Transformers can learn pairwise—but not three-wise—functions

Clayton Sanford (Columbia Computer Science)

Joint work with Daniel Hsu and Matus Telgarsky

Statistical Machine Learning Symposium 2023

Transformer architecture

A **transformer** is a sequence-processing neural network architecture, especially prominent in modern NLP models.

- **Self-attention unit:** For input $X \in \mathbb{R}^{N \times d}$ and parameters matrices $Q, K, V \in \mathbb{R}^{d \times m}$,

$$f(X) = \mathrm{softmax}(XQK^{\mathsf{T}}X^{\mathsf{T}})XV.$$

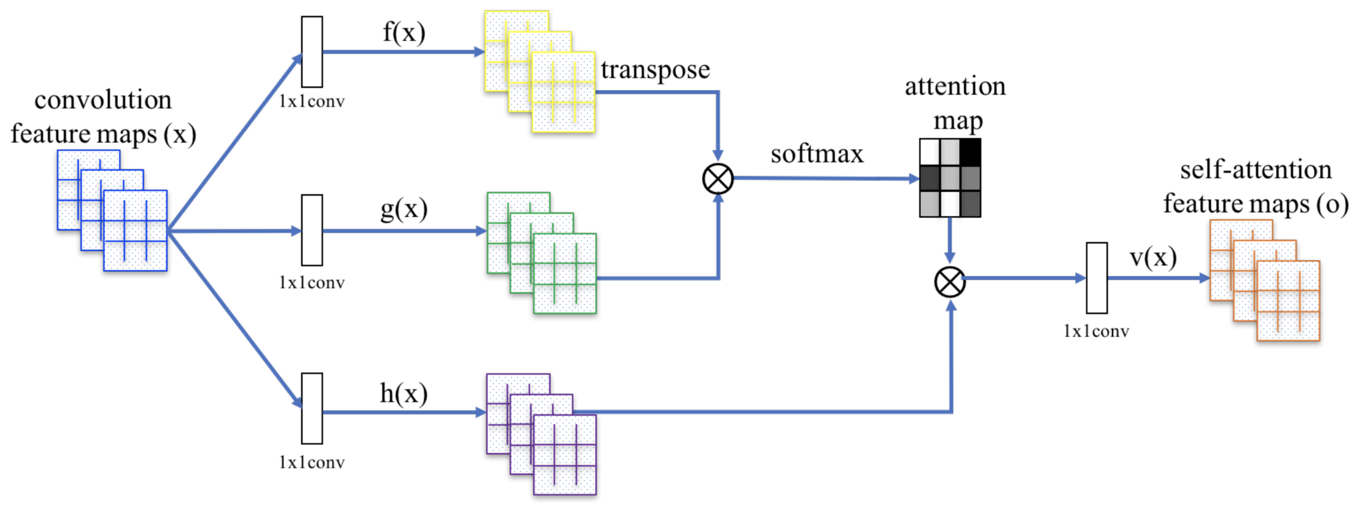

Figure: "Attention? Attention!" *Lil'Log*

- **Multi-headed attention (with skip-level connection):** For self-attention units $f_1, \dots, f_H$:

$$L(X) = X + \sum_{h=1}^{H} f_h(X).$$

- **Element-wise multi-layer perceptron (MLP):** For $\varphi : \mathbb{R}^m \to \mathbb{R}^m$:

$$\varphi(X) := (\varphi(x_1), \dots, \varphi(x_N)).$$

- **Transformer:** For multi-layer attention units $L_1, \dots, L_D$, and MLPs $\varphi_0, \dots, \varphi_D$:

$$T(X) := (\varphi_D \circ L_D \circ \cdots \circ L_1 \circ \varphi_0)(X).$$

Why transformers?

- **Easier to train than RNNs**: simpler parallelization, avoids exploding gradients
- **Attuned to pairwise linguistic structure**: self-attention encodes syntactic and semantic linkages between words
- **Universality:** represents finite-state automata [Liu et al, '22], bounded-depth Dyck languages [Yao et al, '21], Turing machines [Wei et al, '21]

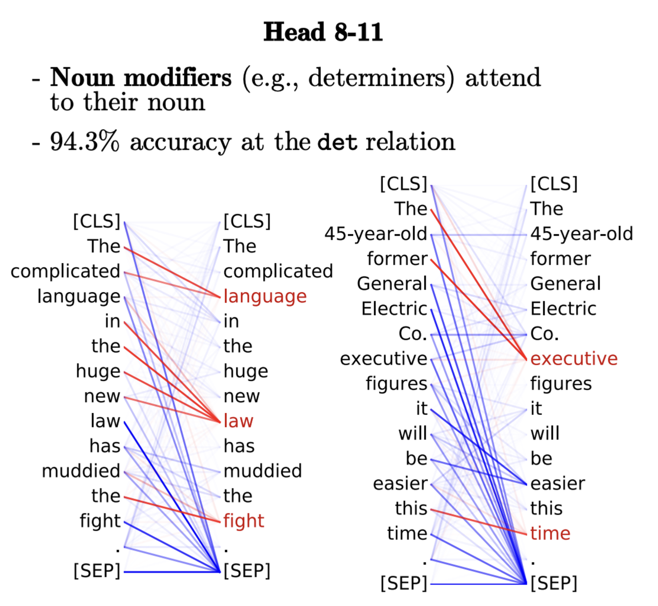

Figure: [Clark, et al '19]

Our questions

- How to mathematically formalize these linkages in target functions?
- How to conceptualize transformer as circuits in theoretical CS language?
- Do those linkages translate to an approximation problem that separates certain transformers architectures from one another?

Modeling decisions

Model	Context length (N)	Depth (D)	Heads (H)	Embedding dimension (m)	MLP parameters (k)
GPT-3	2048	96	96	128	12288
GPT-4	32k	?	?	?	?

- *Observation*: Context length N scales dramatically and $N \gg D, H, m$
  - $\implies$ Sequence model size should be independent of (or at least grow very slowly with) N.
  - $\implies$ Model restricted pairwise computation between elements, governed by low-rank matrices $QX, KX, VX \in \mathbb{R}^{d \times m}$.
- *Observation*: MLP parameter count much larger than self-attention parameters count: $k \gg D, H, m$
  - $\implies$ Model MLPs as universal approximators
  - $\implies$ Unlimited element-wise computational power

Modeling decisions

Functions to approximate:

- $\mathrm{PairID} : X \in [M]^N \mapsto (\mathbb{1}\{\exists j : \; x_i + x_j = 0 \pmod{M}\})_{i \in [N]}$
- $\mathrm{TriID} : X \in [M]^N \mapsto (\mathbb{1}\{\exists j_1, j_2 : \; x_i + x_{j_1} + x_{j_2} = 0 \pmod{M}\})_{i \in [N]}$
- $\mathrm{TriIDGraph} : X \in \{0,1\}^{N \times N} \mapsto (\mathbb{1}\{\exists j_1, j_2 : \; x_{i,j_1} x_{j_1,j_2} x_{j_2,i} = 1\})_{i \in [N]}$

Result	Target	Architecture	Bound
Positive	PairID	Self-attention unit, MLP input & output	$m = O(1)$
Negative	TriID	Multi-headed attention, MLP input & output	$\max(H, m) = N^{\Omega(1)}$
Negative	TriIDGraph	Full transformer with skip-level connections	$\max(D, H, m) = N^{\Omega(1)}$
Positive	TriID	"Three-wise tensor self-attention" unit	$m = O(1)$

Fitting PairID with self-attention

Theorem

For all $\varepsilon > 0$, there exists a self-attention unit f with input and output MLPs with embedding dimension $m = O(1)$ such that

$$\max_{X \in [M]^N} \max_i |f(\varphi(X))_i - \mathrm{PairID}(X)_i| \leq \varepsilon.$$

Proof idea:

- Set φ, Q, K such that $\varphi(X)Q_i = c[\cos(\frac{2\pi x_i}{M}), \sin(\frac{2\pi x_i}{M})]$ and $\varphi(X)K_i = [\cos(\frac{2\pi x_i}{M}), -\sin(\frac{2\pi x_i}{M})]$. This gives

$$(\varphi(X)QK^{\mathsf{T}}\varphi(X)^{\mathsf{T}})_{i,j} = \cos\left(\frac{2\pi(x_i + x_j)}{M}\right) \begin{cases} = c & x_i + x_j = 0 \pmod{M} \\ \leq c(1 - \Omega(\frac{1}{M^2})) & \text{otherwise.} \end{cases}$$

- By choosing sufficiently large c:

$$\mathrm{softmax}(\varphi(X)QK^{\mathsf{T}}\varphi(X)^{\mathsf{T}})_{i,j} \approx \begin{cases} 1 & x_i + x_j = 0 \pmod{M} \\ 0 & \text{otherwise.} \end{cases}$$

Hardness of approximating TriID with multi-headed attention

Theorem

No multi-headed layer with input and output MLPs L with $Hm \leq O(N)$ exists that satisfies

$$\max_{X \in [M]^N} \max_i |L(X)_i - \mathrm{TriID}(X)_i| < \frac{1}{2}.$$

Proof idea:

- Embed instance of set disjointness communication protocol into multi-headed attention. Alice's inputs encoded as $X_1, \dots, X_{N/2}$ and Bob's as the rest.
- Alice and Bob share at most $O(Hm)$ bits by simulating the multi-headed attention together, but must share at least $\Omega(N)$ to solve set disjointness.

Hardness of approximating TriIDGraph with transformer

Theorem

No transformer model T with $DHm \leq O(\frac{N}{\log N})$ exists that satisfies

$$\max_{X \in \{0,1\}^{N \times N}} \max_i |T(X)_i - \mathrm{TriIDGraph}(X)_i| < \frac{1}{2}.$$

Proof idea:

- A Congest communication graph can simulate a multi-layer transformer architecture. Alice and Bob are assigned respective nodes.
- Similar reduction to set disjointness, but more careful embedding scheme.

Fitting TriID with three-wise tensor self-attention unit

A **three-wise tensor self-attention unit** generalizes self-attention to model three-wise interactions by having two key and value transforms and instead computing a tensor product

$$\mathrm{softmax}(XQ \otimes XK_1 \otimes XK_2) \in \mathbb{R}^{N \times N \times N},$$

and multiplying by a value tensor $XV_1 \otimes XV_2 \in \mathbb{R}^{N \times N \times m}$.

Theorem

For all $\varepsilon > 0$, there exists a three-wise self-attention unit f with MLPs with embedding dimension $m = O(1)$ such that

$$\max_{X \in [M]^N} \max_i |f(X)_i - \mathrm{TriID}(X)_i| \leq \varepsilon.$$

Proof idea:

- Same as PairID for self-attention, but instead compute

$$(\varphi(X)Q \otimes \varphi(X)K_1 \otimes \varphi(X)K_2)_{i,j_1,j_2} = \cos\left(\frac{2\pi(x_i + x_{j_1} + x_{j_2})}{M}\right).$$

Open questions and future work

- Strengthen communication complexity lower bounds and extend communication lens to other aspects of transformer learning
- How apt is the "sparse pairwise connectedness" framework for understanding language?
- Are there practical "intrinsically three-wise" learning tasks on which modern transformers fail?

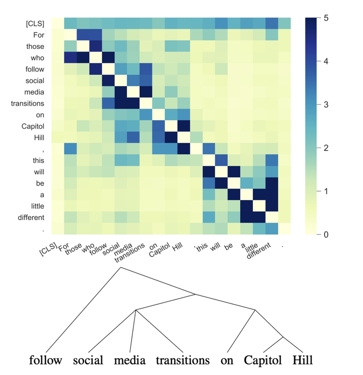

Figure: [Rogers, et al '20]