Group 5 Status Report: Reproduce paper "Memsizer" and extend it

Yaqi Hu yaqihu@usc.edu
Gonglin Chen gonglinc@usc.edu
Libang Xia libangxi@usc.edu
Zhaojin Yin zhaojiny@usc.edu
Luoyuan Zhang luoyuanz@usc.edu

1 Tasks that have been performed by Us

1.1 Task We Currently Doing

For this project, by our proposal, we will deal with three tasks, they are:

1. Reproducing the paper result, the paper contains 6 datasets, we will reproduce three of them. We will compare the result between Memsizer & Transformer
   The alternative plan is : if time is limited, we will only reproduce one of them.

2. Put the model into a new situation to extend its work (A new dataset: OpenOrca)

3. We already have a Q & A system implemented by transformer (Written by our self). We will change the transformer part into Memsizer to see its performance.
   (We hope Memsizer will give a great improvement)

1.2 Current Work result

Based on the three tasks we have given above, we are parallel doing these three tasks.

1.2.1 Reproducing

We have found the three dataset we need to use, and downloaded them. extract, split train/valid/test, and ready to use. The three dataset are:

1. WMT 16 En-De

2. XSUM

3. WikiText-103

Good news is that, we have already get the transformer result on 2 datasets, they are WMT 16 En-De and XSUM. We will get the last dataset result on WikiText-103.
However, We faced the challenge that the Memsizer, author shared code is hard to run. The author provided the .sh file to directly run the whole program, but one of the folder is missing, more detail will be introduced in Risk and Challenges section. Even facing the challenge, we are trying to rewrite the parser of the model and meet all the requirement of the hyper-parameter, more detail will be provided at Our Plan to Risks Mitigate and Challenge address.

1.2.2 Paper Extension

We are successfully download the OpenOrca dataset, split it into train/valid/test, and ready to use. As the reproduce of the Memsizer faced challenge, we are now only getting the transformer result of OpenOrca dataset. We will continuously try to solve the challenge, and get the result.

1.2.3 Replace Q& A Transformer Part With Memsizer

For the replacing module task, we have successfully implement the Q & A system with transformer, and get it result. (Note: the dataset is pretty easy, like given a number, we tell you tenth digit, unit digit of it). We just want to explore how much will the Memsizer improve compared with transformer. After challenge is solved, we will get the final compare result.

2 Risks and Challenges

2.1 Code Execution Difficulties

Our initial attempt into the reproduction project faced a key challenge—the inability to run the code shared by the authors in the provided repository. This hurdle not only stalls our progress but also consumes a significant chunk of our time, which is a scarce resource as the deadline approaches. To be more specific, the author did not provide a complete, clean instruction on the environment executing the code so that it took us a long time on figuring out the environment. At the process of executing the command file, the error shows that

one file, dict.txt and a folder, /revatt are missing. The missing files prevents the code from running normally, so that we are unable to reproduce the results in which the author shows in the paper. We have also asked regarding this problem in the "ISSUES" part in github and sending emails to the author. Non of them have been replied yet.

2.2 Codebase Understanding

A profound understanding of the code structures for both the linearized and traditional transformer is pivotal for the successful substitution of one with the other. Therefore it is one of our priorities to fully understand the underling codes and network logics. However, the intricacies and complexities embedded in the repositories are time-intensive to understand, especially in the absence of comprehensive documentation. Even after we spend time on reading the codes and try to understand them, some details and the training codes are still hard to comprehend which we have to discuss them together.

2.3 Resource Constraints

Resource constraints, be it computational resources or team expertise, could pose a significant risk to the timely completion of our project. Ensuring that adequate resources are allocated and contingency plans are in place is essential to navigate through the challenges that lie ahead. We are currently utilizing the Colab computing resources to perform all the computations. However, the computation units are not free so we have to make the best use of them.

3 Our Plan to Risks Mitigate and Challenge address

We have encountered significant challenges in our attempt reproduce the results of the paper. A critical obstacle is the absence of the essential folder '/Revatt/' and its files. Without these components, reproducing Memsizer becomes a formidable task.

3.1 Running the Code

Aiming to navigate through this challenge, we have strategized a plan articulated as follows:

1. We have contacted original authors reaching out to the original authors, seeking guidance and the pivotal files required for reproduction. However, this effort is yet to bear fruit as responses are pending.

2. Parallelly, our plan includes some more research on this topic. We are hoping that we could find a similar implementation or a way to work around those files. We hope that missing files are common files.

3. If all steps above are not reachable, we also plan to address it by exploring the possibility of crafting our own version of MemSizer. Utilizing the available source code as a foundation, we plan to implement the attention mechanism and hoping this can solve the problem introduced by those missing files.

3.2 Experiments

Once we could successfully reproduce the results of MemSizer, our next step is to train the MemSizer with our previous mentioned datasets. This aims at a comprehensive evaluation of the MemSizer across different domains.
We wants to compares the results and inferencing time between MemSizer and Vanilla transformer. By doing this, we could be able to unveil the improvement of MemSizer.
This experiment seeks to get an idea of guiding potential enhancements in the MemSizer model, therefore, we could explore the possibility of applying MemSizer to practical applications.

4 individual Contributions

Status Report Writing:

1. Tasks that have been performed: Yaqi Hu

2. Risk and Challenge: Zhaojin Yin

3. Our plan to Risks Mitigate and Challenge address: Gonglin Chen, Libang Xia

Project Processing:

1. Reproduce:
   WMT 16 En_De: Libang Xia
   XSUM: Luoyuan Zhang
   WikiText-103: Yaqi Hu

2. Extend Paper: Yaqi Hu

3. Replace Module Work:
   Gonglin Chen, Zhaoyin Jin