Knowledge Distillation for CIFAR-10 Classification in BONN-WideResNet Model

1 Introduction

This project aims to improve CIFAR-10 classification in a binary optimized CNN by using knowledge distillation. Building on work from **here**, where Bayesian methods optimized a 1-bit CNN, we enhance accuracy and speed using a teacher-student framework. Here, a larger, well-trained teacher model guides a smaller student model, increasing accuracy and reducing training time. I presented a Bayesian optimized 1-bit CNN for efficient image classification, focusing on reducing model size while maintaining accuracy. Knowledge distillation, a method where a large model's knowledge is transferred to a smaller model, has shown effectiveness in achieving high accuracy with fewer resources.

2 Methodology

2.1 Dataset

The CIFAR-10 dataset, containing 60,000 images across 10 classes, was split into 50,000 training images and 10,000 test images. Data augmentations like random cropping and flipping were applied to improve generalization.

2.2 Models

The teacher model is a pre-trained ResNet-18 adapted for CIFAR-10. The student model is a modified CNN with reduced parameters on the initial convolution layer.

2.3 Distillation Loss

The distillation loss, $\mathcal{L}_{\text{distill}}$, is a combination of hard and soft loss components:

$$\mathcal{L}_{\text{distill}} = (1 - \alpha) \cdot \mathcal{L}_{\text{hard}} + \alpha \cdot \mathcal{L}_{\text{soft}}$$

where α balances these components, and temperature scaling smoothens the teacher's output to aid student learning.

3 Experimental Setup

3.1 Hyperparameters

The training used a learning rate of 0.01, batch size of 64, and weight decay of 5×10^{-4}. The distillation temperature was set to 2, and α was set to 0.5.

3.2 Training Process

The model was trained for 50 epochs using batch gradient descent with dropout of 0.5. Metrics tracked included training and test accuracy, as well as loss values.

4 Results and Analysis

In Phase 1, using binary convolution, the model achieved approximately 70% test accuracy. In Phase 2, knowledge distillation improved the test accuracy to around 80%. Figure 1 shows the training and distillation loss over epochs, as well as the training and test accuracy.

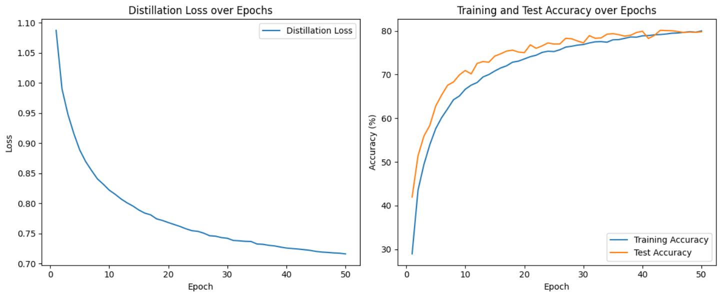

Figure 1: Training and Distillation Loss over Epochs, with Training and Test Accuracy

5 Conclusion

Using knowledge distillation significantly improved the performance of the binary optimized CNN on CIFAR-10, increasing accuracy from 70% to 80%. Future work could explore parameter tuning or apply this method to other datasets.

References

- Gu, J., Shen, Y., Zhao, W., Wang, Y., & Tan, M. (2019). Bayesian Optimized 1-Bit CNNs. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Available at: `https://openaccess.thecvf.com/content_ICCV_2019/papers/Gu_Bayesian_Optimized_1-Bit_CNNs_ICCV_2019_paper.pdf`