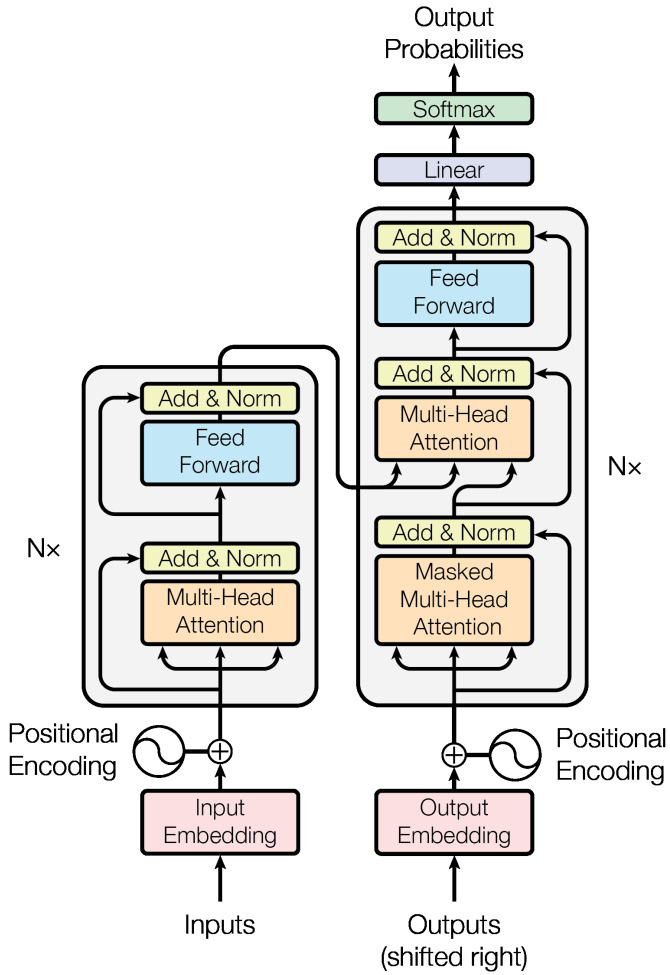
Approximation Rate of the Transformer Architecture for Sequence Modeling

Jiang Haotian

November 9, 2024

Transformer Architecture

1. ChatGPT
2. Gemini
3. Llama
4. Claude
5. ⋮



Transformer Architecture

Main components of Transformer:

$$\cdots \to \text{Attn} \to \text{FFN} \to \text{Attn} \to \text{FFN} \to \dots$$

Hypothesis space: We consider the following one layer Transformer with one attention head

$$\hat{H}_t(\boldsymbol{x}) = \hat{F}\left(\sum_{s=1}^{\tau} \sigma[(W_Q \hat{f}(x(t)))^\top W_K \hat{f}(x(s))] W_V \hat{f}(x(s))\right),$$

where $W_Q, W_K \in \mathbb{R}^{m_h \times n}$, $W_V \in \mathbb{R}^{n \times n}$ and $\hat{F}$ and $\hat{f}$ are the feed-forward components. The approximation budget is determined by n, m_h and m_{FF}.

Transformer approximation results

The following theorem shows a representation of the targets $\boldsymbol{H} \in C(\mathcal{X}, \mathcal{Y})$.

Theorem (Representation of the target space)

Consider d-dimensional, length τ input space $\mathcal{X}$ with position encoding added. Then, for any $\boldsymbol{H} \in C(\mathcal{X}, \mathcal{Y})$, there exists continuous functions $F \in C([0,1]^n, \mathbb{R})$, $f \in C(\mathcal{I}, [0,1]^n)$ and $\rho \in C(\mathcal{I} \times \mathcal{I}, \mathbb{R})$ such that for all $t \in [\tau]$ we have

$$H_t(\boldsymbol{x}) = F\left(\sum_{s=1}^{\tau} \sigma[\rho(x(t), x(\cdot))](s) f(x(s))\right),$$

where $n = 2\tau d + 1$ and σ is the softmax function.

The complexity of this target is determined by F, f and ρ.

Transformer approximation results

Target:

$$H_t(\boldsymbol{x}) = F\left(\sum_{s=1}^{\tau} \sigma[\rho(x(t), x(\cdot))](s) f(x(s))\right),$$

Model:

$$\hat{H}_t(\boldsymbol{x}) = \hat{F}\left(\sum_{s=1}^{\tau} \sigma[(W_Q \hat{f}(x(t)))^\top W_K \hat{f}(x(s))] W_V \hat{f}(x(s))\right).$$

Consider the attention matrix ,

$$\hat{\rho}(x(t), x(s)) = (W_Q \hat{f}(x(t)))^\top W_K \hat{f}(x(s)) = \sum_{k=1}^{m_h} \hat{\phi}_k(x(t)) \hat{\psi}_k(x(s)).$$

Consider the POD decomposition of ρ in the target

$$\rho(t, s) = \sum_{k=1}^{\infty} \sigma_k \phi_k(t) \psi_k(s).$$

Transformer approximation results

Theorem (Jackson-type approximation rates for the Transformer)

Suppose $\sigma_k \leq C\, k^{-\beta}$ and F, f, ρ is approximated by neural networks with rate $\mathcal{O}(m_{FF}^{-\alpha})$, then

$$\inf_{\hat{\boldsymbol{H}} \in \mathcal{H}^{(n, m_h, m_{FF})}} \|\boldsymbol{H} - \hat{\boldsymbol{H}}\| \leq \tau^2 \left(\frac{C_\beta(\boldsymbol{H})}{{m_h}^{2\beta-1}} + \frac{C_\alpha(\boldsymbol{H}) m_h^{\alpha+1}}{m_{FF}^{\alpha}} \right).$$

The target can be efficiently approximated by Transformers if ρ have low POD rank (small C_β) and each components F, G, f can be easily approximated by feed-forward networks (small C_α).

Implications

Temporal ordering is not important for Transformers.

Proposition

Let $\boldsymbol{H} \in \mathcal{C}^{(\alpha,\beta)}$ and p be a fixed permutation. Suppose $\tilde{\boldsymbol{H}}$ is defined by $\tilde{H}_t(\boldsymbol{x} \circ p) = H_t(\boldsymbol{x})$. Then $\tilde{\boldsymbol{H}}$ have same complexity measures with $\boldsymbol{H}$.

Experiment results:

	CIFAR10 (*Acc.*)	ENG-DE (*BLUE*)
Original	0.98	26.85
Altered	0.96	25.91

Table: Numerical results of the Transformer on original target and altered target. The altered target is constructed by permuting the entire input dataset while keeping the output unchanged.

Implications

Temporal mixing affect Transformers.
Suppose $\tilde{\boldsymbol{H}}$ is defined by $\tilde{H}_t(\boldsymbol{x}) = H_t(\theta * \boldsymbol{x})$. In other words we assume there are some mixing of the input tokens in time.

Observation

The complexity measure is affected when temporal mixing exists.

However, it is hard to determine how the complexity measures exactly change. It may increase or decrease depend on the filter θ.

This implies that a convolution on the inputs might potentially improve performance of Transformers.