# Part 1

The paper “Mass-Editing Memory in a Transformer.” and the overview of it were interesting to read. I had a few thoughts as I read them:

* Being able to edit memories in Transformers is a critical need.
* Prior efforts can edit tens to hundreds of memories. MEMIT extends this to thousands. But ultimately, we will need an approach that is unconstrained.
* A key observation is that the Transformer stores memories as Key Value pairs.
* The MEMIT approach is also transparent.

The paper uses the example of “Michael Jordan plays basketball.” being changed to “Michael Jordan now plays baseball.”. In testing they used large sets of facts. They also mention some limits of their approach such as Tim Cook is CEO of Apple needs to be edited separately from The CEO of Apple is Tim Cook.

# Part 2

I opened the file <https://github.com/vijaygwu/SEAS8525/blob/main/Class_4_Transformer_Visualizations.ipynb> in Collab and ran each section. Below are my thoughts:

It uses BERT which is an early implementation of Transformers. It uses a pretrained BERT model, so we are not performing training. I see it tokenizing the input. I do not see positional encoding. I did a search and did not find the term. So, either the data was not positionally encoded, or it is done as a part of one of the other steps.

The visualizations give an interesting view of the attention based on heads, model, and neuron. There is so much information, it is hard to make sense of it all. Two quick observations are that

1. There seems to be a lot of attention paid to the [CLS] and [SEP] tags
2. The lower layers generally pay attention to a lot of tokens where the higher layers seem to be more focused.