On the Optimal Memorization Capacity of Transformers

Tokio Kajitsuka, Issei Sato

The University of Tokyo

ICLR 2025

UTokyo

Expressivity of Transformers

- How to assess the **expressivity** of Transformers?

Universal Approximation Theorem

[Yun et al., ICLR 2020]

$$d_p(f, g) := \left(\int \| f(\boldsymbol{X}) - g(\boldsymbol{X}) \|_p^p \, d\boldsymbol{X} \right)^{1/p} \leq \epsilon$$

Memorization Capacity

[Kim et al., ICLR 2023]

$$f\left(\boldsymbol{X}^{(i)}\right) = \boldsymbol{Y}^{(i)} \quad \forall\, i = 1, \ldots, N$$

Expressivity of Transformers

- How to assess the expressivity of Transformers?

Universal Approximation Theorem

[Yun et al., ICLR 2020]

$$d_p(f, g) := \left(\int \| f(\boldsymbol{X}) - g(\boldsymbol{X}) \|_p^p \, d\boldsymbol{X} \right)^{1/p} \leq \epsilon$$

Approximation of continuous functions

Memorization Capacity

[Kim et al., ICLR 2023]

$$f\left(\boldsymbol{X}^{(i)}\right) = \boldsymbol{Y}^{(i)} \quad \forall\, i = 1, \ldots, N$$

Expressivity of Transformers

- How to assess the expressivity of Transformers?

Universal Approximation Theorem

[Yun et al., ICLR 2020]

$$d_p(f, g) := \left(\int \| f(\boldsymbol{X}) - g(\boldsymbol{X}) \|_p^p \, d\boldsymbol{X} \right)^{1/p} \leq \epsilon$$

Memorization Capacity

[Kim et al., ICLR 2023]

$$f\left(\boldsymbol{X}^{(i)}\right) = \boldsymbol{Y}^{(i)} \quad \forall\, i = 1, \ldots, N$$

Discrete version of UAT

Memorization Capacity

- The minimum size of the model for memorizing finite input-label pairs.
  - **Feed-forward networks**
    - Given $(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), \dots, (\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)}) \subset \mathbb{R}^d \times \mathbb{R}$ and construct a network s.t.

$$f\left(\boldsymbol{X}^{(i)}\right) = \boldsymbol{Y}^{(i)} \quad \forall\, i = 1, \dots, N$$

  - **Transformers:**
    - Given $(\boldsymbol{X}^{(1)}, \boldsymbol{Y}^{(1)}), \dots, (\boldsymbol{X}^{(N)}, \boldsymbol{Y}^{(N)}) \subset \mathbb{R}^{d\times n} \times \mathbb{R}^{d\times n}$ and construct a network s.t.

$$f\left(\boldsymbol{X}^{(i)}\right) = \boldsymbol{Y}^{(i)} \quad \forall\, i = 1, \dots, N$$

Memorization Capacity

- The minimum size of the model for memorizing finite input-label pairs.
  - **Feed-forward networks**
    - Given $(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)}) \subset \mathbb{R}^d \times \mathbb{R}$ and construct a network s.t.

$$f\left(\boldsymbol{X}^{(i)}\right) = \boldsymbol{Y}^{(i)} \quad \forall\, i = 1, \ldots, N$$

  - **Transformers:**
    - Given $(\boldsymbol{X}^{(1)}, \boldsymbol{Y}^{(1)}), \ldots, (\boldsymbol{X}^{(N)}, \boldsymbol{Y}^{(N)}) \subset \mathbb{R}^{d \times n} \times \mathbb{R}^{d \times n}$ and construct a network s.t.

$$f\left(\boldsymbol{X}^{(i)}\right) = \boldsymbol{Y}^{(i)} \quad \forall\, i = 1, \ldots, N$$

Memorization Capacity

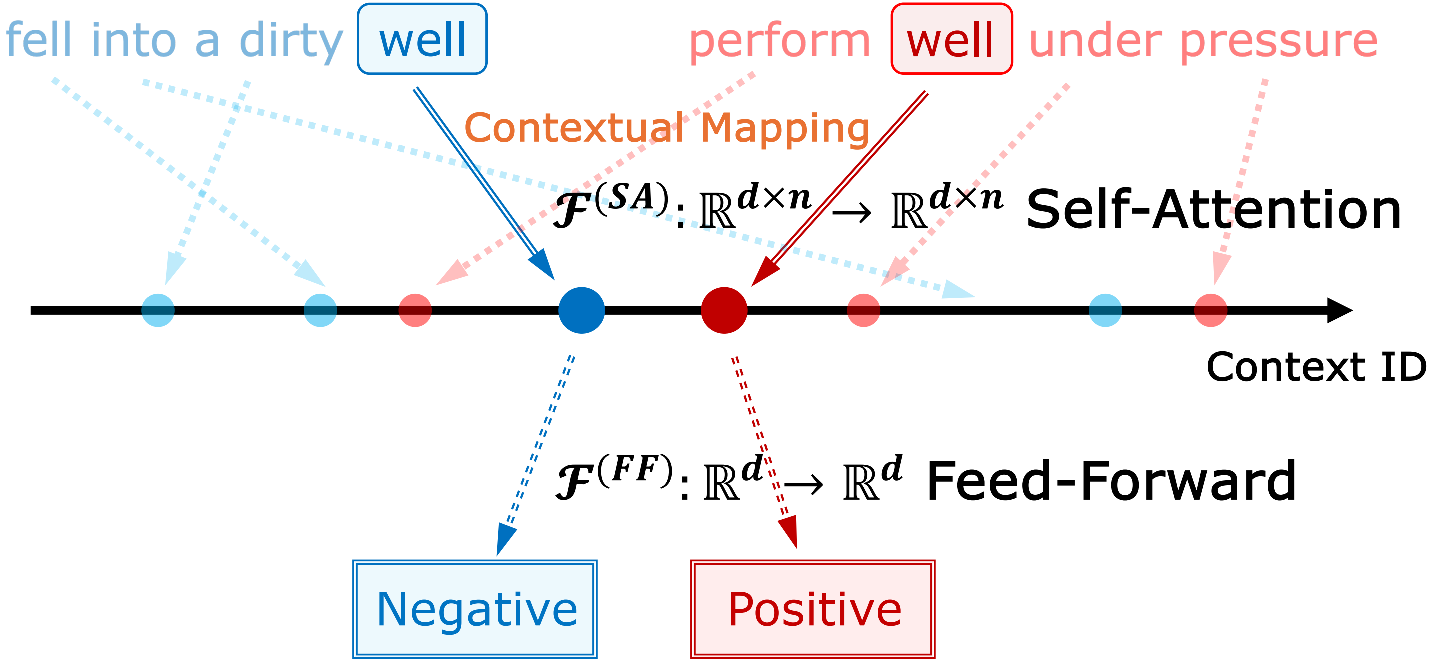
- The minimum size of the model for memorizing finite input-label pairs.
  - Feed-forward networks
    - Given $(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)}) \subset \mathbb{R}^d \times \mathbb{R}$ and construct a network s.t.

$$f\left(\boldsymbol{X}^{(i)}\right) = \boldsymbol{Y}^{(i)} \quad \forall\, i = 1, \ldots, N$$

  - Transformers:
    - Given $(\boldsymbol{X}^{(1)}, \boldsymbol{Y}^{(1)}), \ldots, (\boldsymbol{X}^{(N)}, \boldsymbol{Y}^{(N)}) \subset \mathbb{R}^{d \times n} \times \mathbb{R}^{d \times n}$ and construct a network s.t. (sequence length: n)

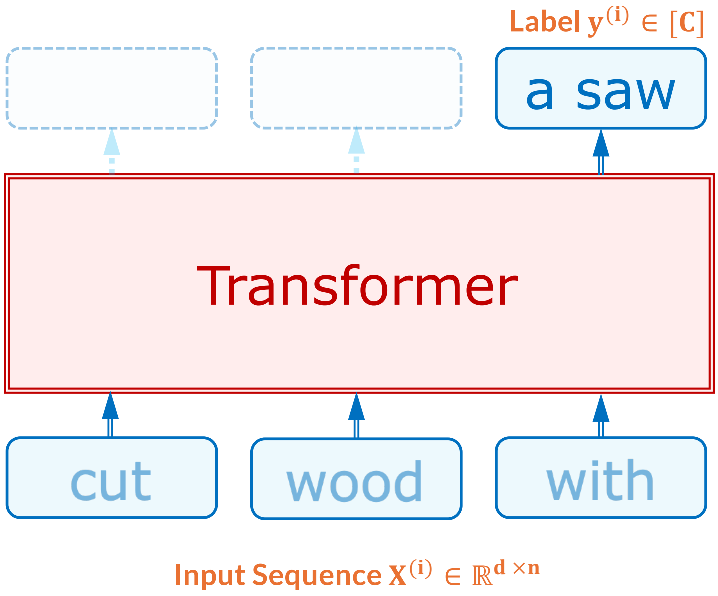
$$f\left(\boldsymbol{X}^{(i)}\right) = \boldsymbol{Y}^{(i)} \quad \forall\, i = 1, \ldots, N$$

Difficulty in Transformers' Memorization

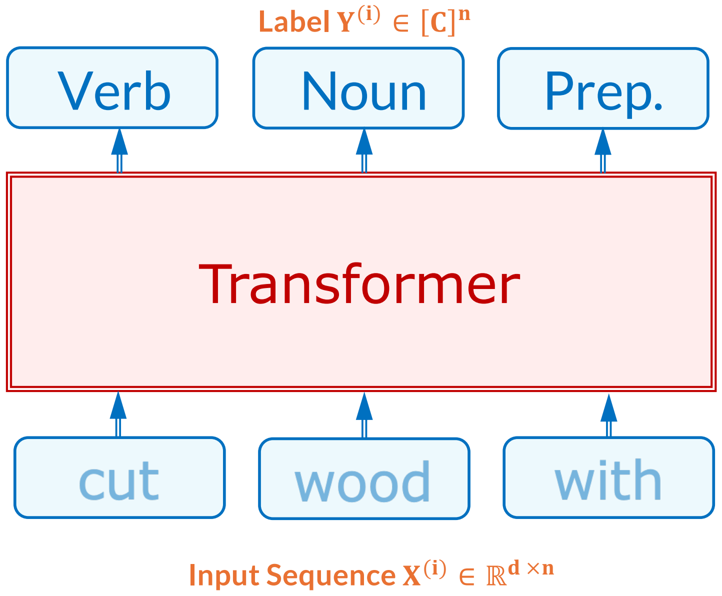


Two Prediction Tasks

Next-token Prediction



Seq-to-seq Prediction



Efficiency of Previous Studies

- Are the constructions of previous studies efficient?

➡ Necessary to investigate the lower bound of the memorization capacity.

Authors	Task	Upper bound	Lower bound
[Kim et al., ICLR 2023]	seq-to-seq	$\tilde{O}(n + \sqrt{nN})$	-
[Mahdavi et al., ICLR 2024]	next-token	$O(d^2 N/n)$	-
[Kajitsuka & Sato, ICLR 2024]	seq-to-seq	$O\left(d(2nN + d)\right)$	-
[Madden et al., 2024]	next-token	$O(\omega N)$	$\Omega(\omega N)$

Analyses of a one-layer Transformer (the last three rows)

($\mathbf{N}$: dataset size, $\mathbf{n}$: input sequence length, $\mathbf{d}$: dim. of input tokens, $\boldsymbol{\omega}$: vocabulary size)

Memorization Capacity in Next-token Prediction

Theorem 1 (Upper bound)

For any dataset $(\boldsymbol{X}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{X}^{(N)}, y^{(N)}) \in \mathbb{R}^{d\times n} \times [C]$, *there is a constant-width Transformer with depth* $\tilde{O}(\sqrt{N})$ *that can memorize the dataset under next-token prediction.*

- **The total number of params is also** $\tilde{O}(\sqrt{N})$**.**

➡ **In next-token prediction, the input sequence length** n **has little impact on the memorization capacity.**

Memorization Capacity in Next-token Prediction

- **Is the construction in Theorem 1 efficient?**

> Theorem 2 (Lower bound)
> *A Transformer that can memorize any dataset of size N $(\boldsymbol{X}^{(1)}, y^{(1)}), \dots, (\boldsymbol{X}^{(N)}, y^{(N)}) \in \mathbb{R}^{d \times n} \times [C]$ under next-token prediction contains at least $\Omega(\sqrt{N})$ parameters.*

➡ **Together with Theorem 1, the memorization capacity in the next-token prediction is of the order of $\sqrt{N}$ up to logarithmic factors**

Memorization Capacity in Seq-to-seq Prediction

- Similar results hold for seq-to-seq prediction as well.

Authors	Task	Upper bound	Lower bound
[Kim et al., ICLR 2023]	seq-to-seq	$\tilde{O}(n+\sqrt{nN})$	-
[Mahdavi et al., ICLR 2024]	next-token	$O(d^2N/n)$	-
[Kajitsuka & Sato, ICLR 2024]	seq-to-seq	$O(d(2nN+d))$	-
[Madden et al., 2024]	next-token	$O(\omega N)$	$\Omega(\omega N)$
[Kajitsuka & Sato, ICLR 2025]	next-token	$\tilde{O}(\sqrt{N})$	$\Omega(\sqrt{N})$
	seq-to-seq	$\tilde{O}(\sqrt{nN})$	$\Omega\left(\sqrt{\frac{nN}{\log(nN)}}\right)$

Analyses of a one-layer Transformer (rows [Mahdavi et al., ICLR 2024] through [Madden et al., 2024])

Implications

- Nearly optimal constructions have been achieved for both next-token prediction and seq-to-seq prediction.
  - Both models consist of a feed-forward + uniform self-attention + feed-forward.

subset of self-attention

$$\mathcal{F}^{(\mathrm{UA})} : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n} \boldsymbol{Z} \mapsto \boldsymbol{Z} + \boldsymbol{W}^{(O)} \boldsymbol{W}^{(V)} \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{Z}_{:,k} \underbrace{(1, \ldots, 1)}_{\in \mathbb{R}^{1 \times n}}$$

  - From a memorization capacity perspective, a single layer of uniform self-attention is sufficient.