Distilling Complex Reasoning Chains in Small LLMs: a Case Study on New York Times Connections's Game

Marco Colangelo[1*]

[1]Department of Computer Science, University of Milan, Milan, Italy.

Corresponding author(s). E-mail(s): marco.colangelo@studenti.unimi.it;

Abstract

This report contains the explanation of the work done for the final project of the Natural Language Processing course. The professor in charge of the course is Professor Alfio Ferrara, University of Milan. The project focused on the application of knowledge distillation, a technique used to compress large language models into smaller and more efficient versions. This is achived by training a smaller model (the student) to replicate the behavior of a larger model (the teacher).

1 Introduction

Language model pre-training, such as BERT, has significantly improved the perfomances of many NLP tasks. However, pre-trained language models are usually computationally expensive, so it is difficult to efficiently execute them on resource-restricted devices. A novel Transformer distillation method, specially designed for knowledge distillation of the Transformer-based models, allows to effectively transfer to a smaller student model called TinyBERT the plenty of knowledge from a large pre-trained BERT model. A two-stage learning framework for TinyBERT which performs Transformer distillation at both the pretraining and task-specific learning stages ensures that TinyBERT can capture the general-domain as well as the task-specific knowledge in BERT. [1]

1.1 State of the art work

So far, two main studies have defined the paradigm of model compression via knowledge transfer. The first, by Hinton et al. [2], introduced the seminal concept of Knowledge Distillation (KD), focusing on transferring the so-called dark knowledge from a large Teacher to a compact Student. Their approach minimizes the distance between the output probability distributions of the two networks, softened by a temperature parameter T. The second and more specialized study, by Jiao et al. [1], proposes a variant named TinyBERT, specifically designed for Transformer-based architectures. While Hinton's method operates on the final output layer, Jiao et al. argue that for complex language models, this is insufficient. Their method leverages a layer-to-layer distillation strategy, forcing the student to mimic not only the teacher's prediction but also its intermediate representations, such as attention matrices and hidden states.

2 Related Work

2.1 Aim of the work

2.2 Data

2.2.1 Preprocessing

2.3 Implementation

2.3.1 Structure of the models

2.3.2 Training phase

For the experiments, a laptop equipped with a Nvidia GTX 1650 Max-Q GPU and 16GB of RAM was used to perform simple tests, while the main training phase involving the TinyBERT model was performed on a desktop PC equipped with a Nvidia RTX 3070 GPU and 32GB of RAM.

2.3.3 Distillation Objectives

3 Results

4 Conclusion

References

[1] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: TinyBERT: Distilling BERT for Natural Language Understanding (2020). https://arxiv.org/abs/1909.10351

[2] Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network (2015). https://arxiv.org/abs/1503.02531