# Birth of a Transformer: A Memory Viewpoint

Johannes Losert

Columbia University

*jal2340@columbia.edu*

October 4, 2024

# Overview

1. Motivation

# Bullet Points

- Memorization vs In-Context Learning
- Associative Memory Hypothesis

# Bigram Language Model

- Unigram Distribution: $\pi_u = p(i)$
- Bigram Distribution: $\pi_b = p(i|j)$ satisfies Markov Property
- $z^n_{1:T}$ is generated with a random walk $z^n_0 \sim \pi_u$, $z_i \sim pi_b(\cdot|z_{i-1})$
- For each of the $N$ sequences $z^1_{1:T} \cdots z^N_{1:T}$
  1. pick $K$ trigger tokens $q_i$
  2. $\forall q_i$ pick a output token $o_i \sim \pi_u$
  3. set $p(o_i|u_i) = 1$
- The paper tests different strategies for picking trigger-output pairs
- For high accuracy model learns $\pi_b$ and how to recognize trigger-output pairs.

# Associative Memory

- orthonormal Embeddings $u_i$, $v_i$, with correlation $\alpha$
- $W = \sum_{i,j} \alpha u_i v_j^T$
- How do we get such embeddings?
  - random Gaussian vectors with variance $1/d$
  - $v_i^T v_i \approx 1$, $v_i^T v_j = O(\frac{1}{\sqrt{d}}) \approx 0$
- Superpositions? What if we want 2-1 mapping? $Wu_{i,j} = Wu_i + Wu_j$ This is why we need MLP's w/ non-linear activation.

# Solving the Modified Bigram Problem w/ Transformers

- Two layer transformer
- $W_Q = I$
- $x_t^{'} := W_O W_V x_{1:t} \sigma(d^{-1/2} x_{1:t}^\top W_k^\top x_t)$
- freeze $W_E, W_U, W_P, W_O^1, W_V^1, W_V^2$ to random initialization. Remap tokens into new tokens, preserve orthogonality.
- pre-softmax attention scores for each index t $(W_k x_{1:t})^\top x_t$
- $W_K^1 = \sum_{t=2}^{T} p_t p_{t-1}^\top$
- $W_K^2 = \sum_{k \in Q} w_E(k)(W_O^1 W_V^1 w_E(k))^\top$ - attend embeddings preceded by triggers
- $W_O^2 = \sum_{k \in N} w_U(k)(W_V^2 w_E(k))^\top$ - output attended token, remapped to output embedding.

# Induction Head Mechanism

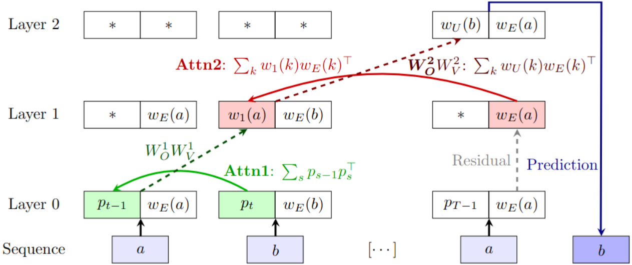


Figure 1: **Induction head mechanism**. Induction heads are a two-layer mechanism that can predict $b$ from a context $[\ldots, a, b, \ldots, a]$. The first layer is a *previous token head*, which attends to the previous token based on positional embeddings ($p_t \to p_{t-1}$) and copies it after a remapping ($w_E(a) \to w_1(a) := W_O^1 W_V^1 w_E(a)$). The second layer is the *induction head*, which attends based on the output of the previous token head ($w_E(a) \to w_1(a)$) and outputs the attended token, remapped to output embeddings ($w_E(b) \to w_U(b)$). Boxes in the diagram represent sets of embeddings in superposition on each token's residual stream, and attention and output associations are shown with the associative memory viewpoint presented in Section 4.

# Associative Memory

- orthonormal Embeddings $u_i$, $v_i$, with correlation $\alpha$
- $W = \sum_{i,j} \alpha u_i v_j^T$
- How do we get such embeddings?
  - random Gaussian vectors with variance $1/d$
  - $v_i^T v_i \approx 1$, $v_i^T v_j = O(\frac{1}{\sqrt{d}}) \approx 0$
- Superpositions? What if we want 2-1 mapping? $Wu_{i,j} = Wu_i + Wu_j$ This is why we need MLP's w/ non-linear activation.

# Conclusions

- Memory recall probes show us theoretical hypothesis is partially correct.

# The End