Project Proposal:
'Can frankenmodels really be a free-lunch?'
Evaluating Layer Duplication in LLaMA

Noga Bregman
nogabregman@mail.tau.ac.il

Sharon Saban
sharonsaban@mail.tau.ac.il

General Description

FrankenModels, a novel approach to enhancing large language models (LLMs), involve repeating and modifying layers within a pre-trained model to assess the impact on performance. This technique has gained traction in the NLP community, with discussions and implementations emerging on platforms like Reddit and GitHub. Our project aims to systematically explore the effectiveness of FrankenModels within the LLaMA family, a state-of-the-art series of language models renowned for their scalability and adaptability.

FrankenModels offer a computationally efficient alternative to training entirely new models. By duplicating specific layers, we hypothesize that it is possible to enhance model depth and capability without significantly increasing computational overhead.

Objective

Our primary objective is to evaluate the performance impact of layer duplication in the LLaMA model family. We aim to:

- Investigate whether repeating layers enhances task-specific performance.
- Compare FrankenModels both to their baseline counterparts (i.e., unmodified LLaMA models) and to other FrankenModels within the same family.
- Ensure that a consistent task is used across experiments while potentially exploring other tasks to verify improvements do not degrade performance on alternative capabilities.

Experiment Description

The experiments will involve:

Layer Duplication: As part of our experiments, we will initially explore which layers to duplicate and the optimal number of duplications to apply. Once these decisions are made, they will remain fixed throughout the experiments to maintain consistency and enable precise comparisons across models. This process will be conducted for each of the chosen models in the LLaMA family. By leveraging this constant setup, we aim to focus on evaluating the broader impacts of layer duplication on model performance rather than continually adjusting these parameters. Tools such as the ExLlamaV2 library, which supports efficient layer repetition during inference, will play a key role in implementing this approach. Further details about ExLlamaV2 will be provided below.

Task Evaluation: We will evaluate model performance after implementing layer duplication by comparing each modified model to its unmodified counterpart and to other FrankenModels within the LLaMA family. This evaluation will be conducted on a specific task to ensure consistency in comparison. We are considering tasks such as text classification and language modeling, as these align with the capabilities of the LLaMA family and provide meaningful insights into generalization and specialization. Task selection will remain flexible as experiments progress, allowing us to adapt to evolving research needs and goals.

Analysis of Results: The benchmarks for analysis will be chosen in accordance with the specific tasks selected for evaluation. Additionally, we may develop our own methods for evaluation depending on the focus of the chosen tasks, providing comprehensive insights into FrankenModel performance.

Third-Party Tools and Resource Usage

Model Selection

We are considering models within the LLaMA family, including:

LLaMA-7B: A lightweight model for resource-constrained tasks.

LLaMA-13B: Balances performance and computational demands.

LLaMA-30B: A more capable model for complex tasks.

We note that these are preliminary choices, and the final selection will depend on experimental results and resource availability.

Libraries and Applications

ExLlamaV2

An optimized Python library for running LLaMA models efficiently on consumer-grade GPUs. This library facilitates the inference process by allowing large models to run on hardware with limited memory. ExLlamaV2 provides a foundational platform for experiments involving FrankenModels, including layer duplication, making it a key resource for our project. More details can be found in its official repository.

SlicedLLaMA

A complementary implementation that extends the capabilities of ExLlamaV2 by enabling finer-grained slicing and manipulation of model layers during inference. This feature enhances the precision of layer duplication experiments by allowing selected parts of a model's architecture to be modified or reused more efficiently. Further details can be found in this GitHub repository and are discussed in this Reddit discussion. This functionality is integral to optimizing inference and providing greater flexibility for our experiments.

Compute Resources

GPU SLURM clusters at Tel Aviv University will provide sufficient computational power for layer duplication and evaluation tasks.

Datasets

Based on the chosen tasks and benchmarks, we will utilize publicly available datasets that align with our evaluation goals. These datasets will be carefully selected to ensure compatibility with the tasks and benchmarks.

Related Work

Papers:

- SelfExtend: Extending Context Windows Efficiently: This paper highlights efficient methods to extend the context window of language models. Conceptually, it aligns with our work by exploring lightweight architectural modifications to enhance performance, potentially informing our approach to scaling.
- SOLAR 10.7B: A Large Language Model Employing Depth Up-Scaling (DUS): This work discusses using Depth Up-Scaling (DUS) to enhance model performance through layer expansion and pretraining. While DUS involves retraining, we drew inspiration from its exploration of scaling techniques to inform our layer duplication strategy.

Open-Source Projects and relevant Community discussions:

- Repeat Layers in ExLlamaV2: A practical implementation of layer duplication in LLaMA models.
- FrankenModels Discussions on Reddit: Community-driven experiments and challenges related to layer repetition.
- Frankenweights: A Stable Diffusion Approach: Explores techniques for model modification, providing conceptual insights that may inform similar approaches in FrankenModels.