Review LLAMA PRO: Progressive LLaMA with Block Expansion

Chengyue Wu, Yukang Gan, Yixiao Ge,
Zeyu Lu,Jiahao Wang, Ye Feng, Ping Luo, Ying Shan

January 17, 2024

Abstract

This is a paper review for LLaMA Pro progressive model as an effective post-pretraining model called *block expansion*, pointing out to the well known obstacle of catastrophic forgetting. The model freezes the blocks inherited from the main model while fine tuning the expended blocks. The goal is to have a LLM with high performance on both general and domain specific tasks.

1 Summary and contributions

Authors of the paper are presenting the strengths and weaknesses of LLMs from the start, and they clearly explain their aim with the paper.

Catastrophic forgetting occurs in ML when a model forgets older information he was trained on, as it learns new information. v

As LLMs are instances of foundation models, the new approach was created on a LLaMA2 model that was trained on massive unlabeled corpus, resulting in a model with strong general capacity. Afterwards, the authors fine tuned the expanded identity blocks using a corpus while freezing the blocks inherited from the base model.

Authors introduce the notion of **block expansion**, and they extend the pre trained LLaMA2 by 8 more blocks, yielding the LLaMA PRO. The new model has 8.3 parameters, and it shows improvement in coding, reasoning and programming.

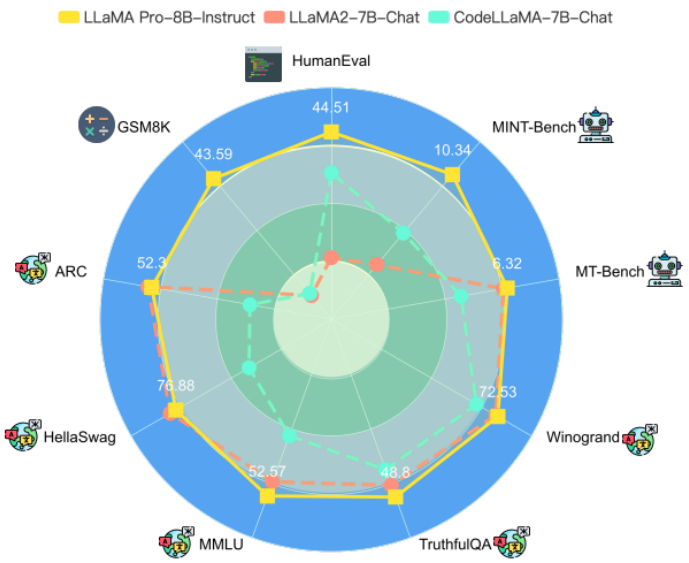

Figure 1: LLAMA PRO - INSTRUCT delivers state-of-the-art performance across a wide variety of tasks, ranging from general language to specific domains, superior to existing models from the LLaMA series.

Figure 1: [1]

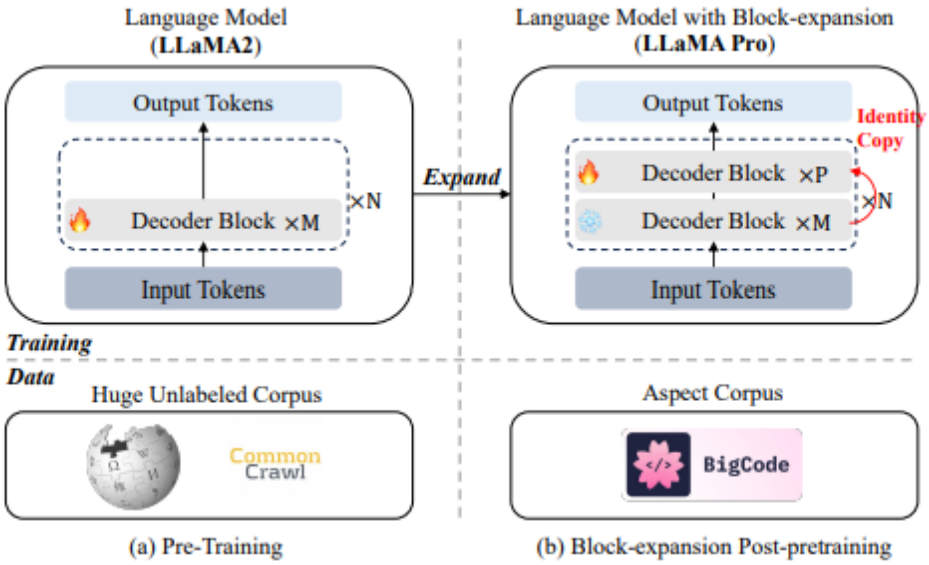

Figure 2: Expansion of LLaMA2 to LLaMA Pro [1]

After being trained for more than a week with expanded blocks on 80B tokens using open source math, coding an reasoning data sets, LLaMA Pro reaches SOTA performance across a broad range of general, code, math and human feedback.

Their study focus on LLMS with performance on great capacity in general and specific tasks, showing superiority over other models from the LLaMA family on both benchmarks and practical applications.

Authors also explained the post-pretraining. Usually, a LM involve an initial general-domain pretraining and then a domain specific training. In their work, they proposed an adaption that combines two already existing strategies, as follows: they combined the continuous training with targeted general capability, without sacrificing the overall performance, as usually the LLMs tend to "forget" the data trained in the beginning.

1.0.1 Method Used

On the left it would be the original layer, and then on the right it will be the newly resulted one. They are making sure that the weight matrix should be 0 so they can give a output to be sure the entire result is a copy, to ensure all is a residual connection of the initialization.

On the training they will allow all weights on the (b) to be updated.

1.1 Conclusions

- Authors proposed a novel method for LLMs, named block expansion, and they introduce the LLaMA PRO and LLaMA Pro PRO - Instruct, that integrates NLP and programming skills as well as excelling in general tasks, programming and mathematics, and last but not least, they demonstrated superiority to other LLMs.

2 Strenghts

- In introduction the authors explain clearly the aim of the paper
- The authors discuss about existing work, the strength and the limitations.
- The authors propose a novel solution for the limitations.

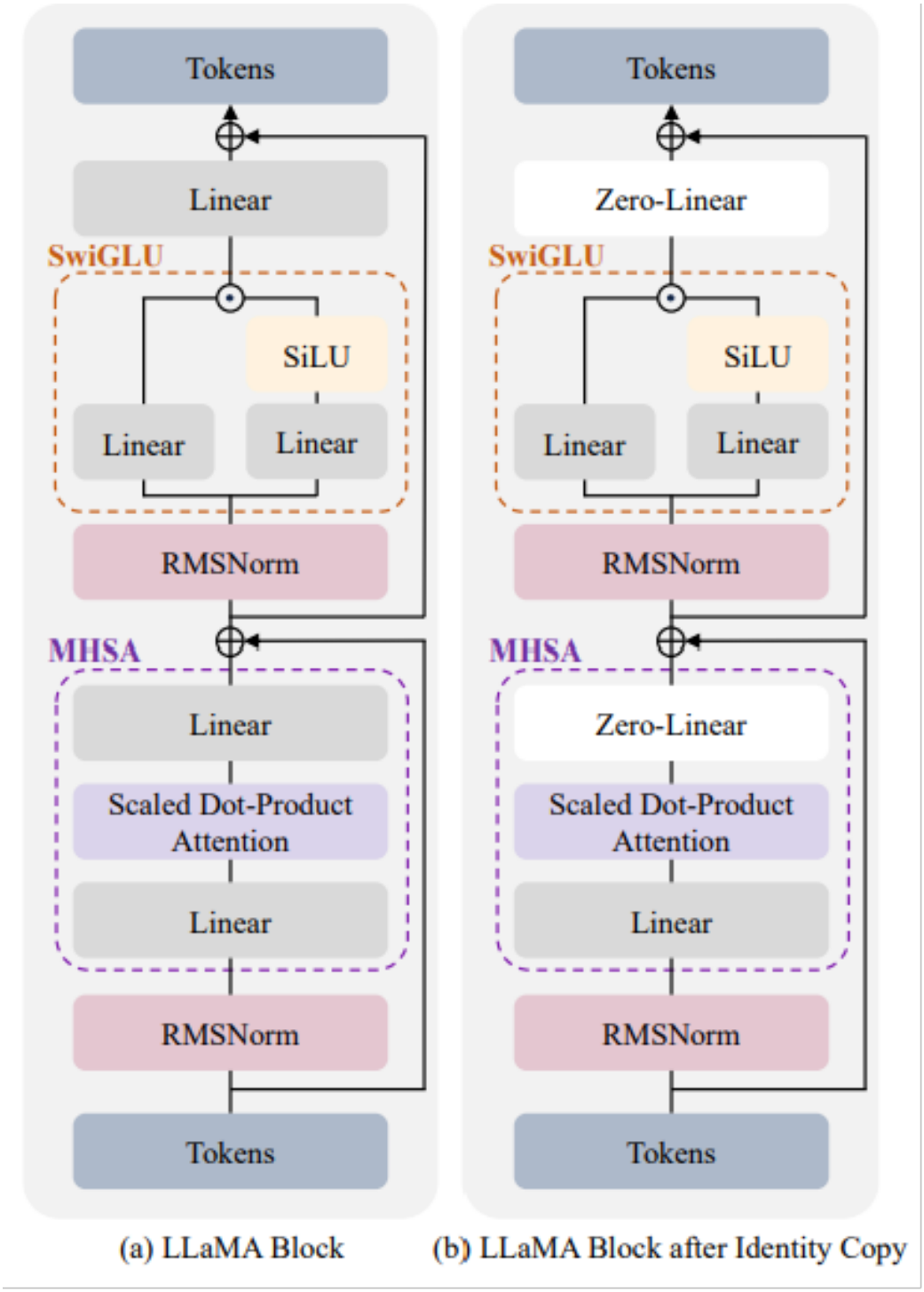

Figure 3: Overview of LLaMA Block [1]

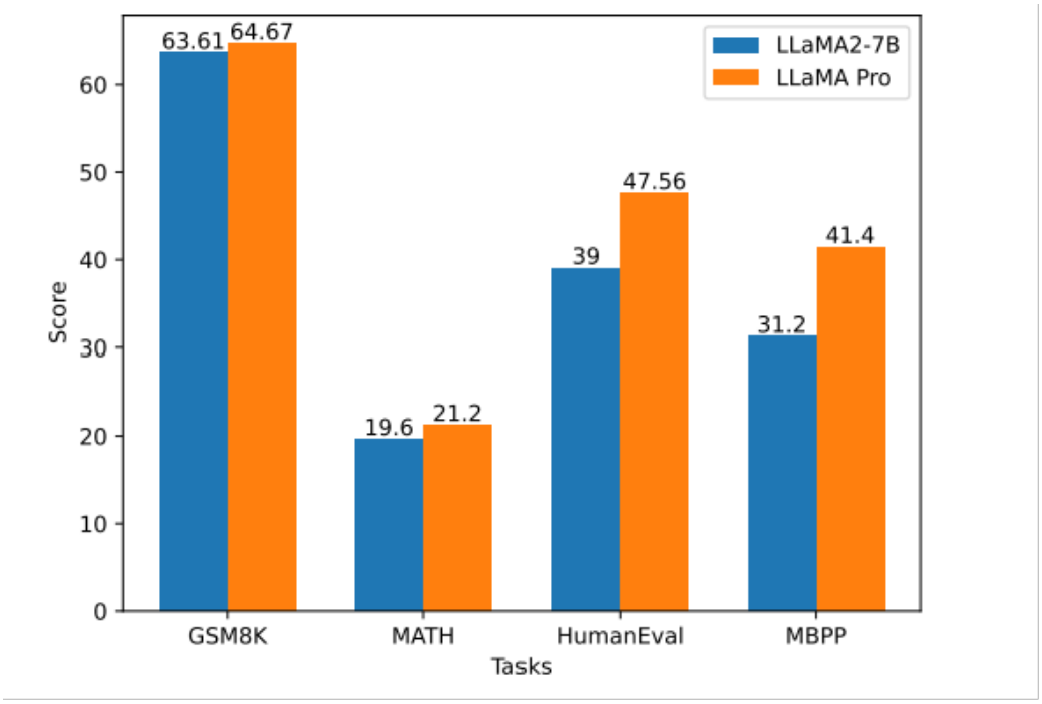

Figure 4: Results [1]

- They are presenting related work, explain how they get the results, and propose for testing the LLaMA PRO, highlighting the applicability in various fields.
- It is explicitly being mentioned the development they did and how it can be used in future research and applications in the area.
- Researchers analyze the impact of the position where the identity blocks are added, either at the bottom or the top of the model, compared to adding them interleaved
- The results are at least as good as the already existing ones

3 Weaknesses

- We learn quite late about the languages between which translation is made. (3.2)
- In 3.3 we find out that they used 7 language pairs from MTTT TED, before we know what are the languages.
- Authors clearly present the packages being used to render text for each language.
- Authors acknowledge that, while effective, it is not clear that sliding window segmentation is optima

4 Correctness

- The datasets chosen for the NLP paper were sourced from reputable repositories

5 Clarity

- Aim of the paper is clear from the beginning
- Clarity on how they trained the model
- Git publicly available

- It is clear how they conducted the study, they included annexes with parameter tuning information, experiments

6 Additional Feedback

As an additional feedback, I would like to mention that it impresses me how authors included detailed explanations of how they conducted the study, steps on how to test the outcomesur and a separate GitHub thread on their work.

7 Overall Score: 4

Grade Criterion:

Comprehensibility (incomprehensible = 1, easy to understand = 3) 5
Information density (very low = 1, very high = 5) 3
Content (many content errors = 1, consistent content = 5) 5
Accuracy (very inaccurate = 1, very accurate = 5) 5
Key message (not recognizable = 1, very clear = 5) 5
Abstract (misses the point = 1, very appropriate = 5) 4
Summary (misses the point = 1, very appropriate = 5) 5
References (not available = 1, very appropriate = 5) 5
Drawings, tables (not available = 1, clear and concise = 5) 5
Length of the paper (far off = 1, within specification = 5) 4

References

[1] Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ping Luo, and Ying Shan. Llama pro: Progressive llama with block expansion, 2024.