Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes

Cheng-Yu Hsieh[1]*, **Chun-Liang Li**[2], **Chih-Kuan Yeh**[3], **Hootan Nakhost**[2], **Yasuhisa Fujii**[3], **Alexander Ratner**[1], **Ranjay Krishna**[1], **Chen-Yu Lee**[2], **Tomas Pfister**[2]
[1]University of Washington, [2]Google Cloud AI Research, [3]Google Research
cydhsieh@cs.washington.edu

Abstract

Deploying large language models (LLMs) is challenging because they are memory inefficient and compute-intensive for practical applications. In reaction, researchers train smaller task-specific models by either finetuning with human labels or distilling using LLM-generated labels. However, finetuning and distillation require large amounts of training data to achieve comparable performance to LLMs. We introduce *Distilling step-by-step*, a new mechanism that (a) trains smaller models that outperform LLMs, and (b) achieves so by leveraging less training data needed by finetuning or distillation. Our method extracts LLM rationales as additional supervision for small models within a multi-task training framework. We present three findings across 4 NLP benchmarks: First, compared to both finetuning and distillation, our mechanism achieves better performance with much fewer labeled/unlabeled training examples. Second, compared to LLMs, we achieve better performance using substantially smaller model sizes. Third, we reduce both the model size and the amount of data required to outperform LLMs; our 770M T5 model outperforms the 540B PaLM model using only 80% of available data on a benchmark task.

1 Introduction

Despite the impressive few-shot ability offered by large language models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Thoppilan et al., 2022; Hoffmann et al., 2022; Smith et al., 2022b; Zhang et al., 2022), these models are challenging to deploy in real world applications due to their sheer size. Serving a single 175 billion LLM requires at least 350GB GPU memory using specialized infrastructure (Zheng et al., 2022). To make matters worse, today's state-of-the-art LLMs are composed of over 500B parameters (Chowdhery et al., 2022), requiring significantly more memory and compute. Such computational requirements are far beyond affordable for most product teams, especially for applications that require low latency performance.

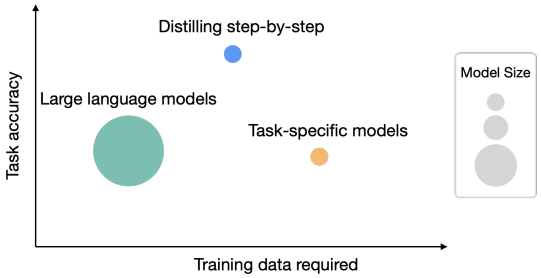

Figure 1: While large language models (LLMs) offer strong zero/few-shot performance, they are challenging to serve in practice. Traditional ways of training small task-specific models, on the other hand, requires large amount of training data. We propose Distilling step-by-step, a new paradigm that extracts rationales from LLMs as informative task knowledge into training small models, which reduces both the deployed model size as well as the data required for training.

To circumvent these deployment challenges of large models, practitioners often choose to deploy smaller specialized models instead. These smaller models are trained using one of two common paradigms: *finetuning* or *distillation*. Finetuning updates a pretrained smaller model (e.g. BERT (Devlin et al., 2018) or T5 (Raffel et al., 2020)) using downstream human annotated data (Howard and Ruder, 2018). Distillation trains the same smaller models with labels generated by a larger LLM (Tang et al., 2019; Wang et al., 2021; Smith et al., 2022a; Arora et al., 2022). Unfortunately, these paradigms reduce model size at a cost: to achieve comparable performance to LLMs, finetuning requires expensive human labels, and distillation requires large amounts of unlabeled data which can be hard to obtain (Tang et al., 2019; Liang et al., 2020).

In this work, we introduce **Distilling step-by-step**, a new simple mechanism for training smaller

*Work done while the author was a student researcher at Google Cloud AI Research.