1 Angle-wise distillation loss

Given a triplet of examples, an angle-wise relational potential measures the angle formed by the three examples in the output representation space:
Idea 1:

$$\begin{aligned}\psi(t_i, t_j, t_k) &= \cos \angle t_i t_j t_k + \frac{\lambda}{2} \sum_{q \in \{i,j,k\}} \|t_q - \bar{t}\|_2^2 \\ &= \langle e^{ij}, e^{jk} \rangle + \frac{\lambda}{2} \sum_{q \in \{i,j,k\}} \|t_q - \bar{t}\|_2^2, \end{aligned} \tag{1.1}$$

Idea 2:

$$\psi(t_i, t_j, t_k) = \alpha \cos \angle t_i t_j t_k + (1 - \alpha) \sum_{q_1, q_2 \in \{i,j,k\}} \|t_{q_1} - t_{q_2}\|_2^2, \tag{1.2}$$

where

$$\alpha \in [0, 1], \ \bar{t} = \frac{t_i + t_j + t_k}{3} \text{ and } e^{ij} = \frac{t_i - t_j}{\|t_i - t_j\|_2}.$$

The angle-wise distillation loss transfers the relationship of training example embeddings by penalizing angular differences. Since an angle is a higher-order property than a distance, it may be able to transfer relational information more effectively, giving more flexibility to the student in training. In our experiments, we observed that the angle-wise loss often allows for faster convergence and better performance.

Since distillation attempts to match the distance-wise potentials between the teacher and the student, this mini-batch distance normalization is useful particularly when there is a significant difference in scales between teacher distances $\|t_i - t_j\|_2$ and student distances $\|s_i - s_j\|_2$, e.g., due to the difference in output dimensions. In our experiments, we observed that the normalization provides more stable and faster convergence in training. Using the distance-wise potentials measured in both the teacher and the student, a distance-wise distillation loss is defined as

$$\mathcal{L}_{RKD-D} = \sum_{(x_i, x_j) \in \mathcal{X}^2} l_\delta \left(\psi_D(t_i, t_j), \psi_D(s_i, s_j) \right),$$

where l_δ is generalized Huber loss, which is defined as

$$l_\delta(x, y) = \begin{cases} \frac{1}{2}|x - y|^\gamma & \text{for } |x - y| \leq 1; \\ |x - y| - \frac{1}{2} & \text{otherwise.} \end{cases} \tag{1.3}$$

where $\gamma \geq 1$ is a tuning parameter. If $\gamma \in [1, 2)$, the loss function tolerates more on the difference between teacher and student than the classical Huber loss function.

References