Memorization in Attention-only Transformers

Léo Dana

June 13, 2024

Memorization in Transformers

It is traditionnaly thought that MLP store information in Transformers.

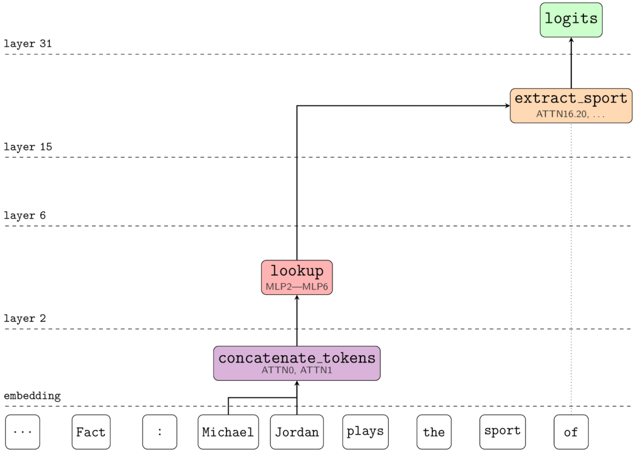


Figure: From *Fact Finding: Attempting to Reverse-Engineer Factual Recall on the Neuron Level*

Memorization in Transformers

We want to complete this view: maybe Attention layers can remember information.

The questions we have begun to answer are:

- Can an Attention-only Transformer memorize ? How well can it do ?
- How much can it memorize ?
- How does it memorize ? What algorithm is implemented ?

Associative memory

In *Birth of a Transformer: A Memory Viewpoint. Bietti et al. 2023*, the Associative memory framework is introduced to Transformers.

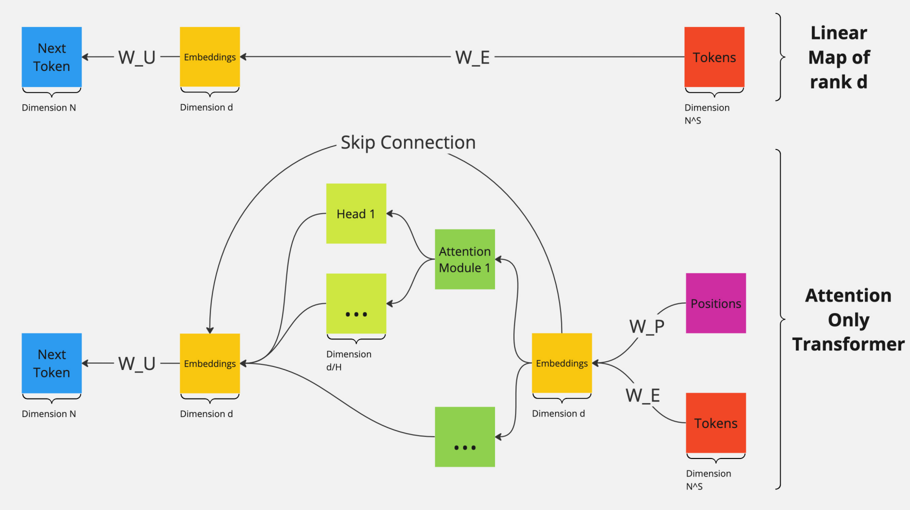
Goal: remember a mapping $g : [N] \to [N]$ using the argmax of $W_U W W_E$, where $W_U, W_E \in \mathbb{R}^{N,d}$ random embedding matrices, and $W \in \mathbb{R}^{d,d}$ is to learn.

Using

$$W = \sum_i w_U(g(i))^T w_e(i)$$

they show the model can remember association, but only d of them.

Attention-only Transformers



Attention-only Transformers

Goal: memorize a distribution $\pi(t_{S+1}|t_S, ..., t_1)$ for all sequences.

Theorem

Let $\varepsilon \geq 0$, N_ε the smallest number of questions whose cumulative probability is greater than $1 - \varepsilon$, and f_{W^*} the optimal linear mapping of rank d.
There exist a transformer T^* with embedding dimension d, $\lceil \frac{N_\varepsilon}{d} \rceil$ total parallel attention module such that whose divergence with π is

$$|d_{KL}(\pi, f_{W^*}) - d_{KL}(\pi, T^*)| \leq \varepsilon \sigma_1(f_{W^*}) C(d, N, k, N_\varepsilon)$$

Optimal Linear Mapping

Now we want to bound the quantity $d_{KL}(\pi, f_{W^*})$ to obtain an upper bound on the best Transformer possible.

$$d_{KL}(\pi, T^*) \leq d_{KL}(\pi, f_{W^*}) + \varepsilon\sigma_1(f_{W^*})C(d, N, k, N_\varepsilon)$$

We can acheive 0 divergence if

$$w_U(t_{S+1})^T w_E(t_{1:S}) = \log(\pi(t_{S+1}|t_{1:S}))$$

so if $L = \log(\pi)$ has rank d.

Optimal Linear Mapping

Now we want to bound the quantity $d_{KL}(\pi, f_{W^*})$ to obtain an upper bound on the best Transformer possible.

$$d_{KL}(\pi, T^*) \leq d_{KL}(\pi, f_{W^*}) + \varepsilon\sigma_1(f_{W^*})C(d, N, k, N_\varepsilon)$$

We can acheive 0 divergence if

$$w_U(t_{S+1})^T w_E(t_{1:S}) = \log(\pi(t_{S+1}|t_{1:S}))$$

so if $L = \log(\pi)$ has rank d.
If $d \geq N - 1$ this is always the case. Otherwise we look at special cases.

Almost rank d

When the distribution to predict has low rank, a Taylor on the divergence reveals two possible optimal mappings:

Almost rank d

When the distribution to predict has low rank, a Taylor on the divergence reveals two possible optimal mappings:

- f_{ls} solution to the least-square problem

$$||W_{ls} - L||_2$$

$$d_{KL}(\pi, f_{ls}) \leq N||\pi||_{+\infty}^2 \sigma_{d+1}^2 + (N||\pi||_{+\infty})^{\frac{3}{2}} \sigma_{d+1}^3$$

Almost rank d

When the distribution to predict has low rank, a Taylor on the divergence reveals two possible optimal mappings:

- f_{ls} solution to the least-square problem

$$||W_{ls} - L||_2$$

$$d_{KL}(\pi, f_{ls}) \leq N||\pi||_{+\infty}^2 \sigma_{d+1}^2 + (N||\pi||_{+\infty})^{\frac{3}{2}} \sigma_{d+1}^3$$

- f_{wls} solution to the weighted least-square problem

$$||I_{\sqrt{\pi}}(W_{wls} - L)||_2$$

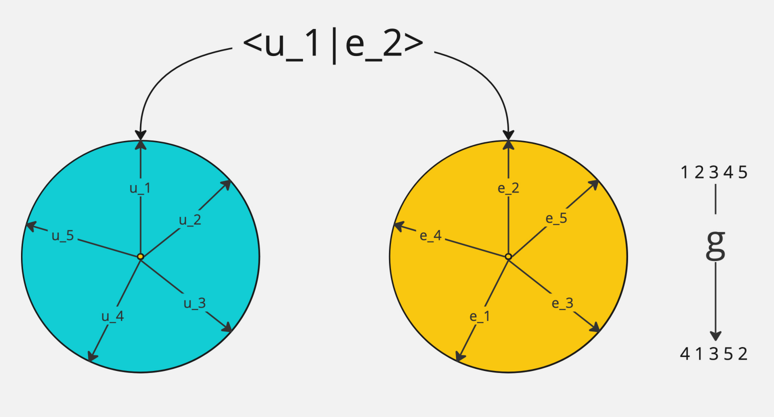
$$d_{KL}(\pi, f_{wls}) \leq ||\pi||_{+\infty} \sigma_{d+1}^2 + \left(\frac{||\pi||_{+\infty}}{||\pi||_{-\infty}} \right)^{\frac{3}{2}} \sigma_{d+1}^3$$

Low Entropy

For low entropy, or look-up table, there exist a function $g : [N^S] \to [N]$ such that $\pi(t|t_{1:S}) \simeq \delta_{t=g(t_{1:S})}$.

Low Entropy

For low entropy, or look-up table, there exist a function $g : [N^S] \to [N]$ such that $\pi(t|t_{1:S}) \simeq \delta_{t=g(t_{1:S})}$.



Bound on the divergence

Theorem

For this choice of f and $C = \sqrt{\frac{32\log(N+1)}{d}}$*, we have the bound*

$$d_{KL}(\pi, f) \leq \mathbb{E}_{t_{1:S}}\left[(1-\pi(g(t_{1:S})|t_{1:S}))\log\left(\frac{N-1}{e^{-H(\pi_{t_{1:S}})}-1}\right)\right]\left(\frac{1+4C}{1-2C}\right)$$

Conclusion

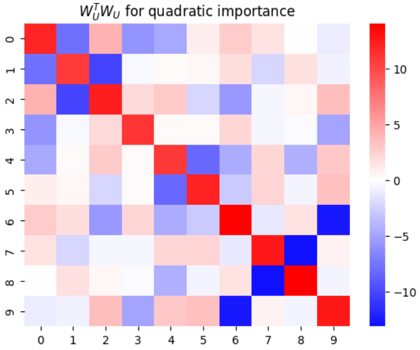
We showed that

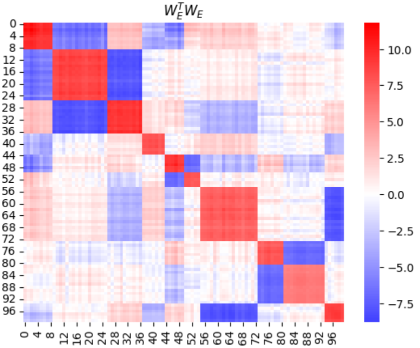
- the attention mechanism in Transformers can memorize,
- we can quantify the error in some best cases,

Limitations:

- We do not know how memorization occurs in the sparce regime $dP < N_\varepsilon$,
- We cannot identify memorization behavior in a real Transformer's attention layer,

Experimental validation





No memorization limit in term of quantity, but in term of quality !