Knowledge Distillation

Farouk Soufary

January 2024

ENSEIRB-MATMECA - IS319 Deep Learning

Table of contents

Definitions

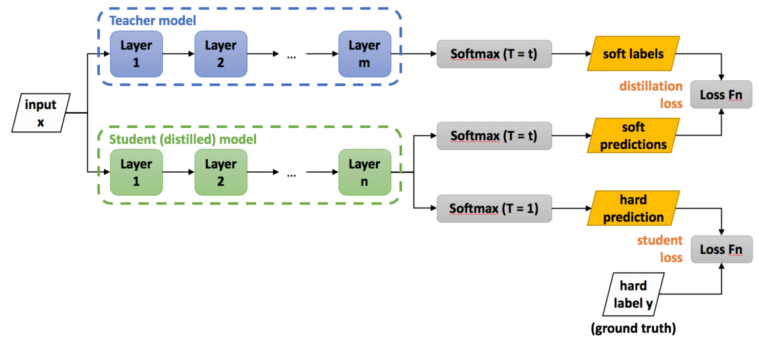
What is knowledge distillation ?

- Model Compression method proposed in 2015 by **Geoffrey Hinton**[1], **Oriol Vinyals** and **Jeff Dean**
- Transfers knowledge from a Large model (Teacher) to a smaller model (Student)
- Enables faster inference for deployment

[1] https://arxiv.org/abs/1503.02531

How does Knowledge Distillation work ?

Idea : transfer the knowledge in the function (teacher model) into a smaller model



- Learn from the output of the teacher (Dark Knowledge)
- Learn from the ground truth label as well

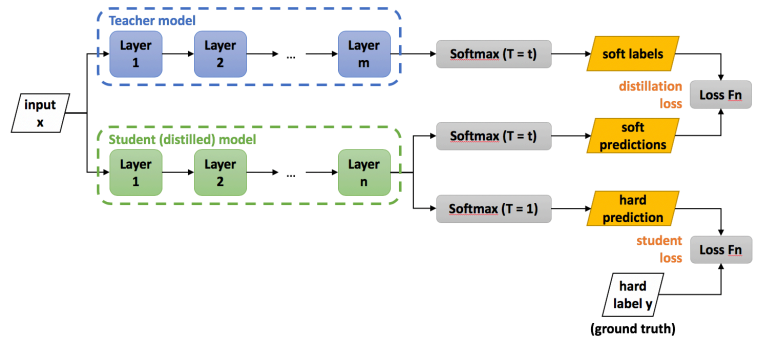
Dark Knowledge

	cow	dog	cat	car
raw logits	11	16	22	2
Softmax	1.6e-5	2.4e-3	9.975	2.0e-9
T = 3	0.02	0.11	0.86	0.001
T = 5	0.07	0.21	0.69	0.01

$$\sigma(y,T)_{(i)} = \frac{e^{(y_i/T)}}{\Sigma_j e^{(y_j/T)}}$$

- Soft Labels reveal the dark knowledge in the teacher model
- Training with soft labels imposes more constraint on parameters
- Works better to fit on soft labels as well as hard targets

How does Knowledge Distillation work ?



$$\sigma(y, T)_{(i)} = \frac{e^{(y_i/T)}}{\Sigma_j e^{(y_j/T)}}$$

$$\mathcal{L}(x, W) = \alpha \times CE(y, \sigma(y_s, T)) + \beta \times CE(\sigma(y_t, T), \sigma(y_s, T))$$

- The CE with hard targets is weighted-down because the derivatives for the soft targets tend to be smaller

Hypothesis on why it works

With a carefully selected temperature value:

- Soft targets prevent the model from being overly sure
- Soft targets add more constraints on the weights during training.
- Soft targets enhance the capacity of the model to generalize since it provides additional information (Resemblance between classes, for example).

Experiments & Results

Hinton's Experiments on MNIST (1)

- Trained a large neural net with two hidden layers of 1200 neurons on MNIST
- Transfered the knowledge to a student network with two hidden layers of 800 neurones (with and without ground truth regularization)
- The Teacher achieved 67 test errors and the unregularized student achieved 146 test errors
- The regularized version of the student achieved 74 test errors

Conclusion : The soft targets can transfer a great amount of information including the knowledge about how to generalize

Hinton's Experiments on MNIST (2)

- Trained a large neural network as a teacher with two hidden layers of 1200 neurons on 60.000 training sample
- Ommitted the examples of the digit 3 from the transfer set before distilling the knowledge
- Transferred the knowledge to the student model
- The student model still achieved an accuracy of 86.8% on class 3 during the test phase despite the fact that it has never seen a 3

Conclusion : The soft targets can transfer a great amount of information including the knowledge about how to generalize

Our experiments on CIFAR10 : The training

- Fine-tuned a Resnet34 on CIFAR10 (21,289,802 parameters) till 0.849 test accuracy
- Created a smaller classifier for the student role (896,522 parameters)
- Transferred knowledge from the teacher to the student on CIFAR10 using the following parameters :

Optimizer	LR	Epochs	Batch size	T	KDL Weight
Adam	0.001	20	128	5	0.8

$$\mathcal{L}_T(y, y_s, y_t) = \alpha * KL\_div(\sigma(y_s, T), \sigma(y_t, T)) * \frac{T^2}{BS} + \beta * CE(y_s, y)$$

with

$$KL\_div(y', y) = \Sigma_i y_i log(\frac{y_i}{y_i'})$$

Our experiments on CIFAR10 : The Student's Architecture

- 3 convolutional layers
- 2 Linear layers
- 896,522 parameters
- 23 times smaller compared to the Teacher model

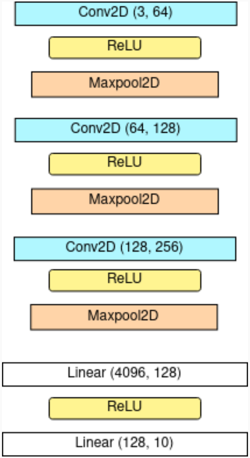


Figure 1: The Student's architecture

Our experiments on CIFAR10 : Convergence

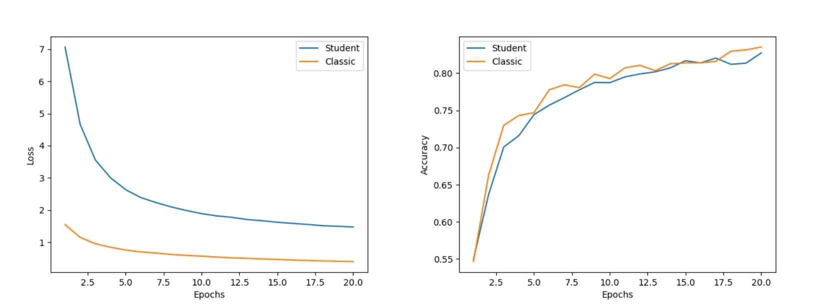


Figure 2: Loss and Test Accuracy : Student vs Classic

- The 2 models were trained under identical conditions
- The Student model achieved a test accuracy of 0.827 (Teacher Test Acc : 0.849)
- The Student model performs as good as the classic model

Our experiments on CIFAR10 : Dark Knowledge Effect

- We tried to reproduce the second experiment on another dataset
- We excluded class 7 (horse class) from CIFAR10 in the transfer dataset
- The student was trained without seeing any sample from class 7
- After train, we added 3.0 to the bias of class 7 in the final layer

Our experiments on CIFAR10 : Dark Knowledge Effect

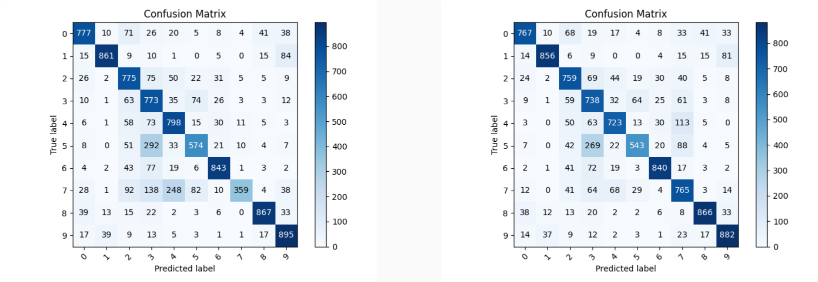


Figure 3: Confusion Matrix : before and after bias increase

Comparison :

	Teacher model	Student model	Classic model
Num Parameters	21,289,802	896,522	896,522
Test Acc	0.849	0.823	0.835
Test Acc[2]	-	0.765	0.0
MFLOPS	149.53	80.20	80.20

- FLOPS were calculated on an input size (1, 3, 32, 32)

[2] Acc : Accuracy on test samples of class 7 (not seen by the model)

Q&A Session