

# Learning “Dark Knowledge” from Teacher

Aditya Choudhary  
21b090001@iitb.ac.in

Sharad Vishwakarma  
200040132@iitb.ac.in

Yash Bhake  
22B2148@iitb.ac.in

## I. INTRODUCTION

As part of our project, we studied the topic of Knowledge Distillation. All our work was done using Google Colab. The corresponding GitHub repository is also linked here

## II. OVERVIEW OF THE TOPIC

This discord between training and test objectives leads to the development of machine learning models that yield good accuracy on curated validation datasets but often fail to meet performance, latency, and throughput benchmarks at the time of inference on real-world test data.

Knowledge distillation helps overcome these challenges by capturing and “distilling” the knowledge in a complex machine learning model or an ensemble of models into a smaller single model that is much easier to deploy without significant loss in performance

The hypothesis is that the student model will learn to mimic the predictions of the teacher model. This can be achieved by using a loss function, termed the distillation loss, that captures the difference between the logits of the student and the teacher model respectively. As this loss is minimized over training, the student model will become better at making the same predictions as the teacher.

## III. IMPLEMENTATION DETAILS

### A. Dataset

We used the CIFAR-10 Dataset that contains 10 classes of images. The dataset was firstly rescaled to 224x224 to match the architecture used by ResNet and also normlaized using the mean and standard deviations of the ImageNet dataset

### B. Teacher Network

The Teacher Network was a ResNet 50 Model trained for 10 Epochs with the Cross Entropy Loss

Epoch	Training Loss	Accuracy
1	0.7513	0.7537
2	0.5895	0.7943
3	0.5631	0.8044
4	0.5494	0.81
5	0.5296	0.8154
6	0.524	0.8189
7	0.5149	0.821
8	0.5151	0.8218
9	0.5114	0.823
10	0.5107	0.8218

TABLE I  
TEACHER MODEL

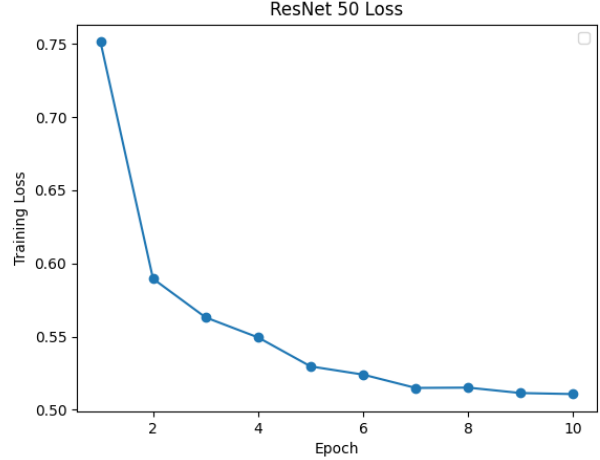


Fig. 1. Teacher Loss

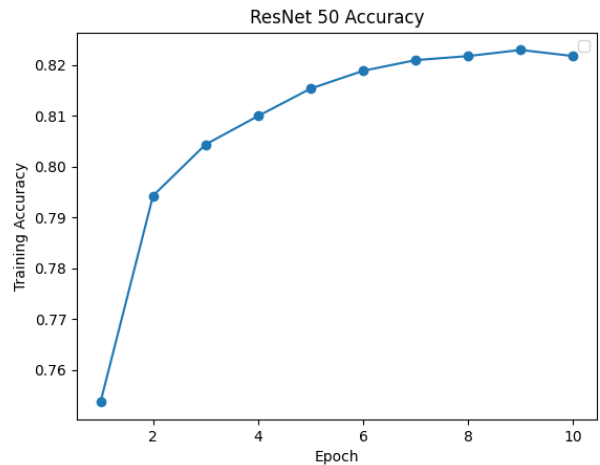


Fig. 2. Teacher Accuracy

### C. Student Network

Student network was a custom CNN built using PyTorch. We trained two networks, one with the CrossEntropyLoss (referred as Without KD from hereon) and one with Knowledge Distillation Loss (referred as With KD from hereon)

$$\mathcal{L}_{KD} = \alpha * T^2 * D_{KL}(P||Q) + (1 - \alpha) * \mathcal{L}_{CrossEntropy}$$

### IV. OUR PROPOSED WORK

We train a network to judge how well a smaller network with lesser parameters could learn and mimic the original network and thereby save us resources. For this purpose, we have trained a CNN with and without the Knowledge Distillation Loss to know how much of a role does KD Loss play in training a newer network using "Dark Knowledge"

We also tried experimenting with a smaller student network however going to smaller architectures (lesser number of parameters) wasn't feasible as training such a network become tough as accuracy over 10s of Epochs stayed close to 0.1 which is to be expected from random guessing as well

### V. RESULTS

It was observed that in the initial epochs KD Loss helped the model learn a lot quickly over its counterpart by having  $\approx 15\%$  higher accuracy. However as the number of epochs increases both the models start to converge to similar results however the model with KD loss still holds an upperhand with 1% better accuracy by the 10th Epoch

This could be attributed to the fact that ResNet 50 is not the most complex model available and was quite similar in architecture of the custom CNN used

Using a much smaller custom CNN, it was observed that using KD Loss did not make a difference as the CNN was too simple to generate usable predictions and hence in either case (with and without KD loss) it was giving accuracy close to 10% which is expected from random guessing as well

Epoch	With KD	Without KD
1	0.4017	0.2597
2	0.4533	0.3762
3	0.49	0.4451
4	0.5192	0.477
5	0.5419	0.5124
6	0.556	0.5312
7	0.5649	0.5448
8	0.5788	0.5586
9	0.5867	0.5747
10	0.5962	0.5855

TABLE II  
STUDENT ACCURACY

Both the models show a similar Loss curve in their training

### REFERENCES

- [1] <https://arxiv.org/abs/1503.02531>
- [2] <https://neptune.ai/blog/knowledge-distillation>
- [3] [https://medium.com/@satya15july\\_11937/understanding-knowledge-distillation-in-simple-steps-3310ef01f5e0](https://medium.com/@satya15july_11937/understanding-knowledge-distillation-in-simple-steps-3310ef01f5e0)

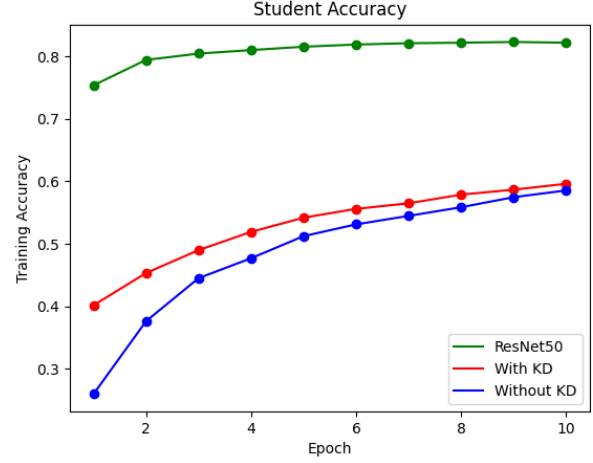


Fig. 3. Model Accuracy

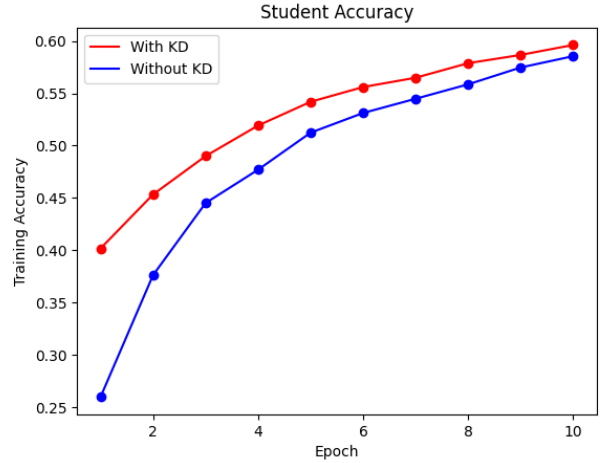


Fig. 4. Student Accuracy

Epoch	With KD	Without KD
1	0.9521	1.9639
2	0.8998	1.6948
3	0.8618	1.5449
4	0.8292	1.4489
5	0.8075	1.3655
6	0.7902	1.3215
7	0.7788	1.2807
8	0.7658	1.2426
9	0.7567	1.2049
10	0.7483	1.1705

TABLE III  
STUDENT LOSS

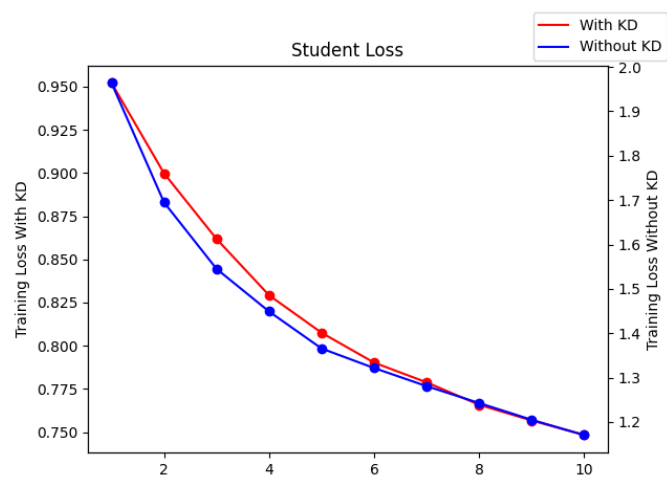


Fig. 5. Student Loss