Knowledge Distillation

Deep Learning (ENGG*6600*01)

School of Engineering,
University of Guelph, ON, Canada

Course Instructor: Benyamin Ghojogh

Summer 2023

Knowledge Distillation

Introduction

- Knowledge Distillation (KD) was proposed in 2015 [1].
- It is used for network compression.
- The large neural network is called the **teacher network**. The smaller version of neural network, i.e., the compressed network, is called the **student network**.
- The student network tries to mimic the behaviour of the teacher network; therefore, it can be considered as the compressed version of the large teacher network.
- Assume the teacher network is already trained on the training dataset. The student network is trained by minimizing the following loss functions:

$$\mathcal{L}_{\text{kl}} = (1-\lambda)\mathcal{L}_{\text{ce}} + \lambda\mathcal{L}_{\text{kd}}, \tag{1}$$

where:

$$\mathcal{L}_{\text{ce}} := \text{CE}\Big(y, \sigma\big(\mathbf{f}_s(\boldsymbol{x})\big)\Big), \tag{2}$$

$$\mathcal{L}_{\text{kl}} := \tau^2 \text{KL}\Big(\sigma\big(\frac{\mathbf{f}_s(\boldsymbol{x})}{\tau}\big) \| \sigma\big(\frac{\mathbf{f}_t(\boldsymbol{x})}{\tau}\big)\Big), \tag{3}$$

where $\sigma(.)$ is the sigmoid activation function, y is the target label for the data $\boldsymbol{x}$, $\tau > 0$ is the temperature, CE and KL denote the cross entropy and KL-divergence functions, respectively. $\mathbf{f}_s(\boldsymbol{x})$ and $\mathbf{f}_t(\boldsymbol{x})$ are the outputs of the student network and the teacher network for the input $\boldsymbol{x}$, respectively.

- CE is for hard labels (target labels) and KL is for soft labels (mimicking).

Knowledge Distillation

- The CE and KL losses are:

$$\mathcal{L}_{\text{ce}} := -\sum_{l=1}^{c} (\boldsymbol{y}_i)_l \log\Big(\sigma\big(\mathbf{f}_s(\boldsymbol{x})\big)_l\Big), \tag{4}$$

$$\mathcal{L}_{\text{kl}} := \tau^2 \sum_{i=1}^{b} \sigma\big(\frac{\mathbf{f}_s(\boldsymbol{x})}{\tau}\big) \log\Big(\frac{\sigma\big(\frac{\mathbf{f}_s(\boldsymbol{x})}{\tau}\big)}{\sigma\big(\frac{\mathbf{f}_t(\boldsymbol{x})}{\tau}\big)}\Big), \tag{5}$$

where target labels are one-hot encoded, i.e., $\boldsymbol{y}_i \in \{0,1\}^c$ (c is the number of classes), $\sigma(.)$ is the sigmoid activation function, and $(\boldsymbol{y}_i)_l$ and $\sigma\big(\mathbf{f}(\boldsymbol{x})\big)_l$ denote the l-th element of $\boldsymbol{y}_i$ and $\sigma\big(\mathbf{f}(\boldsymbol{x})\big)$, respectively.

Annealing in Knowledge Distillation

Annealing in Knowledge Distillation

- We can have two stages [2]:
  - stage 1: gradually mimicking the teacher by the student (learning the soft labels)
  - stage 2: learning the hard labels

  whose loss functions are:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\text{kd}}(i) & \text{stage 1: } 1 \leq \tau_i \leq \tau_{\text{max}} \\ \mathcal{L}_{\text{ce}} & \text{stage 2} \end{cases} \tag{6}$$

where τ_i is the temperature at iteration index i and:

$$\mathcal{L}_{\text{kd}}(i) := \|\mathbf{f}_s(\boldsymbol{x}) - \mathbf{f}_t(\boldsymbol{x})\phi(\tau_i)\|_2^2, \tag{7}$$

$$\phi(\tau_i) = 1 - \frac{\tau_i - 1}{\tau_{\text{max}}}, \quad \tau_i \in \{1, 2, \dots, \tau_{\text{max}}\}. \tag{8}$$

Other Variants of Knowledge Distillation

Other Variants of Knowledge Distillation

- One problem with KD is if the size of teacher and student nets differ significantly, it does not work well. This problem is called the **gap problem**. So, we can have intermediate network(s) called teacher assistant (TA) network (2020) [3]. We can also have a hierarchy of TA networks between the teacher and student networks [3].
- So far, we assumed that the teacher network is fully trained and then the student network is trained. Alternatively, we can train both the teacher and student networks can be trained simultaneously (2021) [4] where the KD loss is used for both. In this way, teacher also learns from the student while the student learns from the teacher.
- One problem with KD is that it has been empirically found out that not necessarily the last iteration of the teacher network is best for training the student network. Some intermediate epoch checkpoint of the teacher may be better to use for training the student network. This needs a checkpoint search in the teacher net. The problem is called the **checkpoint search problem**. Alternatively, we can have two stages where we train the teacher and student together simultaneously in the first stage and in the second stage, we fine tune the student using only CE loss (2021) [5].
- We can mimic the output of every layer of the teacher net for the student net. However, the structures of the two nets differ, so for layer-wise KD, we can use **attention** weights for distilling a linear combination of all layers of the teacher network (2021) [6].
- And many other variants...
- A survey on KD is [7].

References

[1] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[2] A. Jafari, M. Rezagholizadeh, P. Sharma, and A. Ghodsi, “Annealing knowledge distillation,” *arXiv preprint arXiv:2104.07163*, 2021.

[3] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, “Improved knowledge distillation via teacher assistant,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 5191–5198, 2020.

[4] S. Bhardwaj, A. Ghaddar, A. Rashid, K. Bibi, C. Li, A. Ghodsi, P. Langlais, and M. Rezagholizadeh, “Knowledge distillation with noisy labels for natural language understanding,” *arXiv preprint arXiv:2109.10147*, 2021.

[5] M. Rezagholizadeh, A. Jafari, P. Salad, P. Sharma, A. S. Pasand, and A. Ghodsi, “Pro-kd: Progressive distillation by following the footsteps of the teacher,” *arXiv preprint arXiv:2110.08532*, 2021.

[6] Y. Wu, M. Rezagholizadeh, A. Ghaddar, M. A. Haidar, and A. Ghodsi, “Universal-kd: Attention-based output-grounded intermediate layer knowledge distillation,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7649–7661, 2021.

[7] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.