Chapter 7
Can a Student Large Language Model Perform as Well as Its Teacher?

Sia Gholami
Independent Researcher, USA

Marwan Omar
Illinois Institute of Technology, USA

ABSTRACT

The burgeoning complexity of contemporary deep learning models, while achieving unparalleled accuracy, has inadvertently introduced deployment challenges in resource-constrained environments. Through meticulous examination, the authors elucidate the critical determinants of successful distillation, including the architecture of the student model, the caliber of the teacher, and the delicate balance of hyperparameters. While acknowledging its profound advantages, they also delve into the complexities and challenges inherent in the process. The exploration underscores knowledge distillation's potential as a pivotal technique in optimizing the trade-off between model performance and deployment efficiency.

1. INTRODUCTION

In recent years, the landscape of deep learning has been characterized by models that are increasingly large and intricate. While such models, often boasting billions of parameters, consistently set new benchmarks in accuracy, their computational intensity presents deployment challenges, especially in environments with limited computational resources, such as edge devices 9Tao et al., 2020). Knowledge distillation offers a viable solution to this quandary, facilitating the transfer of knowledge from a sophisticated, high-capacity "teacher" model to a more compact "student" model, aiming to retain as much of the performance as possible (Hinton et al., 2015).

DOI: 10.4018/979-8-3693-1906-2.ch007

Central to knowledge distillation is the principle that learning can be enhanced when models are trained not just on hard labels but also on the richer, probabilistic outputs of a teacher model. These soft labels can be perceived as capturing the teacher's confidence distribution across classes, providing nuanced insights which hard labels might overlook (Bucilua et al., 2006)ˇ

A critical component of this approach is temperature scaling, which modulates the granularity of these soft labels. The temperature parameter, introduced by Hinton et al. (2015), plays a pivotal role in controlling the "sharpness" of the teacher's output distributions, thus influencing the quality of the information relayed to the student model.

The training of the student model is then typically guided by a weighted loss function that balances between the conventional cross-entropy loss and the divergence from the teacher's outputs, usually measured using Kullback-Leibler divergence (Lopez-Paz et al., 2015).

However, the process is not without complexities. The optimal architecture of the student model, the quality of the teacher, and the precise balance of hyperparameters are all determining factors in the success of the distillation (Polino et al., 2018). The intricacies of these factors and their interplay remain a focal point of contemporary research.

In conclusion, knowledge distillation emerges as a key technique in the deep learning toolkit, bridging the divide between cutting-edge performance and practical, efficient deployment. Its continued exploration holds the promise of further refining and expanding its applicability across diverse domains.

To use knowledge distillation for creating efficient transformers, the process typically involves the following steps:

1. Train a large, complex transformer model as the teacher model on the task of interest.
2. Generate a dataset of examples for the task, and use the teacher model to generate predictions for each example.
3. Train a smaller, simpler transformer model as the student model on the same task, using the predictions of the teacher model as targets.
4. Use a combination of the original task loss and a distillation loss to train the student model. The distillation loss encourages the student model to mimic the predictions of the teacher model, rather than just trying to optimize the original task loss.

By using knowledge distillation in this way, it is possible to create efficient transformer models that are smaller and faster than the original model, while still achieving comparable or even better performance on the task of interest.

There are several benefits to using knowledge distillation in building efficient transformers:

1. Improved efficiency: Knowledge distillation allows you to create smaller, more efficient Transformer models that require fewer computational resources for training and inference. This enables faster processing and reduced memory usage, making it easier to deploy the models on resource-constrained devices like mobile phones or edge devices.
2. Reduced energy consumption: Smaller models produced through knowledge distillation consume less energy during inference, which is crucial for battery-powered devices and sustainable AI solutions.