Distilling the Knowledge in a Neural Network (2015)

by Geoffrey Hinton, Oriol Vinyals, Jeff Dean

Presenter: Kevin Ivey

Overview

- Motivation for Distillation
- Method of Distillation
- Experiments

Why Distill Networks?

- Ensemble of networks
  - Better performance than a single network
  - Computationally expensive - infeasible to deploy
- Goal: Take performance and generalization of an ensemble and apply to a smaller network through **distillation**

Knowledge in a Network

- Knowledge ≠ learned parameters
  - Knowledge is a mapping from input vectors to output vectors
- Probabilities of incorrect answers have information
  - Which 2s look like 3s and which 2s look like 7s?

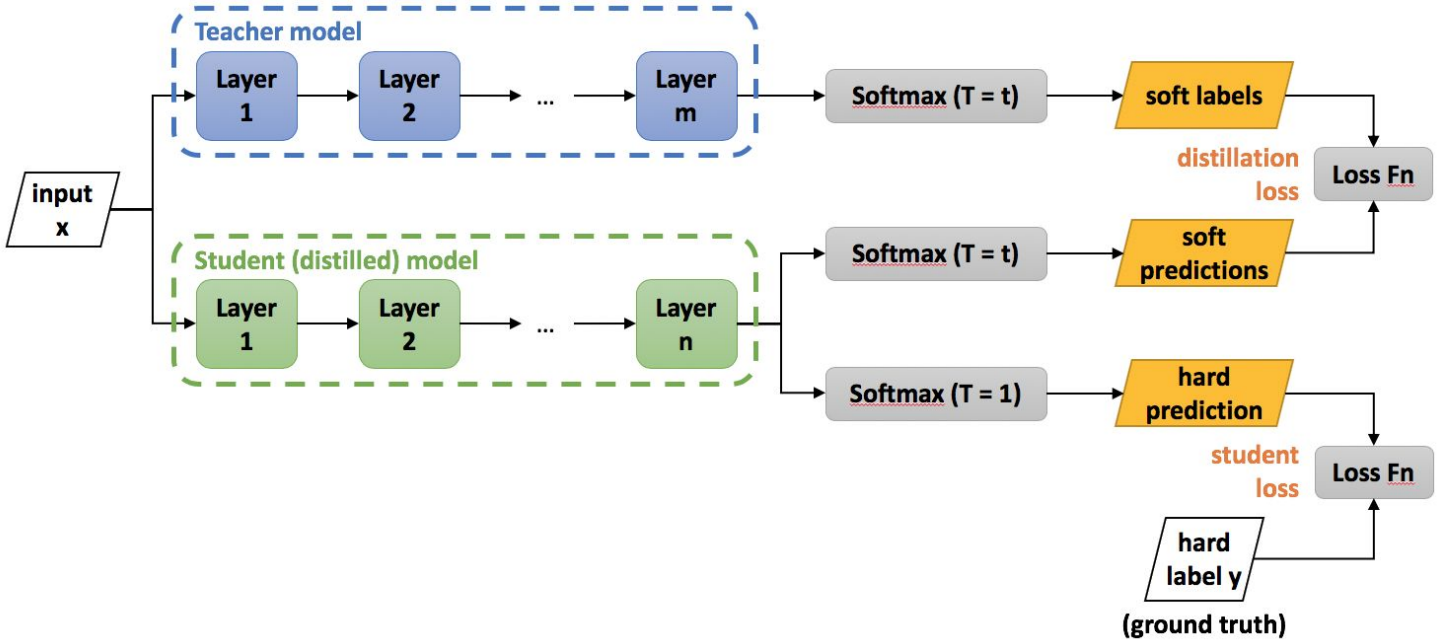
Past Attempt

- Train distilled model using logits (inputs to softmax, z)
  - Special case of new method

Distillation Method

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

- Train distilled model on
  - Transfer set
  - Weighted average of objective functions:
    - Cross entropy of the soft targets (softmax with high temperature, *T*)
    - Cross entropy of the correct label (if known)


Teacher model
Layer 1
Layer 2
...
Layer m
input x
Softmax (T = t)
soft labels
distillation loss
Loss Fn
Student (distilled) model
Layer 1
Layer 2
...
Layer n
Softmax (T = t)
soft predictions
Softmax (T = 1)
hard prediction
student loss
Loss Fn
hard label y
(ground truth)

MNIST Results

Network Architecture	Test Error
Ensemble (2 hidden layers, 1200 ReLU units, dropout)	67
Smaller Network (2 hidden layers, 800 ReLU units, no regularization)	146
Distilled Network (2 hidden layers, 800 ReLU units, regularized by ensemble soft targets (T=20))	74

MNIST Results

- Removing all 3s from the transfer set
  - Distilled model makes 206 test errors, 133 are 3s
  - Increasing the bias by 3.5 drops to 109 errors, 14 are 3s
- **Only** keeping 7s and 8s in the transfer set
  - Distilled model makes 47.3% test errors
  - Increasing the bias by 7.6 for 7s and 8s drops to 13.2% test error

Speech Recognition Task

- Predict the Hidden Markov Model of the 21st frame using features from the wavelength
- Baseline Model: 8 hidden layers, 2,560 ReLU units, softmax with 14,000 labels
    - Old version of the Android voice search

Speech Recognition Results

System	Test Frame Accuracy	Word Error Rate
Baseline	58.9%	10.9%
10xEnsemble	61.1%	10.7%
Distilled Single Model	60.8%	10.7%

JFT Dataset

- 100 million images with 15,000 labels
- Baseline model - deep convolutional network
  - Trained for **6 months** using 2 types of parallelism
  - Ensemble training would prove additional parallelism but needs more cores
    - “Waiting for several years to train an ensemble of models was not an option”
- How to train an ensemble model?

Using Specialists

- Make the ensemble contain a “generalist” model and “specialist” models
- “Specialist” models
  - Trained on confusing classes with a “dustbin” class
  - Initialized with the weights of the “generalist” model

JFT 1:	Tea party; Easter; Bridal shower; Baby shower; Easter Bunny; ...
JFT 2:	Bridge; Cable-stayed bridge; Suspension bridge; Viaduct; Chimney; ...
JFT 3:	Toyota Corolla E100; Opel Signum; Opel Astra; Mazda Familia; ...

Table 2: Example classes from clusters computed by our covariance matrix clustering algorithm

Ensemble Performance

- Find top *n* probable classes using generalist models
- For all specialist models whose subset of classes is in the top *n* probable classes, minimize

$$KL(\mathbf{p}^g, \mathbf{q}) + \sum_{m \in A_k} KL(\mathbf{p}^m, \mathbf{q})$$

- Reduces to the arithmetic or geometric mean when all models produce a single probability for each class

JFT Results

System	Conditional Test Accuracy	Test Accuracy
Baseline	43.1%	25.0%
+ 61 Specialist models	45.9%	26.1%

Table 3: Classification accuracy (top 1) on the JFT development set.

# of specialists covering	# of test examples	delta in top1 correct	relative accuracy change
0	350037	0	0.0%
1	141993	+1421	+3.4%
2	67161	+1572	+7.4%
3	38801	+1124	+8.8%
4	26298	+835	+10.5%
5	16474	+561	+11.1%
6	10682	+362	+11.3%
7	7376	+232	+12.8%
8	4703	+182	+13.6%
9	4706	+208	+16.6%
10 or more	9082	+324	+14.1%

Table 4: Top 1 accuracy improvement by # of specialist models covering correct class on the JFT test set.

Why Soft Targets?

- Soft targets are theorized to have information not encoded in a single hard target (the label)
- Training the previous speech recognition models with soft targets retains information, uses less data, and is less prone to overfitting
- Could potentially be used to decrease overfitting in specialists

System & training set	Train Frame Accuracy	Test Frame Accuracy
Baseline (100% of training set)	63.4%	58.9%
Baseline (3% of training set)	67.3%	44.5%
Soft Targets (3% of training set)	65.4%	57.0%

Table 5: Soft targets allow a new model to generalize well from only 3% of the training set. The soft targets are obtained by training on the full training set.

Similarity to Mixture of Experts

- Mixture of experts uses a gating network which assigns examples to specific experts
  - Both the expert and the assignment of the example are learned
- Mixture of experts is better than the clustering algorithm to find confused classes, but at the cost of less parallelization
  - Limits use on large datasets with clear subsets

Conclusion

- Distillation using soft targets allows for
  - Increased generalizability
  - Decreased size for deployment in production
  - Increased performance in large datasets by means of a specialist