Factual recall in attention vs FF layers

Summary of findings

- Without a sufficiently large feed-forward layer, **attention layers can in fact store and retrieve factual knowledge**.
- From a sufficiently large feed-forward hidden dimension, essentially dim_feedforward = 4 * d_model, **attention layers store significantly fewer facts than the feed-forward layer**.

Background

Large language models have a large corpora of **factual knowledge**. We want to develop methods to tell what a language model knows about the world.

Existing approaches, such as Geva et al (2022), and ROME by Meng et al (2022) find some of these, mostly in the FF layers of the models.

In this work, I intended to check if factual knowledge was present in the attention layers as well, by conducting experiments on a small transformers trained on a factual retrieval task.

Training Task

In the experiments, the task I used is a pure factual retrieval task. With n = 20, pick a random n x n matrix, *M* with entries chosen uniformly from 0..n

To generate the data, choose i, j from 0..n and model the sequence [n+1, i, j, M[i, j]].

The model then learns to predict M[i, j] from i, j. Which, with n=20, consists of 400 facts.

note: n+1 is there as a beginning-of-sentence token, which was left from earlier experiments and might not be necessary

Experiment 1 - hyperparameters

I trained small transformers (layers=1, d_model=8) on the task using the following hyperparameters, and got the following accuracies:

	dim_feedforward = 0	dim_feedforward = 16	dim_feedforward = 32
n_heads = 2	0.46	0.69	**0.90**
n_heads = 4	0.53	0.77	0.87
n_heads = 8	**0.58**	**0.81**	0.86

Experiment 1 - Interpretation

1. Without a feed-forward layer, the model is still able to learn 46%- 58% = ~200 facts, with larger n_heads giving more facts.
2. with dim_feedforward = 2 * d_model, the larger n_heads still has a positive effect on the number of facts that can be stored. So it’s safe to assume the attention still plays a role in fact storage.
3. with dim_feedforward = 4 * d_model, this effect is no longer seen. Suggesting at this point the attention layers have a smaller role in fact storage.

link to experiment 1

Experiment 2 - skipping some layers

here, we directly skip the FF layer, and later both the FF and attention layers (essentially relying only on the embeddings. this is our null hypothesis) and then check for accuracy. here n_heads = 4. See the next slide for explanation of the adapter (this is important!). note that here n=15.

	hidden_size = 0	hidden_size = 16	hidden_size = 32
Accuracy	0.720	0.889	0.996
Accuracy without FF	0.720	0.095	0.115
Accuracy without both	0.121	0.095	0.101
Accuracy with adapter	N/A	0.204	0.185

Experiment 2 - adapter

It’s highly likely that the FF layer we skipped shifted the embeddings, so that even if facts were retrieved in the attention layer, they no longer make sense in embedding space. For this, we freeze the model, and add a residual adapter, with 2 hidden neurons. The goal of this adapter is to translate any information left about the facts back to what the unembed layer expects.

The adapter makes the results improve, which suggests the **attention layer might still have a few factual memories**, but regardless these are **far fewer** than the memories stored by the feed-forward layer.

link to experiment 2