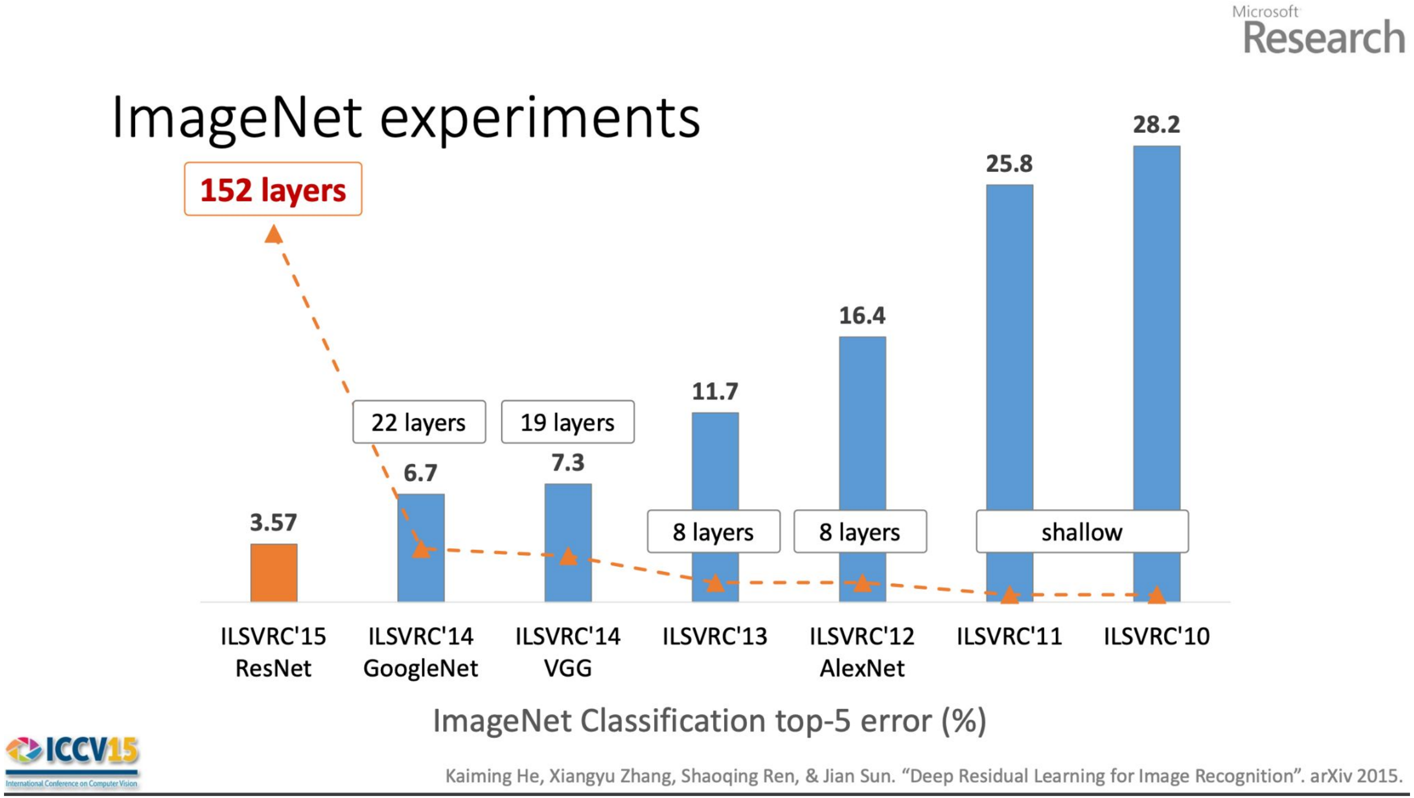
Distilling the Knowledge in a Neural Network
(NIPS 2014 Deep Learning Workshop)

황중원

Historical background



http://kaiminghe.com/ilsvrc15/ilsvrc2015_deep_residual_learning_kaiminghe.pdf

Knowledge Distillation (KD)?

KD

- In order to improve the performance of knowledge distillation,
  - teacher-student network architecture
  - what kind of knowledge is learned from the teacher network
  - where is distilled into the student network.

http://dmqm.korea.ac.kr/activity/seminar/304

The schematic structure of knowledge distillation

Knowledge

- Response-Based
- Feature-Based
- Relation-Based

Distillation

- Offline
- Online
- Self

The generic teacher-student framework for knowledge distillation

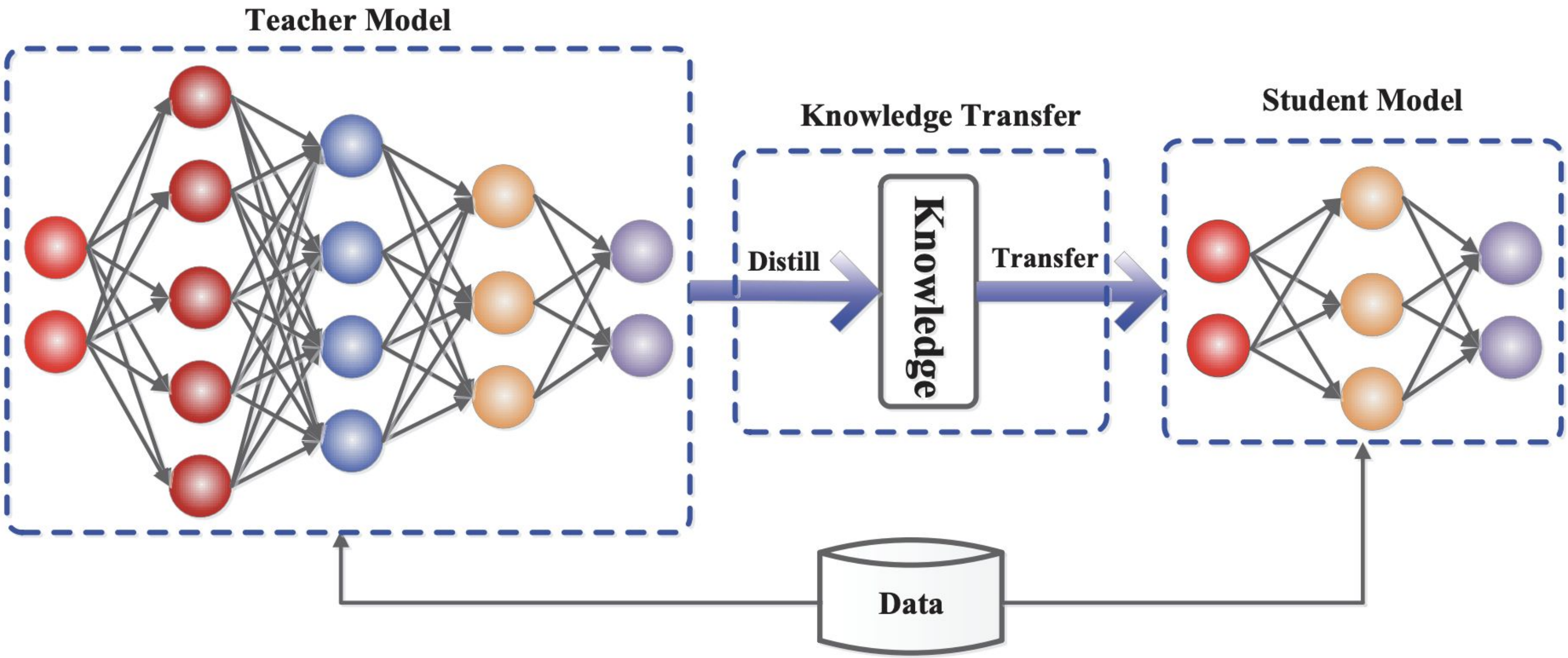


Fig. 1 The generic teacher-student framework for knowledge distillation.

https://arxiv.org/pdf/2006.05525.pdf

Architecture

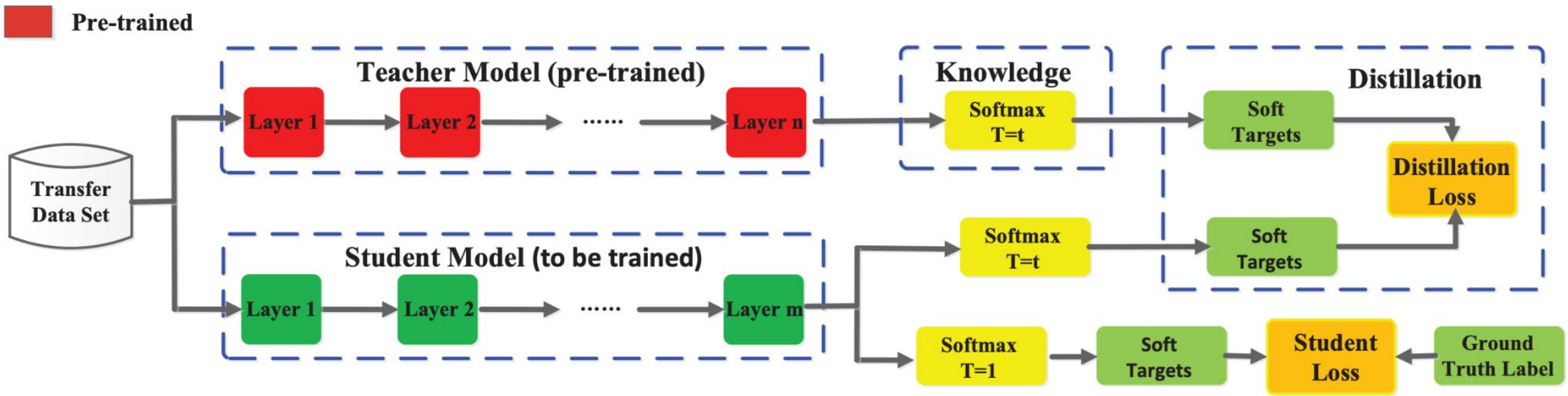


Fig. 5 The specific architecture of the benchmark knowledge distillation (Hinton et al., 2015).

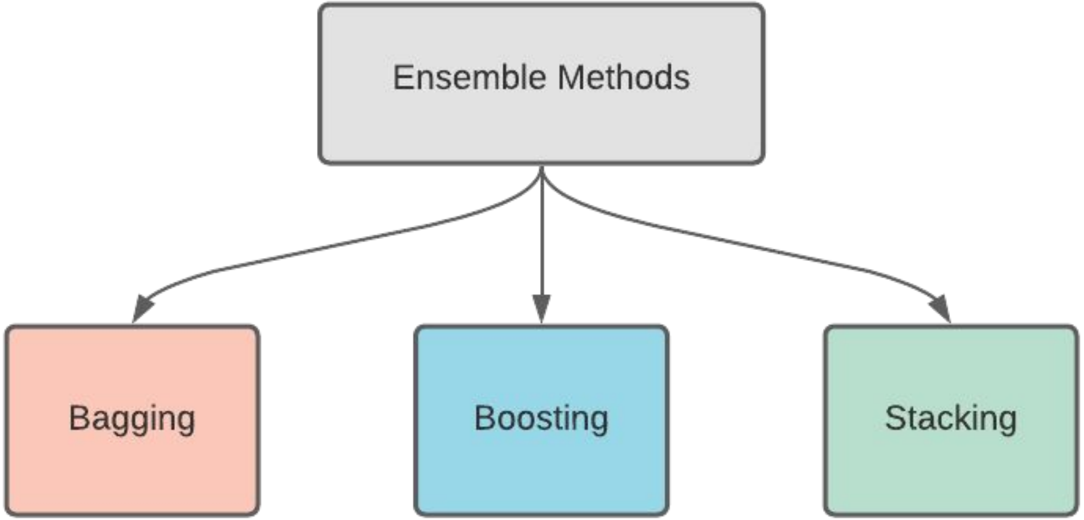
$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \quad (1)$$

https://arxiv.org/pdf/2006.05525.pdf

Abstract

Ensemble method is simple and powerful, but cost expensive

- ensemble model is cumbersome
- Computationally expensive (especially if the individual models are large neural nets)



https://medium.com/luisfredgs/multiple-classifier-systems-a-brief-introduction-71238d9c794f

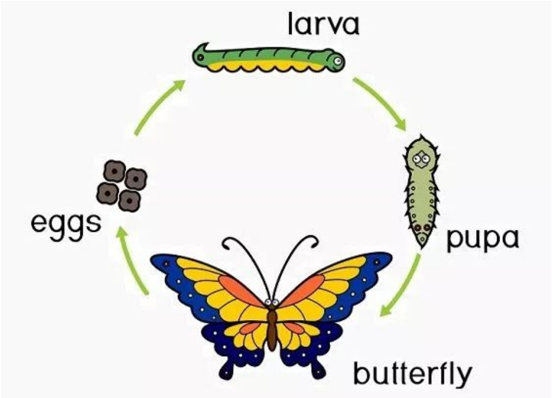
Contribution

- distilling the knowledge in an ensemble of medels into a single model
-

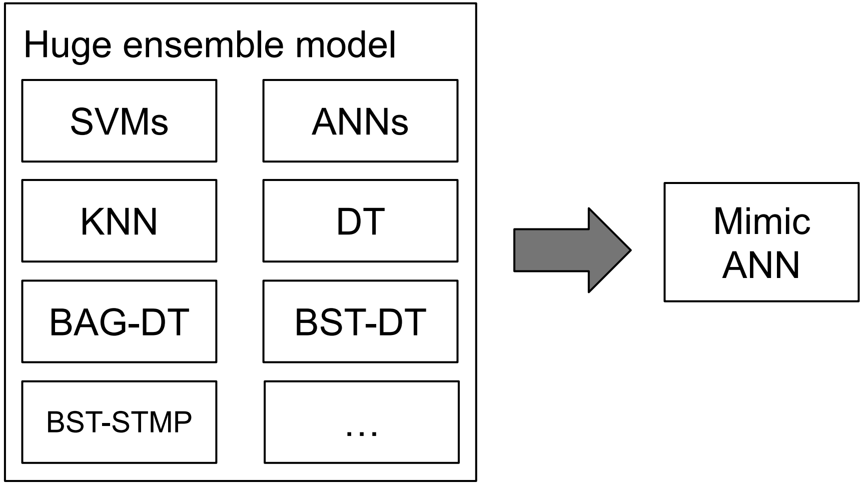
Introduction

Distillation

- to transfer the knowledge from the cumbersome model to a small model that is more suitable for deployment



binary classification problems



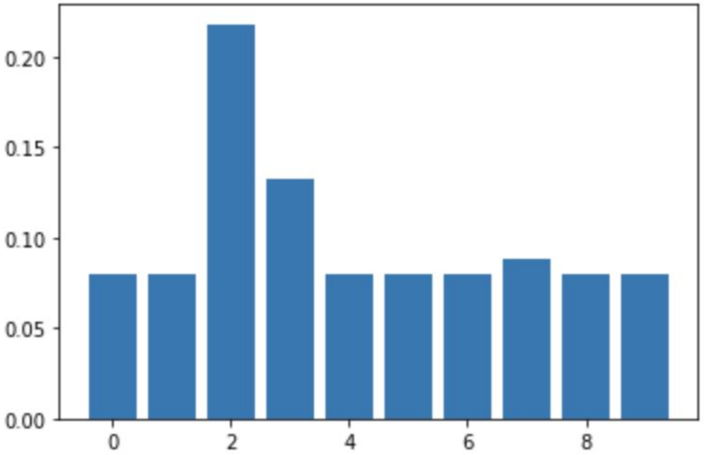
Results on eight test problems show that, on average, the loss in performance due to compression is usually negligible, yet the mimic neural nets are 1000 times **smaller** and 1000 times **faster**.

https://biologydictionary.net/complete-metamorphosis/
https://www.cs.cornell.edu/~caruana/compression.kdd06.pdf

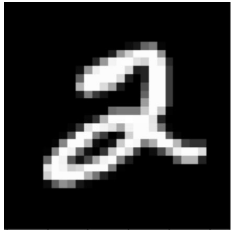
Distillation

- to raise the temperature of the final softmax until the cumbersome model produces a suitabley soft set of targets

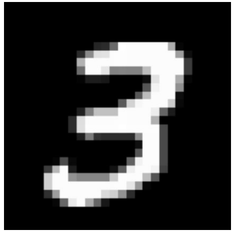
After Softmax



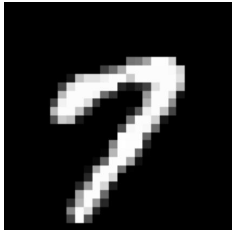
[0.0802, 0.0802, 0.2179, 0.1322, 0.0802, 0.0802, 0.0802, 0.0886, 0.0802, 0.0802]



10^{-1}



50^{-2}



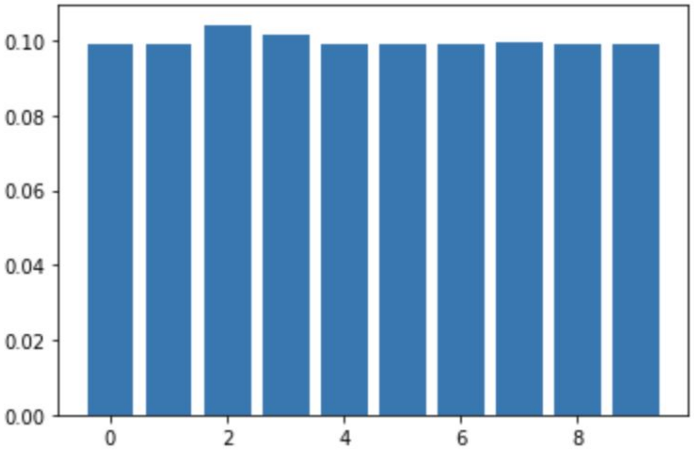
10^{-2}

logits

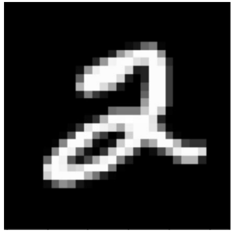
Distillation

- use the same high temperature when training the small model to match these soft targets

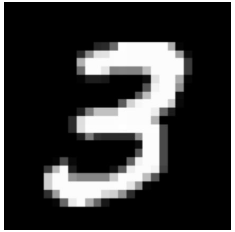
After Softmax with Temperature
T=20



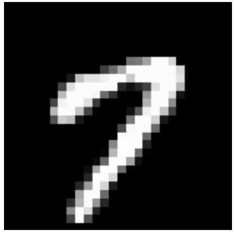
[0.0992, 0.0992, 0.1043, 0.1017, 0.0992, 0.0992, 0.0992, 0.0997, 0.0992, 0.0992]



10^{-1}



50^{-2}



10^{-2}

logits

2 Distillation

Softmax with Temperature scaling

Neural networks typically produce class probabilities by using a "softmax" output layer that converts the logit, z_i, computed for each class into a probability, q_i, by comparing z_i with the other logits.

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \tag{1}$$

where T is a temperature that is normally set to 1. Using a higher value for T produces a softer probability distribution over classes.

3 Preliminary experiments on MNIST

Experiment setup

- Teacher
  - Two hidden layers of 1200 hidden units
  - Relu, dropout
  - 60,000 training cases
  - input images are jittered
  - 67 test errors
- Student
  - Two hidden layers of 800 hidden units
  - Relu
  - 146 test errors
- Distill
  - T = 20
  - 74 test errors
- Etc
  - When 300 hidden units & T>8, results are almost same
  - When 30 hidden units & 4 >= T >= 2.5, performance was dropped

4 Experiments on speech recognition

Results

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.

- Base
  - 8 hiden layers, 2560 hidden units, relu
  - final softmax with 14,000 labels (HMM targets h_t)
- 10xEnsemble
  - = 10 X Base
- Distill
  - T=1,**2**,5,10

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}'} P(h_t|\mathbf{s}_t;\boldsymbol{\theta}')$$

5 Training ensemble of specialists on very big datasets

5 Training ensemble of specialists on very big datasets

- JFT is an internal Google dataset that has 100 million labeled images with 15,000 labels
  - In Google, training CNN model during six months
- Ensemble training is not feasible, so split subsets and train specialist models
  - training 61 specialist models during a few days with 61x300(18,300) classes

6 Soft Targets as Regularizers

Soft targets as regularizers

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.

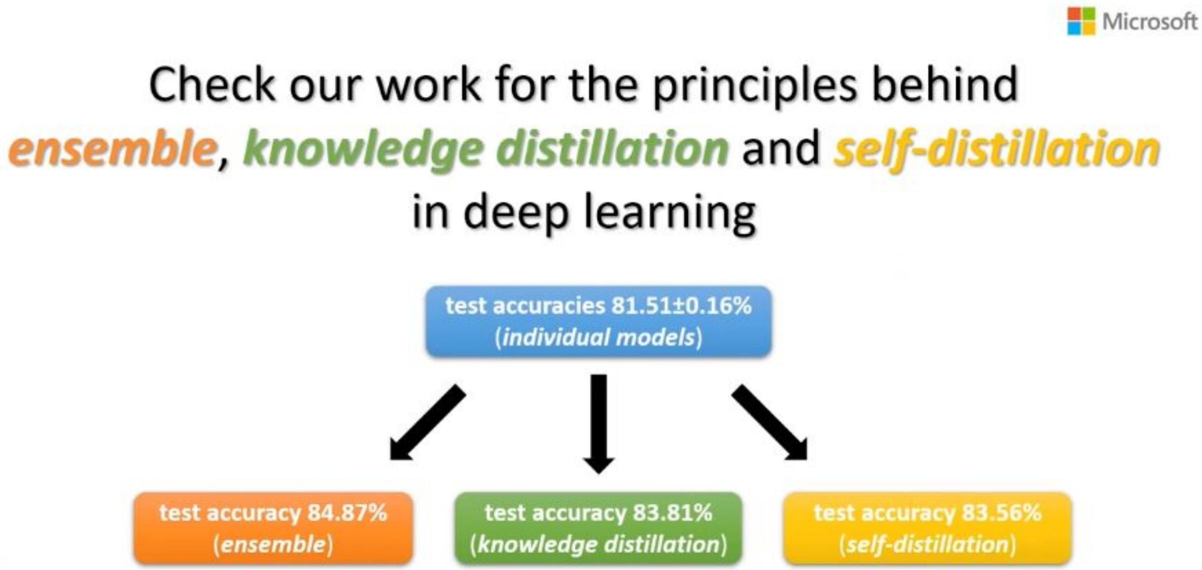
Discussion

Discussion

- We have shown that distilling works very well for transferring knowledge from an ensemble or from a large highly regularized model into a smaller, distilled model.
  - On MNIST, acoustic datasets, JFT
- In order to improve the performance of knowledge distillation,
  - teacher-student network architecture
  - what kind of knowledge is learned from the teacher network
  - where is distilled into the student network.

Appendix

왜 KD를 해야하는가? KD를 하면 뭐가 좋지?



WideResNet-28-10 architecture on the CIFAR-100 dataset 10 times with different random seeds, the mean test accuracy is 81.51% while the standard deviation is only 0.16%.

https://www.microsoft.com/en-us/research/blog/three-mysteries-in-deep-learning-ensemble-knowledge-distillation-and-self-distillation

Model compression (2004, ICML)

Simple Ver

SVMs

ANNs

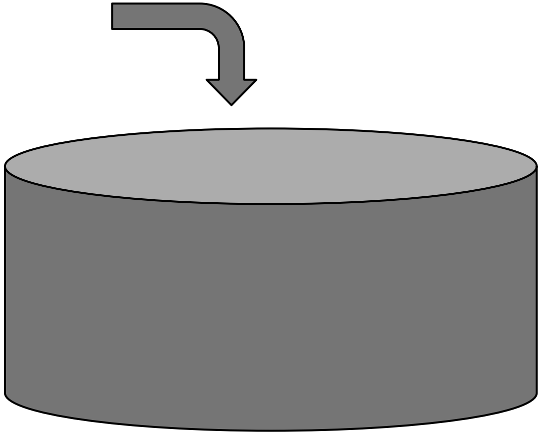
KNN

DT

BAG-DT

BST-DT

BST-STMP

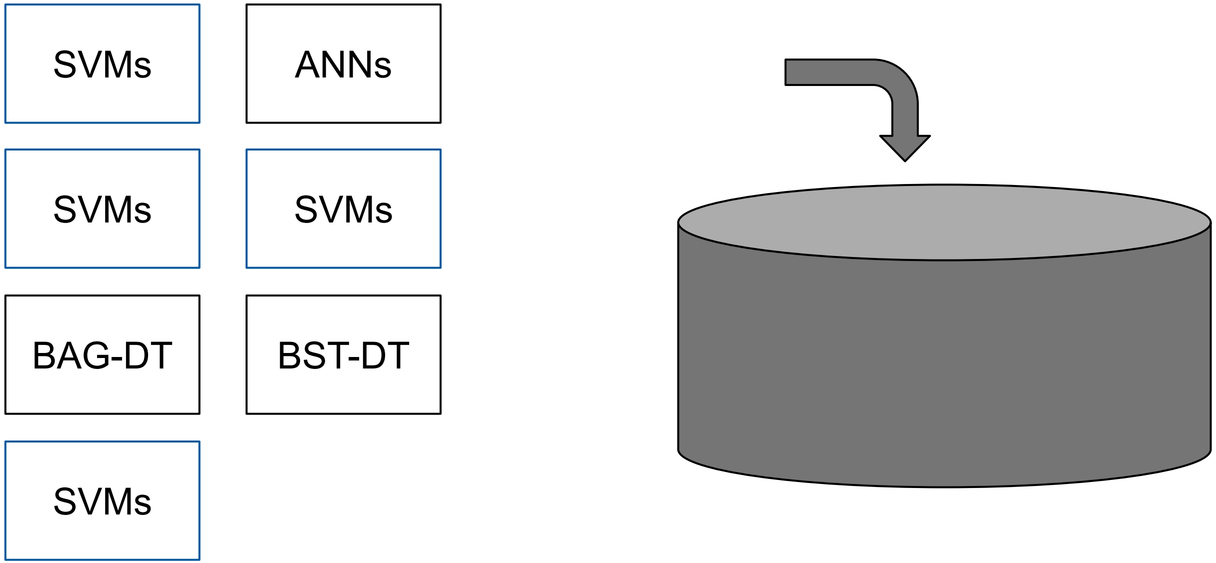


Until maximizes the ensemble's perfomance on a valid dataset

https://www.cs.cornell.edu/~alexn/papers/shotgun.icml04.revised.rev2.pdf
https://www.cs.cornell.edu/~caruana/compression.kdd06.pdf

Model compression (2004, ICML)

Selection with Replacement



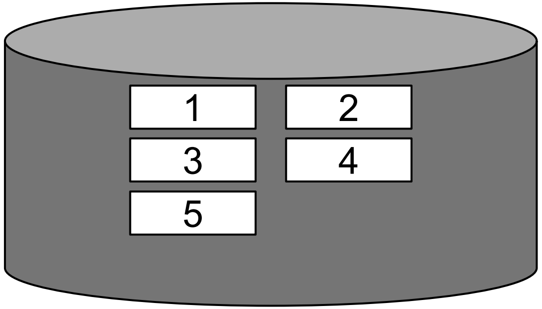
Until maximizes the ensemble's perfomance on a valid dataset

https://www.cs.cornell.edu/~alexn/papers/shotgun.icml04.revised.rev2.pdf
https://www.cs.cornell.edu/~caruana/compression.kdd06.pdf

Model compression (2004, ICML)

Sorted Ensembel Initialization

N = from 5



Until maximizes the ensemble’s perfomance on a valid dataset

https://www.cs.cornell.edu/~alexn/papers/shotgun.icml04.revised.rev2.pdf
https://www.cs.cornell.edu/~caruana/compression.kdd06.pdf