A Comparative Study of CNN and Transformer-based Models for Image Classification on CIFAR-100

Xiang Zheng
21307110169@m.fudan.edu.cn

YunNam Chan
21340980053@m.fudan.edu.cn

Abstract

This report presents a comparative analysis of Convolutional Neural Networks (CNN) and Vision Transformers (ViT) for image classification on the CIFAR-100 dataset. Both models were pretrained on ImageNet and fine-tuned using the same training strategies, including the CutMix data augmentation technique. The ViT model achieved a top-1 accuracy of 91.84% and a top-5 accuracy of 99.07%, outperforming the CNN model, which achieved a top-1 accuracy of 84.73% and a top-5 accuracy of 96.87%. CutMix augmentation improved the generalization capabilities of both models. The results highlight the superior performance of ViT, suggesting its effectiveness for image classification tasks.

Note: You can find the project files in the GitHub repo and the model parameters in Google Drive. For instructions on how to train and test the neural network, please refer to the readme file.

1 Introduction

Image classification is a fundamental task in computer vision, with applications spanning numerous domains such as autonomous driving, medical imaging, and surveillance. Traditionally, **Convolutional Neural Network (CNN)** has been the architecture of choice for image classification due to their ability to capture local spatial hierarchies through convolutional layers. However, recent advancements in deep learning have introduced Transformer-based architectures, initially developed for natural language processing, into the realm of computer vision. **Vision Transformer (ViT)** leverages self-attention mechanisms to model long-range dependencies and global context, offering a novel approach to image classification.

This report presents a comparative study of CNN and ViT on the CIFAR-100 dataset, which contains 100 classes of images. Both models are designed with a nearly equivalent number of parameters and trained using the same strategies, including the CutMix data augmentation technique. CutMix enhances training by combining patches from different images, creating more diverse training samples.

By comparing the performance of CNN and ViT under these conditions, this report aims to provide a clear understanding of their relative strengths and weaknesses in image classification. The insights gained from this comparison will inform the selection of the most appropriate architecture for future image classification tasks.

2 Dataset

The CIFAR-100 dataset is a well-known benchmark in the field of image classification. It comprises 60,000 color images of size 32x32 pixels, categorized into 100 distinct classes. Each class contains 600 images, split into 500 training images and 100 testing images. The 100 classes are organized into 20 superclasses, each containing five fine-grained classes. For the purpose of getting a better comparison between the CNN and ViT, the fine-grained class labels are used, providing a challenging and diverse set of images that cover a wide range of objects.

2.1 Preprocessing

Data preprocessing is a critical step in ensuring models generalize well to unseen data. For this task, the following transformations were applied to the training images step-by-step:

- **Random Cropping and Resizing**: Training images were randomly cropped with a padding of 4 pixels and then resized to 224x224 pixels. This step ensures that the models receive input of a consistent size and helps in making the models invariant to minor translations. Specifically, the images were

cropped to a size of 32x32 pixels and then resized to 224x224 pixels.

- **Random Horizontal Flip and Rotation**: To add variability to the training data and prevent the model from overfitting to position-specific features, random horizontal flips and rotations (up to 10 degrees) were applied.

- **Normalization**: Images were converted to tensors and then normalized using the mean and standard deviation values computed from the CIFAR-100 dataset (mean: 0.5071, 0.4867, 0.4408; standard deviation: 0.2675, 0.2565, 0.2761). This normalization ensures that the pixel values are scaled to a range suitable for neural networks.

- **CutMix**: This advanced augmentation technique involves combining two training images by cutting a region from one image and pasting it onto another. The labels are also mixed proportionally to the area of the cut. CutMix enhances the robustness of the model by encouraging it to learn from a mixture of features from different classes. It's applied in the training process.

For the testing images, they were resized to 224x224 pixels and normalized using the same mean and standard deviation values but were not augmented.

2.2 Dataloader

The training and testing datasets were loaded using the `DataLoader` class. This allows for efficient mini-batch loading, shuffling, and parallel data processing. The DataLoader was configured with a batch size of 64 and used 2 worker threads to handle data loading in parallel, thereby speeding up the training process.

3 Model Architecture

In this task, two different model architectures were used for image classification on the CIFAR-100 dataset: ResNet-50, a Convolutional Neural Network (CNN), and Vision Transformer (ViT), a Transformer-based model. Both models were pretrained on ImageNet, ensuring a strong initialization before fine-tuning on CIFAR-100.

3.1 ResNet-50

ResNet-50, short for Residual Network with 50 layers, is a widely used CNN architecture. It addresses the degradation problem in deep neural networks by introducing residual learning. This is achieved through skip connections, or shortcuts, that bypass one or more layers. The architecture of ResNet-50 is divided into several stages, each containing multiple residual blocks. Each block consists of convolutional layers, batch normalization, and ReLU activation functions. The main characteristics of ResNet-50 include:

- **Depth**: ResNet-50 has 50 layers, including convolutional layers, batch normalization layers, and fully connected layers.

- **Residual Blocks**: The use of identity and convolutional shortcuts to allow gradients to flow through the network more effectively, mitigating the vanishing gradient problem.

- **Pretraining**: The model is pretrained on ImageNet, providing a strong baseline by leveraging knowledge from a large, diverse dataset.

- **Parameters**: The model consists of 23,712,932 parameters, enabling it to learn intricate patterns and features from the data.

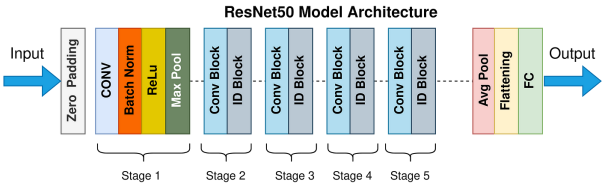

Figure 1: Model Architecture of ResNet-50

The model architecture of ResNet-50 is visualized in Figure 1.

3.2 Vision Transformer (ViT)

The Vision Transformer (ViT) represents a shift from traditional CNN-based approaches by utilizing Transformer architectures, which were originally developed for natural language processing tasks. ViT applies the self-attention mechanism to image patches, enabling the model to capture global context and long-range dependencies. The specific model used in this study is the WinKawaks/vit-small-patch16-224. The key features of ViT include:

- **Patch Embedding**: Images are divided into patches (16x16 pixels in this model), each of which is linearly embedded and then fed into the Transformer.
- **Self-Attention Mechanism**: This mechanism allows the model to weigh the importance of different patches relative to each other, capturing global interactions within the image.
- **Positional Encoding**: Since Transformers lack the inductive biases of convolutional layers (e.g., spatial locality), positional encodings are added to retain spatial information.
- **Pretraining**: The model is pretrained on ImageNet, leveraging extensive pre-learned representations from a large-scale dataset.
- **Parameters**: The model consists of 21,704,164 parameters, providing a balance between model complexity and computational efficiency.

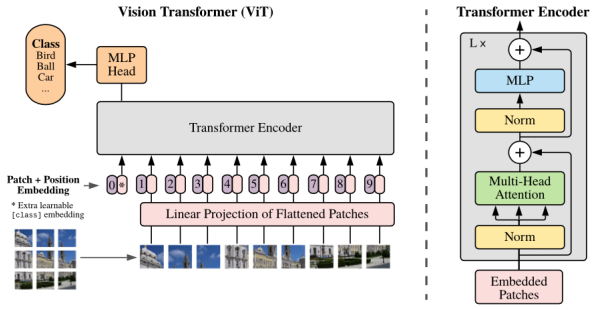

Figure 2: Model Architecture of ViT

The model architecture of ViT is visualized in Figure 2.

4 Train

In this section, we detail the training process of both the ResNet-50 (CNN) and Vision Transformer (ViT) models, including the hyperparameter search and full training phases.

4.1 Hyperparameter Search Results

The hyperparameter search phase involved training both models with various combinations of hyperparameters for a duration of 3 epochs each. The hyperparameters explored included the fine-tuning learning rate, fully connected learning rate, and batch size. Table 1 summarizes the configurations and corresponding accuracies achieved by each model during this phase.

(FT LR, FC LR, Batch Size)	CNN	ViT
(0.00005, 0.00100, 64)	0.5744	0.8779
(0.00005, 0.00100, 128)	0.4150	0.8532
(0.00005, 0.00500, 64)	0.6976	0.8801
(0.00005, 0.00500, 128)	0.6038	0.8633
(0.00005, 0.01000, 64)	0.7187	0.8752
(0.00005, 0.01000, 128)	0.6482	0.8622
(0.00010, 0.00100, 64)	0.6550	0.8852
(0.00010, 0.00100, 128)	0.4598	0.8635
(0.00010, 0.00500, 64)	0.7384	0.8892
(0.00010, 0.00500, 128)	0.6553	0.8765
(0.00010, 0.01000, 64)	0.7521	0.8852
(0.00010, 0.01000, 128)	0.6926	0.8739
(0.00050, 0.00100, 64)	0.7718	0.8973
(0.00050, 0.00100, 128)	0.6741	0.8888
(0.00050, 0.00500, 64)	0.8075	0.9014
(0.00050, 0.00500, 128)	0.7680	0.8981
(0.00050, 0.01000, 64)	**0.8166**	**0.8987**
(0.00050, 0.01000, 128)	0.7848	0.8976

Table 1: Grid search accuracies for ResNet-50 (CNN) and Vision Transformer (ViT) models.

From the results, it is evident that the configuration (0.0005, 0.01, 64) consistently outperforms others, exhibiting higher accuracy for both CNN and ViT models. This configuration will be selected for the subsequent full training phase.

4.2 Full Training

Following the hyperparameter search, both models were trained for 15 epochs using the selected configuration. The training process aimed to further refine the models' weights and improve their performance. The SGD optimizer with a momentum of 0.9 and weight decay of 0.0001 was utilized for both models. Additionally, a learning rate scheduler with a step size of 5 epochs and a gamma value of 0.1 was employed to adjust the learning rate throughout the training process.

During the full training phase, both models were trained with and without CutMix augmentation. CutMix was applied when the CutMix parameter (beta) was set to 1, enhancing the diversity of the training data and improving model generalization. The training progress was monitored using both the training loss and accuracy metrics, with the CrossEntropy loss function utilized for measurement.

After completing the full training phase, the models' performance was evaluated on a separate testing dataset to assess their generalization ability. The evaluation metrics used were top-1 and top-5 accuracies, inherited from the ImageNet dataset since both models were pretrained on this dataset.

The results of the full training phase will be presented and analyzed in the subsequent section.

5 Experiment Results

5.1 Loss

The training and testing loss curves for both the CNN and ViT models are depicted in Figures 3 and 4, respectively.

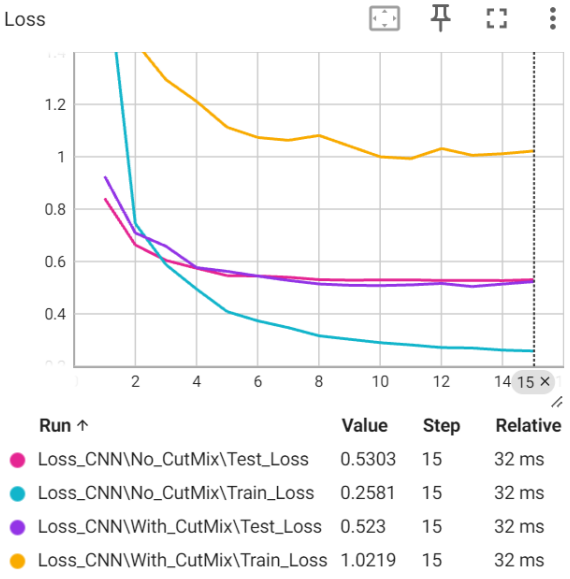

Figure 3: Training and Testing Loss of CNN

Observing the loss curves, it is notable that when CutMix augmentation is applied (beta = 1), the training loss is significantly higher than the testing loss for both models. This discrepancy suggests that the models with CutMix augmentation may be learning more generality from the training data, resulting in higher training loss.

Conversely, when CutMix augmentation is not applied (beta = 0), the training loss is lower than the testing loss. This behavior is typical in training scenarios where the model is able to fit the training data well but struggles to generalize to unseen data.

Interestingly, despite the higher training loss with CutMix augmentation, both sets of models achieve comparable performance in terms of generalization to the testing data, as evidenced by the similar testing loss values.

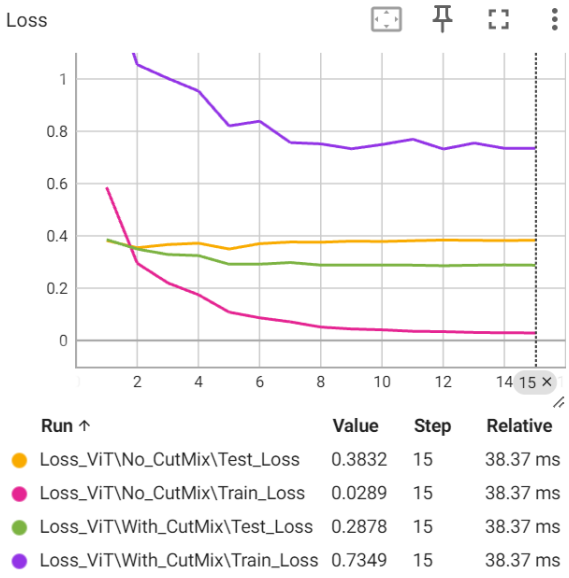

Figure 4: Training and Testing Loss of ViT

This observation suggests that while CutMix augmentation may lead to higher training loss, it does not necessarily hinder the model's ability to generalize to unseen data. Instead, it may promote learning of more general features from the training data, ultimately resulting in comparable performance on the testing data.

This phenomenon holds true for both the CNN and ViT models, indicating that CutMix augmentation affects the training dynamics similarly across different architectures.

5.2 Accuracy

The top-1 and top-5 accuracy results for both the CNN and ViT models are illustrated in Figures 5 and 6, respectively.

For both models, the application of CutMix augmentation results in a slight improvement in both top-1 and top-5 accuracy metrics. Specifically, the ViT model with CutMix achieves a top-1 accuracy of 91.84% and a top-5 accuracy that surpasses its counterpart without CutMix, which has a top-1 accuracy of 90.74%. Similarly, the CNN model with CutMix achieves a top-1 accuracy of 84.73%, compared to 84.21% without CutMix.

Additionally, it is evident from the results that the ViT model consistently outperforms the CNN model in both top-1 and top-5 accuracy metrics, regardless of the CutMix application. This suggests that the self-attention mechanisms and global context modeling capabilities of the ViT model provide a significant advantage over the local spatial

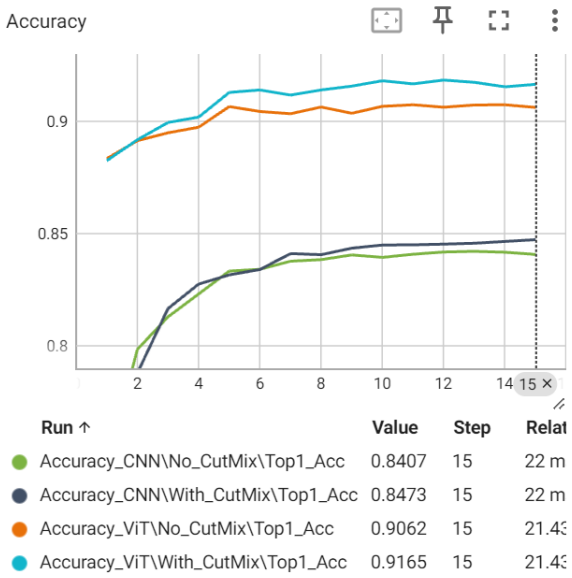

Figure 5: Top 1 Accuracy of CNN and ViT

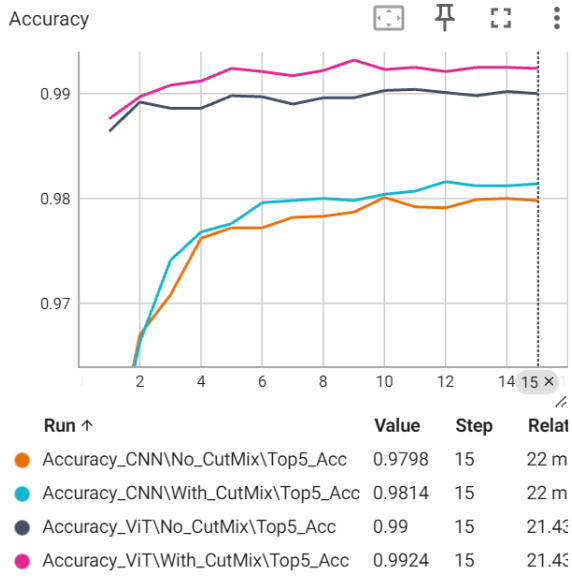

Figure 6: Top 5 Accuracy of CNN and ViT

hierarchies captured by the CNN model.

5.3 Analysis

The results demonstrate the efficacy of CutMix augmentation in enhancing model performance, evident from the higher top-1 and top-5 accuracies achieved. Despite the observed increase in training loss with CutMix, the models' ability to learn more generalized features translates to improved generalization on the testing data.

Moreover, the superior performance of ViT models over CNNs underscores the effectiveness of self-attention mechanisms in capturing long-range dependencies and relevant features. This advantage positions ViT as a compelling choice for image classification tasks, particularly those requiring the modeling of intricate relationships and global context.

The combination of CutMix augmentation with the inherent strengths of the ViT architecture emerges as the most effective approach, yielding the best performance across the experiments. This highlights the potential of integrating advanced data augmentation techniques with state-of-the-art architectures to achieve optimal results in image classification tasks.

6 Discussion & Conclusion

6.1 Performance Comparison

The comparison between Convolutional Neural Network (CNN) and Vision Transformer (ViT) models on the CIFAR-100 dataset illustrates the clear superiority of ViT in terms of both top-1 and top-5 accuracy. With CutMix augmentation, ViT achieves a top-1 accuracy of 91.84%, outperforming the CNN model by a significant margin. Even without CutMix, ViT demonstrates better performance compared to CNN, highlighting its robustness and effectiveness in capturing global context and long-range dependencies.

6.2 Impact of Augmentation

The introduction of CutMix augmentation contributes to a slight improvement in the performance of both models. Despite the associated increase in training loss, CutMix enhances the models' generalization capabilities, leading to improved testing accuracy. This suggests that CutMix aids in learning more generalized features, thereby reducing overfitting tendencies.

6.3 Conclusion

In conclusion, our study underscores the significant advantages of Vision Transformers over traditional CNNs for image classification tasks, particularly on datasets like CIFAR-100. The self-attention mechanisms and global context modeling capabilities inherent in ViT provide a substantial performance boost. Furthermore, the integration of advanced data augmentation techniques such as CutMix further enhances the models' generalization capabilities. These findings emphasize the potential of ViT models, coupled with effective augmentation strategies, as powerful tools for tackling complex image classification challenges.