Deeper Networks for Image Classification

Aditya Iyer

1. Introduction :

This report evaluates the performance of state-of-the-art CNN architectures—VGG, ResNet, and GoogleNet—on the MNIST and CIFAR datasets, aiming to assess their robustness and efficiency in image classification. This comprehensive report starts by reviewing model architectures, followed by an assessment of the training and testing processes. The findings from the experiments are further analysed to explore potential optimisations for improving accuracy and efficiency. The conclusion summarises the key takeaways and suggests future research directions in deep learning for enhancing image classification in domains such as autonomous driving and medical diagnostics.

2. Models involved and their architectures:

Googlenet

GoogLeNet, or Inception V1, is a state-of-the-art image recognition CNN that uses its inception module to apply convolutional filters and pooling operations in parallel, then concatenates the results for efficient multi-scale information capture. This 22-layer network uses 1x1 convolutions for dimensionality reduction, maximising efficiency before larger convolutions. Instead of fully connected layers, it employs global average pooling to lower parameter count and in order to counter overfitting. During training, auxiliary classifiers help diminish the vanishing gradient problem, though only the main classifier is used during inference. With approximately 5.6 million trainable parameters, GoogLeNet won the 2014 ImageNet challenge, revolutionizing deep learning in image recognition.

Inception V3

Inception v3, an extension of GoogLeNet (Inception V1), refines its architecture for finer efficiency in image recognition. This version introduces factorized convolutions, replacing 5x5 filters with successive 3x3 convolutions to reduce computational costs while maintaining spatial pattern capture. It expands the use of 1x1 convolutions for dimensionality reduction and feature recombination, accommodating the network's increased depth and width. Asymmetric convolutions, like 1x7 followed by 7x1, capture spatial features with fewer parameters than larger filters. Inception v3 also optimizes pooling layers for smoother integration with convolutional paths, improving feature extraction. These refinements make the network deeper and broader yet computationally efficient, boosting its image recognition performance.

Resnet 18

ResNet-18, a version of the Residual Network architecture, is customized for efficiency and robustness in image recognition. It comprises 18 layers, including convolutional layers, batch normalization, and ReLU activations. The most striking feature of ResNet-18 is its residual connections or "skip connections," which bypass layers to combat the vanishing gradient problem by maintaining direct gradient flow. Each residual block includes two 3x3 convolutional layers, combining simplicity with the capacity to process a broad range of features without information loss. This structure allows deeper layers to learn from identity functions, maximizing performance and easing training. ResNet-18 strikes a balance between depth and computational efficiency, making it versatile for various deep learning tasks.

Resnet 50

ResNet-50 is a much more refined version of the Residual Network architecture, designed for advanced image recognition tasks. It features 50 layers, including convolutional layers, batch normalization, and ReLU activations. The architecture utilizes bottleneck residual blocks with a 1x1 convolution for dimension reduction, a 3x3 convolution for core processing, and another 1x1 convolution to restore dimensions. These blocks help maintain computational efficiency in deeper networks. ResNet-50's skip connections allow gradients to bypass layers, countering the vanishing gradient problem and ensuring effective learning in deep sections. This structure amps up training and enhances the model's ability to learn complex patterns, making ResNet-50 a standard choice for challenging image processing tasks.

VGG13

VGG13 is a deep convolutional neural network with 13 convolutional layers, known for its simplicity and efficiency in image recognition. It employs 3x3 convolutional filters consistently, complemented by ReLU activations and batch normalization to maximize performance and convergence. Max pooling layers reduce spatial dimensions between convolutional blocks, while fully connected layers followed by a softmax classifier finalize predictions. VGG13's structured, deep architecture allows for detailed feature extraction, making it a solid choice for various visual recognition tasks.

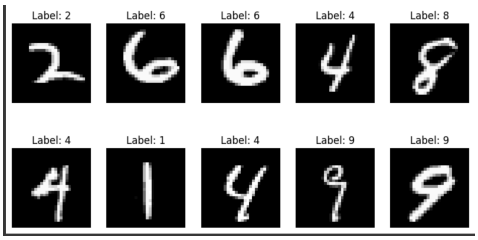
VGG16

VGG16 is a convolutional neural network with 16 layers, including 13 convolutional layers using uniform 3x3 filters, paired with ReLU activations for nonlinear processing. The network uses batch normalization for stability and max pooling to deduct spatial dimensions after convolutional blocks. It concludes with three fully connected layers and a softmax classifier for outputting final predictions. VGG16's deep, expansive architecture enables comprehensive feature extraction, making it highly efficient in a range of image recognition tasks. Its forthright design ensures reliability and rigid performance across diverse visual challenges.

3. Datasets Tested

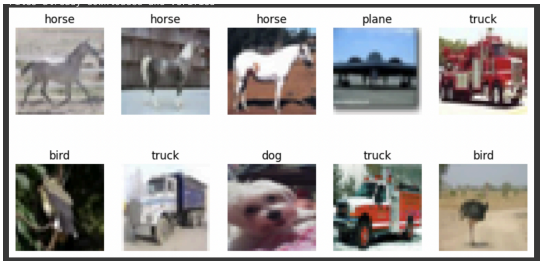
MNIST

The MNIST dataset contains 70,000 grayscale images of handwritten digits (0-9), each 28x28 pixels. It includes 60,000 training images and 10,000 testing images. Mainly used in machine learning, MNIST is a benchmark for evaluating image recognition algorithms, particularly for initial testing of deep learning models.



CIFAR10

The CIFAR-10 dataset consists of 60,000 32x32 color images in 10 classes, with 6,000 images per class. It is divided into 50,000 training images and 10,000 testing images. Commonly used in machine learning and computer vision, CIFAR-10 includes categories like airplanes, cars, birds, and ships. The dataset is ideal for developing and testing algorithms for image classification due to its manageable size and diversity. It serves as a benchmark for evaluating performance of deep learning models before they are scaled to more complex image recognition tasks.



4. Evaluation and Results

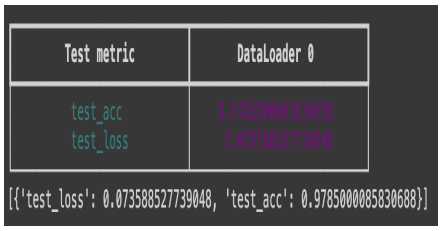
MNIST (Insights and Observations)

In this very thorough research of deep learning applications within image classification, the rapid accomplishment of around 98% accuracy top score by all the models trained namely VGG16, VGG13, ResNet50, ResNet18, GoogLeNet, and Inception V3 on the MNIST dataset after only one epoch of training is a testament to the prominent reciprocity between the simplicity of MNIST and the complexity of these architectures. These models, designed for the more

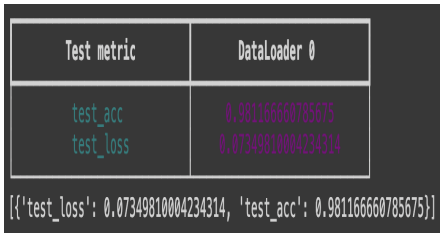
demanding ImageNet classification, possess state-of-the-art feature extraction abilities that seem excessive for MNIST's basic 28x28 grayscale digits. Tasks such as data augmentation and resizing were remarkably weary due to the small size of the dataset and the expansive and deep layers on the architectures. Nonetheless, the use of Adam optimizer and careful preprocessing strategies further facilitated efficient learning and refined generalization. This scenario emphasizes the need for appropriate balance in machine learning between model complexity and task specificity, accentuating how even overfitted architectures can achieve significant breakthroughs when optimally configured with the right optimizers and data preprocessing techniques.

Training and Testing on MNIST

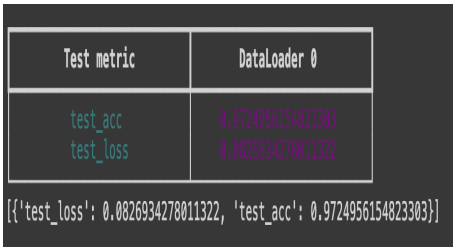
GOOGLENET (INCEPTION V1)



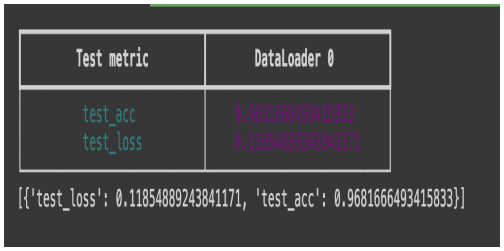
INCEPTION V3



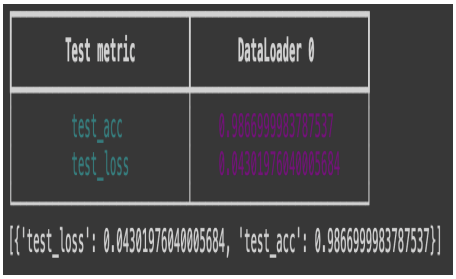
RESNET 18



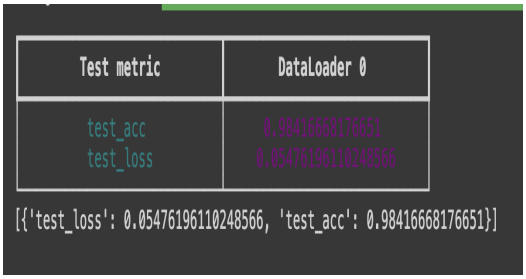
RESNET 50



VGG13



VGG16

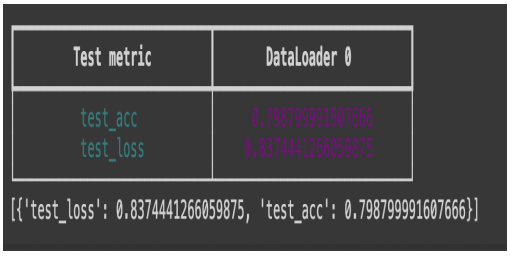


CIFAR10 (Insights and Observations)

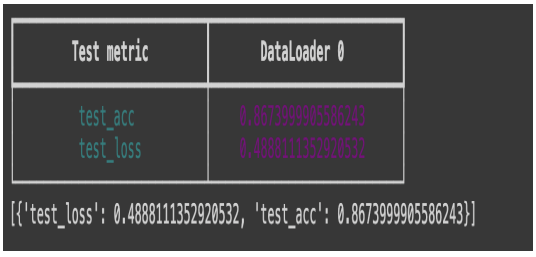
When comparing the performance of very same deep learning models namely VGG16, VGG13, ResNet50, ResNet18, GoogLeNet, and Inception on the MNIST and CIFAR-10 datasets, distinct disparities surfaced due to the datasets' complexity. On MNIST, these models easily achieved around 98% accuracy in just one epoch, emphasizing their potential with simpler 28x28 grayscale images. In contrast, on CIFAR-10, which features more complex 32x32 color images, accuracy hovered around 80% and above, but the training durations also varied due to early stopping based on performance hike. VGG models surprisingly performed optimally with the SGD optimizer, while the others showed better results with Adam. Early stopping played a critical role in CIFAR-10, countering overfitting and ensuring models stopped training when no further improvements in the model learning were evident, optimizing both computational resources and learning outcomes. This comparison highlights the importance of aligning model architecture, optimizer choice, and adaptive training techniques with dataset complexity to maximize image classification performance efficiently.

Training and Testing on CIFAR10

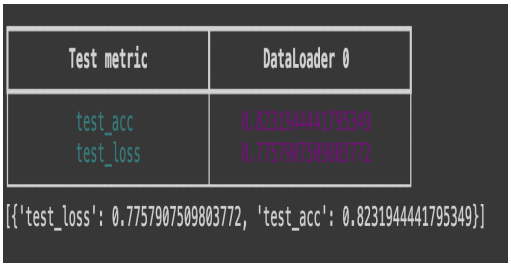
GOOGLENET (INCEPTION V1)



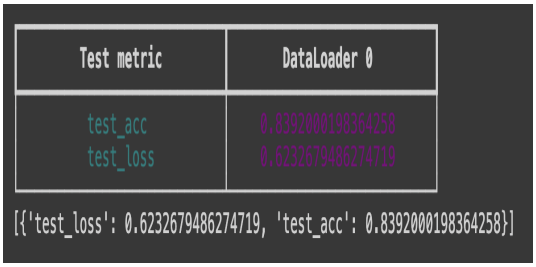
INCEPTION V3 (30 EPOCHS)



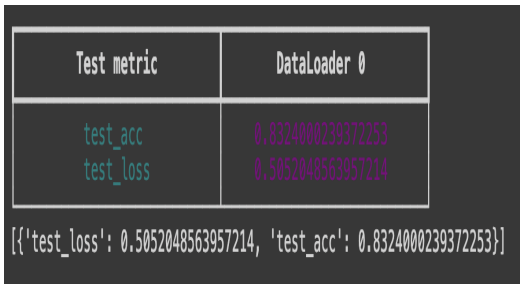
RESNET 18 (15 EPOCHS)



RESNET 50 (Epoch 12)



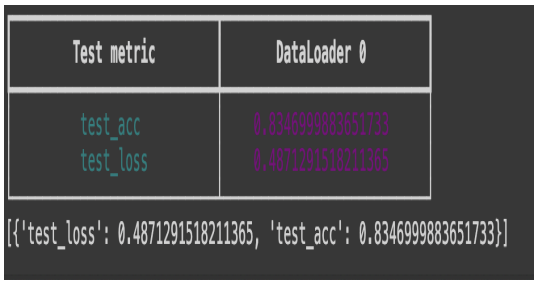
VGG13



Test metric	DataLoader 0
test_acc	0.832400023937253
test_loss	0.505204856395721 4

[{'test_loss': 0.505204856395721 4, 'test_acc': 0.832400023937253}]

VGG16



Test metric	DataLoader 0
test_acc	0.8346999883651733
test_loss	0.4871291518211365

[{'test_loss': 0.4871291518211365, 'test_acc': 0.8346999883651733}]

Conclusion:

This in-depth analysis of VGG16, VGG13, ResNet50, ResNet18, GoogLeNet, and Inception on MNIST and CIFAR-10 datasets reveals some intriguing insights into model performance across varying complexities. On MNIST, all models quickly reached about 98% accuracy in a single epoch, exhibiting their efficiency and refinement with simpler data. For CIFAR-10's diverse 32x32 colour images, accuracy stabilized around 80% and above, with training times powered by early stopping to prevent overfitting. VGG models were more effective with the SGD optimizer, while the other models excelled using Adam. This variation accentuates the significance of matching the optimizer and training strategy to the dataset's complexity. These findings signify the need for strategic choices in model architecture, optimizer selection, and training adjustments to optimize performance and efficiency in image classification tasks, ensuring models are appropriately scaled and tuned for the data they process.