the VGG network. Previous models such as AlexNet and ZFNet have already begun utilizing deeper network structures with successful outcomes; however, their network designs also exhibit certain irregularities (e.g., inconsistent convolution kernel sizes and stride lengths). In contrast, VGG networks explore the impact of depth on performance through several key design decisions.

- **The use of consistent convolutional kernel sizes**: The VGG network maintains the use of 3x3 convolutional kernels across all convolutional layers. This smallest kernel size enables the capturing of spatial relationships between pixels. Employing these small kernels allows for increased network depth while controlling the number of parameters and computational complexity.
- **Increasing network depth:** By repetitively stacking convolutional layers with small kernels,

the VGG network achieves a deeper network structure, with the option to select the number of layers, up to a maximum of 19 layers. This depth significantly enhances the network's learning capacity, enabling the model to capture more complex image features.

- **Regular inter-layer structure:** After several convolutional layers, the VGG network employs 2x2 max pooling layers for feature downsampling, aiding in reducing computational load and preventing overfitting, while maintaining the regularity of the network structure.

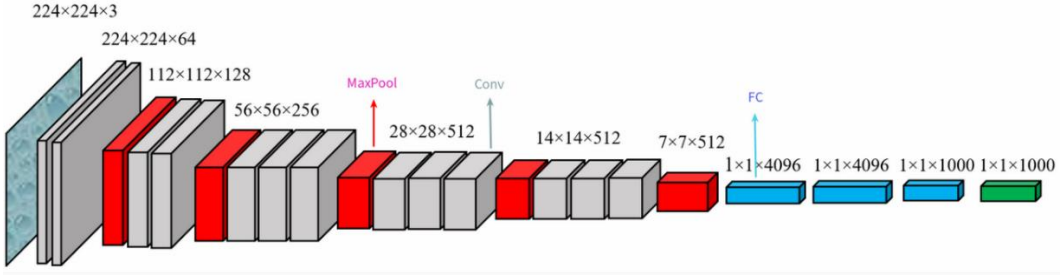

Figure 2: VGG network structure diagram

The VGG-13 model consists of 13 convolutional layers and 3 fully connected layers. The model starts with two 3x3 convolutional layers, each followed by a ReLU activation function. Typically, after every two convolutional layers, there is a 2x2 max pooling layer to reduce the size of feature maps. This pattern repeats five times, with the number of channels in the convolutional layers being 64, 128, 256, 256, 512, and 512, respectively. Following the convolutional layers, the network includes three fully connected layers, with each of the first two fully connected layers containing 4096 units, and the last fully connected layer containing 1000 units (corresponding to the output for 1000 classes). Finally, there is a softmax layer for classification output. The VGG-16 and VGG-19 models have a similar structure to VGG-13, with an increased number of convolutional layers, pooling layers, and fully connected layers accordingly.

4. Results

The purpose of the experimental results section is to analyze the differences in recognition accuracy among different VGGNet models on the same dataset. This helps in determining which VGGNet model to choose when dealing with a simple or complex dataset. By comparing the recognition accuracy of VGG-13, VGG-16, and VGG-19 networks on the MNIST dataset (a simple dataset) and the CIFAR-10 dataset (a complex dataset), we can evaluate and analyze the performance of different VGGNet models on simple and complex datasets. The accuracy is calculated using equation (2), where the number of correctly recognized instances within a single epoch is divided by the total number of instances to obtain the accuracy for that epoch.

$$testAcc = \frac{testAcc.Epoch}{testNum} \tag{2}$$

4.1 Validation on the mnist dataset

Initial experiments were conducted on the MNIST dataset, a commonly used handwriting recognition dataset in the fields of deep learning and machine learning. The MNIST dataset consists of 60,000 training samples and 10,000 test samples. Each sample comprises a 28×28 pixel grayscale image representing one digit from 0 to 9. Figure 3 illustrates a selection of these digits.

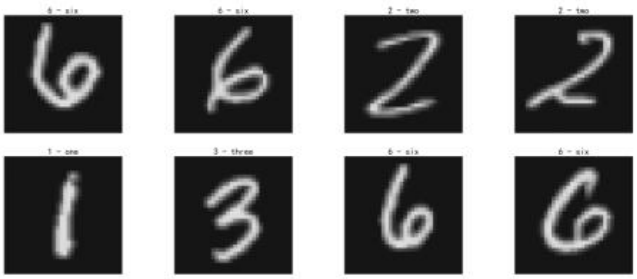

Figure 3: MNIST data set partial sample

On this dataset, the VGG-13, VGG-16, and VGG-19 networks, all with a learning rate of 0.0001, were trained over 15 epochs. Accuracy was then tested on the test dataset, with the results depicted in Figure 4. Specific accuracy values are provided in Table 1.

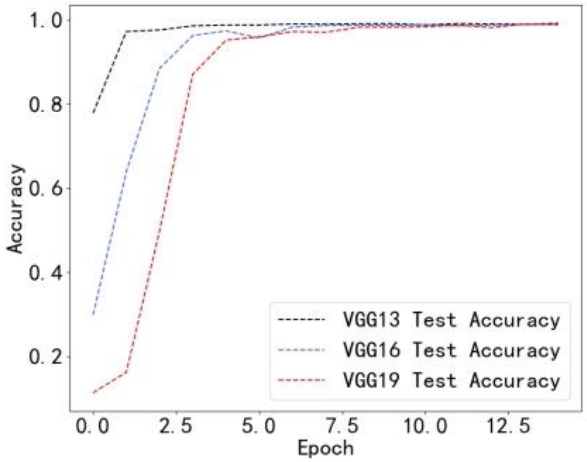

Figure 4: Test set accuracy diagram

Table 1: Test set accuracy

Type of network	Epoch=2 accuracy rate	Epoch=5 accuracy rate	Epoch=10 accuracy rate	Epoch=15 accuracy rate
VGG-13	57.3%	96.7%	98.9%	99.2%
VGG-16	90.7%	96.8%	99.1%	99.3%
VGG-19	98.8%	99.5%	99.4%	99.4%

The graphical and tabular data presented indicate that all three networks achieved excellent results on the test set with the same number of training epochs, reaching accuracies above 99% by the fifteenth epoch. However, the deeper VGG-19 network did not demonstrate superiority on this dataset; in fact, its accuracy occasionally fell below that of the shallower VGG-13 network. This suggests that for simple dataset recognition tasks, the complexity of VGG-19 might be excessive. A simpler VGG-13 network may suffice to capture essential features, whereas the more complex VGG-19 network, due to its intricacies, might struggle with optimization issues such as overfitting, vanishing gradients, and exploding gradients—problems that are particularly pronounced in the early epochs.

4.2 Validation on the CIFAR-10 dataset

Continuing with the CIFAR-10 dataset, it consists of 60,000 color images distributed across 10 categories, with each category containing 6,000 images. This dataset is divided into 50,000 training

images and 10,000 testing images. Each image has a resolution of 32 ×32 pixels. A selection of these images is displayed in Figure 5.

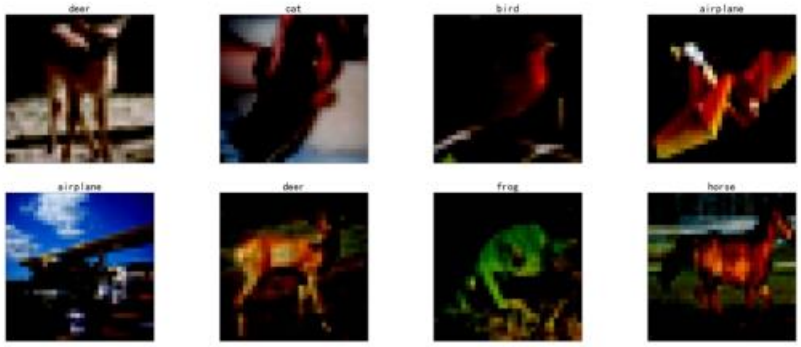

Figure 5: CIFAR-10 data set partial sample

Maintaining the same learning rate, the CIFAR-10 dataset's complexity was addressed by training the VGG-13, VGG-16, and VGG-19 networks over 30 epochs. The accuracy performance on the test set is depicted in Figure 6, with specific accuracy values detailed in Table 2.

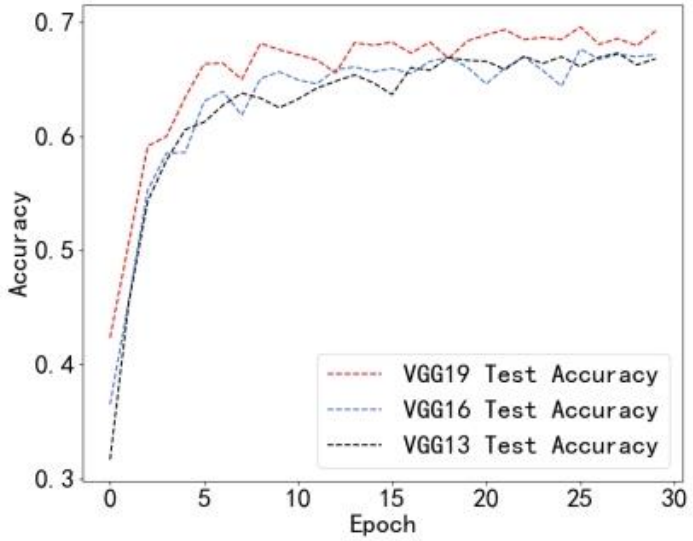

Figure 6: Test set accuracy diagram

Table 2: Test set accuracy

Type of network	Epoch=2 accuracy rate	Epoch=5 accuracy rate	Epoch=10 accuracy rate	Epoch=20 accuracy rate	Epoch=30 accuracy rate
VGG-13	45.5%	60.6%	62.5%	66.6%	66.8%
VGG-16	45.6%	58.6%	65.6%	66.0%	67.1%
VGG-19	50.7%	63.4%	67.6%	68.3%	69.2%

The results illustrated in the graph and table reveal that the VGG-19 network performed the best on the test set, while the performances of VGG-16 and VGG-13 were relatively close, with VGG-16 still outperforming VGG-13 overall. This outcome can be attributed to the additional convolutional and fully connected layers in the VGG-19 and VGG-16 networks, which allow them to capture more critical image features more effectively.

5. Conclusions

Based on the experimental results across different datasets using the same network architectures, it is evident that network depth significantly enhances accuracy, underscoring the importance of depth for model performance. Deeper networks can more effectively capture and represent complex data features, thereby delivering superior performance on more challenging datasets. However, when testing the relatively simple MNIST dataset, the VGG-19 did not show a significant advantage over VGG-13; in fact, in the early stages of training, VGG-19's performance was inferior to that of VGG-13. This observation suggests that for simpler datasets, overly deep networks may lead to overfitting and optimization difficulties, adversely affecting the model's practical performance. Therefore, when dealing with simpler datasets, opting for shallower network architectures may be more appropriate to avoid overfitting and the wastage of computational resources.

References

[1] Li D, Deng L, Cai Z .Research on image classification method based on convolutional neural network[J].Neural Computing and Applications, 2020, 33(14):1-11.

[2] Fajri D , Mahmudy W , Yulianti T .Detection of Disease and Pest of Kenaf Plant Based on Image Recognition with VGGNet19 [J]. Knowl. Eng. Data Sci. 2021.DOI:10.17977/um018v4i12021p55-68.

[3] Umang P, Shruti B, Avik H .Enhancing cross-domain transferability of black-box adversarial attacks on speaker recognition systems using linearized backpropagation[J].Pattern Analysis and Applications, 2024, 27(2)

[4] Song X, Fan L. Human Posture Recognition and Estimation Method Based on 3D Multiview Basketball Sports Dataset [J].Complexity, 2021. DOI:10.1155/2021/6697697.

[5] Srivastava N, Hinton E G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. [J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.

[6] Liu Aili, Chen Zhixiang. ReLU depth network structure and activation function approximation [J]. Journal of shaoxing liberal art &science college, 2024, 44 (02): 58-68.

[7] Liao Qing-Jiang, Liu Ting, Zhang Xing-Yue, et al. For facial expression recognition based on VGGNet study [J]. Journal of software engineering, 2023, 26 (11): 59-62.

[8] Simonyan K, Zisserman A .Very Deep Convolutional Networks for Large-Scale Image Recognition. [J]. CoRR, 2014, abs/1409.1556