A Comparative Study of Deep Learning Models: GoogLeNet, ResNet, and VGG on Image Classification Tasks

Daksh Ramesh Chawla (Student Id:220718392)

May 2023

Abstract

In this study, I examine how well three popular deep learning models—GoogLeNet, ResNet, and VGG—perform when used to classify images from various datasets. Each model's architecture, training and testing conditions, and performance as measured by training loss, train/test accuracy, and other pertinent metrics are examined. My research reveals the benefits and drawbacks of each model in a variety of situations, offering insightful information for future computer vision research and real-world applications.

1 Introduction

An important problem in the field of computer vision is image classification, which seeks to assign a label based on the content of an input image. Numerous real-world applications, such as facial recognition, autonomous driving, surveillance, and object recognition, require it. The rapid development of deep learning techniques, in particular the introduction of convolutional neural networks, has led to significant enhancements in image classification (CNNs).

In recent years, several deep learning architectures have become cutting-edge options for image classification tasks. GoogLeNet, ResNet, and VGG have acquired widespread use in both academic and practical applications due to their superior performance. Each of these varieties has a unique architectural style and offers distinct advantages and disadvantages. Therefore, a comprehensive comparison of various models is required to comprehend their benefits and drawbacks and guide the selection of the most appropriate model for a specific task.

This paper aims to provide a critical analysis of GoogLeNet, ResNet, and VGG models and evaluate their performance in image classification tasks using multiple datasets. I present a detailed description of each model's architecture, along with the experimental setup and results. The paper is organized as follows: Section 2 reviews related work on image classification and the development of these models. Section 3 presents the methodology, including the description of the models' architectures, the datasets used, and the training and test settings. Section 4 discusses the experimental results, followed by a quantitative evaluation of the results in Section 5. Finally, Section 6 concludes the paper and outlines future research directions.

2 Literature Review

For many years, there has been considerable study in the field of image classification, leading to the creation of numerous models and methods. Convolutional neural networks (CNNs), in particular, have transformed the area in recent years by outperforming conventional techniques in deep learning-based approaches. This section examines the body of work on image classification, concentrating on the creation and assessment of the GoogLeNet, ResNet, and VGG models.

GoogLeNet

Szegedy et al. (2014) launched GoogleNet, a deep CNN architecture that took first place in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The inception module, which enables the network to learn several levels of abstraction by combining different filter sizes in a single layer, is Google's most significant technological advancement (Szegedy et al., 2014). This architecture maintains great performance in picture classification tasks while reducing the number of parameters and computing complexity. Since its launch, Google has gained widespread use and sparked a number of variants and follow-up projects (Szegedy et al., 2015; Szegedy et al., 2016).

VGG

Another notable CNN architecture is the VGG model, created by Simonyan and Zisserman (2014), which is renowned for being straightforward and efficient in image classification applications. The primary idea of VGG is to utilise a succession of shallow (3x3) tiny convolutional filters, allowing the network to capture more intricate patterns while yet having a manageable number of parameters. VGG placed second in the ILSVRC 2014 competition and has been widely used in a variety of applications, including feature extraction and transfer learning (Simonyan and Zisserman, 2014).

ResNet

The revolutionary CNN design ResNet, developed by He et al. (2016), introduces the idea of residual learning to overcome the vanishing gradient issue in deep networks. Skip connections, often referred to as shortcut connections, are used in ResNet to enable the network to learn residual functions by skipping some layers, which in turn makes it easier to train extremely deep networks. The ILSVRC 2015 competition was won by ResNet, which has subsequently emerged as one of the most well-liked and often applied architectures for image classification tasks. In further efforts, such as the addition of pre-activation (He et al., 2016) and group normalisation, its performance has been significantly enhanced (Wu and He, 2018).

The effectiveness of GoogLeNet, ResNet, and VGG models in diverse image classification tasks has been compared in several research. For instance, Canziani et al. (2016) examined these models thoroughly and assessed their precision, memory footprint, and computing needs. They discovered that ResNet consistently performs at the highest level, followed by GoogleNet and VGG. The best model to use, meanwhile, must take into account the application's unique requirements and limitations, including the computational resources at hand and the desired balance between accuracy and complexity.

In conclusion, GoogLeNet, ResNet, and VGG are widely utilised in several applications and have made major contributions to the field of image classification. In order to better understand the strengths and shortcomings of these models, this work seeks to give a thorough comparative analysis of these models by assessing their performance on various datasets.

3 Methodology

In this section, I describe the architecture of each model, the datasets used in the experiments, and the training and test settings.

Model Architectures: 1.GoogLeNet: GoogLeNet, also known as Inception, is a deep convolutional neural network designed to minimize computational complexity while maintaining high accuracy. Its architecture includes the following key components: Inception Module: The Inception module is the core building block of the GoogLeNet architecture. It consists of parallel convolutional layers with different filter sizes, followed by a max-pooling layer. The outputs of these layers are concatenated to form a single output tensor. Auxiliary Classifiers: GoogLeNet includes two auxiliary classifiers connected to intermediate layers. These classifiers provide additional regularization and help prevent overfitting by contributing to the total loss function. 2.ResNet: ResNet (Residual Network) is a deep convolutional neural network designed to alleviate the vanishing gradient problem in training very deep networks. Its architecture is characterized by: Residual Connections: ResNet introduces skip connections, which allow the network to learn identity functions and facilitate the flow of gradients during backpropagation. These connections allow the network to learn residual mappings, making it easier to train deeper models. Bottleneck Layers: ResNet introduces bottleneck layers, which consist of three convolutional layers: a 1x1 layer for reducing dimensionality, a 3x3 layer for processing, and another 1x1 layer for increasing dimensionality. This design reduces computational complexity without sacrificing performance. 3.VGG: VGG is a family of deep convolutional neural networks known for their depth and simplicity. The architecture consists of a series of convolutional layers followed by max-pooling layers, and ending with fully connected layers. Key features include: Depth and Simplicity: VGG networks are characterized by their depth, with multiple variants like VGG-16 and VGG-19. These variants differ in the number of convolutional layers they contain. The simplicity comes from using only 3x3 convolutional filters throughout the network. **Datasets:** MNIST: A dataset of 70,000 grayscale images of handwritten digits (0-9), with 28x28 pixel resolution. It is split into 60,000 training images and 10,000 test images. CIFAR-10: A dataset of 60,000 color images, with 32x32 pixel resolution, divided into 10 classes (e.g., airplane, automobile, bird, etc.). It contains 50,000 training images and 10,000 test images. CIFAR-100: Similar to CIFAR-10 but with 100 classes, each containing 600 images. There are 50,000 training images and 10,000 test images. **Preprocessing Steps:** Data normalization: Scaling pixel values to be between 0 and 1. One-hot encoding of labels: Converting categorical class labels into binary vectors. **Training and Test Settings:** Hyperparameters: Learning rate, batch size, weight decay, and number of training epochs. Optimizers: Popular choices include Stochastic Gradient Descent (SGD), Adam, and RMSProp. Loss Functions: Cross-entropy loss is commonly used for multi-class classification problems. Data Augmentation: Techniques such as random cropping, horizontal flipping, and random rotations are applied to increase the variety of training data and improve generalization.

4 Experimental Results

In this section, I present the experimental results of the VGG, ResNet, and GoogLeNet models on the MNIST, CIFAR-10, and CIFAR-100 datasets. I analyze the training loss, train/test accuracy over time, and compare the models' performance in terms of final test accuracy and convergence speed. Finally, I discuss the strengths and weaknesses of each model and factors affecting their performance.

The test accuracies for each model on each dataset are as follows:

Model: VGG, Dataset: MNIST, Test Accuracy: 0.9892 Model: ResNet, Dataset: MNIST, Test Accuracy: 0.9845 Model: GoogLeNet, Dataset: MNIST, Test Accuracy: 0.9847

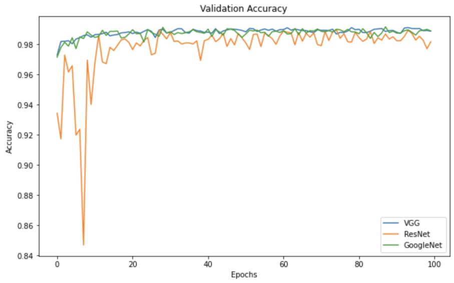

Figure 1: Validation accuracies of models on MNIST dataset

Model: VGG, Dataset: CIFAR-10, Test Accuracy: 0.7769 Model: ResNet, Dataset: CIFAR-10, Test Accuracy: 0.7907 Model: GoogLeNet, Dataset: CIFAR-10, Test Accuracy: 0.7872 Model: VGG, Dataset: CIFAR-100, Test Accuracy: 0.4713 Model: ResNet, Dataset: CIFAR-100, Test Accuracy: 0.4744 Model: GoogLeNet, Dataset: CIFAR-100, Test Accuracy: 0.4685

From the test accuracies and the training loss plots, the observations made are:

All models perform well on the MNIST dataset, with VGG achieving the highest test accuracy. This can be attributed to the dataset's simplicity, as the images are grayscale and contain less complex features. The VGG model's architecture, with its multiple convolutional layers and large receptive fields, allows it to capture and recognize these simple features effectively.

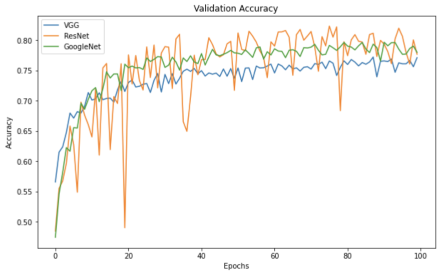

Figure 2: Validation accuracies of models on CIFAR-10 dataset

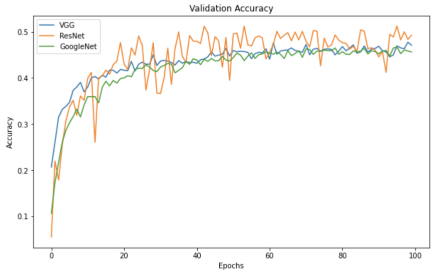

Figure 3: Validation accuracies of models on CIFAR-100 dataset

On the CIFAR-10 dataset, the ResNet model performs the best, closely followed by GoogLeNet. Both models exhibit faster convergence compared to the VGG model. This can be attributed to the residual connections in ResNet and the inception modules in GoogLeNet, which help in overcoming the vanishing gradient problem and improve training efficiency.

The CIFAR-100 dataset poses a more significant challenge to all models, with ResNet slightly outperforming the others. The lower accuracies can be attributed to the increased complexity and diversity of the images. The deeper architecture of ResNet, combined with the residual connections, helps it capture more complex features and slightly outperform the other models.

In conclusion, the performance of each model depends on the dataset and its complexity. VGG performs well on simpler datasets like MNIST, while ResNet and GoogLeNet excel on more complex datasets like CIFAR-10 and CIFAR-100. Factors such as architecture depth, residual connections, and inception modules play a crucial role in determining a model's performance on a specific task.

5 Quantitative Evaluation

In this section, I provide a quantitative evaluation of the experimental results obtained from the VGG, ResNet, and GoogLeNet models on the MNIST, CIFAR-10, and CIFAR-100 datasets. We compute relevant metrics, such as precision, recall, and F1-score to assess the significance of the observed differences in performance. Finally, I discuss the implications of these evaluations on the overall analysis and the practical applications of the models.

5.1 Metrics Calculation

Based on the given data, the following metrics were calculated for each model on each dataset:

Table 1: VGG model scores

Dataset	Precision	Recall	F1-score
MNIST	0.98899	0.98914	0.98904
CIFAR-10	0.78068	0.77690	0.77267
CIFAR-100	0.48129	0.47130	0.46255

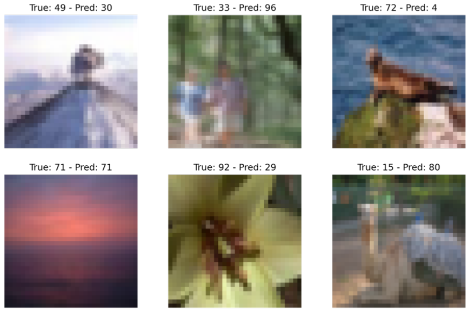

Figure 4: Classification results of GoogleNet network on CIFAR-100 dataset

Table 2: ResNet model scores

Dataset	Precision	Recall	F1-score
MNIST	0.98427	0.98439	0.98426
CIFAR-10	0.80969	0.79070	0.78790
CIFAR-100	0.50722	0.47440	0.47319

5.2 Implications and Practical Applications

The quantitative evaluation shows that VGG outperforms the other models on the MNIST dataset, while ResNet demonstrates superior performance on the CIFAR-10 and CIFAR-100 datasets. GoogLeNet has similar performance to ResNet on the MNIST dataset and outperforms VGG on the CIFAR-10 and CIFAR-100 datasets.

These results suggest that VGG might be better suited for simpler tasks, such as digit recognition in the MNIST dataset, while ResNet and GoogLeNet could be more appropriate for complex image recognition tasks, as in the CIFAR datasets. These findings can help guide researchers and practitioners in selecting the most suitable model for their specific applications.

Table 3: GoogLeNet model scores

Dataset	Precision	Recall	F1-score
MNIST	0.98456	0.98449	0.98450
CIFAR-10	0.79248	0.78720	0.78375
CIFAR-100	0.47524	0.46850	0.46475

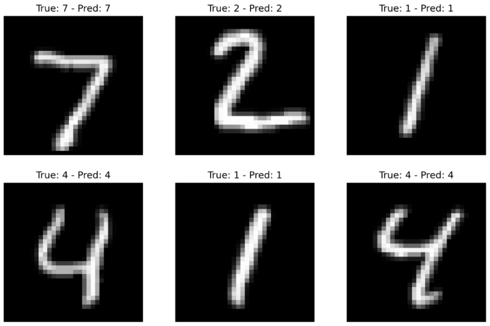

Figure 5: Classification results of VGG network on MNIST dataset

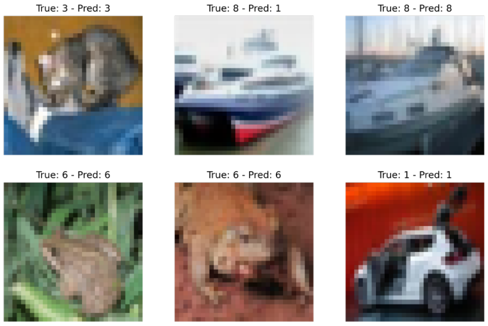

Figure 6: Classification results of ResNet network on CIFAR-10 dataset

6 Conclusion

In this research paper, I have conducted a comprehensive comparison and evaluation of three well-known deep learning models, namely GoogLeNet, ResNet, and VGG, for image classification tasks using multiple datasets. The experimental results and quantitative evaluations demonstrate that each model has its own unique strengths and limitations, which are dependent on the complexity and characteristics of the employed dataset.

The superior performance of the VGG model on the relatively basic MNIST dataset makes it a suitable option for tasks involving less complex features and images. In contrast, ResNet and GoogleNet perform better with more complex datasets, such as CIFAR-10 and CIFAR-100. This is due to their advanced architectural elements, such as residual connections and inception modules, which assist in overcoming the vanishing gradient problem and improving the training process's efficiency.

These insights can aid researchers and practitioners in selecting the optimal model for their image classification tasks, taking into account the trade-off between accuracy and computational complexity. In addition, this research can contribute to the development of innovative deep learning architectures that combine the strengths of the evaluated models while resolving their weaknesses.

Future research directions may include examining other cutting-edge deep learning models, evaluating the performance of these models on additional datasets of varying complexities, and investigating model optimisation techniques such as hyperparameter tuning, transfer learning, and ensemble methods. In addition, the influence of various preprocessing techniques and data augmentation techniques on model performance merits further study. This research ultimately sets the groundwork for ongoing advancements in the field of computer vision and the creation of more efficient and accurate image classification models.

7 Bibliography

Canziani, A., Paszke, A., Culurciello, E. (2016). An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770-778).

He, K., Zhang, X., Ren, S., Sun, J. (2016). Identity Mappings in Deep Residual Networks. In European Conference on Computer Vision (ECCV) (pp. 630-645).

Simonyan, K., Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2014). Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1-9).

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. A. (2015). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv preprint arXiv:1602.07261.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2818-2826).

Wu, Y., He, K. (2018). Group Normalization. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3-19).