Optimizing Wide-ResNet Architecture with Binary Operations and Knowledge Distillation for Improved CIFAR-10 Classification

Richik Majumder

Department of Electrical Engineering, IIT Kanpur

Introduction

This project focuses on improving the CIFAR-10 classification accuracy using a modified Wide-ResNet architecture. In Phase 1, binary operations are applied to the initial convolution layer of the model, inspired by BONN and XNOR-Net. In Phase 2, knowledge distillation is applied to boost accuracy, with a reduced CNN model acting as the student. This approach led to significant improvement in classification performance.

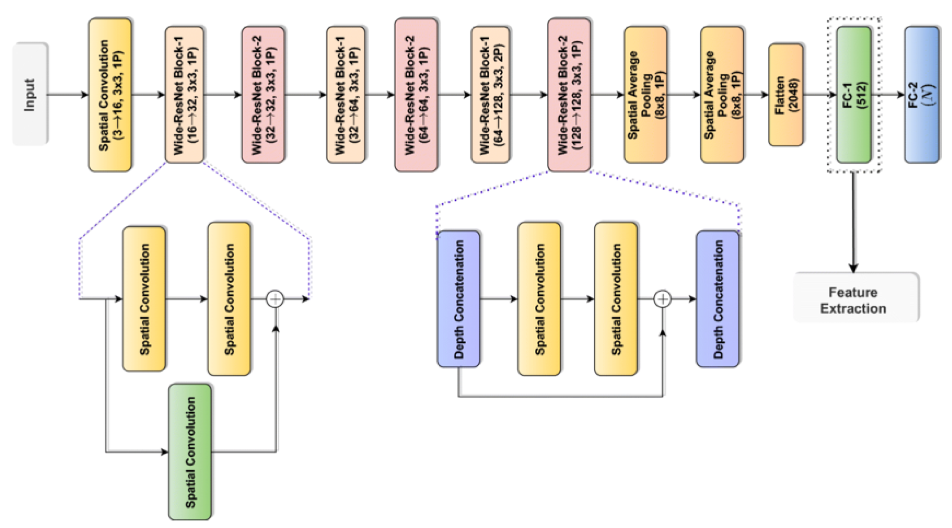

Figure 1. Architecture of the BONN-WideResNet model with binary operations on the initial convolution layer.

This approach led to a significant improvement in CIFAR-10 classification accuracy by applying the aforementioned strategies, combining efficiency with high performance.

Key Contributions

This research highlights several important contributions:

- **Binary operations on initial convolution layer:** Reduced computational complexity while maintaining classification accuracy.
- **Knowledge distillation:** Improved CIFAR-10 classification performance by transferring knowledge from a larger model (teacher) to a smaller model (student).
- **Phase 1 vs Phase 2 performance:** Achieved a significant accuracy improvement from 70% in Phase 1 to 80% in Phase 2 through the combination of binary operations and knowledge distillation.

These approaches showcase the potential to reduce the computational cost without sacrificing performance, offering a practical solution for resource-constrained environments.

Key Insights

This section highlights critical observations that drive the project's success.

The integration of binary operations and knowledge distillation significantly enhances both the efficiency and accuracy of the model. By reducing the computational cost through binary operations on the initial convolution layer, the model performs better while being more resource-efficient. Knowledge distillation, on the other hand, enables the student model to outperform its size through learning from the teacher model.

- **Improved accuracy through binary operations:** By binarizing the initial convolution layer, the model achieves similar performance at a fraction of the computational cost.
- **Performance boost with knowledge distillation:** The transfer of knowledge from the teacher model to the student model enhances overall classification accuracy.
- **Significant improvement from Phase 1 to Phase 2:** Accuracy jumps from 70% to 80% as a result of the combined techniques in Phase 2.

Steps Involved in the Approach

The process of improving the CIFAR-10 classification accuracy through binary operations and knowledge distillation involves several key steps, each contributing to the final model's success.

1. **Phase 1: Binary Operations**, focusing on the modification of the initial convolution layer using binary weights to reduce computational complexity while maintaining performance.
2. **Phase 2: Knowledge Distillation**, where a smaller student model learns from a larger teacher model, improving its accuracy and generalization capabilities.
3. **Performance Evaluation**, involving training and testing the models, followed by a thorough comparison between the results of Phase 1 and Phase 2 to assess the impact of the optimization techniques.

CIFAR-10 Classification Results

The experimental setup involved training models on the CIFAR-10 dataset. The binary convolution and knowledge distillation techniques were applied to improve the model's accuracy and reduce computational complexity. The following figures compare the results of Phase 1 and Phase 2 in terms of training and testing loss.

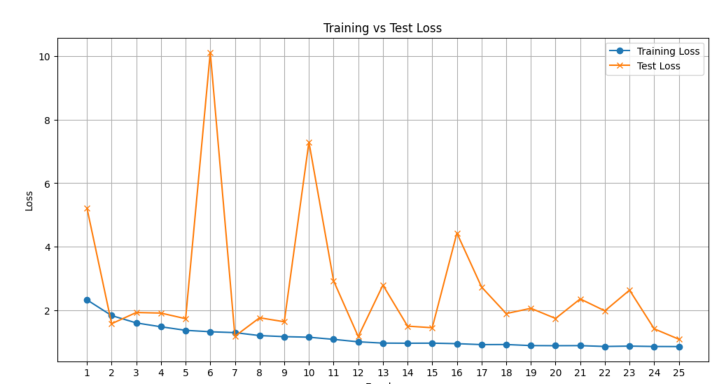

Figure 2. Training and Testing Loss for Phase 1.

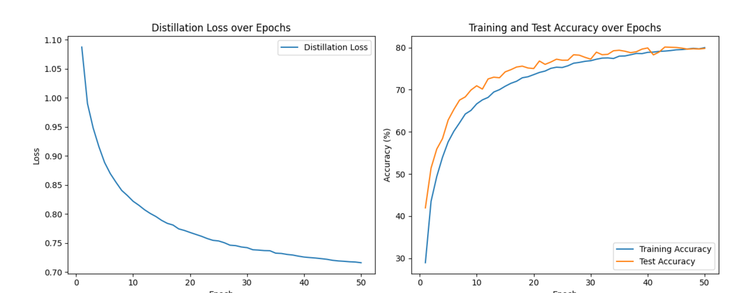

Figure 3. Training and Testing Loss for Phase 2.

CIFAR-10 Classification with Binary Operations and Knowledge Distillation

In this research, we applied binary operations and knowledge distillation to improve the classification performance on the CIFAR-10 dataset. The model architecture used was a Wide-ResNet with modifications to the convolution layer, inspired by BONN and XNOR-Net. We compared two phases: Phase 1, which involves binary convolution operations, and Phase 2, where knowledge distillation was applied to enhance accuracy. The results showed significant improvement in performance, with accuracy increasing from 70% to 80%.

In Phase 1, the binary operations reduced computational complexity by binarizing the weights in the initial convolution layers. In Phase 2, we applied knowledge distillation, where the smaller student model learned from a larger teacher model, leading to better performance.

The following key points summarize the main findings:

- **Phase 1:** Binary operations were used to reduce computational cost without sacrificing accuracy.
  - The initial convolution layer was binarized using a binary modulation vector, reducing the model's overall complexity.
  - This phase achieved an accuracy of 70% on the CIFAR-10 test set.
- **Phase 2:** Knowledge distillation was used to further improve accuracy.
  - The smaller student model learned from a larger teacher model, resulting in improved performance.
  - The accuracy increased to 80% with this approach.
- **Future Work:** Test on more datasets and explore additional optimizations.

A highlighted block containing some math

This block demonstrates the application of optimization in machine learning.

Model Architecture

The model architecture is based on a modified Wide-ResNet with binary operations on the initial convolution layer. This approach reduces computational complexity while maintaining accuracy, as demonstrated by the following equation:

$$\text{Binary Convolution}(W, X) = W \cdot \text{sign}(X)$$

Knowledge Distillation

The knowledge distillation technique involves transferring knowledge from a larger teacher model to a smaller student model. This allows the student to learn better features, improving its performance. We can represent this transfer of knowledge as:

$$L_{\text{KD}} = \alpha \cdot L_{\text{CE}} + (1 - \alpha) \cdot L_{\text{KL}}$$

Model Comparison and Results

This block compares the performance of the models used in Phase 1 (BON-WideResNet) and Phase 2 (Distillation Model) on key metrics.

Model	Accuracy (%)	Parameters
BON-WideResNet	70.2	8.4M
ResNet	75.5	11.2M
Distillation (Student)	80.1	4.6M

Table 1. Model comparison in Phase 1 and Phase 2.

The confusion matrix for these models is shown below:

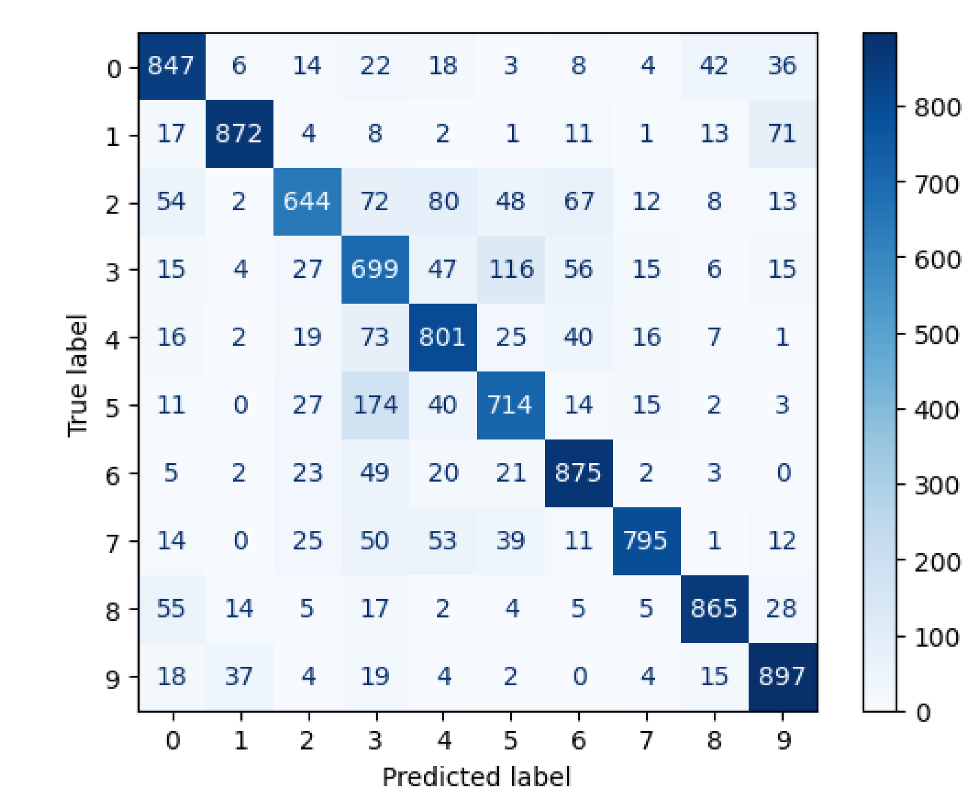

Figure 4. Confusion Matrix for Phase 1 and Phase 2 Models.

Key Observations:

- BON-WideResNet: 70% accuracy, high computational cost.
- ResNet: Improved accuracy, larger model.
- Distillation: 80% accuracy with reduced model size.

These results show that knowledge distillation improves accuracy and reduces complexity.

- **Reference:** Gu, S., al. (2019). Bayesian Optimized 1-Bit CNNs. `https://openaccess.thecvf.com/content_ICCV_2019/papers/Gu_Bayesian_Optimized_1-Bit_CNNs_ICCV_2019_paper.pdf`