Distilling the Knowledge in a Neural Network

Geoffrey Hinton, Oriol Vinyals and Jeff Dean

Guilherme NUNES TROFINO

March 19, 2025

ENSTA Paris

1. Introduction

Introduction

In ”Distilling the Knowledge in a Neural Network” Hilton et al. introduces a technique called **Knowledge Distillation**.

Introduction

In "Distilling the Knowledge in a Neural Network" Hilton et al. introduces a technique called **Knowledge Distillation**.

Knowledge Distillation

Aims to **transfer** knowledge from a large complex model, or an ensemble of models, into a more efficient smaller model that is easier to deploy.

Introduction

In "Distilling the Knowledge in a Neural Network" Hilton et al. introduces a technique called **Knowledge Distillation**.

Knowledge Distillation

Aims to **transfer** knowledge from a large complex model, or an ensemble of models, into a more efficient smaller model that is easier to deploy.

The key idea is to use probabilistic outputs produced by the large model to train the smaller model rather than relying solely on the ground truth from the training data.

1. Introduction

1.1. Knowledge Distillation

Main features of the proposed approach are:

Main features of the proposed approach are:

Soft Targets

Probabilistic outputs produced by a large model, containing more information than hard labels as they capture the relative probabilities of incorrect classes.

Main features of the proposed approach are:

Soft Targets

Probabilistic outputs produced by a large model, containing more information than hard labels as they capture the relative probabilities of incorrect classes.

Temperature Scaling

Raise `softmax` temperature in the large model to produce soften probabilities, which are then used to train the smaller model.

Main features of the proposed approach are:

Soft Targets

Probabilistic outputs produced by a large model, containing more information than hard labels as they capture the relative probabilities of incorrect classes.

Temperature Scaling

Raise `softmax` temperature in the large model to produce soften probabilities, which are then used to train the smaller model.

This way a small model is trained to match the soft targets:

$$\boxed{q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}} \qquad \text{with } T \text{ being temperature} \tag{1}$$

2. Experiments and Results

2. Experiments and Results

2.1. MNIST

Experiments and Results, MNIST

The effectiveness of **Knowledge Distillation** was demonstrated on the MNIST dataset using a single Neural Network (NN) model with 2 hidden layers of 1200 rectified linear hidden units on 60.000 training cases as the large model.

The effectiveness of **Knowledge Distillation** was demonstrated on the MNIST dataset using a single Neural Network (NN) model with 2 hidden layers of 1200 rectified linear hidden units on 60.000 training cases as the large model.

Results

model	errors*
baseline	67

Experiments and Results, MNIST

The effectiveness of **Knowledge Distillation** was demonstrated on the MNIST dataset using a single Neural Network (NN) model with 2 hidden layers of 1200 rectified linear hidden units on 60.000 training cases as the large model.

Results

model	errors*
baseline	67
distilled (without regularization)	146

Experiments and Results, MNIST

The effectiveness of **Knowledge Distillation** was demonstrated on the MNIST dataset using a single Neural Network (NN) model with 2 hidden layers of 1200 rectified linear hidden units on 60.000 training cases as the large model.

Results

model	errors*
baseline	67
distilled (without regularization)	146
distilled (with regularization)	**74**

Experiments and Results, MNIST

The effectiveness of **Knowledge Distillation** was demonstrated on the MNIST dataset using a single Neural Network (NN) model with 2 hidden layers of 1200 rectified linear hidden units on 60.000 training cases as the large model.

Results

model	errors*
baseline	67
distilled (without regularization)	146
distilled (with regularization)	**74**

Insights

The small model could even classify **unseen** digits based on the large's generalization.

2. Experiments and Results

2.2. Speech Recognition

Experiments and Results, Speech Recognition

The distillation strategy was also evaluated with an ensemble of 10 Deep Neural Network (DNN) acoustic models used in Automatic Speech Recognition (ASR) trained on about 2000 hours of spoken English data.

Experiments and Results, Speech Recognition

The distillation strategy was also evaluated with an ensemble of 10 Deep Neural Network (DNN) acoustic models used in Automatic Speech Recognition (ASR) trained on about 2000 hours of spoken English data.

Reults

model	Frame Accuracy	Word Error Rate
baseline	58.9%	10.9%

Experiments and Results, Speech Recognition

The distillation strategy was also evaluated with an ensemble of 10 Deep Neural Network (DNN) acoustic models used in Automatic Speech Recognition (ASR) trained on about 2000 hours of spoken English data.

Reults

model	Frame Accuracy	Word Error Rate
baseline	58.9%	10.9%
ensemble of DNN	61.1%	10.7%

The distillation strategy was also evaluated with an ensemble of 10 Deep Neural Network (DNN) acoustic models used in Automatic Speech Recognition (ASR) trained on about 2000 hours of spoken English data.

Reults

model	Frame Accuracy	Word Error Rate
baseline	58.9%	10.9%
ensemble of DNN	61.1%	10.7%
distilled	**60.8%**	**10.7%**

Experiments and Results, Speech Recognition

The distillation strategy was also evaluated with an ensemble of 10 Deep Neural Network (DNN) acoustic models used in Automatic Speech Recognition (ASR) trained on about 2000 hours of spoken English data.

Reults

model	Frame Accuracy	Word Error Rate
baseline	58.9%	10.9%
ensemble of DNN	61.1%	10.7%
distilled	**60.8%**	**10.7%**

Insights

Distilled model captured **80%** of the ensemble's improvement, being easier to deploy.

3. Extensions

3. Extensions

3.1. Specialist Models

Large datasets (e.g., JFT with 100M images, 15,000 classes) make full ensemble training computationally infeasible. Instead, train specialist models focused on **confusable subsets** of classes, initialized from a generalist model.

Extensions, Specialist Models

Large datasets (e.g., JFT with 100M images, 15,000 classes) make full ensemble training computationally infeasible. Instead, train specialist models focused on **confusable subsets** of classes, initialized from a generalist model.

Results

Training 61 specialist models on clusters of 300 classes each resulted in a **4.4% improvement** in test accuracy. These specialists were trained independently and efficiently in parallel, with larger accuracy gains observed.

Large datasets (e.g., JFT with 100M images, 15,000 classes) make full ensemble training computationally infeasible. Instead, train specialist models focused on **confusable subsets** of classes, initialized from a generalist model.

Results

Training 61 specialist models on clusters of 300 classes each resulted in a **4.4% improvement** in test accuracy. These specialists were trained independently and efficiently in parallel, with larger accuracy gains observed.

Insights

Specialist models, while effective, are prone to overfitting as they are trained on biased subsets of classes. Incorporating soft targets from the generalist model **mitigates this risk** by acting as regularizers, ensuring the specialists retain generalization capabilities.

3. Extensions

3.2. Soft Targets

Extensions, Soft Targets

Training deep models on **limited data** often leads to severe overfitting. Instead of using only hard labels, **soft targets** from a pre-trained model help retain generalization.

Extensions, Soft Targets

Training deep models on **limited data** often leads to severe overfitting. Instead of using only hard labels, **soft targets** from a pre-trained model help retain generalization.

Results

Training an acoustic model with only 3% of the speech dataset demonstrated the effectiveness of soft targets in mitigating overfitting. While hard targets resulted in 44.5% test accuracy due to severe overfitting, soft targets achieved **57.0%** test accuracy.

Extensions, Soft Targets

Training deep models on **limited data** often leads to severe overfitting. Instead of using only hard labels, **soft targets** from a pre-trained model help retain generalization.

Results

Training an acoustic model with only 3% of the speech dataset demonstrated the effectiveness of soft targets in mitigating overfitting. While hard targets resulted in 44.5% test accuracy due to severe overfitting, soft targets achieved **57.0%** test accuracy.

Insights

Soft targets encode valuable class relationships, enhancing generalization and acting as natural regularizers. They enable models trained on limited data to mimic the behavior of models trained on full datasets, effectively **mitigating overfitting**.

4. Conclusion

4. Conclusion

4.1. Takeaways

Conclusion, Takeaways

Model Compression

Distillation enables **compressing large models or ensembles** into smaller, more efficient models, reducing significantly deployment costs while retaining performance.

Conclusion, Takeaways

Model Compression

Distillation enables **compressing large models or ensembles** into smaller, more efficient models, reducing significantly deployment costs while retaining performance.

Soft Targets

Soft targets **capture rich class relationships** beyond hard labels. They allow small models to mimic the generalization ability of larger models, improving test accuracy.

Conclusion, Takeaways

Model Compression

Distillation enables **compressing large models or ensembles** into smaller, more efficient models, reducing significantly deployment costs while retaining performance.

Soft Targets

Soft targets **capture rich class relationships** beyond hard labels. They allow small models to mimic the generalization ability of larger models, improving test accuracy.

Regularization

Training **specialist models** on confusable classes improves accuracy in large-scale datasets. Soft targets also act as **natural regularizers**, preventing overfitting.

4. Conclusion

4.2. Q&A

Conclusion, Q&A

Merci!

Distilling the Knowledge in a Neural Network

Geoffrey Hinton, Oriol Vinyals and Jeff Dean

Guilherme NUNES TROFINO

March 19, 2025

ENSTA Paris