Hard Gate Knowledge Distillation - Leverage Calibration for a Robust and Reliable Language Model

Dongkyu Lee[1,3*] **Zhiliang Tian**[2†] **Yingxiu Zhao**[1]
Ka Chun Cheung[3] **Nevin L. Zhang**[1]
[1]Department of Computer Science and Engineering, HKUST
[2]College of Computer, National University of Defense Technology
[3]NVIDIA AI Technology Center, NVIDIA
[1]{dleear, yzhaocx, lzhang}@cse.ust.hk
[2]tianzhilianghit@gmail.com [3]chcheung@nvidia.com

Abstract

In knowledge distillation, a student model is trained with supervisions from both knowledge from a teacher and observations drawn from a training data distribution. Knowledge of a teacher is considered a subject that holds inter-class relations which send a meaningful supervision to a student; hence, much effort has been put to find such knowledge to be distilled. In this paper, we explore a question that has been given little attention: "*when to distill such knowledge*." The question is answered in our work with the concept of model calibration; we view a teacher model not only as a source of knowledge but also as a gauge to detect miscalibration of a student. This simple and yet novel view leads to a hard gate knowledge distillation scheme that switches between learning from a teacher model and training data. We verify the gating mechanism in the context of natural language generation at both the token-level and the sentence-level. Empirical comparisons with strong baselines show that hard gate knowledge distillation not only improves model generalization, but also significantly lowers model calibration error.

1 Introduction

In recent years, the deep learning community has achieved marked performance gains across a variety of tasks (Brown et al., 2020; Devlin et al., 2018). In the meantime, some deep learning models have become excessively large, limiting their applicability in some scenarios. To cope with the issue, Hinton et al. (2015) proposed knowledge distillation (KD), in which knowledge of a large network, called a teacher network, is transferred to a relatively small model, called a student model.

The benefits of KD have been widely witnessed across multiple domains (Romero et al., 2015; Jiao et al., 2020). Recently, it has been observed that KD can be used in both reducing model size and improving model generalization (Tang et al., 2021; Furlanello et al., 2018). Hinton et al. (2015) argue that a distribution, defined by a teacher, holds inter-class relations, commonly referred to as the *dark knowledge*, and that such distribution brings a meaningful supervision to a student. Therefore, a large body of research in KD has viewed a teacher as a source of knowledge and has focused on *finding a meaningful knowledge* to be transferred (Romero et al., 2015; Bulò et al., 2016; Park et al., 2019; Yuan et al., 2020; Kim et al., 2021).

In this work, we focus on *when to distill knowledge of a teacher*. This is a central question to ask, as a model can benefit from the adaptive control of supervision between ground truth and a teacher; When a model is trained to increase the predictive score of a prediction, a one-hot encoded supervision, without incorporating teacher model, sends a direct signal in increasing the score (Müller et al., 2019). In another case, when a model is trained to learn knowledge of a teacher, a teacher's output without fusing a ground truth sends more direct signal in minimizing the knowledge gap between the student and the teacher. However, the question of "when" has not been answered. For this reason, previous works choose to learn from both of the supervisions.

We give an answer to the question from the perspective of model calibration. Model calibration refers to how well a predicted probability of a model reflects the true accuracy. Therefore, a well-calibrated predictive score represents the **likelihood of correctness of a prediction** (Guo et al., 2017). In this light, such score can be viewed as a gauge to detect a miscalibration of a student in training; when a student makes a prediction with a probability mass that is higher than the expected accuracy of the prediction (overconfidence), a student model is trained with only supervision from a

*This work was done while Dongkyu was an intern at NVIDIA

†Corresponding author

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9793 - 9803