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

Clayton Sanford

Joint work with Daniel Hsu and Matus Telgarsky

Transformer architecture

What is it?

Transformer architecture

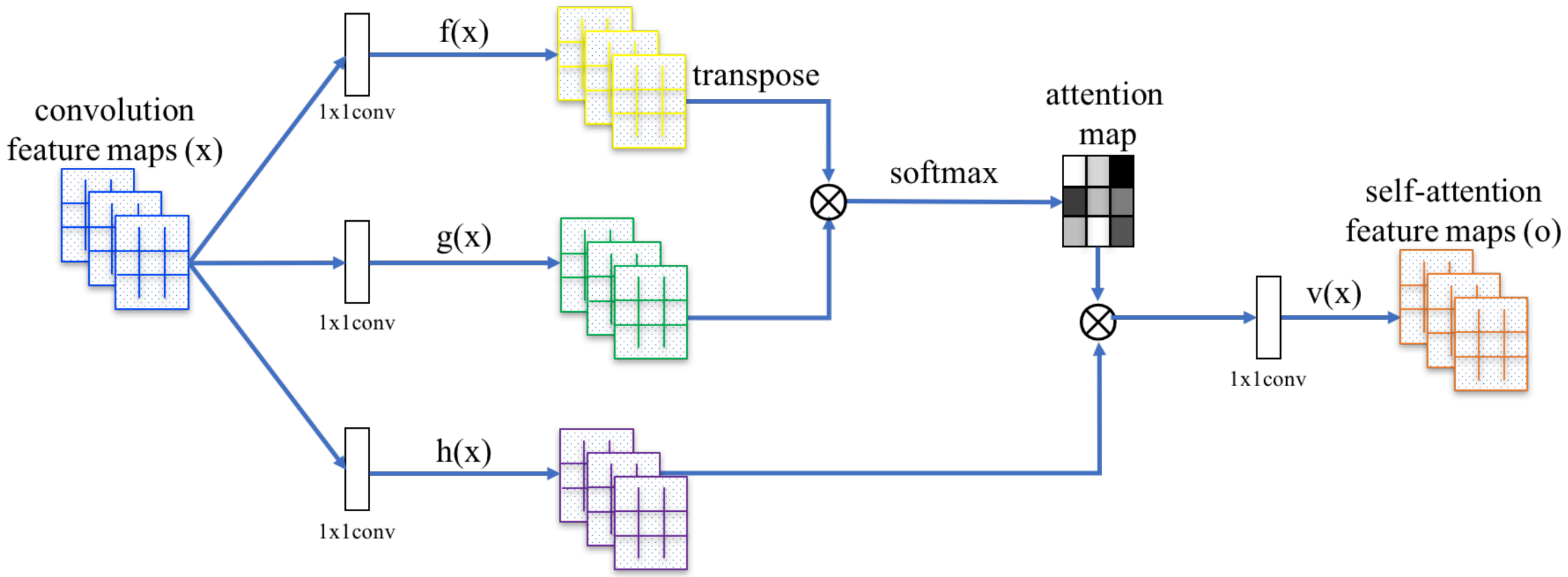
What is it?

- **Self-attention unit:**
  $f(X) = \mathrm{softmax}(XQK^T X^T)XV$ for input $X \in \mathbb{R}^{N \times d}$, model parameters $Q, K, V \in \mathbb{R}^{d \times m}$.

Transformer architecture

What is it?

- **Self-attention unit:**
  $f(X) = \mathrm{softmax}(XQK^TX^T)XV$ for input $X \in \mathbb{R}^{N \times d}$, model parameters $Q, K, V \in \mathbb{R}^{d \times m}$.



Source: https://lilianweng.github.io/posts/2018-06-24-attention/

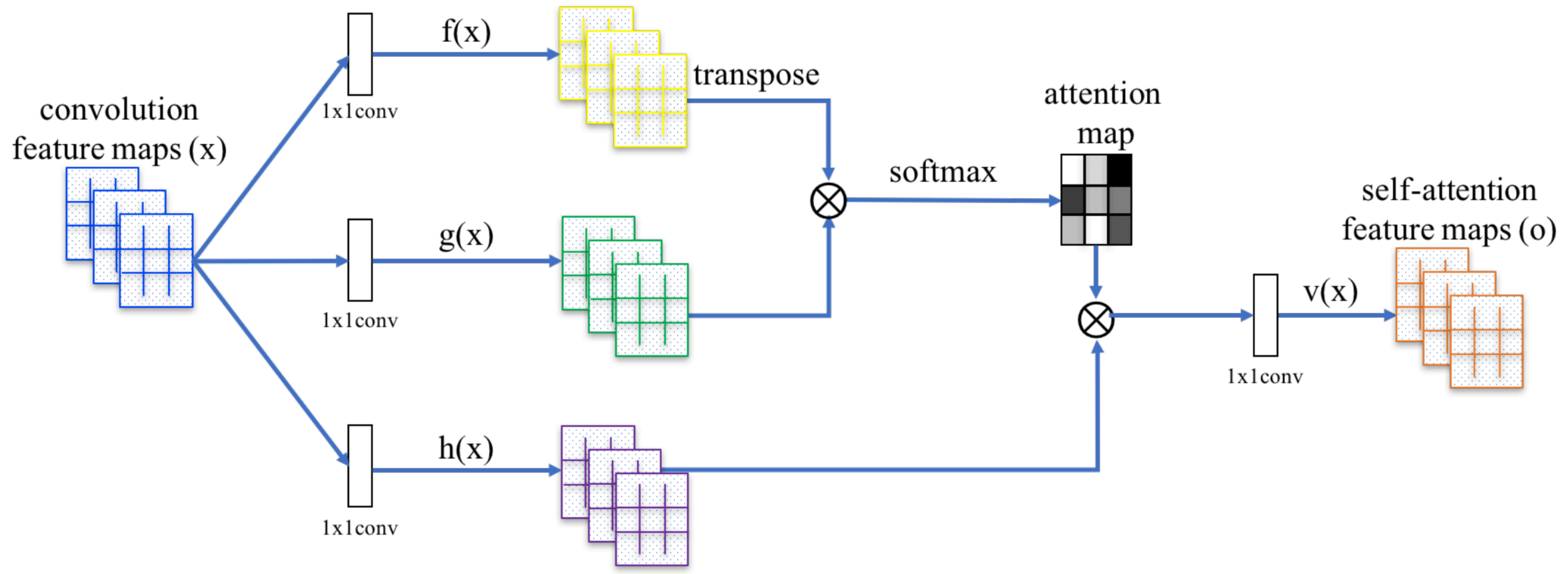
Transformer architecture

What is it?

- **Self-attention unit:** $f(X) = \text{softmax}(XQK^T X^T)XV$ for input $X \in \mathbb{R}^{N\times d}$, model parameters $Q, K, V \in \mathbb{R}^{d\times m}$.

- **Multi-headed attention:**

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



Source: https://lilianweng.github.io/posts/2018-06-24-attention/

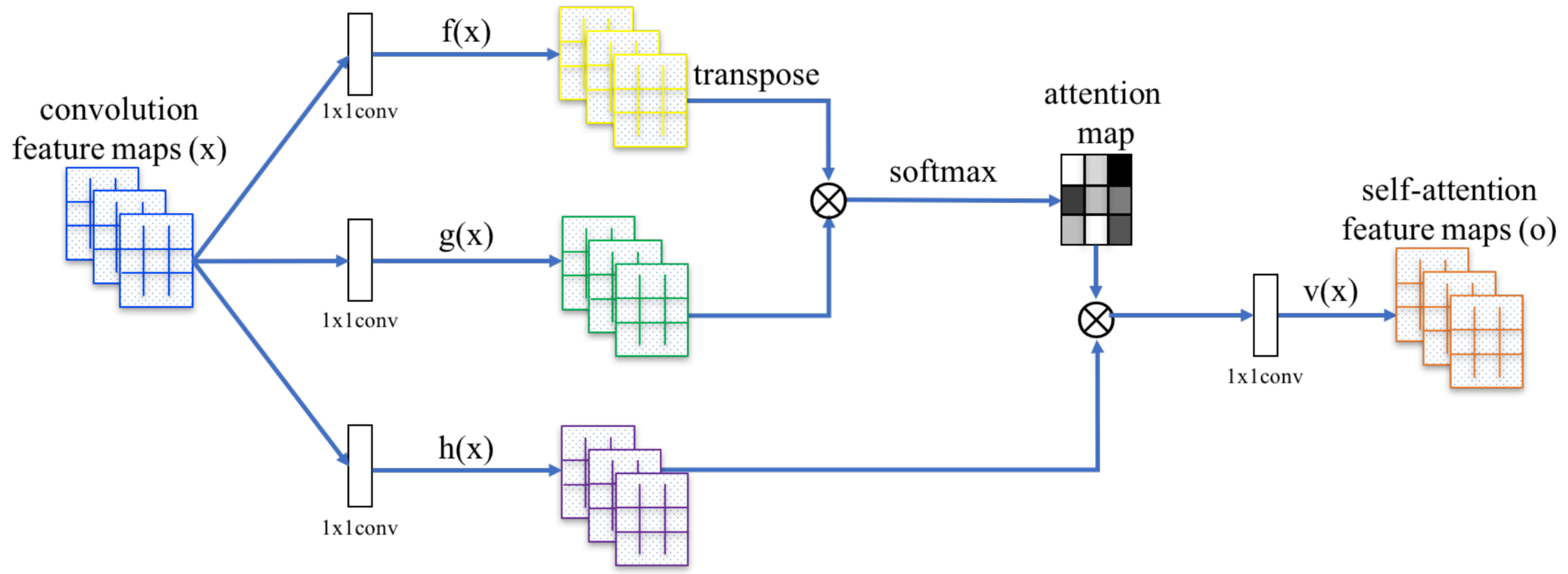
Transformer architecture

What is it?

- **Self-attention unit:**
  $f(X) = \mathrm{softmax}(XQK^TX^T)XV$ for input $X \in \mathbb{R}^{N\times d}$, model parameters $Q, K, V \in \mathbb{R}^{d\times m}$.

- **Multi-headed attention:**
  $$L(X) = X + \sum_{h=1}^{H} f_h(X)$$

- **Element-wise multi-layer perceptron (MLP):**
  $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$



Source: https://lilianweng.github.io/posts/2018-06-24-attention/

Transformer architecture

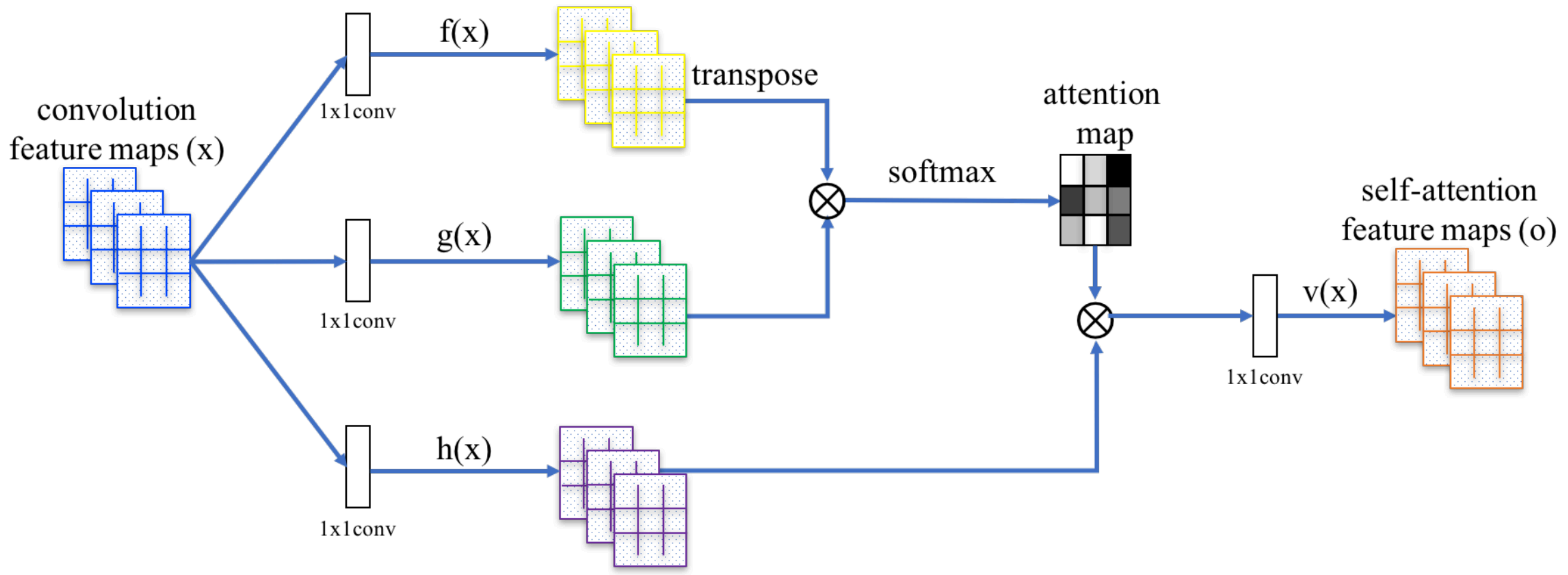
What is it?

- **Self-attention unit:**
  $f(X) = \mathrm{softmax}(XQK^T X^T)XV$ for input $X \in \mathbb{R}^{N\times d}$, model parameters $Q, K, V \in \mathbb{R}^{d\times m}$.

- **Multi-headed attention:**
  $$L(X) = X + \sum_{h=1}^{H} f_h(X)$$

- **Element-wise multi-layer perceptron (MLP):**
  $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$

- **Full transformer:**
  $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$



Source: https://lilianweng.github.io/posts/2018-06-24-attention/

Transformer architecture

What is it?

- **Self-attention unit:** $f(X) = \mathrm{softmax}(XQK^T X^T)XV$ for input $X \in \mathbb{R}^{N\times d}$, model parameters $Q, K, V \in \mathbb{R}^{d\times m}$.
- **Multi-headed attention:**
$$L(X) = X + \sum_{h=1}^{H} f_h(X)$$
- **Element-wise multi-layer perceptron (MLP):** $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$
- **Full transformer:** $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$

Pairwise structure

Transformer architecture

What is it?

- **Self-attention unit:** $f(X) = \mathrm{softmax}(XQK^TX^T)XV$ for input $X \in \mathbb{R}^{N\times d}$, model parameters $Q, K, V \in \mathbb{R}^{d\times m}$.
- **Multi-headed attention:**
$$L(X) = X + \sum_{h=1}^{H} f_h(X)$$
- **Element-wise multi-layer perceptron (MLP):** $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$
- **Full transformer:** $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$

Pairwise structure

- **Attuned to pairwise linguistic structure**: self-attention encodes syntactic and semantic linkages between words*

*I am not a linguist!

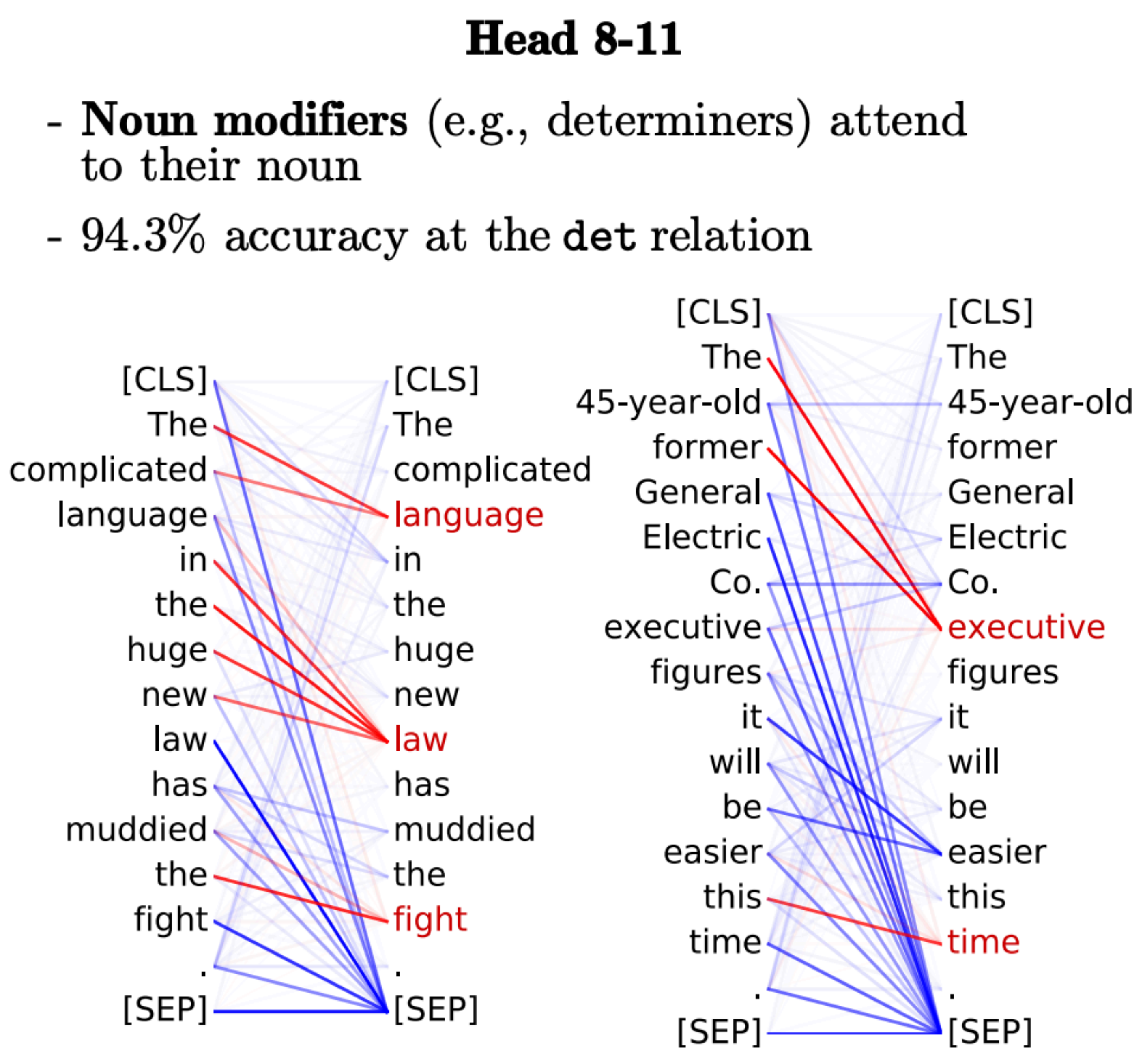
Transformer architecture

What is it?

- **Self-attention unit:** $f(X) = \mathrm{softmax}(XQK^TX^T)XV$ for input $X \in \mathbb{R}^{N\times d}$, model parameters $Q, K, V \in \mathbb{R}^{d\times m}$.
- **Multi-headed attention:**
$$L(X) = X + \sum_{h=1}^{H} f_h(X)$$
- **Element-wise multi-layer perceptron (MLP):** $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$
- **Full transformer:** $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$

Pairwise structure

- **Attuned to pairwise linguistic structure**: self-attention encodes syntactic and semantic linkages between words*



*I am not a linguist!

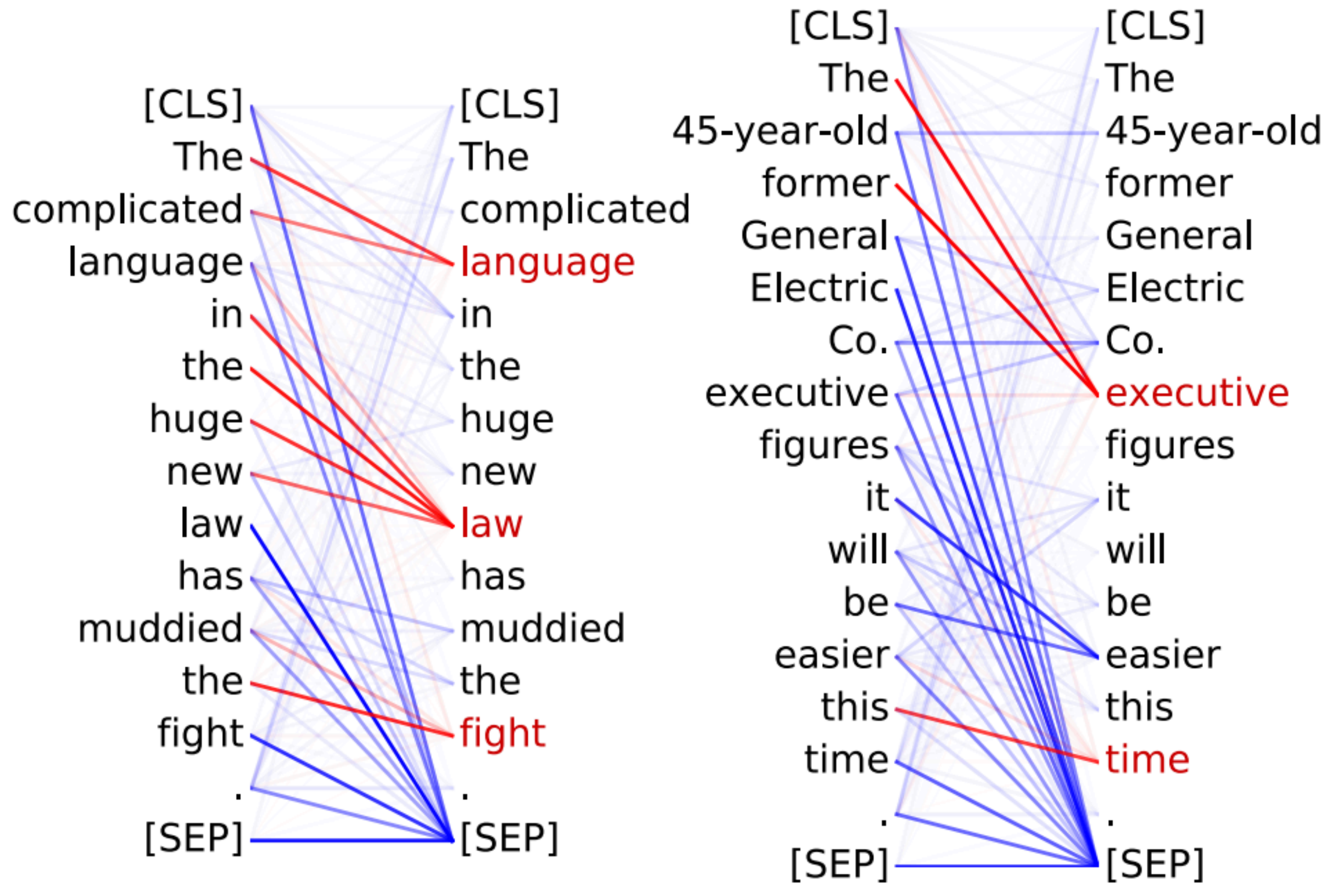
Transformer architecture

Pairwise structure

- **Attuned to pairwise linguistic structure**: self-attention encodes syntactic and semantic linkages between words*

Head 8-11

- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the `det` relation



Our question

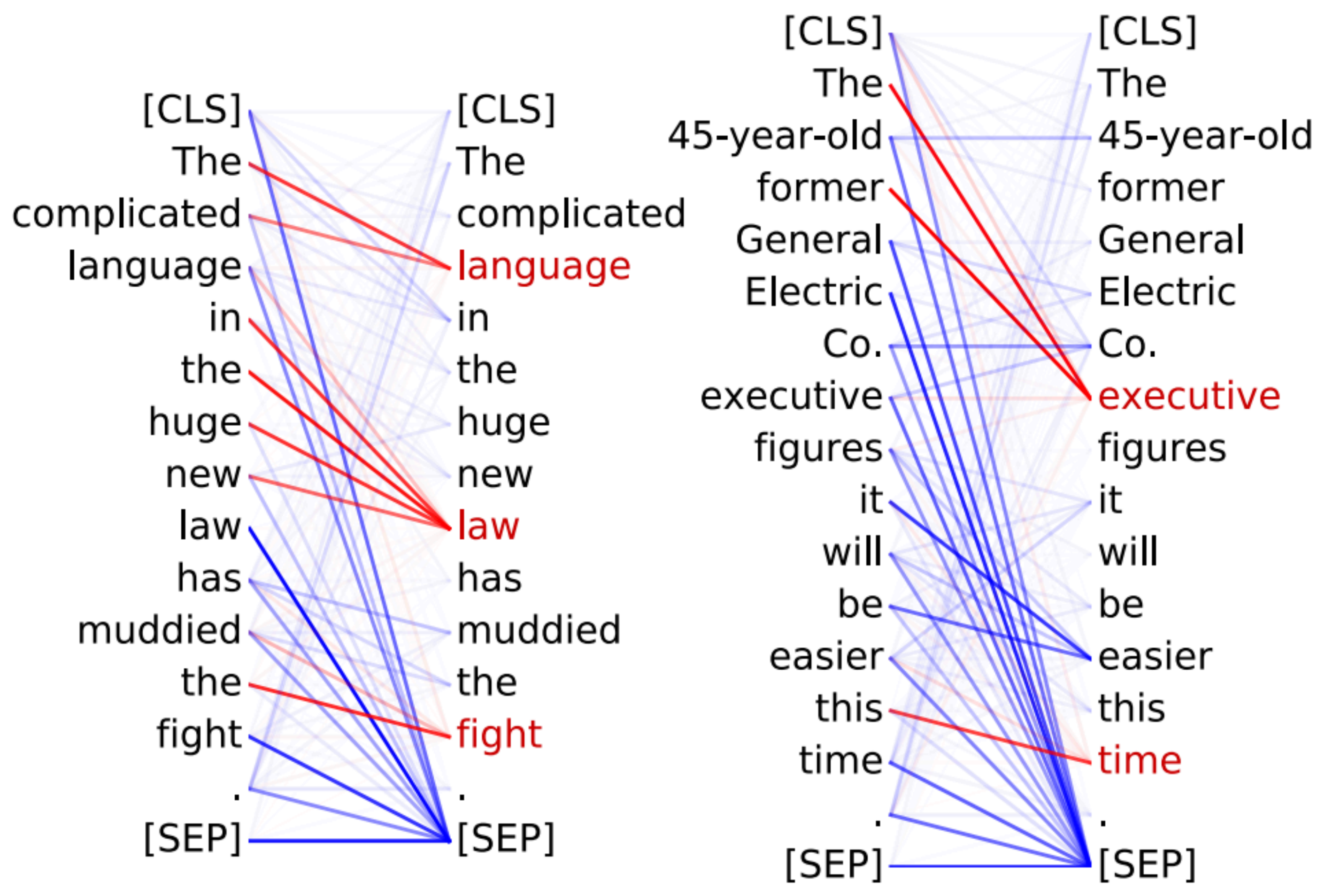
Transformer architecture

Pairwise structure

- **Attuned to pairwise linguistic structure**: self-attention encodes syntactic and semantic linkages between words*

Head 8-11

- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the `det` relation



Our question

How do we formalize these linkages as target functions that elucidate capabilities and limitations of transformers?

Transformer architecture

Our question

How do we formalize these linkages as target functions that elucidate capabilities and limitations of transformers?

Modeling decisions:

Transformer architecture

Our question

How do we formalize these linkages as target functions that elucidate capabilities and limitations of transformers?

Modeling decisions:

Model	Context length (N)	#layers (D)	#heads (H)	*#param self-attn* (m)	*#param MLP* (k)
GPT-3	2048	96	96	128	12288

Transformer architecture

Our question

How do we formalize these linkages as target functions that elucidate capabilities and limitations of transformers?

Modeling decisions:

Model	Context length (N)	#layers (D)	#heads (H)	*#param self-attn* (m)	*#param MLP* (k)
GPT-3	2048	96	96	128	12288
GPT-4	32k	🙃	🙃	🙃	🙃

Transformer architecture

Our question

How do we formalize these linkages as target functions that elucidate capabilities and limitations of transformers?

Modeling decisions:

Model	Context length (N)	#layers (D)	#heads (H)	*#param self-attn* (m)	*#param MLP* (k)
GPT-3	2048	96	96	128	12288
GPT-4	32k	🙃	🙃	🙃	🙃

- Context length $N \gg$ #params in self-attention unit (depth D, heads H, and embedding dim m)

Transformer architecture

Our question

How do we formalize these linkages as target functions that elucidate capabilities and limitations of transformers?

Modeling decisions:

Model	Context length (N)	#layers (D)	#heads (H)	*#param self-attn* (m)	*#param MLP* (k)
GPT-3	2048	96	96	128	12288
GPT-4	32k	🙃	🙃	🙃	🙃

- Context length $N \gg$ #params in self-attention unit (depth D, heads H, and embedding dim m)

  $\Longrightarrow$ **restricted pairwise computation between elements, model size independent of N**

Transformer architecture

Our question

How do we formalize these linkages as target functions that elucidate capabilities and limitations of transformers?

Modeling decisions:

Model	Context length (N)	#layers (D)	#heads (H)	*#param self-attn* (m)	*#param MLP* (k)
GPT-3	2048	96	96	128	12288
GPT-4	32k	🙃	🙃	🙃	🙃

- Context length $N \gg$ #params in self-attention unit (depth D, heads H, and embedding dim m)

  $\Longrightarrow$ **restricted pairwise computation between elements, model size independent of N**

- #params in MLP $k \gg$ #params in self-attention

Transformer architecture

Our question

How do we formalize these linkages as target functions that elucidate capabilities and limitations of transformers?

Modeling decisions:

Model	Context length (*N*)	#layers (*D*)	#heads (*H*)	*#param self-attn (m)*	*#param MLP (k)*
GPT-3	2048	96	96	128	12288
GPT-4	32k	🙃	🙃	🙃	🙃

- Context length $N \gg$ #params in self-attention unit (depth D, heads H, and embedding dim m)

  $\Longrightarrow$ **restricted pairwise computation between elements, model size independent of N**

- #params in MLP $k \gg$ #params in self-attention

  $\Longrightarrow$ **unlimited element-wise computational power**

Our Results

Formulation & bounds

Architecture

- **Self-attention unit:** $f(X) = \text{softmax}(XQK^TX^T)XV$ for input $X \in \mathbb{R}^{N\times d}$, model parameters $Q, K, V \in \mathbb{R}^{d\times m}$.
- **Multi-headed attention:** $L(X) = X + \sum_{h=1}^{H} f_h(X)$
- **Element-wise multi-layer perceptron (MLP):** $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$
- **Full transformer:** $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$

Our Results

Formulation & bounds

- $\text{PairID}(X) = 1\{\exists j : \; x_i + x_j = 0\}_{i \in [N]}$

Architecture

- **Self-attention unit:** $f(X) = \text{softmax}(XQK^T X^T)XV$ for input $X \in \mathbb{R}^{N \times d}$, model parameters $Q, K, V \in \mathbb{R}^{d \times m}$.
- **Multi-headed attention:** $L(X) = X + \sum_{h=1}^{H} f_h(X)$
- **Element-wise multi-layer perceptron (MLP):** $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$
- **Full transformer:** $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$

Our Results

Formulation & bounds

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

Architecture

- **Self-attention unit:** $f(X) = \mathrm{softmax}(XQK^T X^T)XV$ for input $X \in \mathbb{R}^{N \times d}$, model parameters $Q, K, V \in \mathbb{R}^{d \times m}$.
- **Multi-headed attention:** $L(X) = X + \sum_{h=1}^{H} f_h(X)$
- **Element-wise multi-layer perceptron (MLP):** $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$
- **Full transformer:** $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$

Our Results

Formulation & bounds

- $\text{PairID}(X) = 1\{\exists j : \; x_i + x_j = 0\}_{i \in [N]}$
- $\text{TriID}(X) = 1\{\exists j_1, j_2 : \; x_i + x_{j_1} + x_{j_2} = 0\}_{i \in [N]}$

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$

Architecture

- **Self-attention unit:** $f(X) = \text{softmax}(XQK^T X^T)XV$ for input $X \in \mathbb{R}^{N \times d}$, model parameters $Q, K, V \in \mathbb{R}^{d \times m}$.
- **Multi-headed attention:** $L(X) = X + \sum_{h=1}^{H} f_h(X)$
- **Element-wise multi-layer perceptron (MLP):** $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$
- **Full transformer:** $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$

Our Results

Formulation & bounds

- $\text{PairID}(X) = 1\{\exists j: \ x_i + x_j = 0\}_{i \in [N]}$
- $\text{TriID}(X) = 1\{\exists j_1, j_2: \ x_i + x_{j_1} + x_{j_2} = 0\}_{i \in [N]}$

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq N^{\Omega(1)}$

Architecture

- **Self-attention unit:** $f(X) = \text{softmax}(XQK^T X^T)XV$ for input $X \in \mathbb{R}^{N \times d}$, model parameters $Q, K, V \in \mathbb{R}^{d \times m}$.
- **Multi-headed attention:** $L(X) = X + \sum_{h=1}^{H} f_h(X)$
- **Element-wise multi-layer perceptron (MLP):** $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$
- **Full transformer:** $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$

Our Results

Formulation & bounds

- $\mathrm{PairID}(X) = 1\{\exists j: \ x_i + x_j = 0\}_{i\in[N]}$
- $\mathrm{TriID}(X) = 1\{\exists j_1, j_2: \ x_i + x_{j_1} + x_{j_2} = 0\}_{i\in[N]}$

Result	**Target**	**Architecture**	**Bound**
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq N^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq N^{\Omega(1)}$

Architecture

- **Self-attention unit:** $f(X) = \mathrm{softmax}(XQK^T X^T)XV$ for input $X \in \mathbb{R}^{N\times d}$, model parameters $Q, K, V \in \mathbb{R}^{d\times m}$.
- **Multi-headed attention:** $L(X) = X + \sum_{h=1}^{H} f_h(X)$
- **Element-wise multi-layer perceptron (MLP):** $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$
- **Full transformer:** $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$

Our Results

Formulation & bounds

- $\mathrm{PairID}(X) = 1\{\exists j: \ x_i + x_j = 0\}_{i \in [N]}$
- $\mathrm{TriID}(X) = 1\{\exists j_1, j_2: \ x_i + x_{j_1} + x_{j_2} = 0\}_{i \in [N]}$

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq N^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq N^{\Omega(1)}$
[+]	TriID	"Three-wise tensor self-attention unit"	$m = O(1)$

Architecture

- **Self-attention unit:** $f(X) = \mathrm{softmax}(XQK^TX^T)XV$ for input $X \in \mathbb{R}^{N \times d}$, model parameters $Q, K, V \in \mathbb{R}^{d \times m}$.
- **Multi-headed attention:** $L(X) = X + \sum_{h=1}^{H} f_h(X)$
- **Element-wise multi-layer perceptron (MLP):** $\phi(X) = (\phi(x_1), \ldots, \phi(x_N))$
- **Full transformer:** $T(X) = (\phi_D \circ L_D \circ \ldots \circ L_1 \circ \phi_0)(X)$

Our Results

Formulation & bounds

- $\text{PairID}(X) = 1\{\exists j: \ x_i + x_j = 0\}_{i \in [N]}$
- $\text{TriID}(X) = 1\{\exists j_1, j_2: \ x_i + x_{j_1} + x_{j_2} = 0\}_{i \in [N]}$

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq N^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq N^{\Omega(1)}$
[+]	TriID	"Three-wise tensor self-attention unit"	$m = O(1)$

Methodology

Our Results

Formulation & bounds

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

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq N^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq N^{\Omega(1)}$
[+]	TriID	"Three-wise tensor self-attention unit"	$m = O(1)$

Methodology

[+] Simple constructions with trigonometric embeddings

Our Results

Formulation & bounds

- $\text{PairID}(X) = 1\{\exists j: \ x_i + x_j = 0\}_{i \in [N]}$
- $\text{TriID}(X) = 1\{\exists j_1, j_2: \ x_i + x_{j_1} + x_{j_2} = 0\}_{i \in [N]}$

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq \mathrm{N}^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq \mathrm{N}^{\Omega(1)}$
[+]	TriID	“Three-wise tensor self-attention unit”	$m = O(1)$

Methodology

[+] Simple constructions with trigonometric embeddings

[-] Reduction to set disjointness communication complexity

Our Results

Formulation & bounds

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

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq N^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq N^{\Omega(1)}$
[+]	TriID	“Three-wise tensor self-attention unit”	$m = O(1)$

Methodology

[+] Simple constructions with trigonometric embeddings

[-] Reduction to set disjointness communication complexity

- #MLP params $\gg$ #self-attention params $\Longrightarrow$ key representational bottleneck as limitations on pairwise communication

Our Results

Formulation & bounds

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

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq N^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq N^{\Omega(1)}$
[+]	TriID	“Three-wise tensor self-attention unit”	$m = O(1)$

Further work

Our Results

Formulation & bounds

- $\text{PairID}(X) = 1\{\, \exists j : \ x_i + x_j = 0\}_{i \in [N]}$
- $\text{TriID}(X) = 1\{\, \exists j_1, j_2 : \ x_i + x_{j_1} + x_{j_2} = 0\}_{i \in [N]}$

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq \mathrm{N}^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq \mathrm{N}^{\Omega(1)}$
[+]	TriID	“Three-wise tensor self-attention unit”	$m = O(1)$

Further work

- Apply communication complexity to obtain matching bounds

Our Results

Formulation & bounds

- $\text{PairID}(X) = 1\{\exists j : \ x_i + x_j = 0\}_{i \in [N]}$
- $\text{TriID}(X) = 1\{\exists j_1, j_2 : \ x_i + x_{j_1} + x_{j_2} = 0\}_{i \in [N]}$

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq \mathrm{N}^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq \mathrm{N}^{\Omega(1)}$
[+]	TriID	"Three-wise tensor self-attention unit"	$m = O(1)$

Further work

- Apply communication complexity to obtain matching bounds
- How apt is the "sparse pairwise connectedness" framework for understanding language?

Our Results

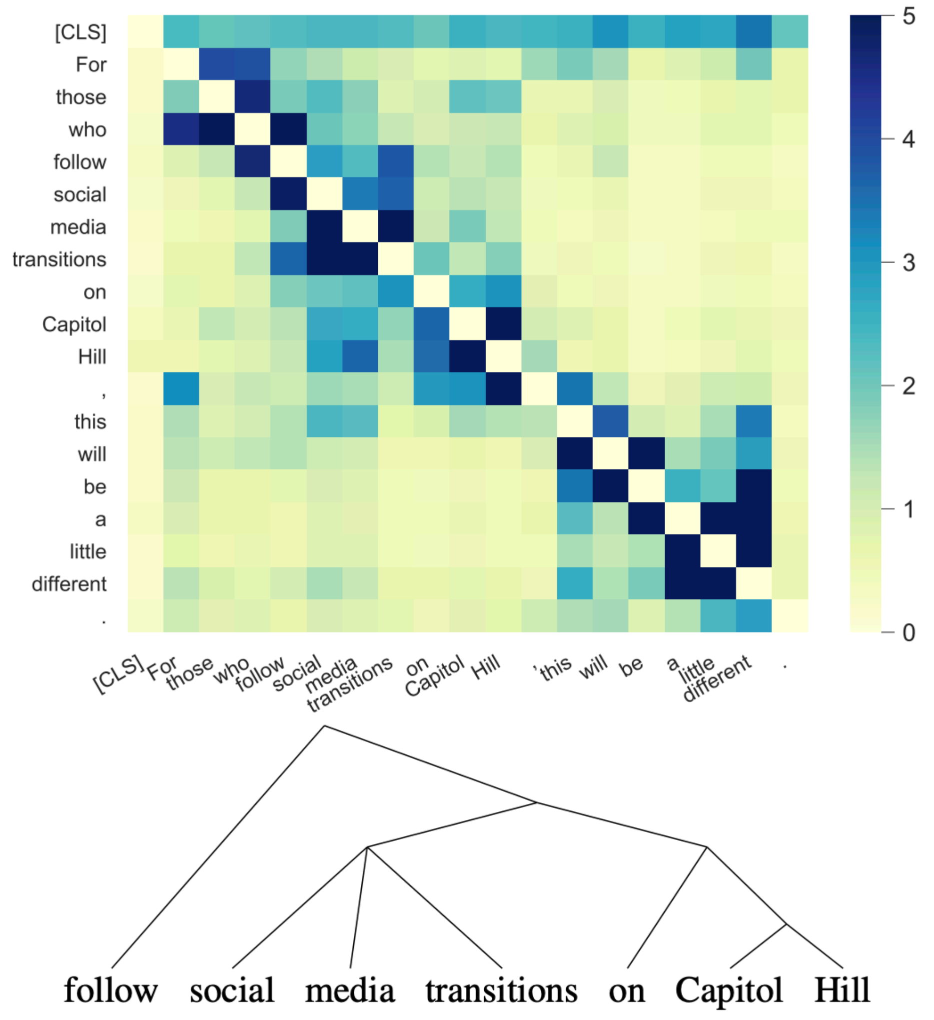
Formulation & bounds

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

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq N^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq N^{\Omega(1)}$
[+]	TriID	"Three-wise tensor self-attention unit"	$m = O(1)$

Further work

- Apply communication complexity to obtain matching bounds
- How apt is the "sparse pairwise connectedness" framework for understanding language?



Our Results

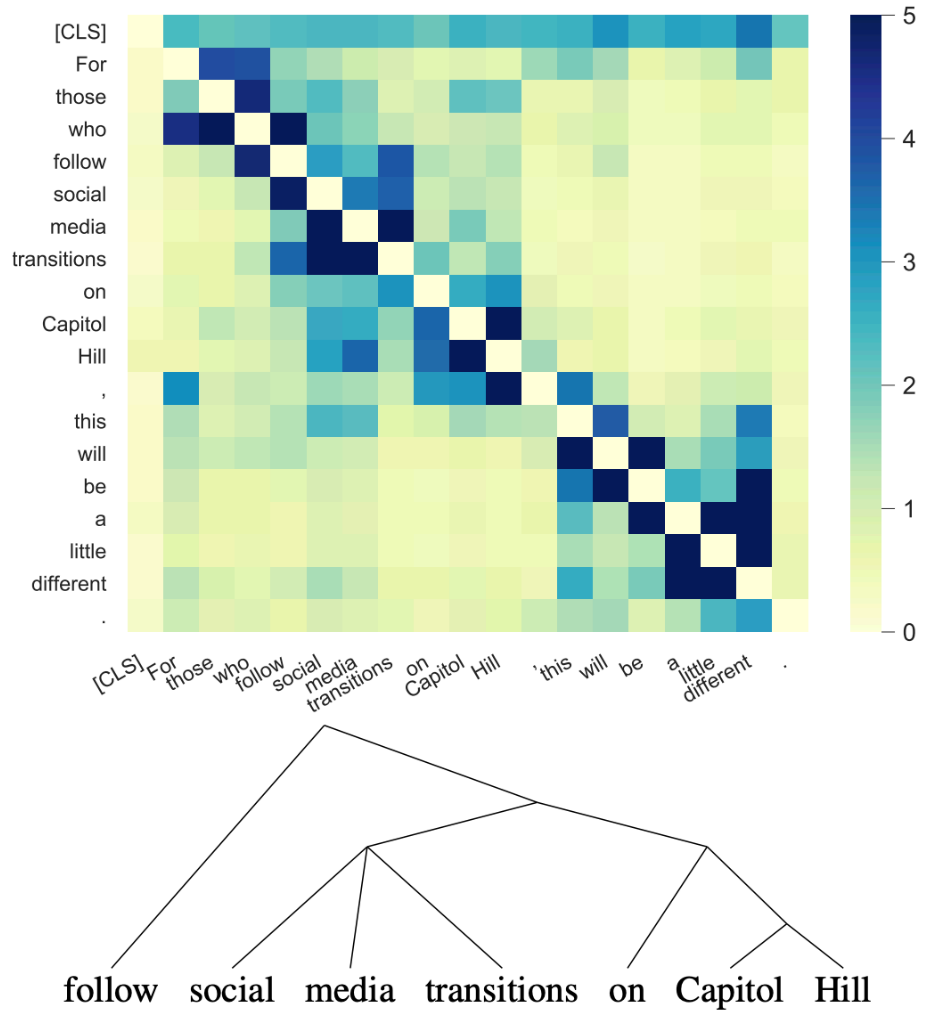
Formulation & bounds

- $\mathrm{PairID}(X) = 1\{\, \exists j : \ x_i + x_j = 0\}_{i \in [N]}$
- $\mathrm{TriID}(X) = 1\{\, \exists j_1, j_2 : \ x_i + x_{j_1} + x_{j_2} = 0\}_{i \in [N]}$

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq N^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq N^{\Omega(1)}$
[+]	TriID	“Three-wise tensor self-attention unit”	$m = O(1)$

Further work

- Apply communication complexity to obtain matching bounds
- How apt is the “sparse pairwise connectedness” framework for understanding language?



Our Results

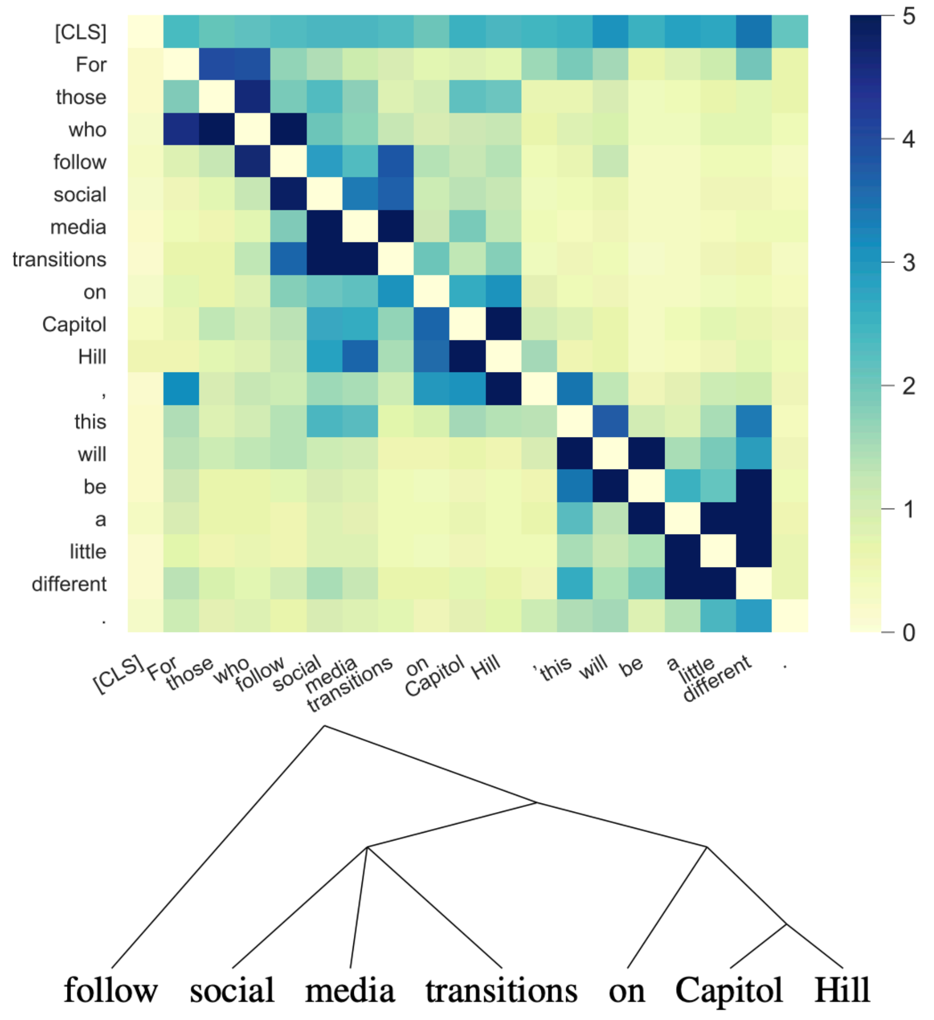
Formulation & bounds

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

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq N^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq N^{\Omega(1)}$
[+]	TriID	“Three-wise tensor self-attention unit”	$m = O(1)$

Further work

- Apply communication complexity to obtain matching bounds
- How apt is the “sparse pairwise connectedness” framework for understanding language?



- Are there practical “intrinsically three-wise” learning tasks where modern transformers fail?

Our Results

Formulation & bounds

- $\mathrm{PairID}(X) = 1\{\exists j: \ x_i + x_j = 0\}_{i \in [N]}$
- $\mathrm{TriID}(X) = 1\{\exists j_1, j_2: \ x_i + x_{j_1} + x_{j_2} = 0\}_{i \in [N]}$

Result	Target	Architecture	Bound
[+]	PairID	Self-attention unit, MLP input	$m = O(1)$
[-]	TriID	Multi-headed attention, MLP input	$\max(H, m) \geq N^{\Omega(1)}$
[-]	Modified TriID	Full transformer	$\max(D, H, m) \geq N^{\Omega(1)}$
[+]	TriID	"Three-wise tensor self-attention unit"	$m = O(1)$

Thank you!

Want to discuss or learn more?
Check out the poster at 2pm.