SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling

Dahyun Kim*, Chanjun Park*†, Sanghoon Kim*†, Wonsung Lee*†, Wonho Song*
Yunsu Kim*, Hyeonwoo Kim*, Yungi Kim, Hyeonju Lee, Jihoo Kim
Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim
Mikyoung Cha, Hwalsuk Lee†, Sunghun Kim†

Upstage AI, South Korea

{kdahyun, chanjun.park, limerobot, wonsung.lee, hwalsuk.lee, hunkim}@upstage.ai

Abstract

We introduce SOLAR 10.7B, a large language model (LLM) with 10.7 billion parameters, demonstrating superior performance in various natural language processing (NLP) tasks. Inspired by recent efforts to efficiently up-scale LLMs, we present a method for scaling LLMs called depth up-scaling (DUS), which encompasses depthwise scaling and continued pretraining. In contrast to other LLM up-scaling methods that use mixture-of-experts, DUS does not require complex changes to train and inference efficiently. We show experimentally that DUS is simple yet effective in scaling up high-performance LLMs from small ones. Building on the DUS model, we additionally present SOLAR 10.7B-Instruct, a variant fine-tuned for instruction-following capabilities, surpassing Mixtral-8x7B-Instruct. SOLAR 10.7B is publicly available under the Apache 2.0 license, promoting broad access and application in the LLM field [1].

1 Introduction

The field of natural language processing (NLP) has been significantly transformed by the introduction of large language models (LLMs), which have enhanced our understanding and interaction with human language (Zhao et al., 2023). These advancements bring challenges such as the increased need to train ever larger models (Rae et al., 2021; Wang et al., 2023; Pan et al., 2023; Lian, 2023; Yao et al., 2023; Gesmundo and Maile, 2023) owing to the performance scaling law (Kaplan et al., 2020; Hernandez et al., 2021; Anil et al., 2023; Kaddour et al., 2023). To efficiently tackle the above, recent works in scaling language models such as a mixture of experts (MoE) (Shazeer et al., 2017; Komatsuzaki et al., 2022) have been proposed. While those approaches are able to efficiently and effectively scale-up LLMs, they often require non-trivial changes to the training and inference framework (Gale et al., 2023), which hinders widespread applicability. Effectively and efficiently scaling up LLMs whilst also retaining the *simplicity* for ease of use is an important problem (Alberts et al., 2023; Fraiwan and Khasawneh, 2023; Sallam et al., 2023; Bahrini et al., 2023).

Inspired by Komatsuzaki et al. (2022), we present depth up-scaling (DUS), an effective and efficient method to up-scale LLMs whilst also remaining straightforward to use. DUS consists of scaling the number of layers in the base model and continually pretraining the scaled model. Unlike (Komatsuzaki et al., 2022), DUS does not scale the model using MoE and rather use a depthwise scaling method analogous to Tan and Le (2019) which is adapted for the LLM architecture. Thus, there are no additional modules or dynamism as with MoE, making DUS immediately compatible with easy-to-use LLM frameworks such as HuggingFace (Wolf et al., 2019) with no changes to the training or inference framework for maximal efficiency. Furthermore, DUS is applicable to all transformer architectures, opening up new gateways to effectively and efficiently scale-up LLMs in a simple manner. Using DUS, we release SOLAR 10.7B, an LLM with 10.7 billion parameters, that outperforms existing models like Llama 2 (Touvron et al., 2023) and Mistral 7B (Jiang et al., 2023) in various benchmarks.

We have also developed SOLAR 10.7B-Instruct, a variant fine-tuned for tasks requiring strict adherence to complex instructions. It significantly outperforms the Mixtral-8x7B-Instruct model across various evaluation metrics, evidencing an advanced proficiency that exceeds the capabilities of even larger models in terms of benchmark performance.

By releasing SOLAR 10.7B under the Apache 2.0 license, we aim to promote collaboration and innovation in NLP. This open-source approach allows

*Equal Contribution † Corresponding Author

[1] https://huggingface.co/upstage/SOLAR-10.7B-v1.0