MCIT
Modeling, control &
information technologies

Distillation of Large Language Models for Text Simplification

https://doi.org/10.31713/MCIT.2023.071

Oleksandr Skurzhanskyi
Department of Computer Science and Cybernetics
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
oleksandr.skurzhanskyi@gmail.com

Abstract — **This work presents a comprehensive methodology for harnessing the capabilities of Large Language Models to address specific Natural Language Processing tasks, with a focus on Text Simplification. While LLMs have demonstrated their prowess in tackling a wide range of NLP challenges, their demanding computational requirements can render them impractical for real-time online inference. In response to this limitation, we suggest the concept of text distillation, a technique aimed at effectively transferring the knowledge stored within LLMs to more compact and computationally efficient neural networks.**

Keywords — artificial intelligence; natural language processing; large language models; text simplication.

I. Introduction

In recent years, the field of Natural Language Processing (NLP) has witnessed a remarkable transformation, primarily driven by the advent of Large Language Models (LLMs), such as InstructGPT [1]. These models, often comprising billions of parameters, have exhibited unprecedented proficiency in solving a multitude of NLP tasks, ranging from machine translation [2] to sentiment analysis. However, their widespread adoption has been somewhat hindered by their substantial computational demands, which can impede their practicality for applications that require real-time processing.

This paper addresses the challenge of making LLMs more accessible for NLP tasks that demand efficiency, with a particular focus on Text Simplification. Text Simplification constitutes a vital aspect of NLP, aiming to make complex texts more comprehensible while preserving their core meaning. To tackle this challenge efficiently, we employ a data distillation approach. The process begins with the selection of a relevant prompt tailored to the nature of the tasks. Subsequently, we generate examples using a Generative AI approach, LLM. Finally, we train a fast, small transformer-based model using this distilled data.

II. Related works

Several strategies have been proposed to mitigate the computational burden associated with deploying Large Language Models (LLMs) for online inference. One prominent approach, as introduced by Geoffrey Hinton, Oriol Vinyals, and Jeff Dean in their seminal work "Distilling the Knowledge in a Neural Network" [3], involves the utilization of text distillation. This technique seeks to distill the knowledge contained within LLMs into smaller, more streamlined neural networks.

In their paper, Hinton et al. presented the concept of model distillation, demonstrating that it is possible to transfer the knowledge stored within a complex neural network (called "teacher") to a smaller and faster network ("student"). This process involves training the student network to mimic the teacher network's behavior, effectively transferring its knowledge while reducing computational demands. While their work primarily focused on traditional neural networks, the principles of knowledge distillation have found widespread application in the context of LLMs, contributing to the development of efficient NLP solutions.

In addition to distillation techniques, researchers have explored model quantization, where LLMs are pruned or quantized to reduce their memory and computational requirements [4]. Model quantization involves compressing the parameters of large models into a smaller representation while maintaining performance. This approach has gained traction due to its ability to significantly reduce the memory footprint and inference latency of LLMs, making them more suitable for resource-constrained environments.

Furthermore, efforts have been made to create domain-specific LLMs, which are tailored to excel in particular NLP tasks while maintaining a more modest model size. These domain-adapted LLMs are fine-tuned on specific datasets and tasks, allowing them to achieve competitive performance with reduced computational demands. Such specialization enhances the efficiency of LLMs for task-specific applications, addressing the challenge of adapting these powerful models to real-world scenarios with limited computational resources.

III. Approach

Our methodology for achieving efficient Text Simplification using Large Language Models (LLMs) revolves around the concept of data distillation. Rather than directly applying knowledge distillation by altering the loss function (just like in the Hinton's paper), we find it more practical specifically for text generation models to generate data and utilize this data for training — a process we refer to as data distillation. The challenge lies in adapting loss modifications while preserving token probabilities, a task complicated by the autoregressive nature of text generation.

A. *Prompt Selection*

The data distillation process begins with the careful selection of a relevant prompt tailored to the nature of the Text Simplification task at hand. These prompts act as specific instructions to the LLM, guiding it to generate simplified versions of complex texts while preserving their essential meaning. Crafting effective prompts is a crucial step in ensuring that the generated examples align with the objectives of Text Simplification.

For instance, prompts may instruct the model to "simplify the following passage while maintaining clarity" or "paraphrase the text to enhance readability." In our experiments prompt “You are a helpful AI assistant with extensive linguistic knowledge. Rewrite the following text to make it simpler and more straightforward.”

B. *Dataset generation*

Next, we harness the capabilities of a Generative AI model, ChatGPT [5], to generate a substantial dataset of training examples. ChatGPT is a versatile language model capable of producing coherent and contextually relevant text based on given prompts. In our case, we design prompts specifically to elicit simplified versions of complex texts, thereby aligning the examples with the objectives of Text Simplification. ChatGPT's proficiency in text generation makes it a valuable tool for generating high-quality training data for our task. For example, given a complex medical document, ChatGPT can be prompted to simplify medical jargon and convoluted sentences, resulting in more accessible content.

C. *Training compact model*

Once we've generated a dataset of simplified text examples through data distillation, we proceed to train a compact transformer-based model designed for efficiency, featuring a reduced parameter count compared to the original LLMs.

For selecting the transformer-based model, we prioritize size without sacrificing quality. We also recommend employing a reduced number of decoder layers, a practice that has demonstrated a significant acceleration in inference time [6]. The training process involves fine-tuning the model on the generated examples while optimizing for both performance and computational efficiency.

Our primary objective is to develop a model capable of efficiently producing simplified text without compromising quality. In pursuit of this goal, our training process may incorporate techniques such as model quantization to further minimize the model's memory footprint while preserving performance. Specifically, floating-point 16 (fp16) quantization is a noteworthy approach, as it typically maintains performance levels while reducing memory usage by a factor of two.

IV. Conclusions

In this paper, we have presented a comprehensive methodology for achieving efficient Text Simplification using Large Language Models (LLMs). Our approach revolves around the concept of data distillation, which leverages the strengths of LLMs to generate high-quality training examples for a compact transformer-based model. This methodology offers a practical solution to the computational challenges associated with LLMs, making them more accessible for Text Simplification and similar Natural Language Processing (NLP) tasks.

Our research has highlighted several key insights:

- **Data Distillation for Improved Efficiency**: Data distillation, as an alternative to traditional knowledge distillation, proves to be a highly effective approach for Text Simplification. By generating simplified text examples through ChatGPT and using them for training, we achieve both efficiency and quality in the final simplified outputs;
- **The Significance of Prompting**: Carefully crafted prompts play a crucial role in guiding LLMs to generate relevant and high-quality simplifications. The choice of prompts directly impacts the success of the data distillation process.
- **Compact Transformer-Based Models**: Our choice of a compact transformer-based model, tailored for efficiency, demonstrates the practicality of reducing parameter counts while maintaining performance. This model choice, coupled with techniques like model quantization, enables efficient inference without compromising the quality of the simplified text.

References

[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al., “Training language models to follow instructions with human feedback,” in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 27730–27744.

[2] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in Advances in neural information processing systems, vol. 27, 2014.

[3] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” arXiv preprint arXiv:1503.02531, 2015.

[4] A. Polino, R. Pascanu, and D. Alistarh, “Model compression via distillation and quantization,” arXiv preprint arXiv:1802.05668, 2018.

[5] “ChatGPT.” [Online]. Available: https://chat.openai.com/. [Accessed: August 24, 2023].

[6] J. Kasai, N. Pappas, H. Peng, J. Cross, and N. A. Smith, “Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation,” in ICLR 2021, 2021.