Convolution Neural Networks Architecture Performance Comparison on the CIFAR-10 Dataset

Matthew Axell
University of Adelaide
Adelaide SA 5005
a1900421@adelaide.edu.au

Abstract

Advancements in computer vision have demonstrated the remarkable efficacy of Convolutional Neural Networks (CNNs) in object and animal recognition tasks from images. The proliferation of CNN architectures over the past decade has yielded numerous sophisticated approaches to visual pattern recognition. This study presents a comprehensive comparative analysis of three distinct CNN architectures: our own custom-designed baseline CNN model, the ResNet-18 architecture incorporating residual connections, and the pioneering AlexNet architecture. The performance evaluation is conducted using the CIFAR-10 dataset, a standardized benchmark in computer vision tasks. The best model architecture will then be further hyper-parameter tuned to find its best performance on the CIFAR-10 dataset. Through empirical investigation, this research aims to quantify and analyze the relative strengths and performance characteristics of these architectures, contributing to the understanding of their effectiveness in image classification tasks. After a best model have been found, further hyper parameter tuning will be done to further experiment if more performance could be gained.

1. Introduction

Recent advancement in computer vision have led to significant breakthroughs in object detection and recognition capabilities. The field has witnessed substantial innovations in model architectures, resulting in continuous improvements in performance metrics and accuracy on complicated tasks. Among the diverse applications of computer vision, object and animal recognition from static images remains a fundamental challenge that continues to drive research and development forward in the field.

This study presents a comparative analysis of various neural network architectures for image-based object and animal recognition tasks. Given the computational constraints and the prevalence of established architectures in the literature, we evaluate three distinct models: a custom-designed simple Convolutional Neural Network (CNN), an 18-layer Residual Network (ResNet-18), and AlexNet architecture. To ensure a fair comparison, all models are trained from scratch without pre-trained weights on the CIFAR-10 dataset. The primary objective is to assess their respective capabilities in classifying images into ten distinct categories of objects and animals.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature and theoretical foundations, Section 3 details the methodology and experimental setup, Section 4 presents our results and analysis, Section 5 reveals where the source code can be read, and Section 6 concludes with implications and future research directions.

2. Background and Related Studies

Recent advancements in Convolutional Neural Network (CNN) architectures have led to significant improvements in computer vision capabilities, particularly in object recognition tasks. A notable innovation in this field is the development of Residual Networks (ResNets), which introduced a novel approach to deep network training in CNNs. Unlike traditional CNNs, ResNets are designed to learn residual functions with reference to the layer inputs, rather than learning unreferenced functions [2]. This architectural innovation enables the training of substantially deeper networks while maintaining robust performance through a mechanism where each layer processes both the input and output from previous layers to learn the residual function between them.

The fundamental component of ResNet architecture is the residual block, illustrated in Fig. 1. These blocks can be systematically stacked to create networks of varying depths, such as ResNet-50, which incorporates 50 convolutional layers through this modular design. The effectiveness of this architecture has been demonstrated in various domains, including medical image classification. For instance, re-

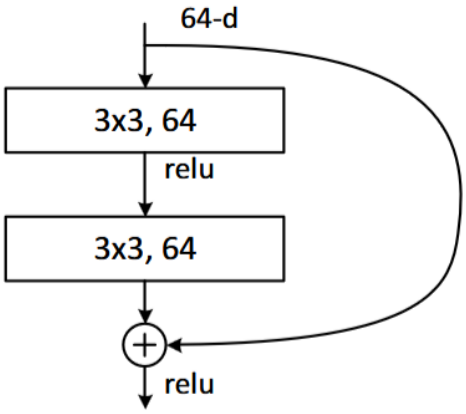

Figure 1. Residual Block Architecture [2]

searchers achieved remarkable performance metrics using ResNet-50, obtaining an accuracy of 94.72% and an F1-score of 97.09% on the NCT-CRC-HE-100K dataset [2, 3].

Another significant contribution to the field of computer vision is AlexNet, a pioneering deep convolutional neural network architecture. The network's design incorporates a strategic combination of convolutional layers, max pooling layers, and fully connected (dense) layers. With approximately 60 million trainable parameters, AlexNet achieved a notable accuracy of 60.3% on the ImageNet dataset [7], representing a substantial improvement over previous state-of-the-art results when it was introduced in 2012 [4]. This breakthrough performance not only demonstrated the potential of deep learning in computer vision tasks but also laid the foundation for subsequent future architectural innovations in the field.

3. Methodology

This section depicts the methodological framework, tools and architectural specifications employed in our comparative analysis of the CNN architectures. In accordance with principles of scientific reproducibility, we provide a comprehensive documentation of model architectures, and implementation protocols.

3.1. CIFAR-10 Dataset

The CIFAR-10 (Canadian Institute For Advanced Research) dataset, introduced by Krizhevsky et al. [1], represents a curated subset of the 80 Million Tiny Images dataset. This benchmark dataset comprises 60,000 RGB images, each image with dimensions of 32×32 pixels. Samples of these images can be seen in Fig. 2. The dataset is organized into 10 mutually exclusive classes, with a uniform distribution of 6,000 images per class as seen in Fig. 3. The

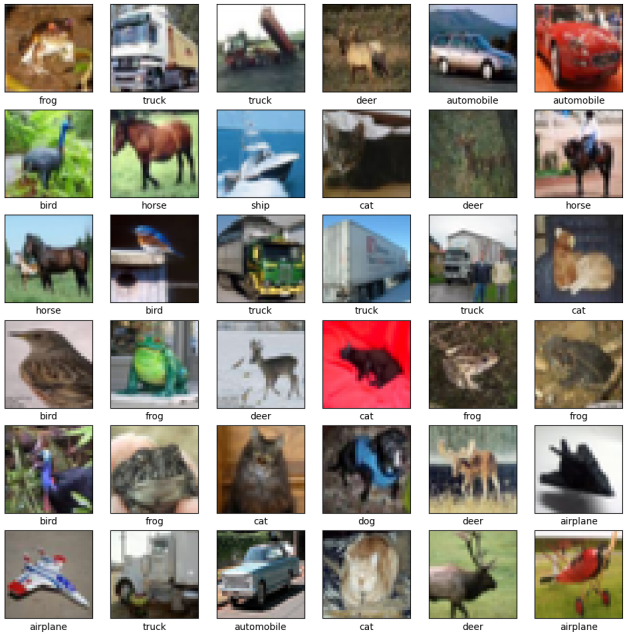

Figure 2. CIFAR-10 Image Data Samples

classes are as follows: "airplane", "automobile", "bird", "cat", "deer", "dog", "frog", "horse", "shop", and "truck".

It is noteworthy that the dataset maintains strict categorical separation between similar classes. Specifically, the "automobile" category encompasses consumer vehicles such as sedans and SUVs, while the "truck" category is restricted to large commercial vehicles. Both categories explicitly exclude pickup trucks to maintain categorical distinctness.

The CIFAR-10 dataset was selected for this comparative analysis due to its established position as a canonical benchmark in numerous computer vision research, enabling direct comparative analysis with existing literature. The dataset's scale (N=60,000) presents an optimal balance between computational requirements and model complexity, while maintaining sufficient task challenge for any computer vision architectures. Several inherent characteristics of CIFAR-10 further justify its selection: the dataset exhibits balanced class distribution (6,000 images per class), eliminating the necessity for class weighting strategies; the images undergo rigorous quality control and verification processes for its labels, minimizing training noise; and the structured categorization provides unambiguous classification targets. These attributes collectively establish CIFAR-10 as an ideal testbed for evaluating and comparing CNN architectural innovations while maintaining experimental reproducibility.

Tab. 1 shows a detailed specification of the CIFAR-10 image dataset for reference.

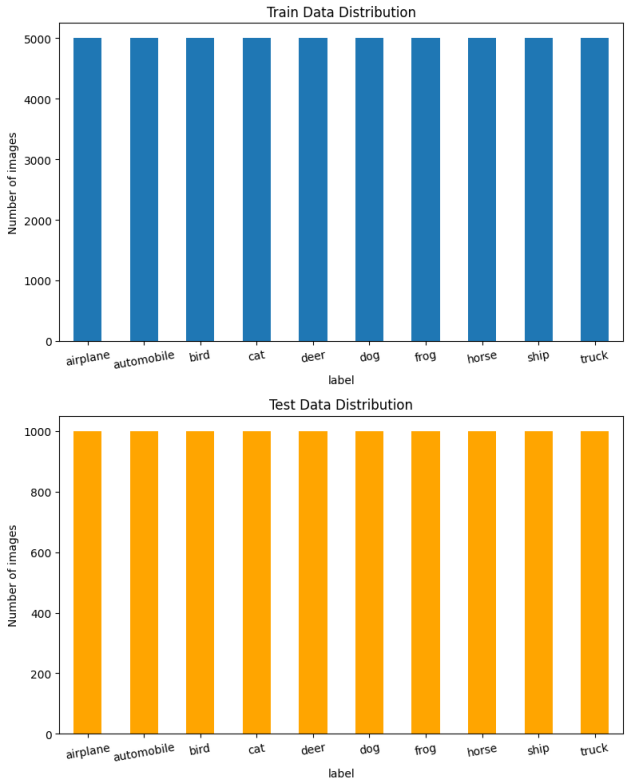

Figure 3. CIFAR-10 Dataset Class Distribution. **Top**: Training Set Class Distribution. **Bottom**: Test Set Class Distribution.

Parameter	Specification
Total Images	60,000
Image Dimensions	32×32 pixels
Color Channels	3 (RGB)
Classes	10
Images per Class	6,000
Bit Depth	24-bit color

Table 1. CIFAR-10 Dataset Specifications

3.2. Tools

This paper's main programming language is Python, using the following libraries: SciKit-Learn was used for the gridsearch, metric-evaluation methods, and in-built data splitting; TensorFlow and Keras were used for its dataset and prebuilt layers for the model architecture building; MatPlotLib and Seaborn to produce data visualization.

3.3. Preprocessing

For image preprocessing in object recognition tasks where the dataset comprises solely images and their corresponding labels, a simple streamlined approach suffices. The primary and only preprocessing step involves normalizing pixel values across all color channels to the interval [0,1] through division by 255. This normalization procedure has been demonstrated to facilitate more efficient training by reducing the occurrence of zero gradients during backpropagation, thereby accelerating convergence and potentially enhancing model accuracy [5].

3.4. Data Split

This paper will partition the CIFAR-10 dataset into the standard training, validation and testing set. The original dataset structure is comprised of training and testing sets in a 5:1 ratio. To establish a solid validation set while preserving the data quantity on the training set, the testing set was chosen to be further subdivided into validation and testing subsets at a 2:3 ratio, respectively. Although the dataset exhibited balanced class distribution, stratification was implemented as a methodological precaution to maintain consistent class representation across all partitions.

The training subset served as the primary learning corpus for the CNN architectures during the model fitting phase. The validation subset functioned as an independent evaluation platform, facilitating comparative analysis of architectural variants and enabling objective assessment of model performance during development, while avoiding bias during the best model architecture selection. Both the training and validation subsets were subsequently utilized in a grid-search paradigm for hyper-parameter optimization of the best model architecture. This iterative optimization process facilitated the identification of optimal model parameters.

The test subset remained isolated throughout the developmental phase, maintaining its integrity as an objective evaluation corpus. This subset was exclusively employed for the final performance assessment of the optimal CNN architecture with its refined hyper-parameters, thereby providing an unbiased estimate of the model's generalization capabilities on previously unseen data.

3.5. Model Architectures

As previously mentioned, this study will primarily focus on doing a comprehensive comparison between our own custom-designed CNN baseline model, the ResNet-18 architecture and the AlexNet architecture

3.5.1 Baseline CNN

The initial architecture in our comparative analysis is a simple custom-designed convolutional neural network (CNN), which serves as an excellent simple baseline model for evaluating more sophisticated architectures. The network employs a sequential structure, as illustrated in Fig. 4, comprising three convolutional layers interspersed with max pooling layers. The feature maps from the final convolutional layer undergo flattening before being processed through

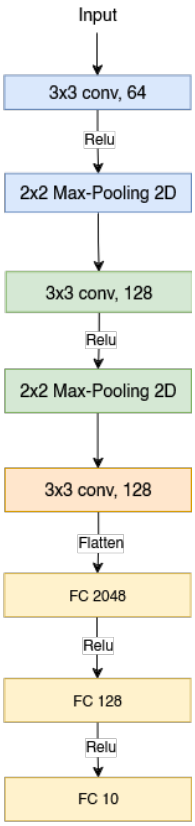

Figure 4. Custom-Designed CNN Architecture Representation

three fully connected layers. The network culminates in a softmax activation function for multi-class probability distribution across the predicted classes.

3.5.2 ResNet-18

The second architecture under evaluation is the Residual Network (ResNet-18), which exemplifies the advantages of residual learning through skip connections. Given computational resource constraints and our objective to train the models from scratch, we selected the 18-layer variant of ResNet as it presents a balance between model complexity and computational efficiency. This architecture maintains the fundamental benefits of residual learning while remaining tractable for training without pre-trained weights.

As seen in Fig. 5, the ResNet-18 architecture consists of multiple residual blocks, each containing two 3×3 convolutional layers with batch normalization and ReLU activation functions. The skip connections, fundamental to the residual learning framework, facilitate improved gradient flow during training and mitigate the vanishing gradient problem commonly encountered in deeper networks.

While larger variants of ResNet (such as ResNet-50 and

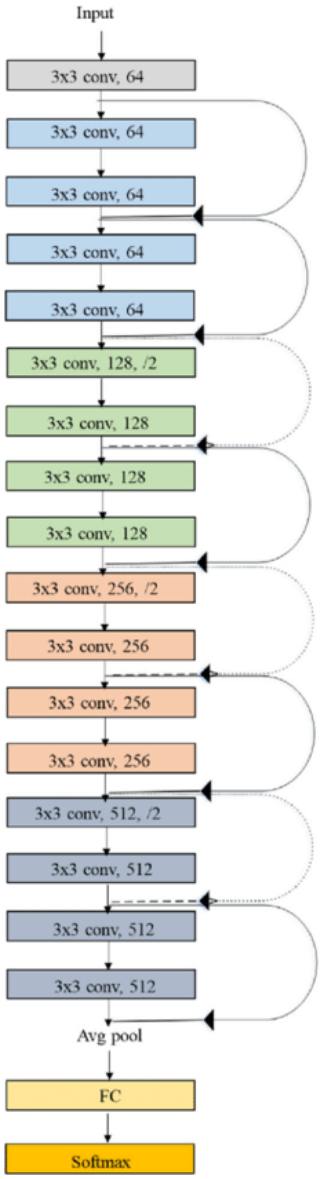

Figure 5. ResNet-18 Model Architecture Representation [8]

ResNet-101) have demonstrated superior performance on various computer vision tasks, ResNet-18 provides several advantages for our comparative study. First, its relatively shallow depth allows for efficient training while maintaining competitive performance. Second, the model's architecture incorporates sufficient complexity to demonstrate the benefits of residual learning while remaining computationally feasible for training from scratch. Additionally, the reduced parameter count compared to deeper variants makes it more suitable for applications with limited computational resources while still maintaining robust feature extraction capabilities.

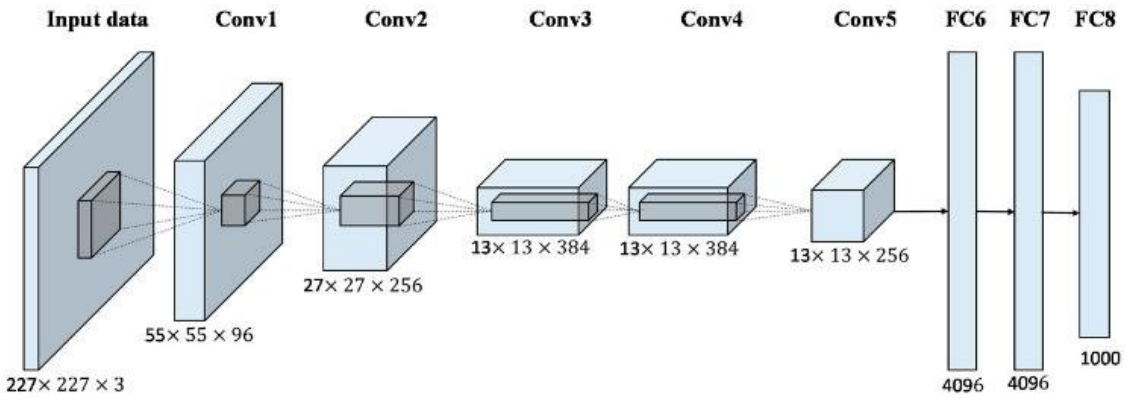

Figure 6. AlexNet Model Architecture Representation [6]

3.5.3 AlexNet

The third architecture in our comparative analysis is AlexNet, a pioneering deep convolutional neural network that significantly influenced the field of computer vision. While historically notable for its groundbreaking performance on the ImageNet dataset, we evaluate its capabilities in a similar context when trained from scratch on the CIFAR-10 dataset, which is a subset in the ImageNet dataset.

In Fig. 6, the AlexNet's architecture comprises five convolutional layers followed by three fully connected layers, incorporating several key design elements that contributed to its historical success. The network begins with large kernel sizes (11×11 in the first convolutional layer, followed by 5×5, and 3×3 in subsequent layers) to capture broad spatial hierarchies in the input features. Local Response Normalization (LRN) and max pooling layers are strategically placed throughout the network to enhance feature selectivity and reduce spatial dimensions, respectively. The architecture employs ReLU activation functions, which helped mitigate the vanishing gradient problem common in deeper networks of its era.

Despite its relatively simple architecture by contemporary standards, AlexNet offers several advantages for our comparative study. First, its architectural design represents a fundamental approach to convolutional neural networks, providing a complex CNN model for understanding architectural evolution in the field. Second, with approximately 60 million parameters, it presents an interesting middle ground between our custom CNN and ResNet-18 in terms of model complexity. Furthermore, AlexNet's historical significance and continued relevance in computer vision applications make it a valuable reference point for evaluating modern architectural innovations.

Model	Accuracy	Precision	Recall	F1-score
Basline CNN	73.28	73.77	73.28	73.34
ResNet-18	**76.85**	**77.54**	**76.85**	**76.72**
AlexNet	62.63	64.39	62.63	62.14

Table 2. Model Architecture Evaluation Metric Comparison on Validation Data in %

4. Experimental Analysis

This section describes an experiment aimed at identifying the optimal model architecture for object recognition on the CIFAR-10 dataset. Following the identification of the best model, further parameter tuning will be conducted to enhance its performance. Additionally, data augmentation techniques will be employed to increase the quantity of the training set, thereby theoretically aiding the model in learning more effectively.

4.1. Model Selection

In accordance with our research objectives, the initial phase of experimentation involved a comparative evaluation of the three architectures trained on the CIFAR-10 dataset. To maintain methodological rigor and prevent potential bias in subsequent analyses, as stated before, models were evaluated using a separate validation set. Early stopping was implemented based on the validation loss as a regularization technique to prevent overfitting while preserving the models' generalization capabilities.

The comparative performance metrics presented in Tab. 2 demonstrate that the ResNet-18 architecture achieved superior performance across all evaluation metrics, with an accuracy of 76.85% and an F1-score of 76.72%. This superior performance can be attributed to its advanced architectural features, particularly the residual connections that facilitate effective gradient flow during training. The custom-designed CNN baseline model demonstrated unex-

pectedly robust performance, achieving 73.28% accuracy and a 73.34% F1-score, positioning it as the second-best performing architecture.

Notably, AlexNet, despite its more complex architecture, achieved lower performance metrics (62.63% accuracy, 62.14% F1-score). This apparent under-performance can be attributed to several factors. First, AlexNet was originally designed for the ImageNet dataset, which comprises 1000 classes and higher resolution images, making its architecture potentially over-parameterized for the CIFAR-10 classification task. Second, the relative simplicity of the CIFAR-10 dataset (10 classes, 32×32 pixel images) may favor architectures with fewer parameters, as evidenced by the baseline CNN's performance. This observation aligns with the principle that model complexity should be commensurate with task complexity for optimal performance.

Furthermore, these results suggest that architectural innovation, as exemplified by ResNet-18's residual connections, can be more influential in determining model performance than raw parameter count or model depth. The superior performance of the simpler baseline CNN compared to AlexNet also highlights the importance of architectural alignment with the specific characteristics of the target dataset.

4.2. Hyper-Parameter Tuning

Following the selection of ResNet-18 as the optimal architecture, we conducted a comprehensive hyperparameter optimization study to maximize the model's performance on the CIFAR-10 dataset. A grid search methodology was implemented to systematically explore the following parameter space:

1. Using regularization or not (L2 Regularization)
2. Using Residual Connection or not
3. Using Dropout or not (Dropout rate=0.5)

Hyperparameter optimization was conducted to address the observed disparities between training and validation performance of the ResNet-18 architecture. As illustrated in Fig. 7, the validation loss exhibits early plateauing around epoch 10, while the training loss continues to be able to decrease, indicating potential overfitting behavior. This observation motivated the exploration of various regularization techniques and architectural modifications.

The investigation focused on three primary parameters: L2 regularization, dropout rates, and the impact of residual connections. A grid search methodology was employed to systematically evaluate these parameters. Contrary to initial hypotheses, neither L2 regularization nor dropout mechanisms yielded significant improvements in model performance. However, the experimental results confirmed the positive impact of residual connections on model accuracy as the best model from the gridsearch kept using the residual connection, validating their theoretical advantages in deep architectural designs.

Dataset	Accuracy	Precision	Recall	F1-score
Normal Data	**76.85**	**77.54**	**76.85**	**76.72**
Augmented Data	69.83	72.03	69.83	69.25

Table 3. ResNet-18 Dataset Evaluation Metric Comparison on Validation Data in %

Additionally, a discrete learning rate analysis was conducted across three orders of magnitude [0.01, 0.001, 0.0001]. While this parameter was evaluated independently from the grid search to optimize computational resources, empirical results demonstrated that the default learning rate of 0.001 provided optimal convergence characteristics for this specific classification task. This finding suggests that the initially selected learning rate coincidentally aligned with the optimal parameter space for the CIFAR-10 dataset characteristics and model architecture.

4.3. Data Augmentation Experiment

To explore potential performance improvements, we investigated the impact of data augmentation techniques on the training dataset. The augmentation pipeline comprised three transformations: random horizontal flipping, random rotation with a maximum angle of 0.1 radians, and random zoom with a scale factor of 0.1. These transformations were selected to introduce controlled variations while preserving the semantic content of the images in the training set.

As evidenced in Tab. 3, contrary to our initial hypothesis, the implementation of data augmentation techniques resulted in a degradation of model performance. The augmented dataset yielded performance metrics predominantly below 70%, with only the precision metric maintaining a higher value, but still lower than the alternative. This unexpected outcome suggests that the introduced variations may have interfered with the model's ability to learn discriminative features specific to the CIFAR-10 dataset's object categories.

Several factors may contribute to this performance degradation. First, the CIFAR-10 dataset's low-resolution nature (32×32 pixels) means that geometric transformations could potentially distort critical feature information. Second, the selected augmentation parameters, while moderate in magnitude, may have introduced variations that deviate from the natural distribution of the target classes. Additionally, the object-centric nature of CIFAR-10 images suggests that maintaining precise spatial relationships might be more crucial for accurate classification than introducing artificial variations.

Figure 7. ResNet-18 Training & Validation Loss and Accuracy over Epoch

Consequently, based on these empirical results, data augmentation was excluded from the final optimal model configuration. This finding emphasizes the importance of carefully evaluating augmentation strategies in the context of specific dataset characteristics and resolution constraints.

4.4. Final Test Set Result

Following comprehensive experimentation and hyperparameter optimization, the optimal ResNet-18 model demonstrated robust performance on the test dataset. Analysis of the confusion matrix presented in Fig. 8 reveals both the model's strengths and characteristic misclassification patterns. Notable classification errors primarily occurred between visually similar categories, particularly in cases where low image resolution (32×32 pixels) may have obscured discriminative features. Specifically, the model exhibited confusion between structurally similar object pairs:

1. Trucks were occasionally misclassified as automobiles, likely due to shared geometric characteristics
2. Birds were sometimes incorrectly identified as planes, presumably due to similar silhouettes at low resolution
3. Confusion between cats and dogs persisted, reflecting the challenge of distinguishing between these mammals when fine-grained features are limited by resolution constraints

The comparative analysis between validation and test metrics, as shown in Tab. 4, demonstrates remarkable consistency across both datasets. This performance stability strongly indicates that the model has achieved good generalization capabilities without overfitting to the training data. Such consistency between validation and test performance metrics suggests that the model has successfully learned robust and transferable features for object recognition, despite the inherent challenges posed by the low-resolution nature of the CIFAR-10 dataset.

Model	ResNet-18
Accuracy	75.75
Precision	76.73
Recall	75.75
F1-score	75.69

Table 4. ResNet-18 Evaluation Metric on Test Set in %

These results underscore the effectiveness of our optimization strategy while highlighting the fundamental limitations imposed by image resolution on fine-grained object discrimination. The observed misclassification patterns align with intuitive expectations regarding visual similarity and the information capacity of low-resolution images.

5. Code

The code of this paper can be found in the link: https://github.com/AxellLim00/CIFAR10-CNN-Architecture-Comparison in a file format of a jupyter notebook

6. Conclusion

This paper aims to compare different model architectures and build an optimal model to accomplish the complex task of recognizing various objects and animals. Through a comparative analysis, it was found that among the three model architectures evaluated, the ResNet-18 architecture performed the best on the CIFAR-10 dataset. This superior performance can be attributed to the residual connections in ResNet-18, which have a more significant impact on model performance than raw parameter count or model depth. Further hyperparameter tuning as well as data augmentation did

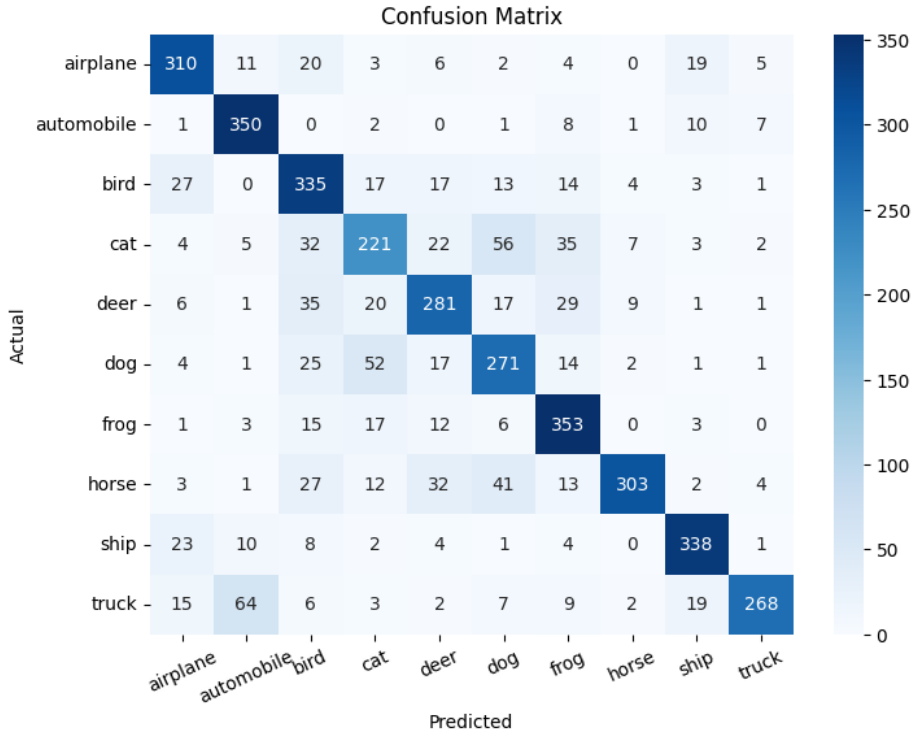

Figure 8. ResNet-18 Confusion Matrix on Test Set

not improve the performance.

However, this study acknowledges certain limitations. Due to constraints in time and computational resources, the exploration of larger models and a broader range of parameters was not feasible. This study only investigates one hyperparameter tuning scenario, leaving open the possibility that other models with optimal parameters might have achieved higher performance than the current optimized model.

Future work could involve the aforementioned experiments, including the evaluation of larger and more complex models, as well as extensive hyperparameter tuning across multiple models. Additionally, experiments could explore transfer learning from existing models. Given that many pre-trained models are based on the ImageNet dataset (contains CIFAR-10), it would be necessary to use another dataset to compare the performance of pre-trained and non-pre-trained models.

References

[1] Krizhevsky Alex. Learning multiple layers of features from tiny images. *https://www. cs. toronto. edu/kriz/learning-features-2009-TR. pdf*, 2009. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[3] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, Apr. 2018. 2

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[5] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. *Efficient BackProp*, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. 3

[6] Farheen Ramzan, Muhammad Usman Khan, Asim Rehmat, Sajid Iqbal, Tanzila Saba, Amjad Rehman, and Zahid Mehmood. A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks. *Journal of Medical Systems*, 44, 12 2019. 5

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2

[8] Shamoil Shaees, Muhammad Rashid Naeem, Hamad Naeem, Hamza Syed, Muhammad Arslan, and Hamza Aldabbas. Facial emotion recognition using transfer learning. 09 2020. 4