Deep Neural Network With Massive Learned Knowledge

By Abhilasha

Teacher :

$$q^*(\boldsymbol{Y}) \propto p_\theta(\boldsymbol{Y}|\boldsymbol{X}) \exp\left\{C \sum_l \lambda_l f_l(\boldsymbol{X}, \boldsymbol{Y})\right\},$$

Student :

$$\boldsymbol{\theta}^{(t+1)} = \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} (1-\pi)\ell(\boldsymbol{y}_n, \boldsymbol{\sigma}_\theta(\boldsymbol{x}_n)) + \pi\ell(\boldsymbol{s}_n^{(t)}, \boldsymbol{\sigma}_\theta(\boldsymbol{x}_n)),$$

1. f_l(X,Y)) = r_l(X,Y) - 1, where r_l(X,Y) in [0,1]
2. r_l(X,Y) is fully specified a priori and fixed throughout training.
3. Lamda_l has to be manually set.

1. To substantially extend the scope of knowledge used in the framework, we introduce learnable modules φ in the knowledge expression denoted as fφ.
2. We aim to learn the knowledge by determining φ from data.
3. As any meaningful knowledge is expected to be consistent with the observations, a straightforward way is then to directly optimize against the training data:

$$\Phi^* = argmax_\phi(\frac{1}{N}\sum_n f_\phi(x_n, y_n))$$

$$\boldsymbol{\phi}^{(t+1)} = \arg\max_{\phi \in \Phi} \frac{1}{N} \sum_{n=1}^{N} (1 - \pi') h_{\phi}(\boldsymbol{x}_n, \boldsymbol{y}_n) + \pi' \mathbb{E}_{q^{(t)}(\boldsymbol{y})} [h_{\phi}(\boldsymbol{x}_n, \boldsymbol{y})]$$

$$\boldsymbol{\lambda}^{(t+1)} = \arg\max_{\boldsymbol{\lambda} \geq 0} \frac{1}{N} \sum_{n=1}^{N} q_{\lambda}(\boldsymbol{y}_n)$$

Algorithm 1 Mutual Distillation

Input: Training data $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$,
Initial knowledge constraints $\mathcal{F} = \{f_{\phi,l}\}_{l=1}^{L}$,
Initial neural network p_θ,
Parameters: π, π' – imitation parameters
C – regularization parameters

1: Initialize neural network parameters $\boldsymbol{\theta}$
2: Initialize knowledge parameters $\boldsymbol{\phi}$ and weights $\boldsymbol{\lambda}$
3: **while** not converged **do**
4: Sample a minibatch $(\boldsymbol{X}, \boldsymbol{Y}) \subset \mathcal{D}$
5: Build the teacher model q with Eq.(2) and Eq.(6)
6: Update p_θ with distillation objective Eq.(3)
7: Update f_l $(l = 1, \ldots, L)$ with distillation objective Eq.(5)
8: **end while**

Output: Learned network p, knowledge modules $\mathcal{F}$, and the joint teacher network q