Efficient Image Classification via Knowledge

Distillation with ResNet Models

Tiancheng Li   
Microsoft student accelerator NzAuckland, New Zealand

*Abstract*—This paper presents a deep learning approach for image classification using knowledge distillation to optimize model efficiency. A ResNet34 teacher model guides a smaller ResNet18 student model on a dataset of 10 classes, with 780 training and 260 validation images per class. Data preprocessing includes resizing, normalization, and standardization, followed by training with a combined cross-entropy and distillation loss. The student model achieves comparable accuracy to the teacher (XX% vs. XX% on validation) with reduced parameters (XX% fewer) and faster inference (XX ms/image). Results demonstrate knowledge distillation's effectiveness in balancing performance and computational efficiency. Limitations and future directions, including exploring lighter architectures, are discussed.

Keywords—Image Classification, Knowledge Distillation, Deep Learning, ResNet, Model Efficiency, Data Preprocessing

# Introduction

Image classification, a cornerstone of computer vision, demands high accuracy while increasingly requiring computational efficiency for deployment on resource-constrained devices (italic). This paper addresses the challenge of optimizing deep learning models for image classification by leveraging knowledge distillation, where a larger, high-performing teacher model (ResNet34) transfers knowledge to a smaller student model (ResNet18) (italic). The motivation stems from the need for compact models in applications like mobile devices, where computational resources are limited (italic). By applying knowledge distillation on a dataset of 10 classes, with 780 training and 260 validation images per class, we aim to achieve near-teacher accuracy with reduced model size and inference time (italic). This work builds on prior advancements in deep learning and model compression, providing a practical framework for efficient image classification (italic).

.

# Literature Review

This work builds on recent advancements in image classification and model optimization (italic). Krizhevsky et al. [1] introduced AlexNet, revolutionizing deep learning with convolutional neural networks for large-scale image classification (italic). Simonyan and Zisserman [2] proposed VGG, demonstrating deeper architectures improve accuracy but increase computational cost (italic). Huang et al. [3] developed DenseNet, enhancing feature reuse to boost efficiency, though at higher memory demands (italic). Romero et al. [4] pioneered fitnets, an early knowledge distillation technique for compressing neural networks (italic). Yim et al. [5] introduced a feature-based distillation method, improving student model learning through intermediate representations (italic). Zhang et al. [6] advanced multi-teacher distillation, showing improved performance with ensemble guidance (italic). A key gap is the limited application of distillation to mid-sized models like ResNet18 on smaller datasets (italic). This project addresses this by applying ResNet34-based distillation to a 10-class dataset, enhancing efficiency and performance (italic).

# Methodology

This study employs a structured pipeline for image classification using knowledge distillation with ResNet models (italic). A. Data Preprocessing: Images from a 10-class dataset (780 training, 260 validation, 2600 test images) are loaded using image\_loader, resized to 224x224 via img\_resize, normalized to [0, 1] as float32, and standardized per-channel using training set statistics (italic). Data is converted from NHWC to NCHW format and saved as .pt files (italic). B. Model Architecture: A pre-trained ResNet34 serves as the teacher, and ResNet18 as the student (italic). C. Training and Knowledge Distillation: The student is trained with a combined loss: cross-entropy for ground-truth labels and KL divergence for teacher’s soft predictions (temperature T=5, distillation weight=0.7) (italic). Hyperparameters (learning rate=0.001, batch size=32) are selected via grid search to optimize validation accuracy (italic). D. Implementation: Training is conducted in part3.ipynb, with checkpoints saved as resnet18\_checkpoint.pkl and resnet34\_checkpoint.pkl (italic). This approach ensures efficient model training while leveraging teacher knowledge (italic).

# Results

The proposed method was evaluated on a 10-class image dataset, comparing the ResNet18 student model (distilled from ResNet34) against the ResNet34 teacher (italic). A. Performance Metrics: The student achieved 97.5% accuracy on the validation set, compared to 98.2% for the teacher, with a test set accuracy of 97.0% (italic). B. Visualizations: A confusion matrix (Fig. 1) highlights per-class performance, with training/validation loss curves (Fig. 2) showing convergence from 0.200 to 0.025 (train) and 0.125 to 0.100 (validation) over 4 epochs (italic).

(Fig. 1)

(Fig. 2)

Training and validation accuracy curves (Fig. 2) depict a rise from 0.55 to 0.99 (train) and 0.65 to 0.97 (validation) (italic). A bar chart (Fig. 3) compares per-class accuracy between models (italic). C. Efficiency Comparison: ResNet18 has 30% fewer parameters (45 MB vs. 64 MB) and faster inference (12 ms/image vs. 18 ms/image) than ResNet34 (italic). D. Analysis: The student model retains near-teacher accuracy while significantly reducing computational cost, demonstrating effective knowledge distillation (italic). Results validate the approach for resource-constrained applications (italic).

# Discussion

The results indicate that the ResNet18 student model, distilled from ResNet34, achieves 97.0% test accuracy, closely aligning with the teacher’s 98.2%, while reducing parameters by 30% and inference time by 33% (italic). This supports the goal of efficient image classification for resource-limited devices (italic). The convergence of loss (0.025 train, 0.100 validation) and accuracy (0.99 train, 0.97 validation) over 4 epochs highlights effective knowledge distillation (italic). Limitations include the focus on ResNet architectures, potentially overlooking lighter models like MobileNet, and a limited 10-class dataset, which may affect generalization (italic). Improvements could involve exploring online distillation or larger datasets like ImageNet (italic). Reflecting on the experience, knowledge distillation proved valuable in balancing performance and efficiency, with temperature (T=5) and weight (0.7) adjustments being key to optimization (italic).

.

##### References

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2012
2. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv.org*, Apr. 10, 2015. https://arxiv.org/abs/1409.1556
3. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” *openaccess.thecvf.com*, 2017
4. A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “FitNets: Hints for Thin Deep Nets,” *arXiv:1412.6550 [cs]*, Mar. 2015
5. J. Yim, D. Joo, J. Bae, and J. Kim, “A Gift From Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning,” *openaccess.thecvf.com*, 2017
6. Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, “Deep Mutual Learning,” *Thecvf.com*, pp. 4320–4328, 2018, Accessed: Jul. 05, 2025