A NEW PHYSICAL ARCHITECTURE THAT GENERAL INTELLIGENCE NEEDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformer-type models face high computational overhead due to quadratic complexity when handling ultra-long contexts. Existing sparse attention methods lose global dependency and perform poorly on long-text inference tasks. This paper proposes a new physical architecture that supports continuous learning, and does not require large computing power to obtain super long attention, compressing computational load while obtaining global information, reduces inference complexity from quadratic growth with sequence length to linear growth with sequence length. Under the current Transformer architecture, the supported context length is strictly limited by computing power. The current common order of magnitude is 1 million. The new physical architecture in this paper increases effective context length by four orders of magnitude under current computing and storage conditions. This architecture may help realize artificial general intelligence.

1 INTRODUCTION

The main form of Artificial Intelligence technology currently is Transformer-based Vaswani et al. (2023) large language models (LLMs), such models currently can't support continuous learning well, and require large computing power to obtain super long attention.

Existing continual learning approaches for large language models can be grouped into three categories. The first category is full-parameter retraining, or full fine-tuning. The second consists of parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA)Hu et al. (2021). The third category is Retrieval-Augmented Generation (RAG). LoRA is a mainstream technique, when the amount of newly acquired knowledge is large, there are correlations among the knowledge, and cross-knowledge fusion reasoning is required, multiple adapters of current LoRA cannot naturally integrate the knowledge and cannot achieve unified representation of the knowledge.

Popular long-context optimization techniques, including sliding-window attention, sparse attention, and KV (Key-Value) cache quantization compression, only cut computation and memory costs at inference time and are purely acceleration-oriented optimizations. Sliding-window attention discards remote KV information beyond the window boundary, while quantization compression introduces representation noise and degrades semantic embedding quality.

In this work we propose a new physical architecture, it uses a switcher mechanism to innovates the method of continuous learning, our architecture can be simply referred to as a Switcher,It makes technical improvements to the existing low-rank adapter fine-tuning (such as LoRA parameter-efficient fine-tuning schemes), but its implementation is not limited by LoRA, and has addressed the shortcomings of the existing technologies.It adopts the continuous addition of the trained text data and the user input text as parameters of the large model. The added parameters can coexist harmoniously with the original parameters. Under the same computing power and storage conditions, by adding the context as parameters of the large model, the effective context length can be increased by four orders of magnitude, achieving a relationship where the inference complexity is approximately linearly proportional to the effective context memory length, thus addressing the shortcomings of current models, this architecture may help to achieve artificial general intelligence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 BACKGROUND

Current Low-Rank Adaptation (LoRA) is a mainstream technique of continuous learning, the standard for parameter-efficient fine-tuning, freezing the pretrained backbone W and training low-rank matrices A and B. However, in continual and multi-task settings, it exhibits fundamental limitations.

Although it freezes the pretrained backbone and trains only low-rank matrices, it does not inherently protect pretrained knowledge. Its updates can cause catastrophic forgetting by overwriting pretrained spectral structure and introducing intruder dimensions that degrade general capabilities. LoRA lacks explicit knowledge boundaries to distinguish known, uncertain, and unknown. Thus, LoRA provides parameter-efficient plasticity but lacks the stability and isolation required for genuinely continual learning systems.

In this work, we make technical improvements to the existing LoRA, continuous learning and super long attention is combined into one, if the new architecture can achieve continuous learning, super long attention automatically is achieved.

The new architecture stems from inspiration: If we can identify the boundaries of the knowledge in large models, that is, be able to effectively distinguish which knowledge is known by the large model and which is unknown or uncertain, then we can utilize this for targeted updates. We can take such steps: Use an adapter to work in conjunction with the original model. The knowledge that the original large model already knows will be handled by the original large model. This avoids the intervention of the adapter, which would cause interference with the old knowledge of the original large model. The knowledge that the original large model is unknown or uncertain will be processed by the adapter, allowing new knowledge and old knowledge to coexist harmoniously.

3 MODEL ARCHITECTURE

The key point is to effectively distinguish the knowledge boundaries of the model. The method we adopt is to utilize word-element connections. The associations between word-elements are discrete and explicit. We use this relationship structure to determine the knowledge boundaries. Word-element co-occurrence is the most obvious type of word-element connection. For example, ”Titanium density 4.506 g/cm³” appears together, representing the density information of Titanium.

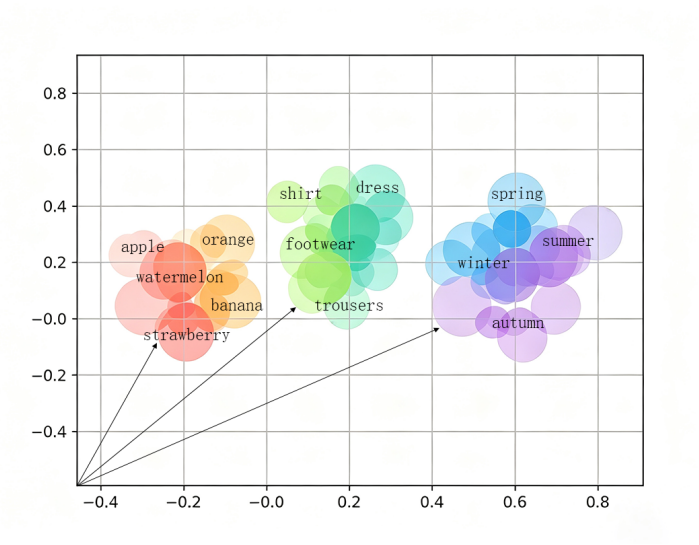

Figure 1: Intuitive display of the spatial distribution of the vectors

We use the co-occurrence of word-element to find connection, co-occurrence serve as a signal of association, or connection: if two words often appear simultaneously in the same context, it is

considered that they are related. By using the dimensionality reduction visualization method, high-dimensional word vectors can be projected onto a two-dimensional plane, enabling an intuitive display of the spatial distribution of the vectors. Words with similar semantics will form clusters on the plane based on their semantic similarity.

Titanium density represents a piece of meta-knowledge, if we can distinguish the knowledge boundaries ideally, it is reasonable to identify each piece of meta-knowledge, for example, the original model does not include Titanium density, so when handling request related to Titanium density, the original model will fail. If the updated model includes Titanium density, such request should be handled by updated model. When each piece of meta-knowledge is ideally distinguished, request tasks are handled by the right parameter of the model, a seamless continuous learning can be achieved.

3.1 MAP OF KNOWLEDGE BOUNDARIES

Word-element connections are utilized in this work, it can further develop into the Meta-knowledge network, so we can check whether an inference is supported by evidence, this is useful for artificial general intelligence. This information forms a map, the map simply the road to continuous learning, it has boundaries for known knowledge and unknown knowledge, as shown in Figure 2.

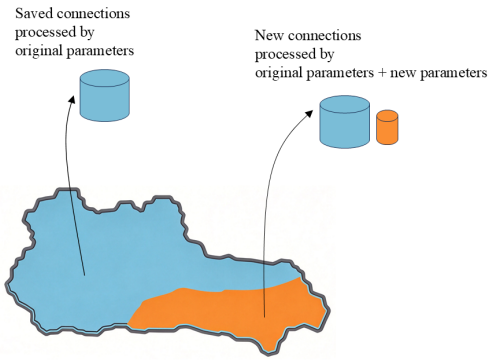

Figure 2: Map of knowledge boundaries

We ask us a question: if we hope to reach continuous learning, then it is correct that the first thing is to distinguish the Map of known knowledge boundaries ? it is reasonable that this is correct.

3.2 FEATURE ANALYSIS

DeepSeek-V3 adopts a Byte-level Byte-Pair Encoding (BBPE) tokenizer with a vocabulary size of 129,280, each word has one connection to each other, the total number of connections is 16,713,318,400, the memory shall allocate room to store connections and connection belongings.

If super long context length is transferred as model parameter, we make an analysis, with 10,240 Megabyte of raw context length, stored as the same size as model parameter is reasonable, such effective context length is affordable under current memory situation.

Super long effective context length is quite helpful for LLM-assisted programming, a large system or software has more than 100 million lines of code.

3.3 FRAMEWORK DIAGRAM

The key elements of the architecture are Knowledge boundaries determination and Switch.

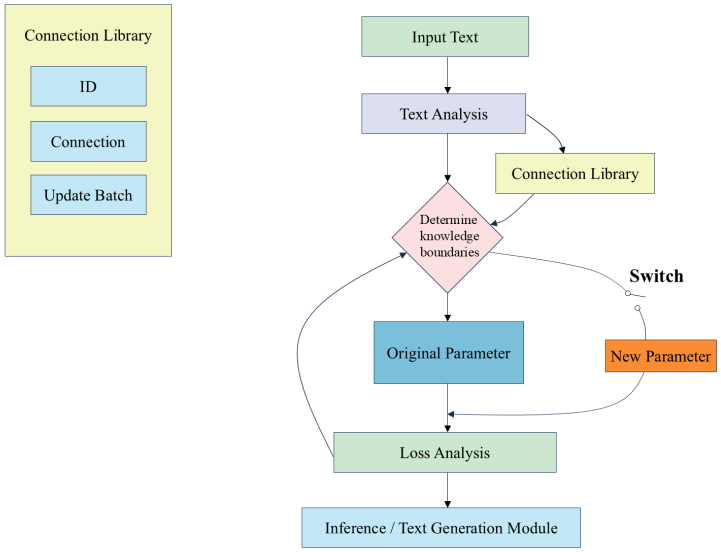

Figure 3: Framework diagram of this new architecture

"Concepts are purely differential and defined not by their positive content but negatively by their relations with the other terms of the system."

– Saussure, 1916

Starting from the fundamental principles of physics, if we are going to design the architecture of brain, which is the correct way ?

The way with the lowest energy should be: each piece of Word-element connection, if ever saved in the system in the historical training, it will turn the switch of new parameters to close, the inference process is just handled by original parameters, no interference is introduced to original parameters.

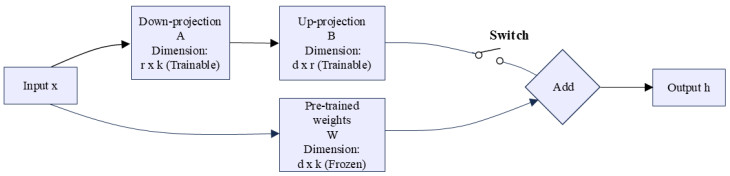

Figure 4: Flow diagram of update of weight matrix parameters

4 Conclusion

This paper proposes a new architecture that supports continuous learning, seamless continuous learning is quite challenging, currently there is no thoughtful solution to handle this challenge well. The work brings effective solution, and it is helpful to resolve the "stability-plasticity dilemma" - the model needs to be able to quickly absorb new knowledge while also permanently retaining all the previously acquired old knowledge without any performance degradation.

The addition of new knowledge only involves adding new parameters, without changing the parameters corresponding to the old knowledge, achieving a harmonious coexistence of new and old parameters, and suppressing catastrophic forgetting and representation drift from the root of the parameters.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not raise any ethical concerns.

REFERENCES

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.