Comparative Evaluation of Small CNN, Fine-Tuned ImageNet Models, and Knowledge Distilled Classifiers on CIFAR-10

Sakib Bin Faruque Rusho
ID : 2010776109

Contents

1 Introduction

The CIFAR-10 dataset consists of 60,000 color images (50,000 training and 10,000 test) divided into 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Each image is 32×32 pixels in size.

The goal of this study is to evaluate and compare the performance of five different classifiers:

- **Small CNN**: Small CNN trained from scratch.
- **ResNet50**: Fine-tuned ResNet50 pre-trained on ImageNet.
- **VGG16**: Fine-tuned VGG16 pre-trained on ImageNet.
- **KD ResNet50**: Small CNN trained using knowledge distillation (KD) from ResNet50.
- **KD Ensemble(ResNet50+VGG16)**: Small CNN trained using KD from an ensemble of ResNet50 and VGG16.

The main objectives are:

1. Assess the effectiveness of fine-tuning large pre-trained models on a small resolution dataset (CIFAR-10).
2. Evaluate the impact of knowledge distillation on improving small CNN performance.
3. Investigate whether ensemble-based distillation enhances generalization further.

2 Methodology

2.1 Dataset Preparation

The CIFAR-10 dataset was used with the following preprocessing:

- Normalized pixel values to [0, 1].
- One-hot encoded the class labels.
- Split: 50,000 training images and 10,000 testing images.

2.2 Model Architectures

2.2.1 Small CNN

- Conv2D (32 filters, 3×3) $\rightarrow$ MaxPooling (2×2) $\rightarrow$ Dropout (0.25)
- Conv2D (64 filters, 3×3) $\rightarrow$ MaxPooling (2×2) $\rightarrow$ Dropout (0.25)
- Flatten $\rightarrow$ Dense (256 units, ReLU) $\rightarrow$ Dropout (0.5) $\rightarrow$ Dense (10, Softmax)

Optimizer: Adam (learning rate = 0.001). Trained for 20 epochs.

2.2.2 Fine-Tuned ResNet50

- Pre-trained ResNet50 (ImageNet) with last 2 layers unfrozen.
- Global Average Pooling $\rightarrow$ Dense (10, Softmax).

Optimizer: Adam (learning rate = 0.0001).

2.2.3 Fine-Tuned VGG16

- Pre-trained VGG16 (ImageNet) with last 2 layers unfrozen.
- Flatten $\rightarrow$ Dense (256, ReLU) $\rightarrow$ Dense (10, Softmax).

Optimizer: Adam (learning rate = 0.0001).

2.2.4 Knowledge Distillation Models

KD uses a temperature $T = 3.0$ and a loss function:

$$\mathcal{L} = \alpha \times \text{KLDiv(soft targets)} + (1 - \alpha) \times \text{CrossEntropy(hard labels)},$$

where $\alpha = 0.5$.

- **KD ResNet50**: Teacher = ResNet50.
- **KD Ensemble**: Teacher = Ensemble of ResNet50 and VGG16 (logits averaged).

3 Training and Evaluation

- Batch size: 64
- Epochs: 20 (Early stopping with patience = 3)
- Metrics: Test Accuracy, Per-Class Accuracy, Confusion Matrix, Precision, Recall, and F1-score.

4 Results and Analysis

4.1 Test Accuracy Summary

Model	**Test Accuracy (%)**
Small CNN	70.57
Fine-tuned ResNet50	42.81
Fine-tuned VGG16	61.24
KD from ResNet50	68.79
KD from ResNet50 + VGG16	66.13

Table 1: Test accuracies for all models.

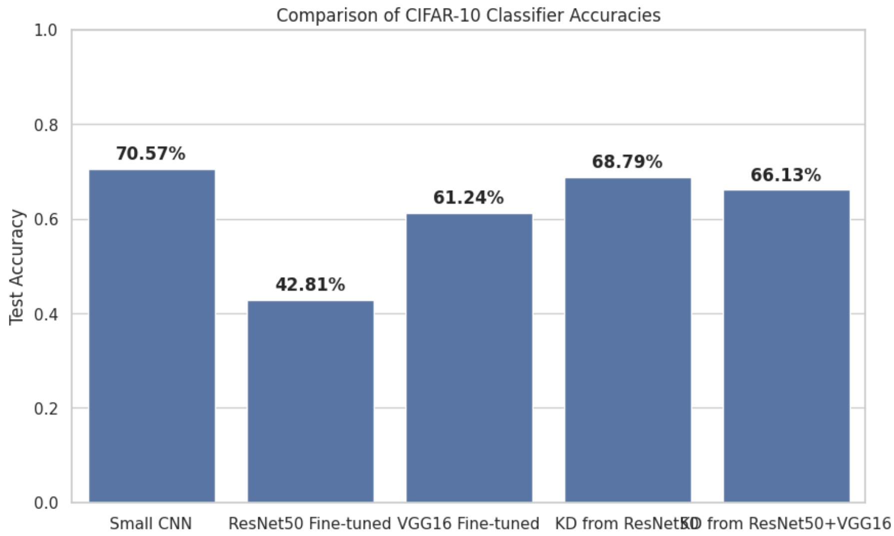


Figure 1: Overall test accuracy comparison across models.

4.2 Detailed Per-Class Metrics

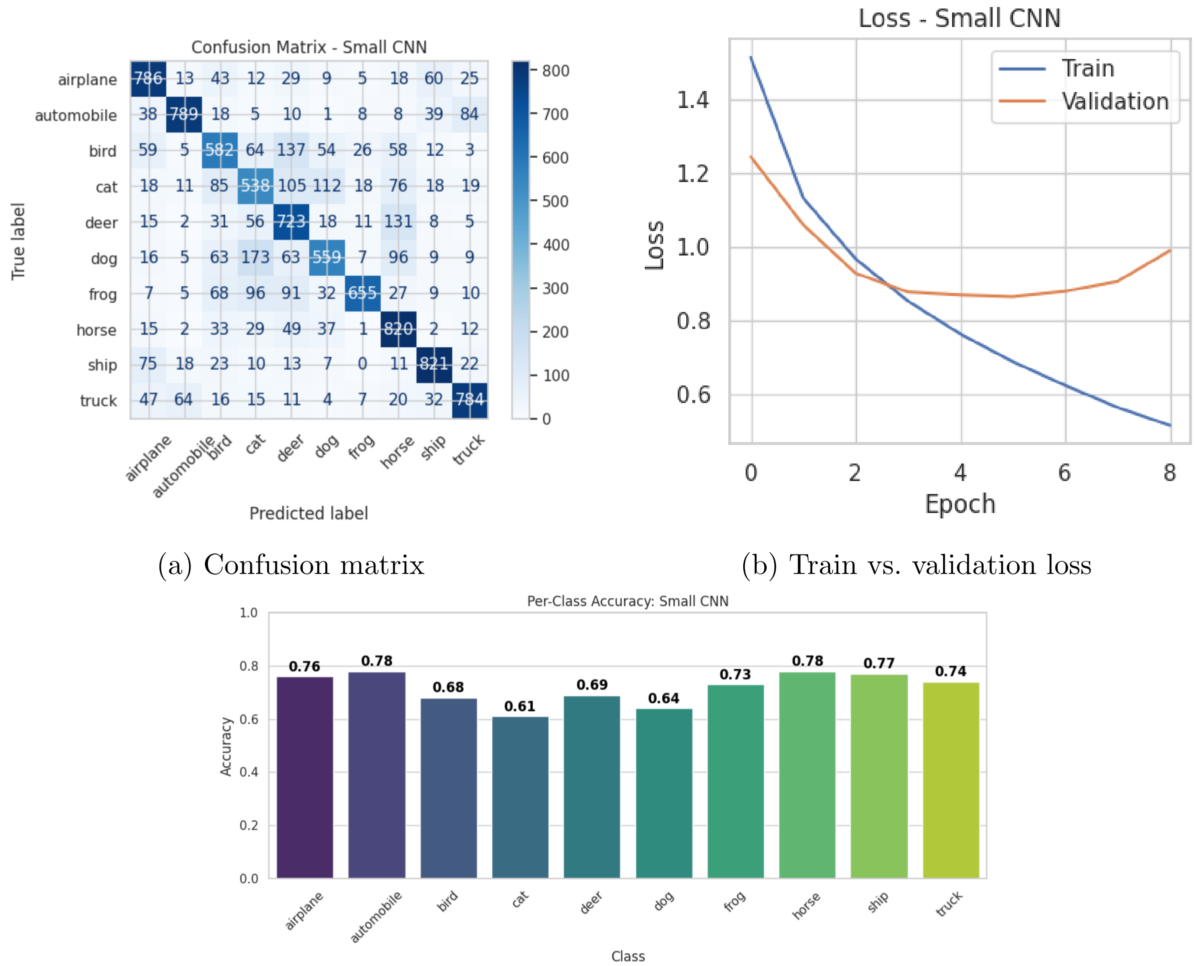
Class	**Small CNN**				**ResNet50**				**VGG16**				**KD ResNet50**				**KD Ensemble**			
	Prec.	Recall	F1-score	Acc(%)	Prec.	Recall	F1-score	Acc(%)	Prec.	Recall	F1-score	Acc(%)	Prec.	Recall	F1-score	Acc(%)	Prec.	Recall	F1-score	Acc(%)
airplane	0.73	0.79	0.76	78.6	0.49	0.52	0.50	52.2	0.63	0.74	0.68	74.3	0.71	0.79	0.75	79.1	0.67	0.72	0.70	70.0
automobile	0.86	0.79	0.82	78.9	0.39	0.64	0.48	64.4	0.65	0.71	0.68	71.3	0.80	0.81	0.80	84.0	0.77	0.82	0.79	75.7
bird	0.60	0.58	0.59	58.2	0.41	0.13	0.20	13.2	0.58	0.47	0.52	46.6	0.59	0.59	0.59	52.7	0.58	0.59	0.59	48.4
cat	0.54	0.54	0.54	53.8	0.29	0.26	0.27	26.1	0.45	0.47	0.46	46.5	0.57	0.42	0.48	37.4	0.50	0.54	0.52	36.0
deer	0.59	0.72	0.65	72.3	0.37	0.43	0.40	43.5	0.54	0.56	0.55	56.3	0.62	0.69	0.65	74.8	0.67	0.59	0.63	75.0
dog	0.67	0.56	0.61	55.9	0.38	0.43	0.41	43.0	0.56	0.52	0.54	51.9	0.66	0.61	0.64	39.9	0.57	0.61	0.59	52.9
frog	0.89	0.66	0.75	65.5	0.42	0.52	0.46	52.4	0.69	0.59	0.64	59.1	0.80	0.76	0.78	80.2	0.74	0.76	0.75	73.2
horse	0.65	0.82	0.72	82.0	0.53	0.44	0.48	44.1	0.70	0.68	0.69	67.8	0.74	0.79	0.77	80.4	0.79	0.66	0.72	79.9
ship	0.81	0.82	0.82	82.1	0.53	0.52	0.52	52.1	0.69	0.76	0.72	75.8	0.71	0.85	0.78	80.5	0.80	0.77	0.79	75.3
truck	0.81	0.78	0.79	78.4	0.56	0.37	0.45	37.1	0.62	0.63	0.63	62.8	0.81	0.73	0.77	78.9	0.75	0.76	0.76	74.9

Table 2: Per-class Precision, Recall, F1-score, and Accuracy (%) for all models on CIFAR-10 test set.

4.3 Small CNN

The small CNN achieved the highest accuracy of 70.57%.

Observation. The architecture is lightweight and better optimized for small input images (32 × 32). It avoids the inductive bias and capacity mismatch of very deep ImageNet models at this resolution, yielding better bias–variance trade-off on CIFAR-10.



(a) Confusion matrix

(b) Train vs. validation loss

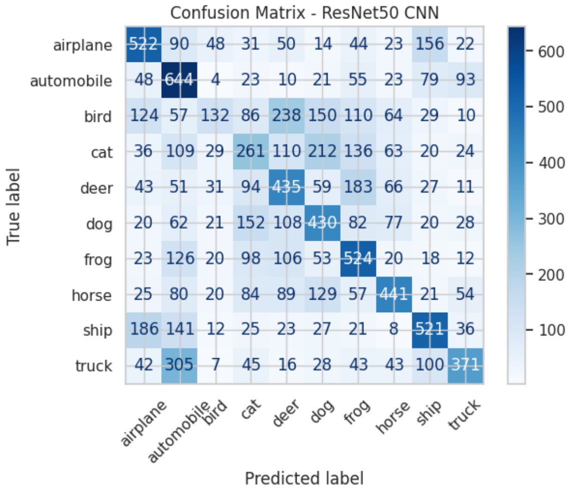
(c) Per-class accuracy

Figure 2: Diagnostics for Small CNN.

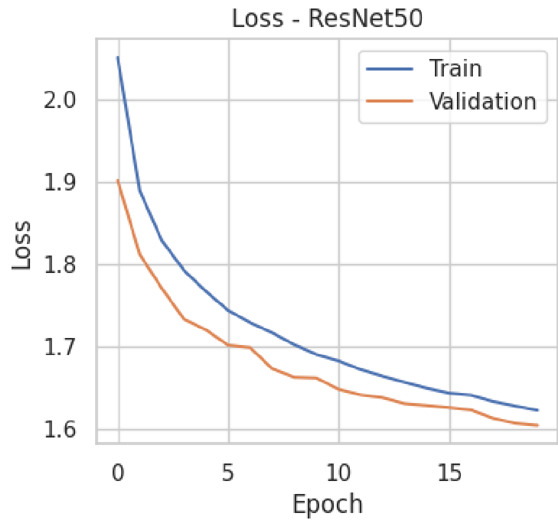
4.4 Fine-Tuned ResNet50

ResNet50 achieved only 42.81% accuracy.

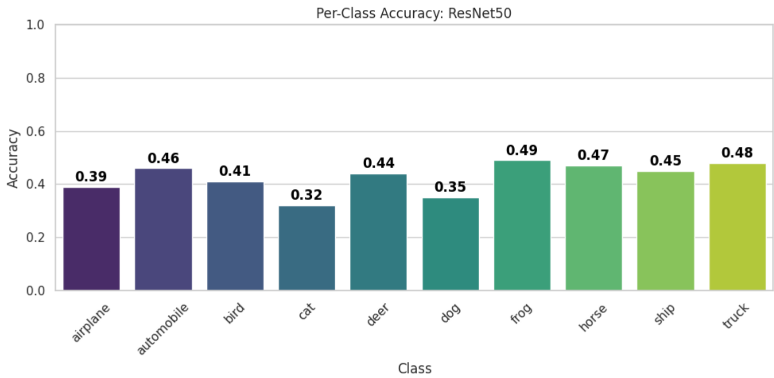
Observation. Fine-tuning large models on low-resolution images is sensitive: the stem and early blocks of ResNet50 are optimized for 224×224 inputs and rich textures. With limited unfreezing (last 1–2 layers) and small images, the learned features transfer poorly; additionally, overfitting can occur due to high capacity relative to the dataset without sufficient augmentation or longer schedule.



(a) Confusion matrix



(b) Train vs. validation loss



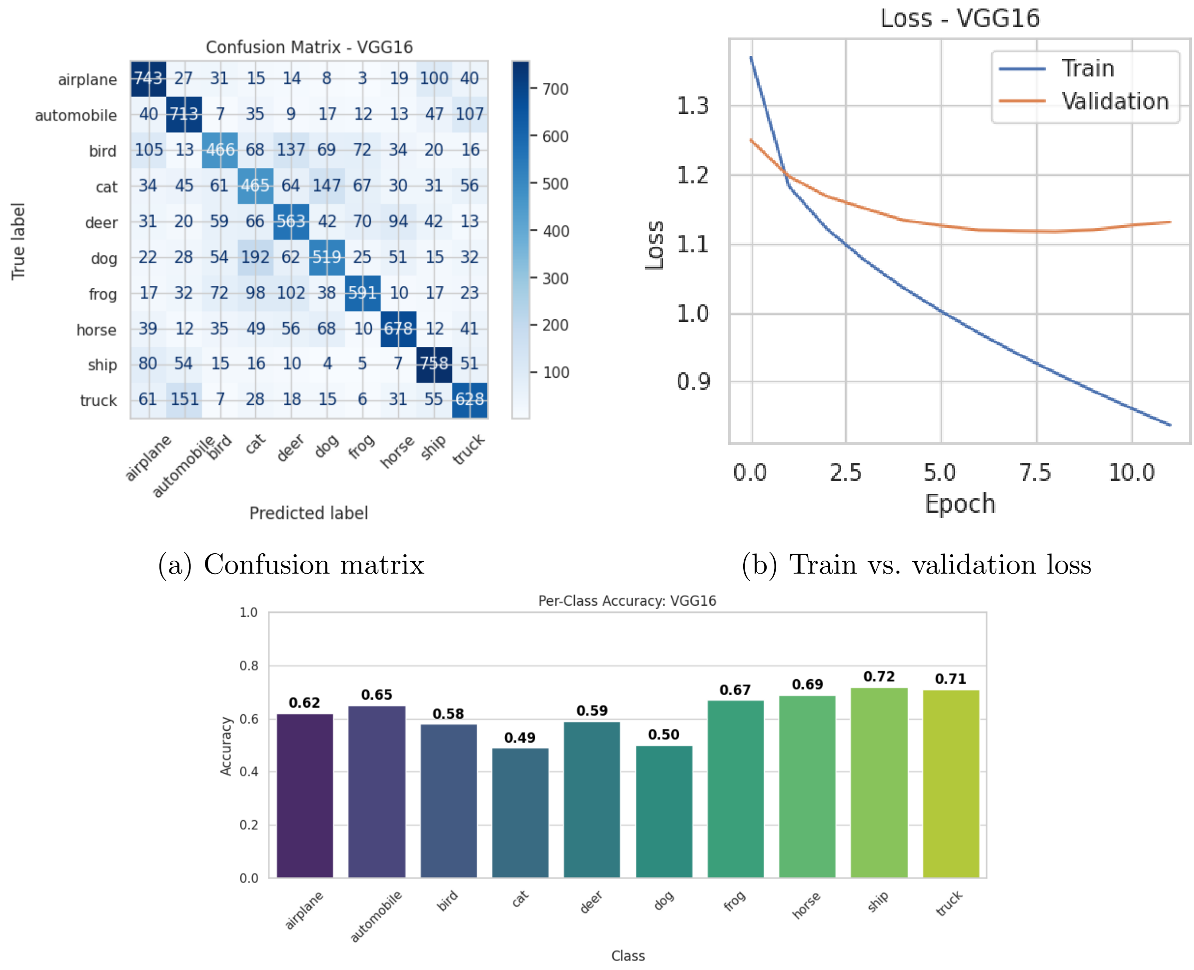
(c) Per-class accuracy

Figure 3: Diagnostics for ResNet50 fine-tuned.

4.5 Fine-Tuned VGG16

VGG16 achieved 61.24% accuracy.

Observation. VGG16's simpler, more local-receptive-field features adapt better to 32×32 images. Although still constrained by partial unfreezing and the mismatch to ImageNet scale, its shallow early-stage filters transfer more robustly than ResNet50's residual blocks under the same fine-tuning budget.



(a) Confusion matrix

(b) Train vs. validation loss

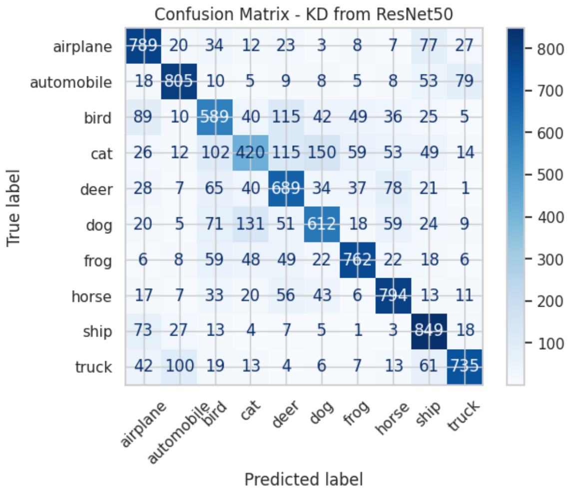
(c) Per-class accuracy

Figure 4: Diagnostics for VGG16 fine-tuned.

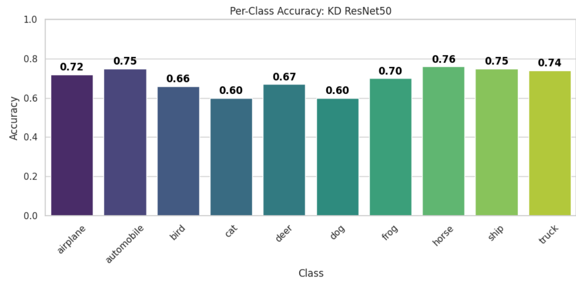
4.6 KD from ResNet50

Knowledge distillation (teacher: ResNet50) boosted the small CNN to 68.79%.

Observation. Soft targets encode inter-class similarities ("dark knowledge"), guiding the student to less overconfident boundaries and better calibration. Despite the teacher's modest top-line accuracy, its logits still carry useful relational structure that improves the student's generalization nearly to the scratch-trained baseline.



(a) Confusion matrix



(b) Train vs. validation loss

(c) Per-class accuracy

Figure 5: Diagnostics for KD from ResNet50.

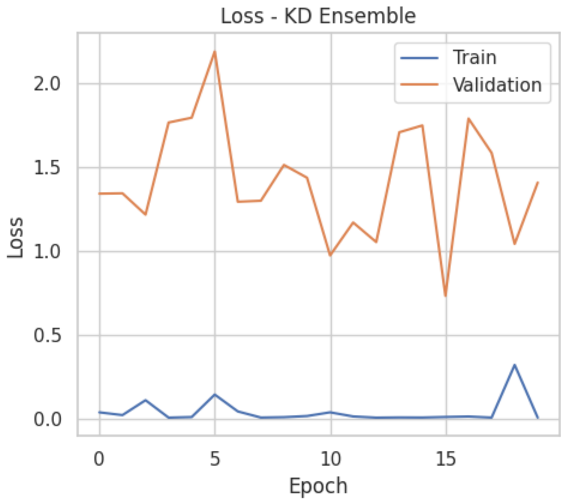
4.7 KD from ResNet50 + VGG16 Ensemble

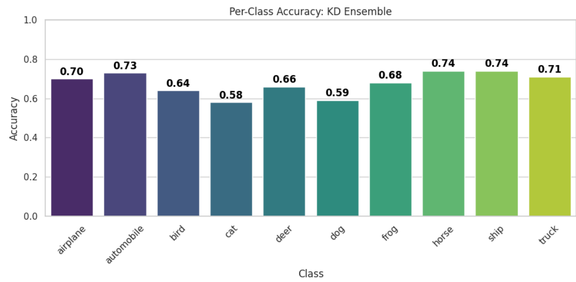
KD from the ensemble teacher yielded 66.13%.

Observation. Although the ensemble teacher has better top-1 accuracy, distillation with multiple teachers (via logits averaging) gave slightly lower student accuracy than the single ResNet50 teacher distillation. Possible reasons:

- Averaged logits smooth inter-class differences excessively, causing loss of useful discriminative signal.

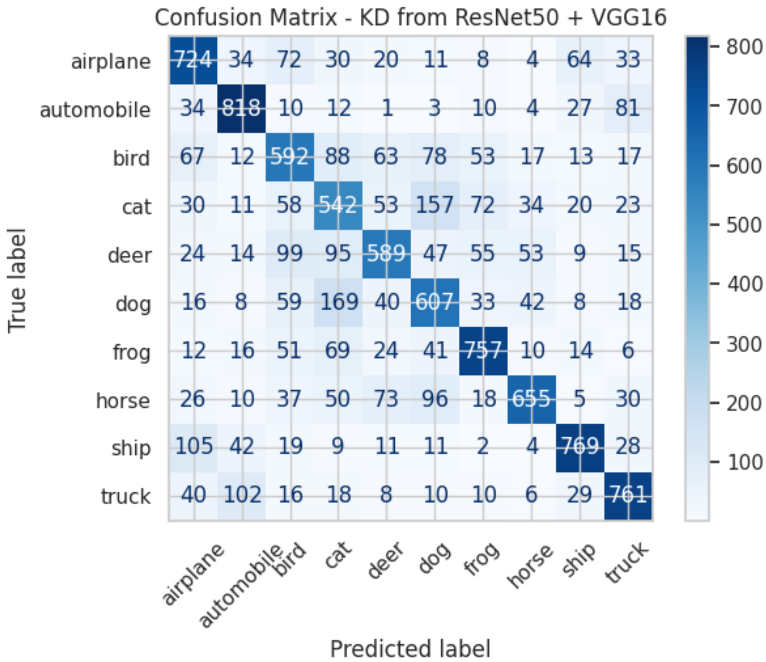
- Optimization complexity or hyperparameter mismatch.





(a) Confusion matrix

(b) Train vs. validation loss



(c) Per-class accuracy

Figure 6: Diagnostics for KD from ResNet50 + VGG16 ensemble.

5 Conclusions

- The small CNN trained from scratch performs best on CIFAR-10 at 70.57%.
- Fine-tuning large ImageNet models with limited unfreezing and small inputs is challenging; ResNet50 suffers from severe performance degradation.
- VGG16 fine-tuning is more effective, reaching 61.24%.

- Knowledge distillation improves small CNN performance by transferring teacher knowledge, nearly matching the baseline.
- Distillation from an ensemble teacher did not improve student accuracy beyond the single-teacher KD, highlighting the need for better ensemble distillation techniques.

6 GitHub Repository

The complete code, trained models, and additional resources for this project are available at:

`https://github.com/sbfrusho/Deep-Learning---CSE4261.git`