Distilling the Knowledge in a Neural Network

Authors : Geoffrey Hinton, Oriol Vinyals, Jeff Dean

Problem Statement

Although we can increase the performance of a given machine learning algorithm by taking the same data and training an ensemble of models and finally averaging their predictions. Unfortunately training ensemble models becomes cumbersome and not feasible in GPU poor setting.

Hence this work answers the question “Can we train small model which can perform on par with a large model on the same data?” in other words can we transfer knowledge from a larger model to a smaller model and expect the small model to perform on par with the larger model?

Experiment Details

- The authors apply this method on two different tasks, Image Classification and Speech Recognition.
- Dataset used for image classification is MNIST dataset and JFT dataset (this is an internal dataset used by Google with 100M images and 15K labels)
- For speech recognition they use 2K hours of spoken English data which produces 700M training examples.
- For experiments with MNIST dataset they trained large model with 2 hidden layers and 1200 ReLU units. For the smaller model had 2 hidden layers and the number neuron changed from 800, 300,30 units with varying temperatures.
- For speech, the authors use an architecture with 8 hidden layers each with 2560 hidden units and final softmax layer with 14K labels.

Results for Image task

- The larger model achieved 67 test errors where as the small model with 800 units achieved 146 error.
- When the smaller model was regularized only with soft labels generated by the larger model with temperature 20, it achieved 74 test errors. This clearly shows that there is transfer of knowledge.
- When the distilled net had 300 or more units in each of its two hidden layers, all temperatures above 8 gave fairly similar results.
- But when this was radically reduced to 30 units per layer, temperatures in the range 2.5 to 4 worked significantly better than higher or lower temperatures.
- Further more experiment results are presented in the paper.

Results on Speech task

System	Test Frame Accuracy	WER
Baseline	58.9%	10.9%
10xEnsemble	61.1%	10.7%
Distilled Single model	60.8%	10.7%

Table 1: Frame classification accuracy and WER showing that the distilled single model performs about as well as the averaged predictions of 10 models that were used to create the soft targets.

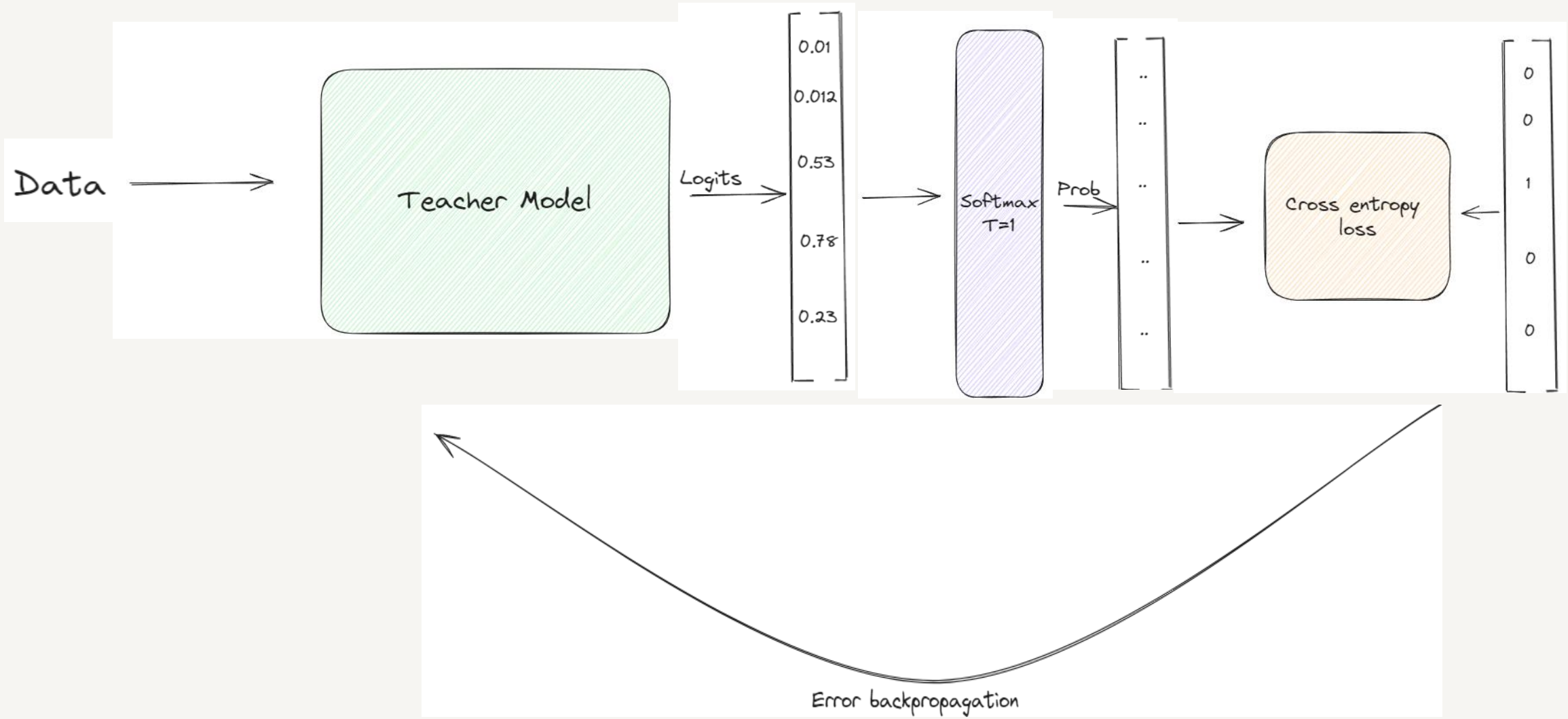
Methodology in detail

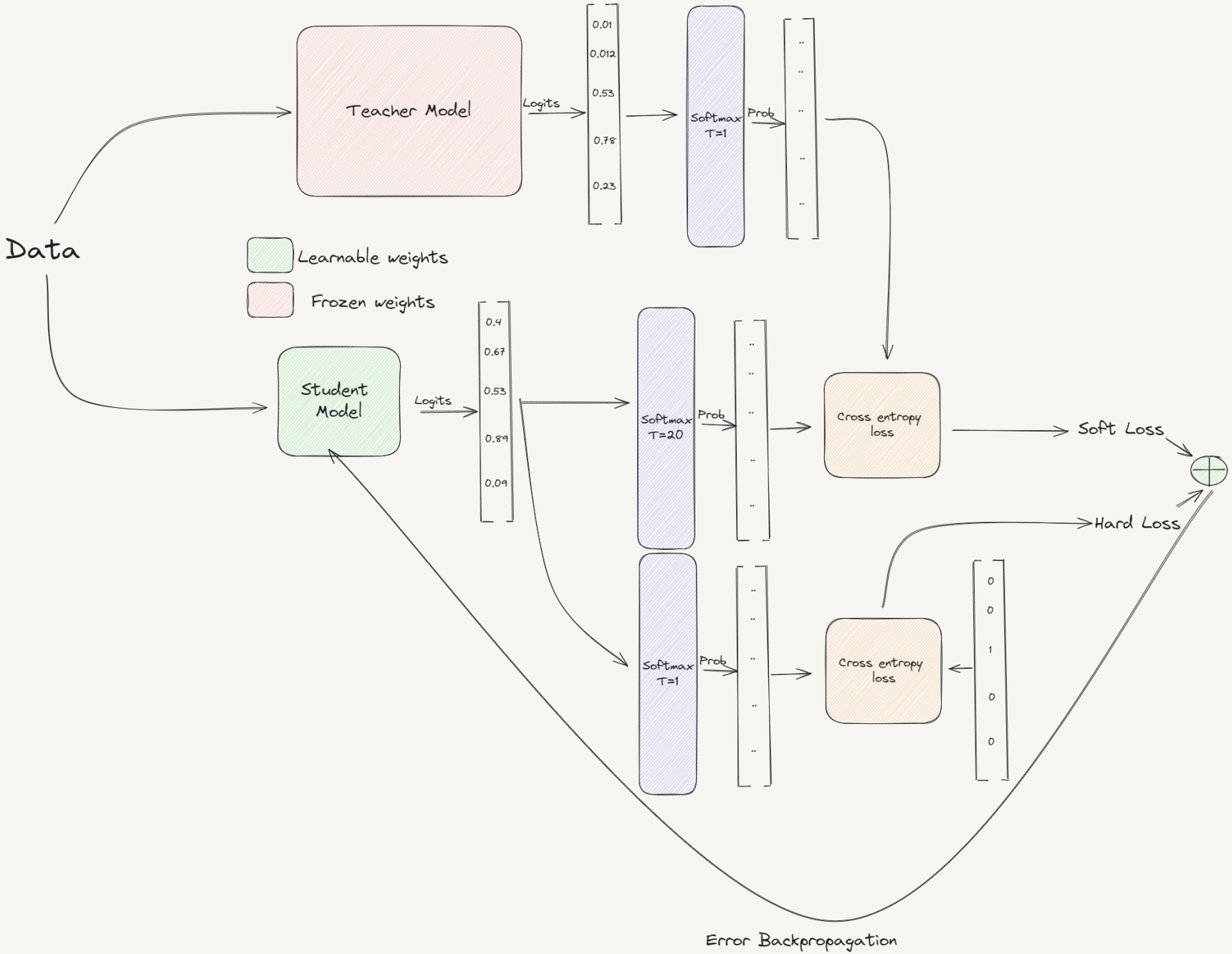
Softmax and effect of temperature

- Softmax is a function which converts the output logits to a probability distribution, and is governed by the equation given below,

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

To understand the effect of the temperature on these probability lets visit the collab notebook.


Data
Teacher Model
Logits
0.01
0.012
0.53
0.78
0.23
Softmax
T=1
Prob
Cross entropy
loss
0
0
1
0
0
Error backpropagation


Data
Teacher Model
Logits
0.01
0.012
0.53
0.78
0.23
Softmax T=1
Prob
Learnable weights
Frozen weights
Student Model
Logits
0.4
0.67
0.53
0.89
0.09
Softmax T=20
Prob
Cross entropy loss
Soft Loss
Hard Loss
Softmax T=1
Prob
Cross entropy loss
0
0
1
0
0
Error Backpropagation

Thank you