Impact of Positional Encoding: Clean and Adversarial Rademacher Complexity for Transformers under In-Context Regression

Weiyi He, Yue Xing

Michigan State University

Introduction

Motivation:

- Positional encoding (PE) is a core component of Transformers, but its role in generalization and robustness is still not well understood.
- We study this question in a single-layer Transformer under in-context regression and analyze both standard and adversarial Rademacher complexity.

Definitions and Setup:

Notations:

ICL passes a prompt with their labels $\{(x_i, y_i)\}_{i=1,\dots,t}$ to the transformer.
Prompt format:

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_t & x_q \\ y_1 & y_2 & \dots & y_t & 0 \end{pmatrix}^\top \in \mathbb{R}^{(t+1)\times(d+1)}$$

Model Structure:

We consider a single layer and single head Transformer with trainable PE. The final input representation is $H = XW_{\text{in}} + P$.
The attention head is denoted as:

$$H' = \sigma\left(\text{RowSoftmax}\left(\frac{HW_{QK}H^\top}{\sqrt{d_m}}\right)HW_V\right)$$

Data Generation:

In each prompt, the example paires and the query are i.i.d form model:

- The input $x \sim \mathcal{N}(0, I_d)$.
- The output $y = \mu^\top x$.
- The true cofficients $\mu \in \mathbb{R}^d$ are the same for all samples within a prompt but are drawn independently for each prompt from $\mu \sim \mathcal{N}(0, I_d/d)$

Main Theoretical Results

Rademacher Complexity for Transformers without PE

For the no-PE model, the funtion-class complexity is bound by:

$$\text{Rad}_S(\mathcal{F}_{S_t}) \lesssim O\left(\frac{L_f\sqrt{rD}\sqrt{\log t/r}}{\sqrt{mt}}\right)$$

- The bound decays with context length.
- This formalizes the intuition that longer context improves in-context generalization.

Rademacher Complexity with PE

When PE is fully trainable, the complexity bound becomes:

$$\text{Rad}_S(\mathcal{F}_{S_t,\text{PE}}) \lesssim O\left(\frac{L_f\sqrt{rD}\sqrt{\log t/r}}{\sqrt{mt}}\right) + \frac{C_{PE}\sqrt{d}}{\sqrt{m}}$$

- Trainable PE adds an extra additive term, which acts as an upward shift of the clean complexity bound.
- At the context length grows, the context-dependent term decays, but the PE-induced term remains.
- This shows that the price of learnable positional information is a larger generalization gap.

Adversarial Rademacher Complexity for Transformers

To handle adversarial loss, we upper-bound it with a surrogate loss and derive an Adversarial Rademacher Complexity bound. For the no-PE model,

$$\text{Rad}_{S'}(\tilde{\mathcal{L}}_{\mathcal{F}_{S_t'}}) \lesssim O\left(\frac{L_h L_f\sqrt{rD}\sqrt{\log t/r}}{\sqrt{mt}}\right)\cdot\Phi(\varepsilon,t,d) \quad \text{with} \quad \Phi(\varepsilon,t,d) = \frac{1}{1-\sqrt{d/t}-\varepsilon/\sqrt{t}}$$

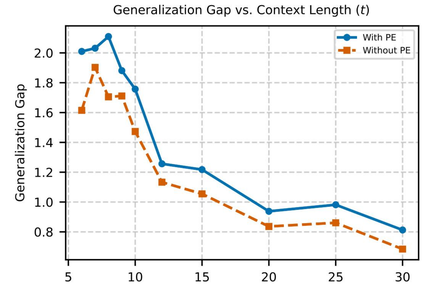
Under adversarial perturbations, the trainable-PE model incurs an additional penalty: $\frac{C_{PE}\sqrt{d}}{\sqrt{m}} + \frac{C_{PE}\varepsilon}{\sqrt{m}}$.

- In the clean setting, the PE gap is $\Delta_{\text{clean}} \approx C_{PE}\frac{\sqrt{d}}{\sqrt{m}}$, while under attack, the gap becomes $\Delta_{\text{adv}} \approx C_{PE}\frac{\sqrt{d}}{\sqrt{m}} + C_{PE}\frac{\varepsilon}{\sqrt{m}}$
- Therefore, trainable PE is not only costly in clean generalization, but even more costly under attack.

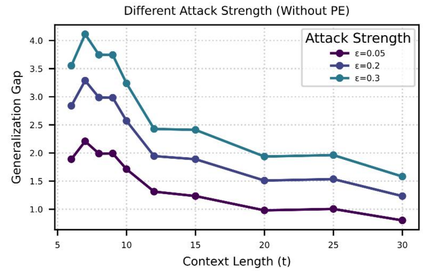
Proof Idea

- Use Dudley's entropy integral to bound clean Rademacher Complexity.
- Reduce function-class complexity to the geometry of an effective parameter space.
- Connect the Transformer to an effective linear predictor under ICL.
- In the adversarial setting, introduce a surrogate loss and quantify attack amplification through $\Phi(\varepsilon,t,d)$.

Experimental Results



Models with trainable PE consistently exhibit a larger generalization gap than models without PE. This matches the clean RC theory: trainable PE adds an upward complexity shift, while both cures still improve with larger context length.



As attack strength increases, the generalization gap increases; as context length increases, the gap decreases. Stronger attacks amplify vulnerability, while longer context still improves robustness even in the adversarial setting.

Acknowledgement

Weiyi He is supported by Open Philanthropy. Yue Xing is supported by NSF DMS 2515194, Open Philanthropy, NVIDIA Academic Grant Program and Google Cloud Research Credit.