Birth of a Transformer: A Memory Viewpoint

Damien Bouet – Gabrielle Caillaud

CentraleSupélec

Context and Objectives

- Study decoder-only auto-regressive architecture (typical of GPT-like models [3])
- Analyse how global and contextual knowledge is learnt
- Demonstrate how an *associative memory* is created within the weight matrices
- Give some theoretical results on the gradients and the learning

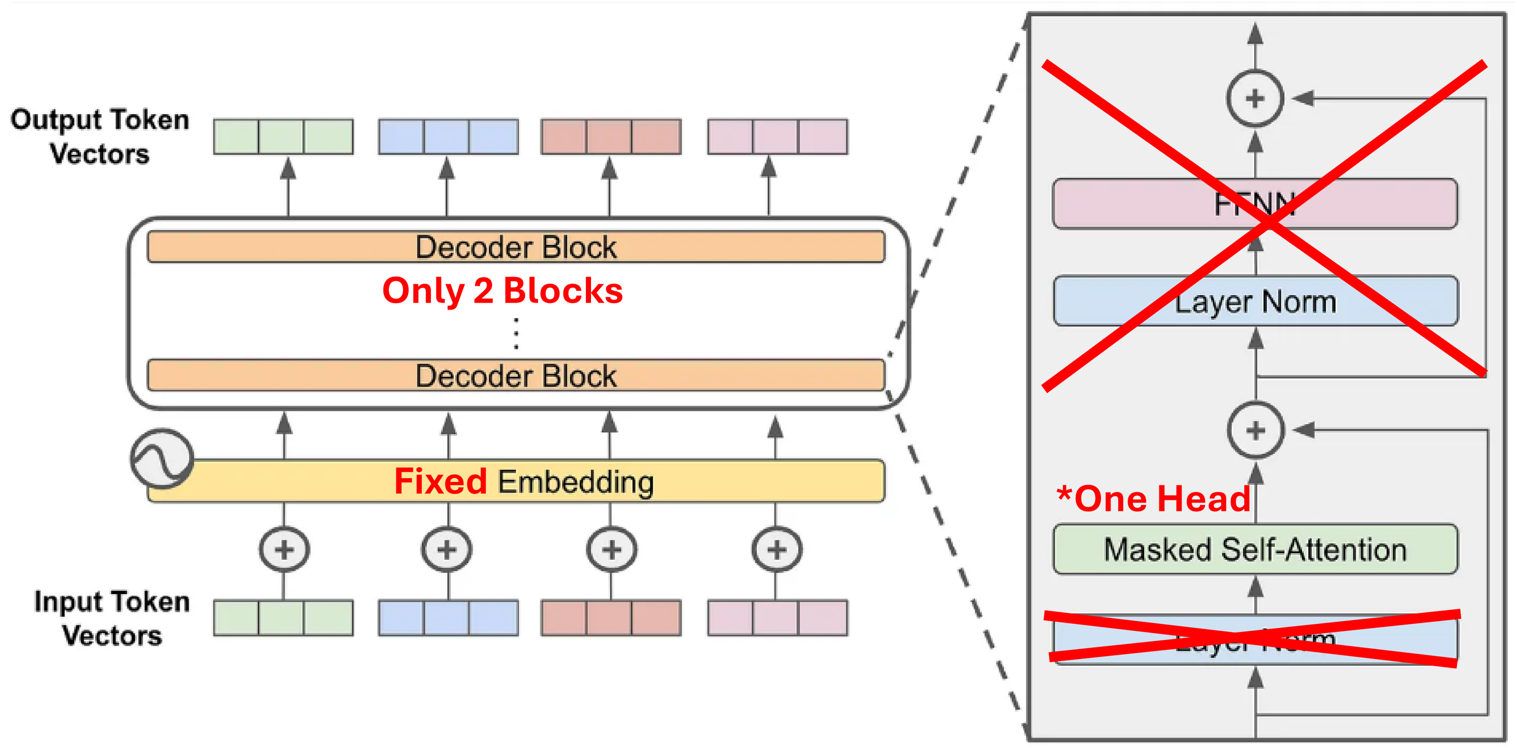
Associative Memory

- Simplified Attention block : $W_{Key},\ W_{Query}, W_{Value},\ W_{Output}$
- It computes $x_t' = W_O\, W_V\, \sigma(x_{1:t}^T W x_t)$ avec $W = W_K^T W_Q$ (we use $W_Q = Id$)
- If key vector u and query vector v are orthogonal (almost the case in high dimension):

  Then, $W = \sum_{i,j} \alpha_{i,j} v_j u_i^T$ and $v_j^T W u_i = \alpha_{i,j}$ which is the relevance of the (I, j) pair.

Simplified Transformer Setup

- We simplify the usual transformer architecture [2] to be able to have theoretical results



Theoretical Lemma

1. With the loss (CE): $L(W) = \mathbb{E}_{(z,y)\sim p}[\ell(y, W_U W w_E(z))]$

The learnt matrix is: $\hat{W} = \alpha_0 W_0 + \sum \alpha_{ij} w_U(j) w_E(i)^\top$

- Combination of the outer-products = associative memory

2. This associative memory can filter noise from the input

- It can result in perfect association accuracy, even with noise (ex: positional embedding or residual connection)

Discussion and Limitations

- Simplified transformer model don't match real architectures
- Only 2 layers here, GPT-2 has 12: potentially more complex behaviours that the associative memory described with 2 layers
- Real-life languages cannot be modeled by a bigram dataset
- But the approach can still help to understand transformers

Simplified Dataset Setup

Synthetic bigram dataset :

- Data sampled from a base distribution **(general knowledge)**
- At each sequence, some predefined **trigger tokens** are **always followed** by some defined **output tokens** (sequence specific pattern, simulate local/prompt knowledge)

Example:

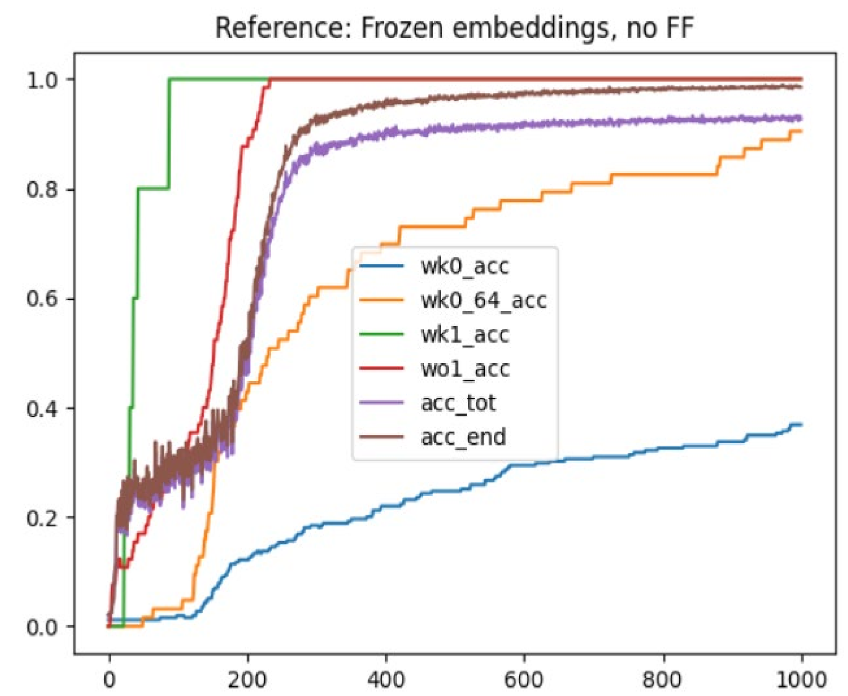
- We sample random letters based on English frequency rate **(general knowledge)**
- but every "a" is always followed by "p" **(specific knowledge)**

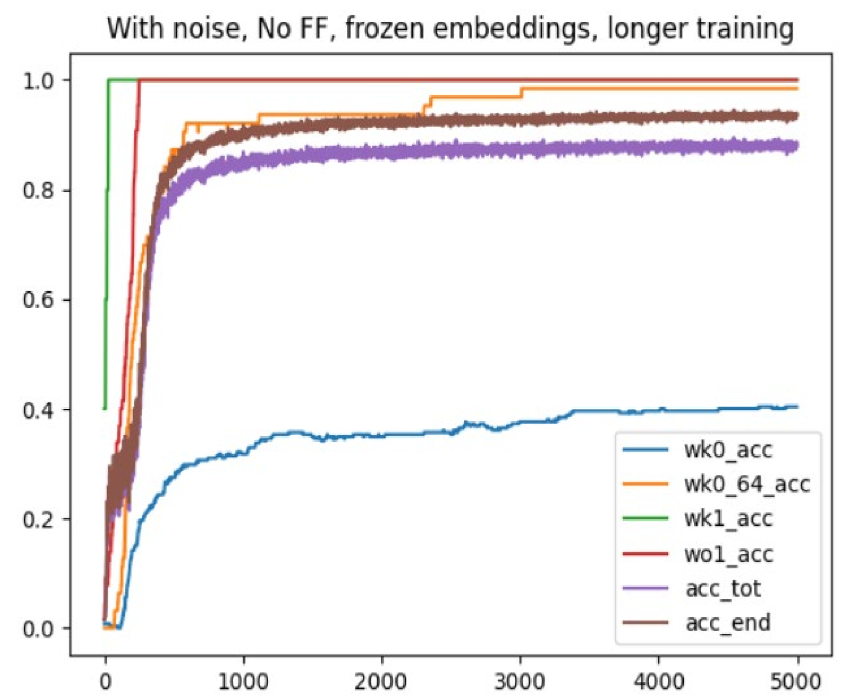
Experiments and Results

- Accuracy evaluated using:

$$attention\_scores = Q \cdot K^T$$

- wk0 : first layer
- wk0_64 : first layer on 64 first tokens of the sequence
- wk1 : second layer
- wo1 : output projection, block 2
- acc_tot : accuracy on all predicted tokens
- acc_end : similar to acc_tot, but computed on longer sequences





- With noise, acc_end lose 10%, even with long training
  - Lemma 2 not exact in practice

Links

Our GitHub

Original paper

References

1. Alberto Bietti et al., "Birth of a Transformer: A Memory Viewpoint".,2023, arXiv: 2306.00802 [stat.ML]
2. Ashish Vaswani et al., "Attention Is All You Need", 2023, arXiv: 1706.03762 [cs.CL]
3. Tom B. Brown et al., "Language Models are Few-Shot Learners", 2020, arXiv: 2005.14165