Knowledge Distillation of Convolutional Neural Networks

Final Project

Oscar Navarrete Parra

Universidad Autónoma de Yucatán, Facultad de Matemáticas

May, 2024

1. Introduction
2. Training Method
3. Dataset
4. Model

1 Introduction

2 Training Method

3 Dataset

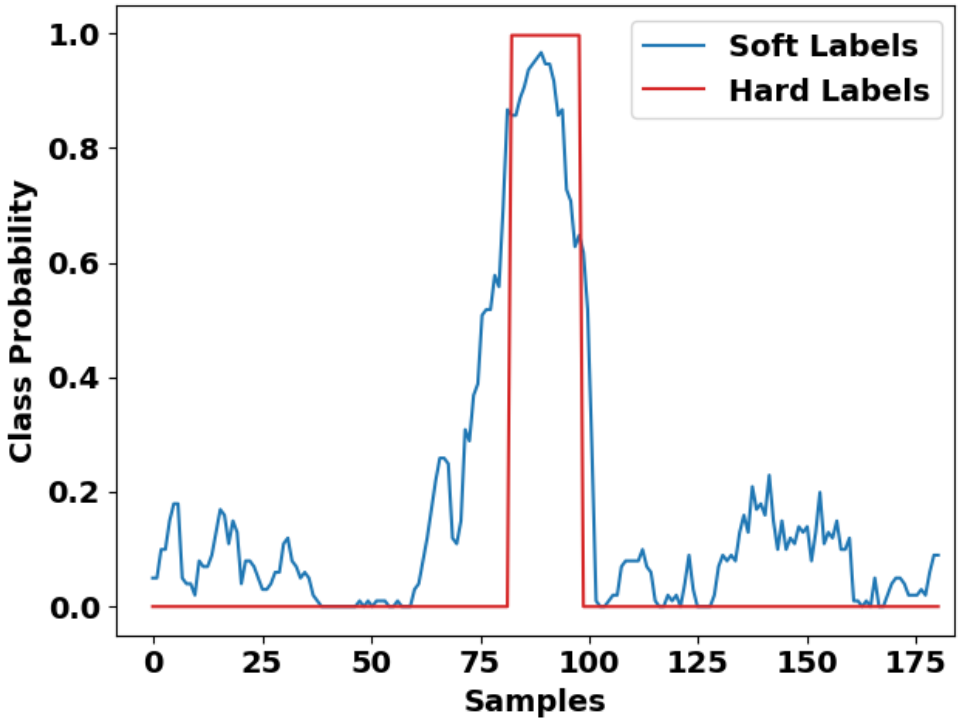
4 Model

Motivation

- In ML, speed and hardware requirements of deployed solutions are important issues.
- In training stage, these constraints are not as strict as in inference.
- After training knowledge is encoded in parameters.
- But model's knowledge can also be interpreted as a learned mapping between input and output vectors.

Soft Targets VS Hard Targets

- Classifiers assigns probabilities of incorrect and correct classes.
- The relative probabilities of incorrect answers tell us a lot about how a model tends to generalize.



1 Introduction

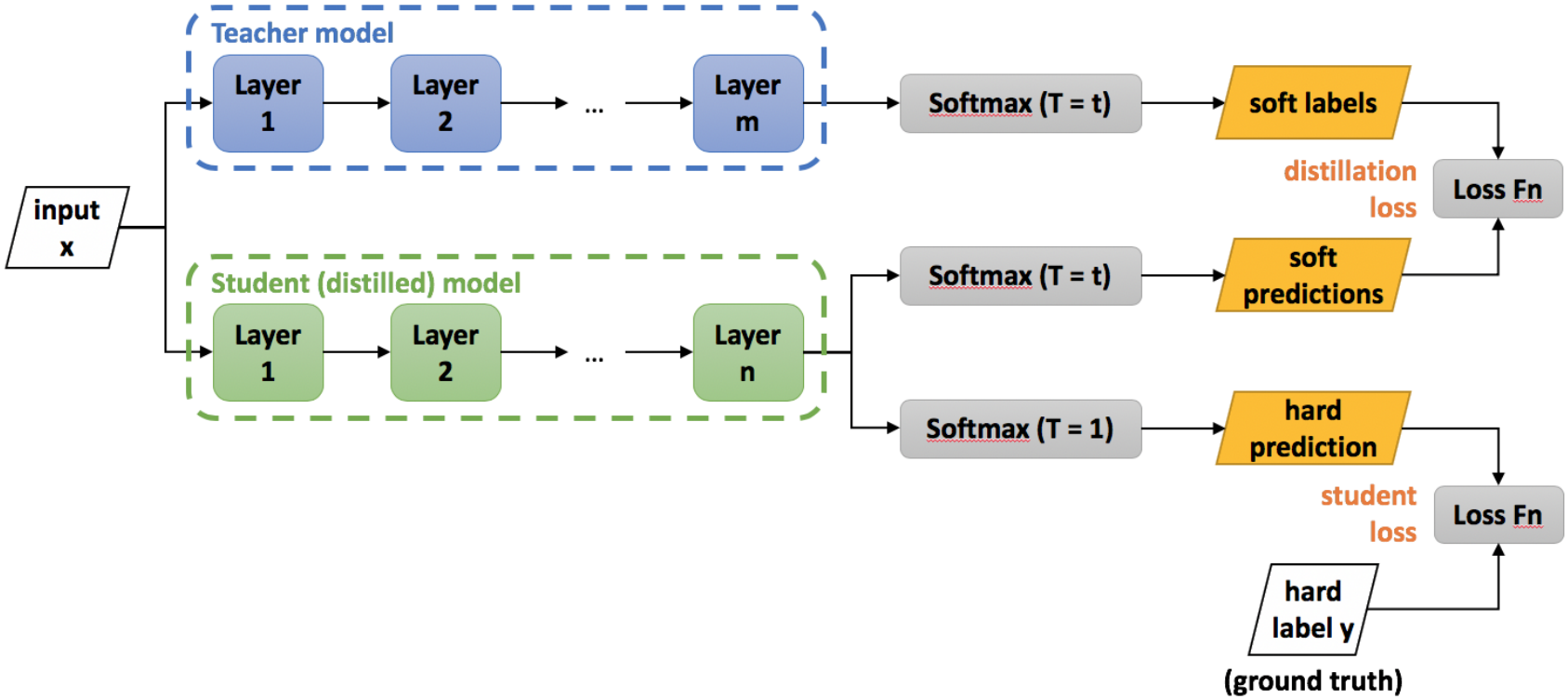
2 Training Method

3 Dataset

4 Model

Knowledge Distillation (KD)

- We use (KD) to transfer the knowledge from a large model, also called the Teacher, to a smaller model, also known as the Student.
- Hence, the Student learns to mimic the output of the Teacher.



1 Introduction

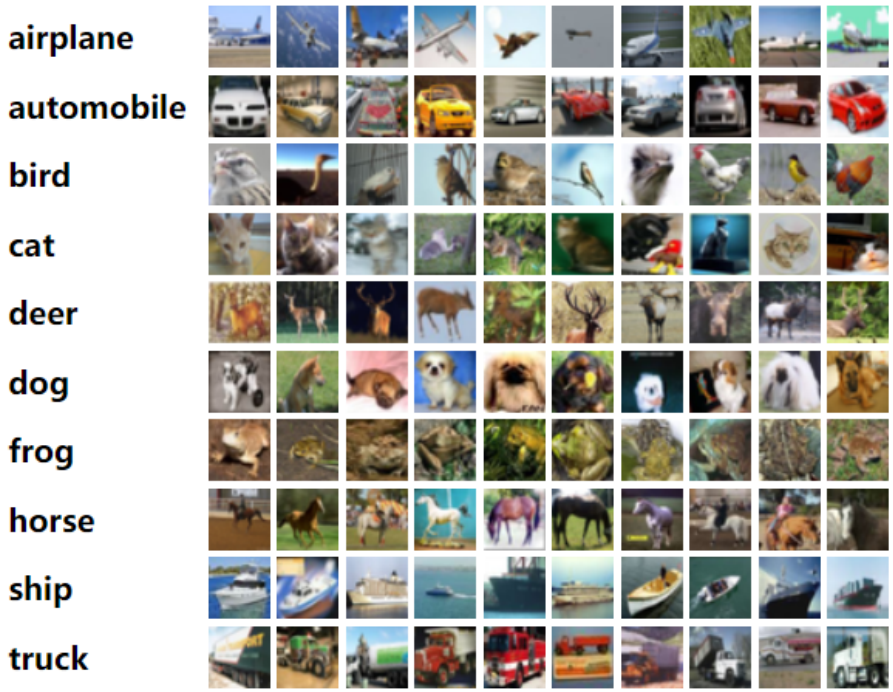
2 Training Method

3 Dataset

4 Model

CIFAR-10/100

- We will train a classifier of different image classes.
- CIFAR-100 would be ideal but is highly expensive to train. An alternative could be CIFAR-10.



1 Introduction

2 Training Method

3 Dataset

4 Model

Student and Teacher Model

- For the teacher model, we need a NN that can provide us with rich image representations (i.e. good probabilities over all our image classes).
- We can choose from a CNN like ResNet to a Vision Transformer model like OpenAI CLIP.
- For our student we will build our own Convolutional Neural Network with a custom architecture small enough to produce accurate predictions and fast inferences.

Evaluation

- Finally, we will evaluate our results on a test split by comparing the accuracy and speed of our distilled model against the teacher model and the student trained only on hard labels.
- We will use precision, recall and F1-score to measure the model's performance and analyze the effectiveness of our approach.