Literature Review: Mass Editing Memory in a Transformer

Andrea Miele | 302925 | andrea.miele@epfl.ch
REAL

1 Summary

This paper introduces MEMIT, a new method for updating factual knowledge in Large Language Models (LLMs) by directly manipulating the transformer's weights (specially in MLPs). The ability to perform multiple updates (10^4) represents an important advancement in the field, addressing the challenge of keeping LLMs accurate and relevant without full retraining or fine-tuning. The problem is reduced by the authors to triplets of (subject, relation, object), or (s_i, r_i, o_i). They evaluate this by giving the model a prompt and analyzing the output. The object is then changed, while the subject and relation remain the same.

The authors demonstrate that MEMIT can effectively handle the simultaneous update of thousands of memories, a scale previously not achievable with previous methods. The approach employs a new technique of making edits across multiple "critical" MLP layers within the transformer, ensuring that each layer contributes to the overall update. This method not only improves the efficiency of memory updates but also maintains the model's performance and stability, which a sign of robustness for the method.

To validate the effectiveness of MEMIT, the authors conducted extensive experiments using datasets (for instance the CounterFact dataset) that test the model's ability to incorporate true and counterfactual information. The results show that MEMIT significantly outperforms traditional methods such as fine-tuning and other recent memory editing techniques, especially in handling large-scale edits. The method's performance is assessed on various metrics, including efficacy, generalization, specificity, fluency, and consistency, demonstrating its superior capability to make numerous edits without degrading the model's performance.

Despite its strengths, MEMIT also presents several challenges. The complexity of its implementation may limit its adaptability to different models or architectures. Additionally, the focus on updating only factual knowledge using subject-relation-object triplets restricts its applicability to other types of knowledge updates, such as procedural or conceptual information. Furthermore, the computational resources required for implementing MEMIT, though less than those needed for retraining, are still substantial (with only thousands of edits possible), potentially limiting its use to those with access to significant computing power/ or within a certain time limit.

Overall, MEMIT represents a solid contribution in the knowledge updating of LLMs. By enabling efficient and scalable edits of model memories, it opens new possibilities for the practical use of LLMs or their interpretability. However, it will require addressing its current limitations and exploring its applicability to broader knowledge types and different model architectures.

2 Strengths

This paper offers several significant contributions to the field of NLP, specially in the field of the knowledge editing of LLMs.

The paper is generally well-written and easy to follow. Authors made efforts to make their work reproducible with details provided in the paper, appendixes, and well-documented source code (for instance well presented README).

This paper presents a new problem proposal called "mass-editing" that is well-motivated by prior literature on the topic of memory editing in transformers.

Also MEMIT is the only mass-memory editing method that works on the proposed new setting. Furthermore, the methodology employed by MEMIT, which involves spreading edits across multiple MLP layers, represents a new approach to memory editing in neural networks and specially in transformers. This is proven by experimental re-

sults which includes a comparison to a fine-tuning baseline. MEMIT demonstrates excellent performance across multiple evaluation metrics, including efficacy, paraphrase, specificity, and fluency, where the technique outperforms baseline methods. The paper provides extensive empirical evidence to support the efficacy of MEMIT.

This works seems to be also a step in the right direction of model interpretability as you understand where are stored the factual knowledge in a transformer model.

The author propose a detailed and clear section to explain the editing algorithm which seems technically valid.

Another key points it that the authors note the limitations of their method (quite slow, doesn't cover temporal reasoning or cover spatial reasoning).

Another important strengths of MEMIT is its scalability. It takes the work from ROME (Meng et al., 2023) that makes an edit on a single fact and scale it to thousands of facts. Therefore, it brings a solid improvement over existing techniques.

MEMIT maintains a high level of specificity and generalization in the experiments. This shows the updates' precision and that the LLM remains robust and reliable after the edits have been made.

In summary, the strengths of MEMIT includes its scalability, precision, innovative methodology, strong empirical support, and commitment to open source which present it as a strong contribution in the field of NLP and ML. This not only shows the method's current capabilities but also its potential to influence future developments in model knowledge updating or interpretability.

3 Weaknesses

While MEMIT brings several solid contributions, it also presents certain limitations and weaknesses that could impact its broader applicability and effectiveness. One of them is that methods are implemented only for the GPT models family (GPT-J and GPT-NeoX), not other transformers models that are different, or open-source models like Llama for instance.

The method focuses mostly on directional (subject-relation-object or (s_i, r_i, o_i)) relations, which could potentially limit its applicability to different or more complex knowledge representations. I wonder in what sense this could be extended or not to other types of knowledge representations.

Another thing is that even if the weight updates are performed in batch, the calculation of key and residuals of the layers is still done iteratively.

Also, even if the methodology and algorithms are detailed and seem correct, they are quite complex, which could be a challenge for practical and broader use. Also, if you compare it to their previous paper, ROME (Meng et al., 2023), this paper almost feels like engineering to adapt ROME (Meng et al., 2023) to multiple facts simultaneously.

To conclude, while MEMIT presents a innovative approach to memory editing in LLMs, its complexity, limited scope of application, high computational demands, and the need for further validation in practical settings represent significant challenges.

References

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt.