Image Classification Using ViT on CIFAR-10: A Dual-Phase Training Approach

Roop H. Vankayalapati
Department of Industrial Engineering
Texas A&M University
College Station, TX 77843
roopharshit@tamu.edu

Abstract

Recent advancements in computer vision have been significantly driven by deep learning architectures, with Transformers emerging as a powerful contender to traditional convolutional neural networks (CNNs). This study explores the application of Vision Transformers (ViT) to the CIFAR-10 dataset, a benchmark in image classification. Unlike CNNs, Transformers utilize self-attention mechanisms to capture global dependencies within the input data, which may enhance learning efficacy on complex image datasets. We implement a dual-phase training approach, initially training on a larger subset of the dataset for extensive feature extraction followed by fine-tuning on a smaller subset to refine these features towards specific classes. This method aims to combine the comprehensive feature-learning ability of Transformers with targeted optimization to improve model generalization on unseen data. Our results indicate that while the Vision Transformer achieves commendable training accuracy, it exhibits a gap in validation performance, suggesting overfitting. We discuss these findings and propose future directions to enhance the generalization capabilities of Vision Transformers.

1 Introduction

Image classification is a fundamental task in the field of machine learning, serving as the backbone for numerous applications that impact our daily lives. From facial recognition systems to automated medical diagnosis, the ability to accurately categorize images is crucial. Traditionally, this task has been dominated by Convolutional Neural Networks (CNNs), which have proven effective due to their ability to learn hierarchical image features through local receptive fields. However, CNNs often require extensive data pre-processing and can struggle with global contextual information due to their inherent architectural constraints.

The transformative introduction of the transformer architecture, initially designed for natural language processing tasks, has opened new avenues in image classification. The Vision Transformer (ViT) model adapts the transformer's self-attention mechanism for visual tasks, allowing it to consider the entire image at once rather than processing parts independently. This global processing capability is particularly advantageous for complex image scenes where contextual understanding is essential.

This project explores the application of Vision Transformers to the CIFAR-10 dataset, a staple in the machine learning community for benchmarking image recognition algorithms. The CIFAR-10 dataset presents a diverse set of challenges that make it an ideal candidate for testing the effectiveness of new model architectures like ViT. Our study focuses on implementing a Vision Transformer and optimizing its performance through a dual-phase training approach. Initially, the model is trained on a large subset of the dataset to learn general features and subsequently fine-tuned on a smaller subset to enhance its precision on more nuanced data. Through this project, we aim to evaluate the

Preprint. Under review.

practicality and efficiency of Vision Transformers in handling common image classification tasks compared to traditional CNNs.

2 Methodology

This section outlines the comprehensive approach adopted in designing and training the Vision Transformer (ViT) model for the CIFAR-10 image classification task. It includes descriptions of the dataset, the model architecture, the training procedure, and the evaluation methods used.

2.1 Dataset Description

The CIFAR-10 dataset comprises 60,000 images distributed across 10 classes, with each class containing 6,000 images. Each image is a 32x32 color image. The dataset is split into a training set of 50,000 images and a test set of 10,000 images. For our experiments, the training set was further subdivided into two segments: 45,000 images for the primary training phase and 5,000 images for the fine-tuning phase to refine the model on more nuanced features.

2.2 Model Architecture

The Vision Transformer (ViT) architecture introduces a novel approach to handling image data by adapting the mechanisms that have led to success in natural language processing. Here's an elaboration of each component of the ViT architecture:

- **Patch Embedding Layer:** In conventional convolutional neural networks (CNNs), the convolution operation extracts local features through a sliding window, capturing spatial hierarchies. In contrast, the Vision Transformer begins by reshaping the input image into a sequence of flattened 2D patches. Each patch is then mapped to a D-dimensional embedding space, akin to how words are embedded in natural language models. This step is crucial as it linearizes the 2D structure, allowing the model to process the image similarly to a 1D sequence of tokens in NLP tasks.
- **Multi-Head Self-Attention:** The heart of the Transformer is the self-attention mechanism, which, unlike pooling operations in CNNs that blend inputs within a fixed neighborhood, allows the model to weigh inputs from the entire image irrespective of their positions. This global perspective enables the model to learn contextual relationships between patches. The multi-head self-attention further expands the model's ability to focus on different positional and representational subspaces, enabling parallel processing of multiple attention "views" at once, thereby capturing a diverse range of relationships in the data.
- **Transformer Encoder:** Each encoder layer within the ViT architecture is composed of a multi-head self-attention mechanism followed by a multilayer perceptron (MLP) block. Layer normalization is performed before every block, and residual connections are applied after every block. The sequential application of these layers enables the model to build a complex understanding of the input by reiterating the attention-driven contextualization process, enriching the patch representations with global image information that accumulates through the layers.
- **Classification Head:** At the culmination of the transformation process, ViT employs a classification head to translate the enriched patch embeddings into final class predictions. This is typically achieved using a linear layer, which takes the transformer encoder's output corresponding to a special classification token ('[CLS]'). This token is prepended to the sequence of embedded patches and serves as an aggregate representation of the entire image after passing through the transformer layers. The classification head maps this representation to the number of target classes, producing logits that correspond to class probabilities after a softmax activation. This design allows ViT to make decisions based on the comprehensive and contextual understanding of the image, encapsulated in the '[CLS]' token representation.

2.3 Training Procedure

The training process was carried out in two main stages:

1. **Initial Training:** The model was first trained on the larger subset of 45,000 images for 250 epochs. This phase employed a One Cycle learning rate policy to handle learning rate adjustments, aiming for a quick convergence initially and then fine-tuning the learning rate to stabilize the model's performance.
2. **Fine Tuning:** After initial training, the model was fine-tuned on a smaller set of 5,000 images for 10 epochs. This stage used a reduced learning rate to make subtle adjustments, enhancing the model's ability to generalize from specific features learned during the initial training.

2.4 Training and Evaluation Protocol

The training process is divided into two phases to enhance model performance:

1. **Initial Training:** The model is initially trained on a large subset of the training data (45,000 images) to learn robust feature representations.
2. **Fine-Tuning:** The model is then fine-tuned on a smaller subset (5,000 images) of the training data. This stage is designed to refine the model's weights to better adapt to the nuances and specific features of the dataset.

2.5 Optimization Techniques

We employ the Adam optimizer with a scheduled learning rate, starting with a higher rate and decreasing it based on the OneCycleLR policy to stabilize training in later phases. This adaptive learning rate helps in converging to a better local minimum efficiently.

By leveraging the described methodologies, this project aims to explore the capabilities and limitations of Vision Transformers in handling relatively low-resolution images in CIFAR-10, providing insights into the potential scalability and applicability of transformer models in broader image processing contexts.

2.6 Evaluation Metrics

The model's performance was assessed using the standard classification accuracy metric, which measures the percentage of correctly predicted labels against the true labels in the test set. This metric was crucial for evaluating the effectiveness of the model in handling unseen data and determining its generalization capability.

3 Implementation

This section provides an overview of the practical aspects of implementing the Vision Transformer model for the CIFAR-10 dataset. It details the environment setup, model configuration, and specific adjustments made to optimize the training and evaluation of the model.

3.1 Environment Setup

The implementation was carried out using Python 3.8, with PyTorch as the primary deep learning framework due to its flexibility and efficient tensor operations. Torchvision was utilized for data loading and transformations, which significantly simplifies image preprocessing and augmentation processes. Libraries such as numpy for numerical operations and tqdm for progress monitoring during training were also integral to the development process.

The model training and evaluation were performed on a system equipped with an NVIDIA RTX 4070 CUDA-enabled GPU, which provided the necessary computational power to handle the intensive calculations required by the Vision Transformer architecture. This setup allowed for rapid iteration and experimentation with different configurations and parameters.

3.2 Model Configuration

The Vision Transformer was configured with specific parameters to suit the CIFAR-10 dataset:

- **Image size:** 32x32 pixels, which is the standard size for CIFAR-10 images.
- **Patch size:** 4x4 pixels, allowing the model to capture fine-grained details by dividing each image into smaller segments.
- **Channels:** 3, corresponding to the RGB color space of the images.
- **Embedding dimensions:** 512, which provides a robust feature representation for each patch.
- **Number of heads:** 8, facilitating the model to focus on different subspaces within the embedding.
- **MLP dimension:** 1024, ensuring sufficient capacity to learn complex patterns.
- **Number of layers:** 4, striking a balance between model complexity and computational efficiency.
- **Dropout rate:** 0.1, used to prevent overfitting by randomly dropping units during the training process.

3.3 Data Handling

Data loaders were implemented to efficiently manage data batching, shuffling, and transformations during training and testing phases. The data augmentation techniques included random cropping, horizontal flipping, rotation, and color jittering to enhance the diversity of the training data and simulate various scenarios that could help improve the model's robustness.

3.4 Training Process

Training involved two phases: initial training on 45,000 images and fine-tuning on 5,000 images. The Adam optimizer was used for its adaptive learning rate capabilities, which are particularly effective in converging deep learning models. The learning rate was initially set high and then reduced during the fine-tuning phase to refine the model's weights. Training loss and accuracy metrics were logged in real-time using the tqdm library, providing immediate feedback on the training progress and facilitating early stopping if the validation accuracy ceased to improve.

3.5 Model Evaluation and Testing

The model was evaluated using the test set of 10,000 images to measure its generalization capability. The evaluation process recorded the classification accuracy, providing insights into the model's performance on unseen data. Additionally, predictions were generated on a separate set of publicly available test images to validate the model's effectiveness in a real-world scenario.

These implementation details encapsulate the technical and practical aspects of deploying a Vision Transformer model on the CIFAR-10 dataset, highlighting the comprehensive efforts undertaken to achieve robust performance in image classification tasks.

4 Results and Discussion

The evaluation of the Vision Transformer model on the CIFAR-10 dataset showcased its efficacy and robustness in handling complex image classification tasks. This section provides a detailed analysis of the training performance, validation effectiveness, and results on the public testing dataset.

4.1 Training and Validation Performance

The model underwent an extensive training regimen, consisting of an initial phase on 45,000 images followed by fine-tuning on a subset of 5,000 images. During the initial training phase, the model reached a high training accuracy of 96.5%, demonstrating its ability to effectively learn detailed features from the CIFAR-10 images. However, the validation accuracy at this stage was approximately 83%, which indicated a potential overfitting issue despite the application of regularization techniques such as dropout and extensive data augmentation.

The subsequent fine-tuning phase aimed to refine the model's generalization abilities, resulting in a slight improvement in validation accuracy to around 84.6%. This phase was critical in adapting the model to the nuances of the dataset, improving its performance on validation data. Fine-tuning proved to be a pivotal strategy in mitigating overfitting and enhancing the model's performance on the validation set. This approach is particularly effective in fine-tuning the model to specific subtleties of the dataset, further enhancing its accuracy and generalizability.

4.2 Discrepancy Between Training and Validation Accuracy

The significant gap between training and validation accuracy throughout the training process suggested a tendency for the model to overfit to the training data. This issue was partially addressed during the fine-tuning phase, which emphasizes the importance of this strategy in enhancing model generalization.

4.3 Future Directions

The promising results achieved by the Vision Transformer on the CIFAR-10 dataset highlight its potential in handling complex image classification tasks. However, the performance of Vision Transformers can be further optimized with appropriate resource allocation and extended training sessions. Below, we explore strategic directions to harness the full capabilities of Vision Transformers:

1. **Resource allocation:** Vision Transformers require significant computational power. Investing in better hardware, like advanced GPUs or TPUs, can reduce training times and allow for more extensive model experimentation.
2. **Extended training:** Longer training periods can help the model converge more effectively. Coupling longer training with robust validation can ensure better generalization to unseen data.
3. **Advanced regularization techniques:** Implementing techniques such as stochastic depth or label smoothing can help prevent overfitting, enhancing the model's ability to generalize across various datasets.
4. **Hybrid architectures:** Merging CNNs with Vision Transformers could combine the local processing of CNNs with the global perspective of transformers, potentially leading to more balanced and efficient models.
5. **Scaling up the model:** Larger Vision Transformer models, given sufficient computational resources, could achieve superior performance, following trends observed in recent research.

5 Conclusion

In conclusion, our exploration into the application of Vision Transformers (ViT) on the CIFAR-10 dataset highlighted the architecture's robustness in handling complex image data through its innovative use of self-attention mechanisms. The dual-phase training strategy, involving initial extensive training followed by targeted fine-tuning, demonstrated potential in enhancing model performance, although it also revealed challenges related to overfitting as evidenced by the disparity between training and validation accuracies.

This study reaffirms the importance of careful hyperparameter tuning and model architecture adjustments to balance between model complexity and generalization capabilities. While ViTs show promise in achieving high accuracy on training data, their performance on unseen data suggests that further research into regularization techniques and training strategies is necessary to fully leverage their capabilities in practical applications.

Future work will focus on integrating more sophisticated data augmentation and regularization techniques to mitigate overfitting, exploring the scalability of ViTs to larger datasets, and comparing their performance with state-of-the-art CNNs under similar conditions. The ultimate goal is to refine the application of Vision Transformers in image recognition tasks, pushing the boundaries of what is achievable with this emerging model architecture.