Knowledge Distillation

Q&A Session

11/24

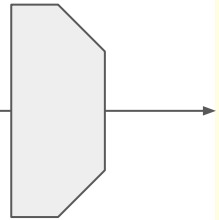
TA : 김성년, 최진환

Covering papers as follows:

- Distilling the Knowledge in a Neural Network, 15 `NIPS
- Knowledge Distillation by On-the-Fly Native Ensemble, 18`NIPS

1. Distilling the Knowledge in a Neural Network, 15 `NIPS

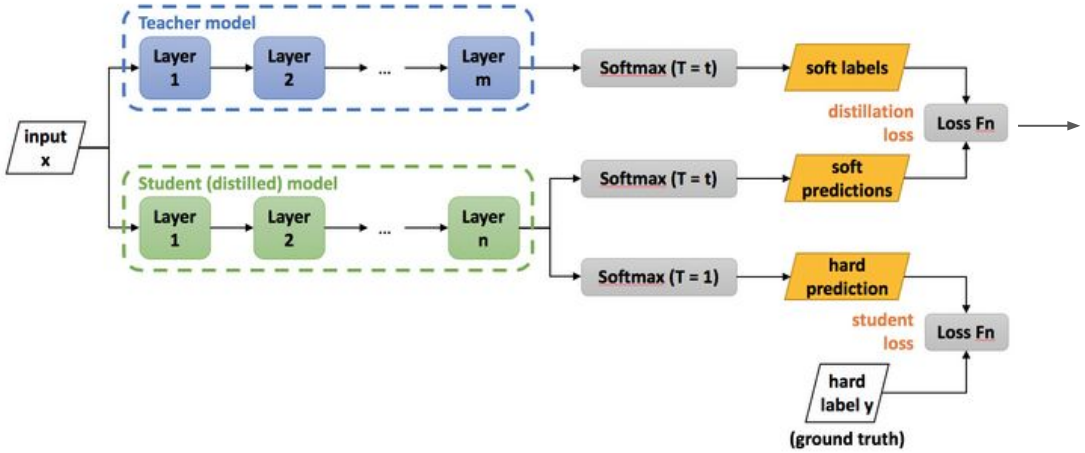
- Summary: 크고 복잡한 네트워크에서 작고 간단한 네트워크로의 지식 전달(knowledge transfer)을 하고자 함. 지식을 전달하는 방법으로, 복잡한 네트워크(teacher model)의 아웃풋인 클래스별 확률분포를 “soft target”으로 써서 간단한 네트워크를 학습하는 데에 이용.
- What is “soft target”?



cow	dog	cat	car	
0	1	0	0	original hard targets
cow	dog	cat	car	
10^{-6}	.9	.1	10^{-9}	output of geometric ensemble
cow	dog	cat	car	
.05	.3	.2	.005	softened output of ensemble

1. Distilling the Knowledge in a Neural Network, 15 `NIPS

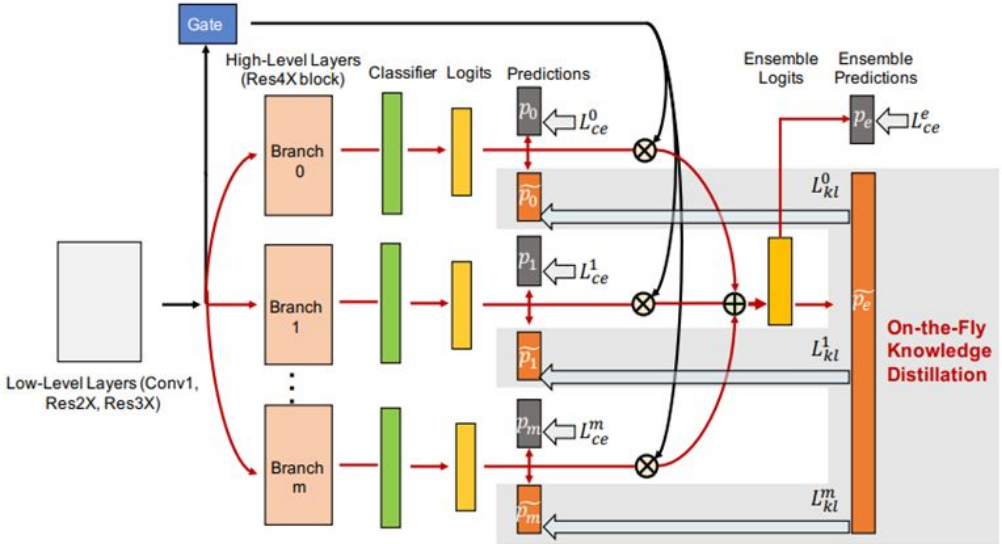
- Teacher model의 예측값을 student model이 따라가도록 하는 것!
- hard target보다 soft target이 데이터에 대한 정보를 더 많이 담고 있다는 가정. 따라서 작은 모델로도 더 좋은 generalization 성능을 얻을 수 있고, 데이터셋 일부만을 사용해도 성능이 크게 떨어지지 않는다.



$$\sum_{x_i \in \mathcal{X}} \mathrm{KL}\left(\mathrm{softmax}\left(\frac{f_T(x_i)}{\tau}\right), \mathrm{softmax}\left(\frac{f_S(x_i)}{\tau}\right)\right)$$

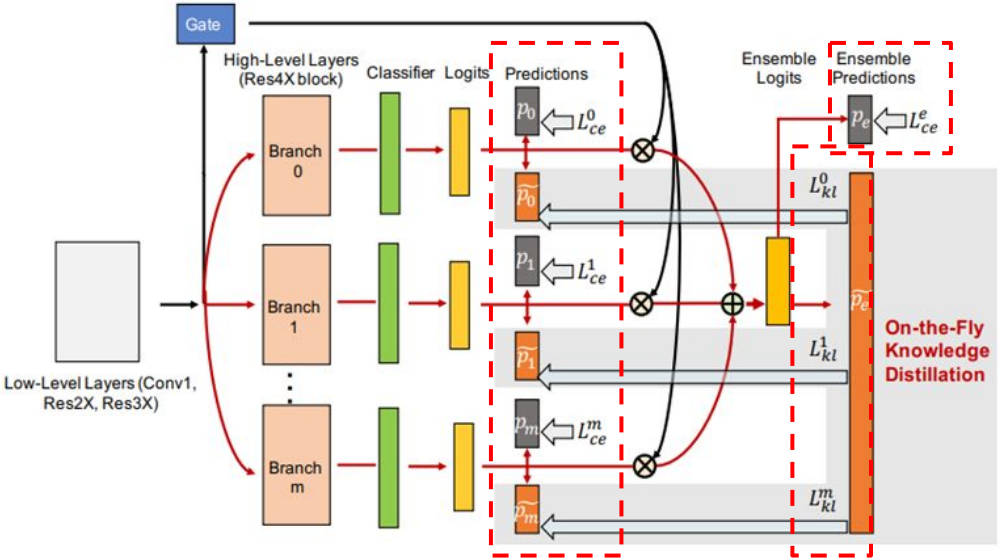
2. Knowledge Distillation by On-the-Fly Native Ensemble, 18`NIPS

- Summary: 이 논문은 일반적으로 KD 분야에서 사용되던 two-phase training에서 벗어나 one-phase distillation training을 제안하여, 기존의 모델들이 가지고 있던 1) computational cost and memory usage 2) complex training procedure 3) Longer training process 의 문제들을 해결합니다.



2. Knowledge Distillation by On-the-Fly Native Ensemble, 18`NIPS

1. Lower level layers를 share한 상태로, Highest level layer를 duplicate하여 여러 개의 branch로 만들어 준다.
2. 각 branch에서의 logit 들의 weighted sum으로 teacher logit을 approximation 한다.
3. Total Loss = 각 branch의 prediction cross-entropy + teacher logit의 prediction cross-entropy + KL divergence (teacher-student)



$$\mathcal{L} = \sum_{i=0}^{m} \mathcal{L}_{\text{ce}}^{i} + \mathcal{L}_{\text{ce}}^{e} + T^2 * \mathcal{L}_{\text{kl}}$$

$$\mathcal{L}_{\text{ce}} = -\sum_{c=1}^{C} \delta_{c,y} \log\Big(p(c|\boldsymbol{x}, \boldsymbol{\theta})\Big)$$

$$\tilde{p}_i(c|\boldsymbol{x}, \boldsymbol{\theta}^i) = \frac{\exp(\boldsymbol{z}_i^c / T)}{\sum_{j=1}^{C} \exp(\boldsymbol{z}_i^j / T)}, c \in \mathcal{Y}$$

$$\mathcal{L}_{\text{kl}} = \sum_{i=0}^{m} \sum_{j=1}^{C} \tilde{p}_e(j|\boldsymbol{x}, \boldsymbol{\theta}^e) \log \frac{\tilde{p}_e(j|\boldsymbol{x}, \boldsymbol{\theta}^e)}{\tilde{p}_i(j|\boldsymbol{x}, \boldsymbol{\theta}^i)}.$$