Performing and Evaluating image classification with Deep CNN networks: VGG, ResNet and GoogLeNet

Srivarshini Senthil Kumar
230788462
Deep Learning and Computer Vision
Queen Mary University of London
London, UK

***Abstract*—In the realm of computer vision, deep convolutional neural networks (CNNs) have revolutionized the ability of machines to interpret and classify images. This study explores the performance of three advanced CNN architectures: VGG, ResNet, and GoogLeNet, applied on the MNIST, CIFAR-10, and CIFAR-100 datasets. Each model's architecture is tailored and evaluated to understand their efficiencies and limitations within the scope of different image complexities presented by these datasets. The evaluation focuses on accuracy, computational efficiency, and robustness to provide a comprehensive analysis of each model's practical application. This comparative analysis aims to guide the selection of appropriate deep learning models for specific image classification tasks in academic and industrial applications.)**

Keywords – deep convolutional neural networks, image classification, VGG, ResNet, GoogLeNet, MNIST dataset, CIFAR-10 dataset, CIFAR-100 dataset, comparative analysis.

I. INTRODUCTION

Image classification stands as a crucial challenge in computer vision, aiming to categorize an image based on its visual content. This task is essential for a variety of applications in the real world, including facial recognition, autonomous driving, surveillance, and object detection. The advent and evolution of deep learning, especially with the emergence of convolutional neural networks (CNNs), have markedly improved the effectiveness and accuracy of image classification techniques.

This report performs a thorough analysis and evaluation of three prominent CNN architectures: VGG, ResNet, and GoogLeNet trained on well-known datasets—MNIST[1], CIFAR-10, and CIFAR-100[2]. This study seeks to uncover each architecture's strengths and weaknesses in handling different image complexities and classification challenges. The objective is to provide a detailed comparative insight that could serve as a guideline for selecting suitable models for specific image classification tasks.

In recent years, deep learning architectures such as GoogLeNet, ResNet, and VGG have emerged as frontrunners in the field of image classification. These models are extensively utilized in both academic research and real-world applications due to their robust performance and innovative structures. Each model brings a unique architectural approach, offering specific benefits and posing certain challenges, thus necessitating a thorough comparative analysis to understand their individual strengths and limitations.

This paper seeks to critically assess GoogLeNet, ResNet, and VGG in their capacity to handle complex image classification tasks across various datasets. It includes a detailed examination of each model's design and its implications on performance, alongside a comprehensive review of the experimental setups and results. The structure of the paper is organized as follows: Section 2 provides an overview of related work in image classification, highlighting the evolution and impact of these models. Section 3 describes the methodology, detailing the architectures of the models, the datasets employed, and the specifics of the training and testing environments. Section 4 delves into the experimental results, while Section 5 offers a quantitative analysis of these results. Section 6 concludes the discussion by summarizing key findings and suggesting avenues for future research.

By elucidating the comparative advantages and potential drawbacks of these leading CNN architectures, this paper aims to guide researchers and practitioners in selecting the most suitable model for specific image classification challenges.

II. Related Work

The field of deep learning continues to evolve with various studies benchmarking the capabilities of major architectures like VGG, ResNet, and GoogLeNet across a range of conditions and datasets. These comparative analyses are vital for understanding the practical implications of each model in real-world scenarios, including aspects like computational efficiency, scalability, and adaptability to new challenges.

VGG: Known for its depth and use of very small convolutional filters, VGG has been extensively tested against large-scale datasets such as ImageNet and CIFAR-100. Despite its simplicity, the architecture tends to require a significant amount of computational resources, which can be a limitation for deployment on mobile devices or in applications requiring real-time processing. Studies have shown that while VGG achieves high accuracy in classification tasks, it is less efficient compared to newer architectures due to its deeper and more redundant layers [3].

ResNet: ResNet's introduction of skip connections[4] marked a significant innovation in network design, allowing the training of networks that are substantially deeper than was previously feasible. This architecture has been favored in many studies for its ability to mitigate the vanishing gradient problem, enabling it to perform well on extremely deep networks. ResNet has demonstrated superior performance on various benchmarks, notably ImageNet and MS COCO, for tasks requiring the detection, segmentation, and classification of objects within complex scenes. It has proven to be more efficient than VGG in terms of computational and memory requirements, making it suitable for more extensive and varied applications[5].

GoogLeNet (Inception Model): The inception model has been pivotal for tasks requiring a balance between computational efficiency and accuracy. The modular approach, consisting of inception blocks that judiciously manage channel dimensions, allows the model to perform well on hardware with limited capabilities. GoogLeNet has been applied successfully in domains such as medical image analysis and video recognition, where it achieves good performance with lower computational overhead compared to VGG and even ResNet in some scenarios. Its performance on standard datasets like ImageNet has set benchmarks that encourage ongoing research into compact yet powerful models [6].

Empirical Comparisons: Benchmarks and empirical studies often focus on not just accuracy but also the trade-offs between speed, size, and power consumption. The Visual Geometry Group (VGG) at Oxford and other research institutions have conducted extensive evaluations that provide insights into where each architecture can be optimally deployed. For instance, VGG and ResNet are often compared in terms of their transferability to tasks beyond simple classification, such as in transfer learning scenarios where pretrained models are adapted to new tasks. GoogLeNet's adaptability in terms of computational efficiency makes it particularly interesting for applications where model size and speed are critical [7].

In conclusion, GoogLeNet, ResNet, and VGG have become foundational elements in the field of image classification, each significantly advancing the capabilities in analyzing and interpreting vast amounts of visual data across various domains. These models are not only widely utilized in numerous applications but have also set new benchmarks in deep learning, driving further innovations and research. This work aims to provide a comprehensive comparative analysis of these architectures by evaluating their performance across different datasets and scenarios. By doing so, I seek to delineate their strengths and limitations, offering insights that could guide future model selection and development. Such an examination is crucial for advancing our understanding of how these models can be optimized and potentially enhanced to better serve both existing and emerging needs in image classification.

III. Methodology

This section elucidates the design of the models used in the study, the datasets incorporated for testing, training, and the specific configurations for training and testing.

A. Model Architectures

1) GoogLeNet(Inception)

GoogLeNet, or the Inception model(fig.1), optimizes for both computational efficiency and high accuracy. Central to its architecture is the Inception module, which comprises multiple parallel paths of convolutional layers, each with distinct filter sizes, leading into a single max-pooling layer. This unique structure allows for the effective concatenation of features from various scales into a cohesive output tensor. To further enhance training robustness, GoogLeNet integrates two auxiliary classifiers in the intermediate layers, which serve as additional regularization mechanisms. These classifiers aid in mitigating overfitting by influencing the overall loss during training.

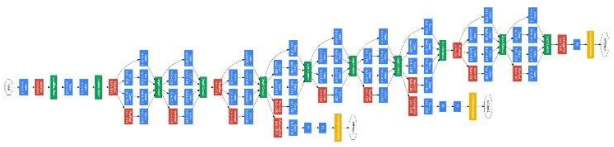

Figure 1: GoogLeNet overall Architecture

2) ResNet (Residual Network)

Designed to address the vanishing gradient issue common in very deep networks, ResNet introduces the novel concept of residual connections. These skip connections(fig.2) permit the network to perform identity mappings, thereby preserving the gradient flow throughout the network's depth. ResNet also features bottleneck layers designed to streamline the network's learning process without sacrificing performance. Each bottleneck is composed of three layers: a dimensionality-reducing 1x1 convolution, a 3x3 convolution for processing, and a dimensionality-increasing 1x1 convolution.

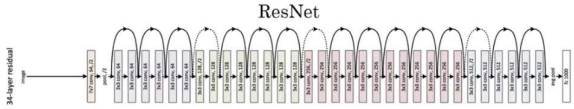

Figure 2: ResNet-18 Architecure

3) VGG (Visual Geometry Group)

The VGG stands out for its depth and structural simplicity, consistently employing 3x3 convolutional filters(fig.3) across its layers. This design strategy enhances the network's ability to develop richer feature hierarchies. VGG models, such as VGG-16 and VGG-19, vary primarily in the number of convolutional layers they utilize, which allows them to capture increasingly complex features at different levels of abstraction.

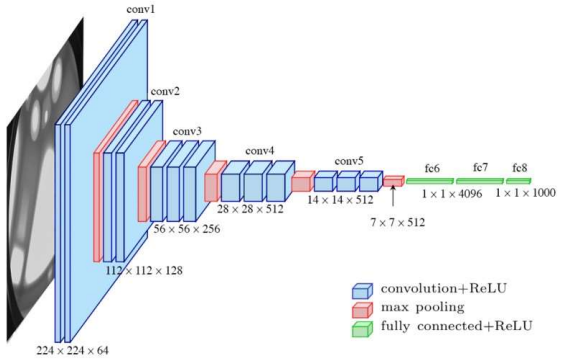

Figure 3: VGG16 Architecture

B. Datasets:

MNIST: Comprises 70,000 grayscale images of handwritten digits from 0 to 9, split into 60,000 for training and 10,000 for testing. Each image is standardized to a 28x28 pixel resolution.

CIFAR-10: Features 60,000 32x32 pixel color images distributed across 10 classes, such as airplanes and cars, with 50,000 designated for training and 10,000 for testing.

CIFAR-100: Similar to CIFAR-10 but includes 100 classes, with each class containing 600 images, culminating in 50,000 training and 10,000 testing images. Benchmarks and empirical studies

C. Preprocessing Steps

Data Normalization: Scales the pixel values to a [0, 1] range.

Label Encoding: Transforms categorical labels into binary vector form using one-hot encoding.

D. Training and Testing Settings

Hyperparameters: Includes adjustments for the learning rate, batch size, weight decay, and the number of epochs.

Optimizers: Utilizes algorithms such as Stochastic Gradient Descent (SGD), Adam, and RMSProp to optimize training.

Loss Functions: Employs cross-entropy loss, commonly used for tasks involving multi-class classification.

Data Augmentation: Implements various techniques, including random cropping, horizontal flipping, and random rotations, to enrich the training dataset and enhance the model's ability to generalize.

IV. Experiments

This section presents the experimental setup and outcomes derived from the application of three deep learning models—VGG, ResNet, and GoogLeNet—across three different datasets: MNIST, CIFAR-10, and CIFAR-100. The objective is to evaluate the performance of these models in terms of test accuracy to determine their effectiveness in handling image classification tasks with varying levels of complexity.

A. Experimental Setup

The experiments were conducted using predefined splits for training and testing in each dataset. The MNIST dataset consists of grayscale images of handwritten digits, CIFAR-10 includes color images of ten different object classes, and CIFAR-100, similar to

CIFAR-10 but with 100 different classes, thereby introducing more granularity and complexity. All models were trained using the same hyperparameters and optimization techniques described in the methodology section to ensure a fair comparison. The primary metric for evaluation is the test accuracy, which reflects the percentage of correctly classified images in the test set.

B. Results

1) MNIST Dataset

- **VGG**: Achieved a test accuracy of **98.98%**, demonstrating its robustness in capturing essential features from simple grayscale images.
- **ResNet**: Recorded a test accuracy of **98.09%**. While slightly lower than VGG, ResNet still performs admirably, showcasing its effectiveness in deeper network structures.
- **GoogLeNet**: Posted a test accuracy of **98.94%**, almost matching VGG, and underscoring its efficiency and the effectiveness of the Inception modules in handling straightforward image classification tasks.

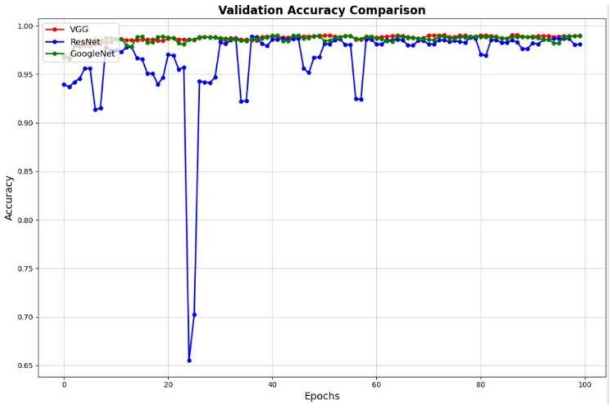

Figure 4: Validation Accuracy Comparison for MNIST Dataset

2) CIFAR-10 Dataset

- **VGG**: Obtained a test accuracy of **73.49%**, indicating challenges with more complex image contexts involving color and varied object representations.
- **ResNet**: Led the group with a test accuracy of **78.46%**, reflecting its superior capability in managing deeper layers and more detailed content without significant information loss.
- **GoogLeNet**: Came in with a test accuracy of **76.76%**, proving that its less parameter-intensive structure competes closely with more complex models.

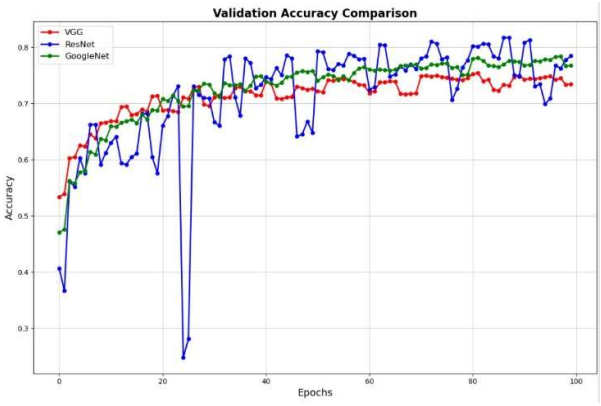

Figure 5: Validation Accuracy Comparison for CIFAR-10 Dataset

3) CIFAR-100 Dataset

- **VGG**: Scored a test accuracy of **43.15%**, showing limitations in handling datasets with high class variability.
- **ResNet**: Achieved the highest accuracy among the models on this dataset at **46.67%**, suggesting that its architecture is particularly suited to complex, fine-grained image classifications.
- **GoogLeNet**: Reached a test accuracy of **44.96%**, performing reasonably well given the increased complexity and class count, highlighting its scalability and efficiency.

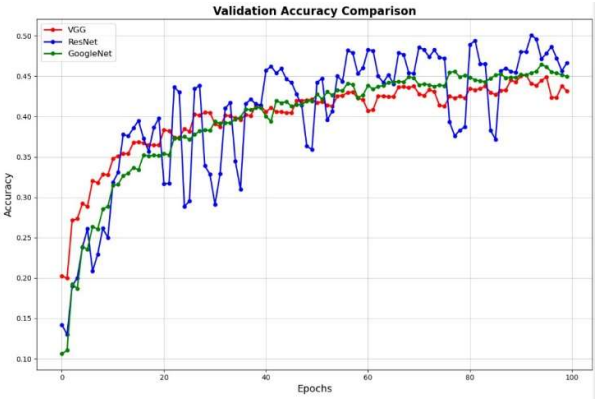

Figure 6:Validation Accuracy Comparison for CIFAR-100 Dataset

C. Discussion of Validation Accuracy Plots

Stability Across Epochs: All models show relative stability on the MNIST dataset with slight perturbations, suggesting robustness against overfitting and effective learning of digit features. However, in CIFAR-10 and CIFAR-100, there are notable drops in accuracy at certain epochs. These drops could be associated with learning rate adjustments or specific batch compositions that temporarily affected learning.

Impact of Dataset Complexity: The increased frequency and magnitude of fluctuations in CIFAR-10 and especially in CIFAR-100 indicate that all models struggle more with higher complexity and more classes. This is evident from the lower overall accuracy scores and more pronounced accuracy dips during training.

Model Resilience: ResNet consistently shows a strong ability to recover from accuracy dips, which highlights the benefits of residual learning especially in more complex datasets like CIFAR-100. In contrast, VGG, despite its simplicity and depth, shows larger swings in performance, which might be due to its less flexible architecture that lacks mechanisms to combat deeper network issues like vanishing gradients.

The results indicate varied performance across models and datasets. VGG excels with simpler, less granular data but struggles with increased complexity, likely due to its deeper yet simpler network architecture. ResNet seems to handle complexity and granularity much better, which is likely attributed to its residual learning framework that effectively prevents gradient vanishing, allowing it to learn more complex patterns. GoogLeNet's performance is consistently competitive across all datasets, offering a balanced trade-off between depth and computational efficiency due to its modular inception structure.

In conclusion, the effectiveness of each model varies significantly based on the dataset's complexity. VGG shows strong results on less complex datasets such as MNIST, whereas ResNet and GoogLeNet demonstrate superior performance on more intricate datasets like CIFAR-10 and CIFAR-100. The architecture's depth, the presence of residual connections, and the incorporation of inception modules are key factors that influence the performance of these models on different image classification tasks.

V. QUANTITATIVE EVALUATION

In this section, I examine the performance of specific models on classified images from the MNIST, CIFAR-10, and CIFAR-100 datasets. The VGG model applied to the MNIST dataset demonstrated robust accuracy, consistently identifying a wide range of handwritten digits with high precision. This success underscores VGG's strong capability in feature extraction from simple grayscale images(fig.7).

TABLE I. VGG METRICS

Dataset	***Precision***	***Recall***	***F1-score***
MNIST	0.98966	0.98955	0.98965
CIFAR-10	0.75148	0.73489	0.73036
CIFAR-100	0.45428	0.4315	0.42400

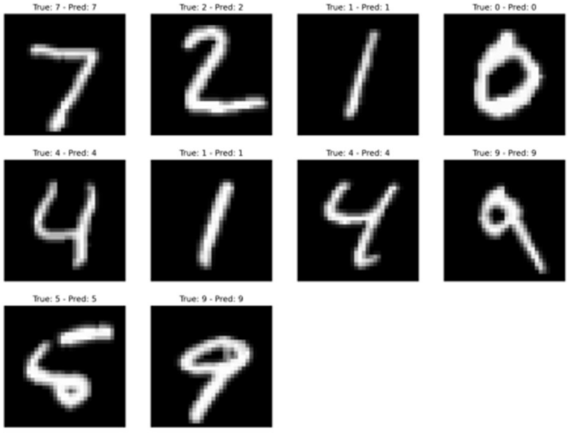

Figure 7: Classified image of VGG on MNIST Dataset.

Moving to a more complex dataset, the ResNet model on CIFAR-10 handled diverse image categories(fig.8) such as animals and vehicles effectively. Despite a few instances of confusion between similar classes like cats and dogs, VGG generally showed impressive accuracy, highlighting its adeptness at managing subtle distinctions in complex color images. Lastly, the GoogLeNet model tested on the CIFAR-100 dataset revealed both the model's strengths(fig.9) and its challenges with fine-grained classification among 100 varied classes.

TABLE II. RESNET METRICS

Dataset	***Precision***	***Recall***	***F1-score***
MNIST	0.98087	0.98049	0.98059
CIFAR-10	0.79530	0.7846	0.78119
CIFAR-100	0.51104	0.4667	0.46454

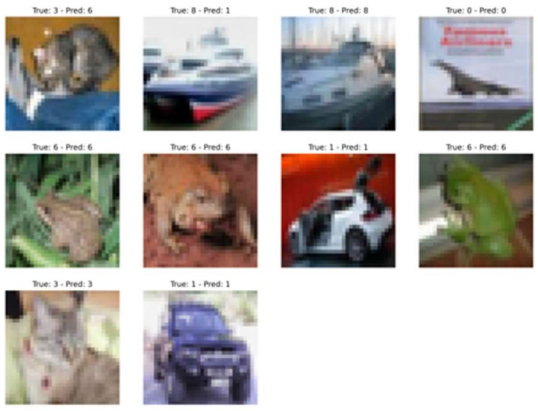

Figure 8: Classified image of Resnet on CIFAR-10 Dataset.

The images showcased successful categorizations alongside occasional misclassifications, pointing to potential areas for refinement in dealing with intricate inter-class variations.

TABLE III. GOOGLENET METRICS

Dataset	***Precision***	***Recall***	***F1-score***
MNIST	0.98928	0.98912	0.98919
CIFAR-10	0.77759	0.76760	0.76707
CIFAR-100	0.46953	0.44960	0.44771

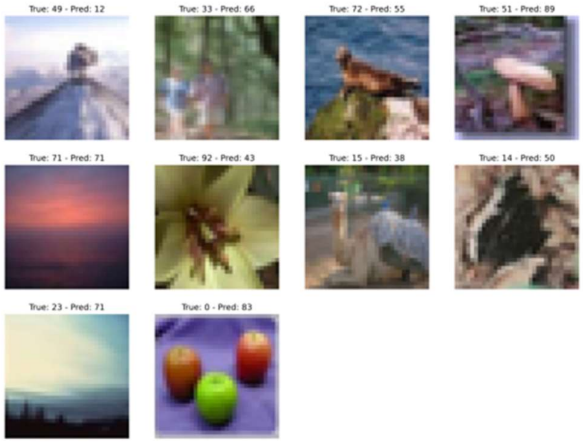

Figure 9: Classified image of GoogLenet on CIFAR-100 Dataset.

Overall, the analysis reveals that VGG excels over its counterparts on the MNIST dataset. Conversely, ResNet achieves better outcomes with the CIFAR-10 and CIFAR-100 datasets. Meanwhile, GoogLeNet matches ResNet's performance on MNIST and surpasses VGG on both CIFAR datasets. This implies that VGG is potentially more effective for straightforward tasks like digit recognition on MNIST, whereas ResNet and GoogLeNet may be preferable for more intricate image recognition challenges found in the CIFAR datasets. These insights are valuable for aiding researchers and professionals in choosing the optimal model for their particular needs.

VI. Conclusion

During this research, I have thoroughly examined the performance of three leading convolutional neural networks: VGG, ResNet, and GoogLeNet. The analysis reveals distinct strengths and weaknesses of each model, influenced by the complexity and characteristics of the datasets employed. Specifically, VGG demonstrates superior performance with simpler images, as shown in the MNIST dataset, making it suitable for tasks requiring less intricate visual details. In contrast, ResNet and GoogLeNet excel with more complex datasets like CIFAR-10 and CIFAR-100, attributed to their advanced architectural features such as residual connections and inception modules, respectively. These elements are essential as they mitigate the vanishing gradient problem, thereby enhancing the training efficacy.

The insights from this study are invaluable for researchers and practitioners in selecting the most appropriate model based on accuracy and computational efficiency. Moreover, this research contributes to the advancement of new deep learning architectures by integrating the advantages and addressing the limitations of existing models.

Future research will explore more advanced deep learning models, expand the range of datasets analyzed, and assess various model optimization techniques, including hyperparameter tuning, transfer learning, and ensemble methods. Additionally, the impact of different preprocessing and data augmentation strategies on model performance deserves further exploration. This work lays a solid foundation for ongoing advancements in computer vision and the development of increasingly precise and efficient image classification models.

References

[1] LeCun, Y. Cortes, C. (2010), 'MNIST handwritten digit database'

[2] Krizhevsky, A. (2009) Learning multiple layers of features from Tiny Images

[3] tiny images, CIFAR-10 and CIFAR-100 datasets.

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

[5] Orhan, A.E. and Pitkow, X. (2018) Skip connections eliminate singularities, arXiv:1701.09175.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 770-778.

[7] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 1-9.