Knowledge Distillation with a Precise Teacher and Prediction with Abstention

Yi Xu, Jian Pu, Hui Zhao

School of Software Engineering, East China Normal University,
Institute of Science and Technology for Brain-inspired Intelligence, Fudan University
Email: 51184501068@stu.ecnu.edu.cn, jianpu@fudan.edu.cn, hzhao@sei.ecnu.edu.cn

Introduction

Knowledge distillation has achieved remarkable results in supervised learning. However, there are two major problems with existing knowledge distillation methods. To address the first issue, we apply knowledge adjustment to correct teachers’ supervision using ground truth. To address the second problem, we use the selective classification framework to train the student model. In particular, the deep gambler loss is adopted to predict with reservation by introducing the extra class.

Method

Knowledge Distillation with Deep Gambler

- We can add a class to stand for abandoning predictions and reservations according to Deep Gambler Loss in selective classification.

\[ W(b(f), p) = \sum (p \cdot f_w(x)) + f_w(x)_{m+1} \]

- We proposed the loss function that utilizes Deep Gambler (DG) loss to the KA method.

\[ L = \sum A(q_i) \log \left( p_i + \frac{1}{p_i} \right) \]

Knowledge Adjustment

- The Knowledge Distillation(KD) loss as:

\[ L_{KD} = \alpha \tau^2 \cdot CE(q_i, p_i) + (1 - \alpha) \cdot CE(Y, p_i) \]

- Swap the incorrect value with the true targets. We only need to operate on the incorrect ones and denote it as an operator A(·). The KD loss becomes:

\[ L_{KD^*} = \tau^2 CE(A(q_i), p_i) \]

Motivation

1. We propose to use knowledge adjustment to revise teacher’s incorrect supervision using the ground truth label.
2. We propose to use the deep gambler loss to train the student network in an end-to-end way.
3. We evaluate the proposed method under two knowledge distillation settings, i.e., knowledge distillation across different network structures and distillation across networks with different depths.

Reference

