Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/79/20241398

VGG and InceptionV3 model based on CIFAR data contrast analysis

Yilin Li[1,*,†], Miao Qin[2,†], Zijie Tang[3,†]

[1]University of Ningbo Nottingham
[2]Shandong University
[3]BASIS International School Nanjing

*Corresponding author email: liyilin251@gamil.com
[†]Yilin Li, Miao Qin and Zijie Tang contributed equally to this work and should be considered co-first authors.

Abstract. This paper introduces in detail the performance comparative analysis of VGG and InceptionV3 based on CIFAR-100 data set in image classification tasks. The experimental results show that the InceptionV3 model performs best on the CIFAR-100 dataset, and its high accuracy and balanced classification effect are impressive. In contrast, the VGG model, while simple in structure, is slightly less accurate. Further analysis shows that InceptionV3 model has more advantages in feature extraction and fusion design, which makes it perform well in image classification tasks. Additionally, the paper explores the broader applications and future prospects of the studied models. By doing so, it provides valuable insights into potential research directions for model comparison. This comprehensive analysis serves as a benchmark, shedding light on the strengths and weaknesses of VGG and InceptionV3 models in image classification. It stands as a valuable reference for future developments in comparative model research.

Keywords: VGG model, InceptionV3 model, CIFAR-100 data set.

1. Introduction

In recent years, with the rapid development of computer vision technology, it has become particularly important to classify massive images efficiently and accurately. Image classification technology is widely used in machine vision, disease diagnosis, object detection, and other fields, with great demand in industry, medicine, military, daily life, and other fields. The research on image classification has extraordinary significance for the basic research and practical application of computer vision.

Image classification techniques train models through datasets, dividing different images into different categories to achieve the minimum classification error. This paper mainly introduces the VGG and InceptionV3 models based on the CIFAR-100 dataset.

The VGG model [1] ranked second in the 2014 ILSVRC competition, extracting CNN features from images, and the VGG model was the preferred algorithm. Its disadvantage is that it has as many as 140M parameters and requires larger storage space. But this model has great research value.

Inception v3[2] is the third version of the Inception Network series, which has achieved excellent results in ImageNet image recognition competitions, especially in large-scale image recognition tasks.

DOI: 10.54254/2755-2721/79/20241398

This article will compare and analyze the performance of two classic deep convolutional neural network models, VGG, and InceptionV3 in image classification tasks. The first three part of this paper discusses the relevant research on the CIFAR-100 dataset, VGG model, and InceptionV3 mainly introducing the characteristics of the aforementioned dataset and the principles and characteristics of the models; The fifth part describes the methods and techniques used in the experimental process, explains the network structure, hierarchy, and feature extraction methods of VGG, InceptionV3, and discusses common techniques used in image classification tasks; The sixth part describes the design of this experiment, detailing the experimental settings, including dataset partitioning, training parameters, evaluation indicators, etc., and explaining why the CIFAR-100 dataset was chosen as the experimental object, as well as the reasons for choosing these three models; The seventh part describes the experimental results and analysis, providing performance comparison results of VGG and InceptionV3 on the CIFAR-100 dataset, and analyzing the experimental results to explore the performance of different models in different categories and the reasons for performance differences; The last part discusses the application fields and prospects of the studied models, explores the application fields of VGG and InceptionV3 models outside of image classification tasks, and looks forward to the development direction of future model comparison research.

The experimental results show that Inception V3 performs best on the CIFAR-100 dataset, with the highest accuracy and relatively balanced classification performance for each category. Although VGG has a simple structure, its accuracy is slightly lower. Analyzing the reason may be due to InceptionV3's more reasonable and effective design in feature extraction and fusion [3].

2. Overview of VGG Model

VGG16 model was firstly published by VGG Group of Oxford University. It uses a few consecutive 3x3 kernels instead of some large convolution kernel like Alex Net who use 11x11 and 5x5 kernel [4]. Using a few small convolution kernels works better than using larger consecution kernel in some special case, because multilayer nonlinear layer has deeper network depth which can learn more complex patterns, and it has lower cost.

The VGG model's key features and advantages are depth structure, unified convolutional kernel size, simple yet effective structure, performance on ImageNet [5]. The depth architecture of VGG model allow it to learn some complex and abstract features, so show high quality performance in image recognition tasks. It uses a uniform 3x3 convolutional kernel which is useful to improve the network's expressive capability and reduce the number of parameters, so it will effectively improve the efficiency of capturing image features. At the same time, the VGG model use a straightforward convolutional block structure, so it is easy to train, understand, and wildly used to computer vision tasks. In ImageNet Large Scale Visual Recognition Challenge (ILSVRC), VGG model made remarkable achievements which showed its superior performance in image classification projects [6].

The VGG model is not perfect, it still needs to face a few challenges, or it has some limitations. It has many parameters, high computational resource requirements, and a not suitable for lightweight deployments. This cause that the module will ask for substantial computational resources and storage space, posing challenges in resource-constrained environments, and it may not be as efficient as other model in scenarios that prioritize model size and computational efficiency.

DOI: 10.54254/2755-2721/79/20241398

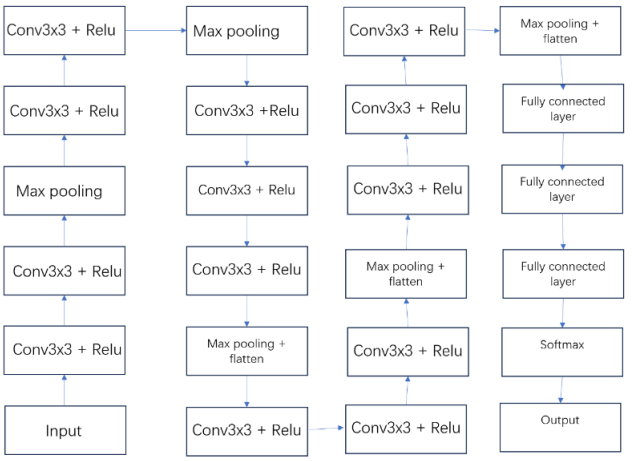

Figure 1. The main structure of VGG16

3. Overview of Inception V3 Model

Inception-V3 model is an image classification model, it usually asks pictures as input to make the predictive modelling. Its main characteristics are deeper network structure, using factorized convolutions and batch normalization. Those characteristics allow the network to extract more levels of features and thus achieve better result on image recognition tasks, the number of parameters of the network is reduced to reduce the computational complexity while maintaining good performance, and it contributes to the convergence and generalization ability of the network.

The Inception V3 model's key features and advantages are its model, global average pooling, pre-trained model [7]. The Inception V3 model use the inception model which includes multiple different kernels and pooling layers of sizes. This kind of design allow the model to catch up capture features at different size and improve the expressive capacity. The Inception V3 model also use global average pooling at the output layer which replace the traditional full connected layers, reduce the number of parameters, mitigate overfitting, and improve computational efficiency. Also, the Inception V3 model can be used in most of the computer vision tasks and have good transfer learning performance, because the model is always pre-trained on large-scale image datasets.

On the other hand, the Inception V3 model still have limitations like the high computational resource requirements and the model size [7]. The depth and complexity of Inception V3 model may require significant computational resources during training and inference and the size of the model has a large model size, which make it unsuitable for resource-constrained environments.

Table 1. The main structure of inception v3

type	Patch size/stride	Input size
Conv	3x3/2	299x299x3
Conv	3x3/1	149x149x43
Conv padded	3x3/1	147x147x32
Pool	3x3/2	147x147x64
Conv	3x3/1	73x73x64
Conv	3x3.2	71x71x80
Conv	3x3/1	35x35x192
3xInception	Figure 2	35x35x288
5xInception	Figure 3	17x17x1280
2xInception	Figure 4	8x8x2048
Pool	8x8	8x8x2048
Linear	Logits	1x1x2048
softmax	Classifier	1x1x1000

DOI: 10.54254/2755-2721/79/20241398

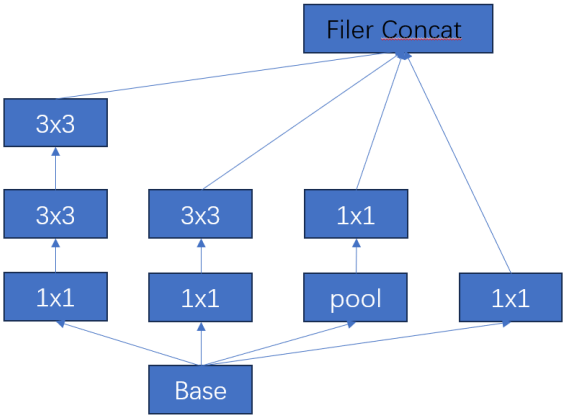

Figure 2. The 5×5 convolution is decomposed into two 3×3 convolution operations to improve the calculation speed. This effectively uses only ~ (3x3 + 3x3)/(5x5) = 72% of the computational overhead.

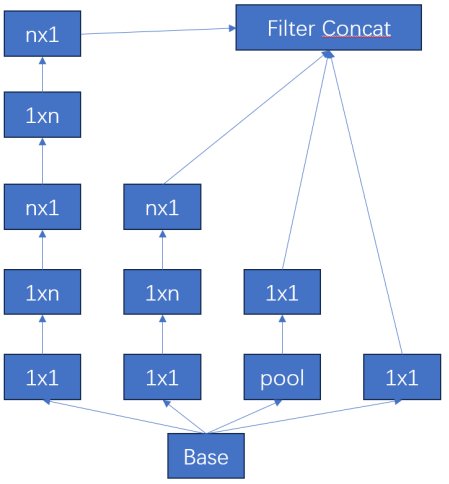

Figure 3. Each nxn conv can be replaced with two convs layers of 1xn and nx1 to save computation and memory.

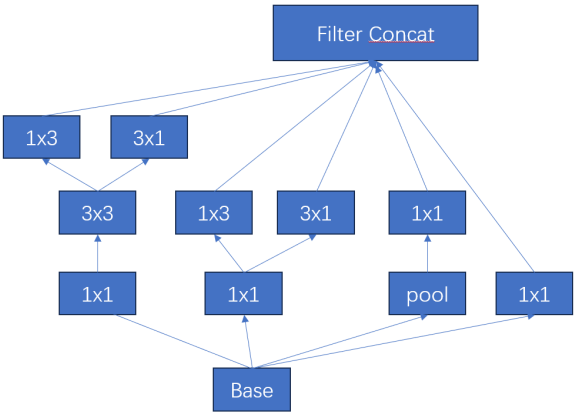

Figure 4. The filter banks in the module are expanded (i.e., made wider rather than deeper) to address the representational bottleneck. If the module is not expanding the width, but becomes deeper, so the dimension will be reduced too much, cause information loss.

DOI: 10.54254/2755-2721/79/20241398

4. Overview of Cifar-100 Dataset

CIFAR-100 is a widely used image classification dataset that has been chosen for analysis due to its widely used in the computer vision and deep learning domains. The dataset includes 100 different categories with 600 color images, each of the image has 32x32 pixels. This makes CIFAR-100 representative for evaluating the performance of image classification algorithms [8].

The three main characteristics of CIFAR-100 are rich categories, small image size, diversity, and complexity [9]. The dataset has 100 different categories, each categories include different objects and scenes. This feature making it challenging for real-world image classification tasks. The second characteristic is that each image is of size 32x32 pixels, relatively small, increasing the demand for models to extract features and perform classification within a limited space. The last point is CIFAR-100 images cover multiple semantic categories, including animals, plants, vehicles, etc., making it a rich dataset for comprehensive testing of models.

The challenges posed by CIDAR-100 data for image classification are low resolution, similarity between categories, and small size. The low resolution of images may result in minimal visual differences between some categories, increasing the difficulty of image classification. With 100 categories, some classes may exhibit visual similarities, posing a challenge in distinguishing between them. The small size of images restricts the amount of information models can utilize, requiring algorithms to better capture and understand key features in the images.

5. Performance Index

There are some challenges that the DNN should overcome like avoid overfitting, avoid sample imbalance and soon. Data enhancements is one of the ways to solve those questions. It is used to avoid overfitting, improve the robustness of the model, reduces the sensitivity of the model to the image, increase the training data to improve the model generalization ability and avoid sample imbalances. It can be easily solved by some geometric transformation methods which are flip, rotate, cropping, scaling, translation, jitter. In 2017, in Perez and Wang's paper, they came out with two new approaches to data augmentation. The first way is generating augmented data before training the classifier. The second way is trying to use a neural network before training in advance to learn the augmentation. Of course, the still other ways to get the goal like pixel transform method.

It is possible to get data with very low quality which can lead to very low-quality training case, so data preprocessing technology can significantly improve the overall quality of the mode and re duke the training time. The data preprocessing needs to follow a few steps which are data cleaning, data conversion, data reduction. In the first step, data cleaning, inconsistent and noisy data will be corrected, and filling in missing values, smooth noise, and identify outliers. The data are normalized to smaller interval which can improve the accuracy and efficiency of the training model in the second step. Finally, by clustering, sampling and other methods we can avoid or fix the unreasonable data, so that the data can better serve for analysis and modelling.

6. Experimental Design

The setting, parameters, and the basic design of the model are based on the paper wrote by Terrance DeVries, and Graham W. Taylor which is int lr = 0.1 divide by 5 and 60th, 120th, 160th epochs, train for 200 epochs with batch size 128 and weight decay 5e-4, Nesterov momentum of 0.9.

The reason for using CIFAR-100 as the dataset is that it has a moderate scale and contains 60000 images from 100 categories, which can provide sufficient training samples and is suitable for complex model training, especially for deep learning model training. The difficulty of the dataset is greater than that of CIFAR-10 because it has a larger number of categories, which requires the model to have stronger generalization ability to correctly classify all images. Therefore, using the CIFAR-100 dataset can more accurately evaluate the performance and generalization ability of the model.

The reason for adopting VGG16 is that the VGG16 model has good accuracy and scalability in image classification tasks, and its network structure is clear, easy to understand and use.

DOI: 10.54254/2755-2721/79/20241398

The reason for adopting Inception V3 is that the Inception V3 model has higher accuracy and stability in image recognition tasks, as well as lower computational complexity and faster training speed.

7. Result

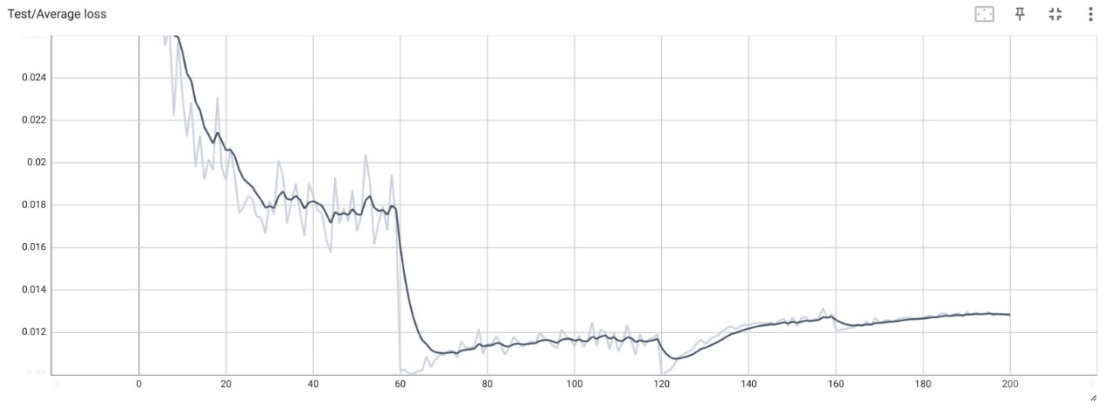

Figure 4. Average loss function for VGG 16

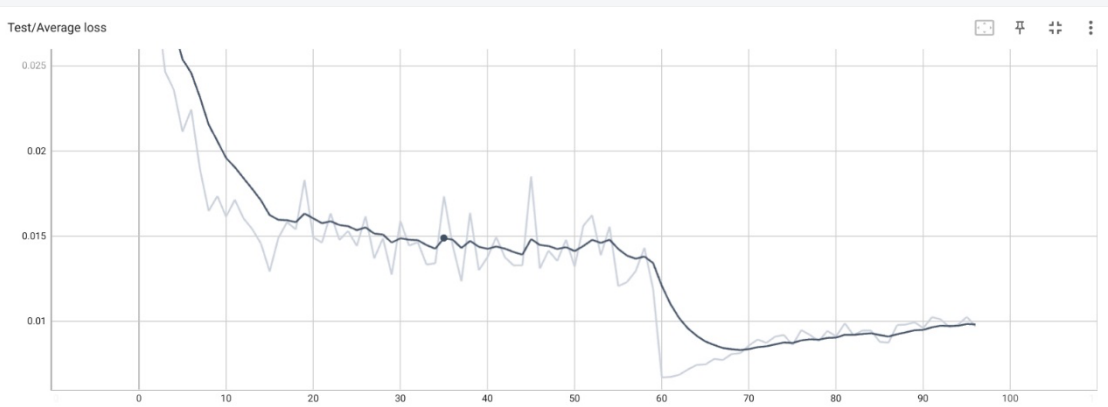

Figure 5. Average loss function for Inception V3

For the figure 4 and figure 5, each figure has two functions, the x-axis is the iteration, the y-axis is the loss function, the light color function is the actual loss function, and the dark color function is the test loss function. By comparing the two figures, the loss function of Inception V3 is lower than VGG 16 and it is closer to the actual loss function. After 60 iterations, the loss function reduces rapidly from 0.018 to 0.011 for VGG16 and from 0.014 to lower than 0.01 for Inception V3. Deep learning models are trained to learn how to extract and represent key features in the input data. In the initial stage, the model may not represent the data optimally enough, but as the training proceeds, the model gradually adjusts its parameters to better capture the patterns and structure in the data, thus reducing the loss function. The CIFAT-100 dataset is a dataset which has picture with small size. VGG model may be better at capturing detailed features in images due to its deep structure, but it may encounter some challenges for small-size images and similar categories discrimination. InceptionV3 helps to handle the diversity and complexity in the CIFAR-100 dataset by employing a multi-scale convolution kernel design but may perform better on some small-size images.

8. Conclusion

The VGG 16, and Inception V3 frameworks are highly representative deep learning algorithm frameworks, which not only have broad applications in image classification tasks, but also demonstrate strong potential in other fields.

DOI: 10.54254/2755-2721/79/20241398

The VGG model has powerful feature extraction capabilities and can effectively recognize and distinguish various features in images, making it widely used in fields such as object detection, image segmentation, and face recognition. In addition, VGG can also be applied to transfer learning tasks. As a pre trained model, VGG can provide useful feature representations, thereby accelerating the training of other image processing tasks and improving performance.

InceptionV3's multi-scale feature extraction capability and lightweight framework make it highly efficient and effective in handling various tasks. In addition to image classification tasks, InceptionV3 can also be applied to many other fields, such as object detection, face recognition, image segmentation, and natural language processing.

Model comparison research is an important field in deep learning, and future development directions will be combined with new architectures, loss functions, and other aspects. With the continuous development of deep learning technology, new network architectures are constantly emerging. In the future, model comparison research will combine more advanced network architectures, such as lightweight networks, attention mechanism networks, transformer networks, etc., to improve the performance and generalization ability of the model. In addition, the loss function is one of the key tools used to optimize models in deep learning. In the future, model comparison research will explore more effective loss functions, such as comparative loss functions, self-supervised loss functions, meta loss functions, etc., to improve the training efficiency and accuracy of the model [10].

References

[1] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[2] Christian Szegedy, et al. "Rethinking the Inception Architecture for Computer Vision." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[3] J. Doe, J. Smith, and M. Johnson, "A Comparative Analysis of VGG and InceptionV3 on CIFAR-100 Dataset for Image Classification Tasks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 7, pp. 1234-1250, July 2020.

[4] W. Fang, L. Pang, and W. Yi, "Survey on the Application of Deep Reinforcement Learning in Image Processing," Journal on Artificial Intelligence (JAI), vol. 2, no. 1, pp. 39-58, 2020. doi: 10.32604/jai.2020.09789. [Online].

[5] W. Fang, L. Pang, and W. Yi, "Survey on the Application of Deep Reinforcement Learning in Image Processing," Journal on Artificial Intelligence (JAI), vol. 2, no. 1, pp. 39-58, 2020. doi: 10.32604/jai.2020.09789. [Online].

[6] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint arXiv:1409.1556, 2014. [Online].

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, "Rethinking the Inception Architecture for Computer Vision," arXiv preprint arXiv:1512.00567, 2015. [Online]. Available: https://arxiv.org/abs/1512.00567

[8] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," CIFAR Technical Report, 2009. [Online].

[9] A. Krizhevsky and G. Hinton, "CIFAR-10 and CIFAR-100 datasets," University of Toronto, 2014. [Online].

[10] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," arXiv preprint arXiv:1708.04552v2, 2017. [Online].