week 3

Knowledge Distillation

Distilling the Knowledge in a Neural Network (NIPS 2014)

Geoffrey Hinton, Oriol Vinyals, Jeff Dean

2022.09.29

곽민지

Introduction

(1) Previous method

Making predictions using a whole ensemble of models.

1. **Train many different models on the same data.**
2. **Average their predictions.**

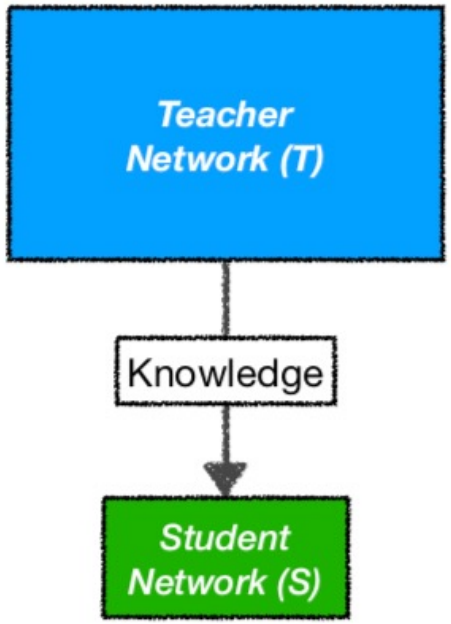
- Cumbersome
- Too computationally expensive to allow deployment to a large number of users.
  (Especially If the individual models are large neural nets.)

Introduction

(2) Proposed method

It is possible to compress the knowledge in an ensemble into a single model.

→ Much easier to deploy.



Method

(1) Softer softmax

- Neural networks typically produce class probabilities by using a “softmax” output layer.
- Using a higher value for T produces a softer probability distribution over classes.
- Temperature 를 사용한 경우가 낮은 입력값의 출력을 더 크게 만들어주고 큰 입력값의 출력은 작게 만들어준다.

$$L = \sum_{(x,y)\in\mathbb{D}} L_{KD}(S(x,\theta_S,\tau), T(x,\theta_T,\tau)) + \lambda L_{CE}(\hat{y}_S, y)$$

cow	dog	cat	car	
0	1	0	0	original hard targets

cow	dog	cat	car	
.05	.3	.2	.005	softened output of ensemble

Method

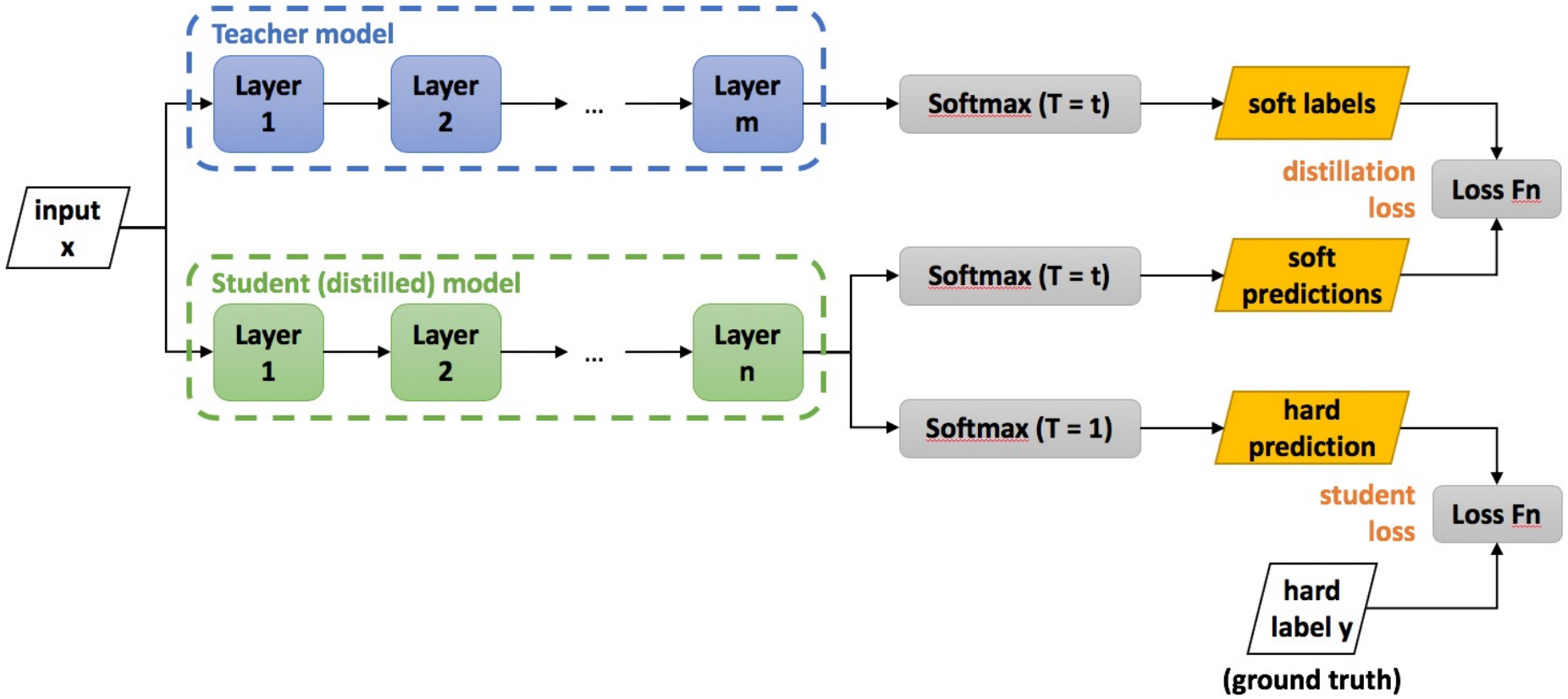
(2) Loss function

$$L = \sum_{(x,y)\in\mathbb{D}} L_{KD}(S(x,\theta_S,\tau),T(x,\theta_T,\tau))+\lambda L_{CE}(\hat{y}_S,y)$$

```
KD_loss = nn.KLDivLoss()(F.log_softmax(outputs/T, dim=1),
                         F.softmax(teacher_outputs/T, dim=1)) * (alpha * T * T) + \
          F.cross_entropy(outputs, labels) * (1. - alpha)
```

Method

(3) Framework



Experiments

(1) MNIST

Train: 60,000 / Test: 10,000

netA: 784 → 800 → 800 → 10 (146 test error)

netB: 784 → 1200 → 1200 → 10 (more parameter than netA, dropout, jittering) (67 test error)

netC: 784 → 800 → 800 → 10 (Distillation)

net B로 original training data에 soft label을 생성 → 원래의 hard label과 soft label을 둘 다 활용

기존 net A 보다 훨씬 성능이 좋은 74 test error

Experiments

(1) MNIST

MNIST without “3”

Conclusion

- Distilling works very well for transferring knowledge from an ensemble or from a large highly regularized model into a smaller, distilled model.

- On MNIST distillation works remarkably well even when the transfer set that is used to train the distilled model lacks any examples of one or more of the classes.

Reference

- https://arxiv.org/abs/1503.02531
- https://intellabs.github.io/distiller/knowledge_distillation.html
- https://baeseongsu.github.io/posts/knowledge-distillation/
- https://blog.lunit.io/2018/03/22/distilling-the-knowledge-in-a-neural-network-nips-2014-workshop/