On the Generalization Ability of Next-Token-Prediction Pretraining

Zhihao Li [1] Xue Jiang [2,3] Liyuan Liu [1] Xuelin Zhang [1] Hong Chen [1] Feng Zhen [2]

[1]Huazhong Agricultural University [2]Southern University of Science and Technology [3]Hong Kong Baptist University

Abstract

Large language models (LLMs) have demonstrated remarkable potential in handling natural language processing (NLP) tasks and beyond. LLMs usually can be categorized as transformer decoder-only models (DOMs), utilizing Next-Token-Prediction (NTP) as their pre-training methodology. Despite their tremendous empirical successes, the theoretical understanding of how NTP pre-training affects the model's generalization behavior is lacking. To fill this gap, we establish the fine-grained generalization analysis for NTP pre-training based on Rademacher complexity, where the dependence between tokens is also addressed. Technically, a novel decomposition of Rademacher complexity is developed to study DOMs from the representation learner and the token predictor, respectively. Furthermore, the upper bounds of covering number are established for multi-layer and multi-head transformer-decoder models under the Frobenius norm, which theoretically pioneers the incorporation of mask matrix within the self-attention mechanism. Our results reveal that the generalization ability of NTP pre-training is affected quantitatively by the number of token sequences N, the maximum length of sequence m, and the count of parameters in the transformer model Θ. Experiments on public datasets verify our theoretical findings.

Background

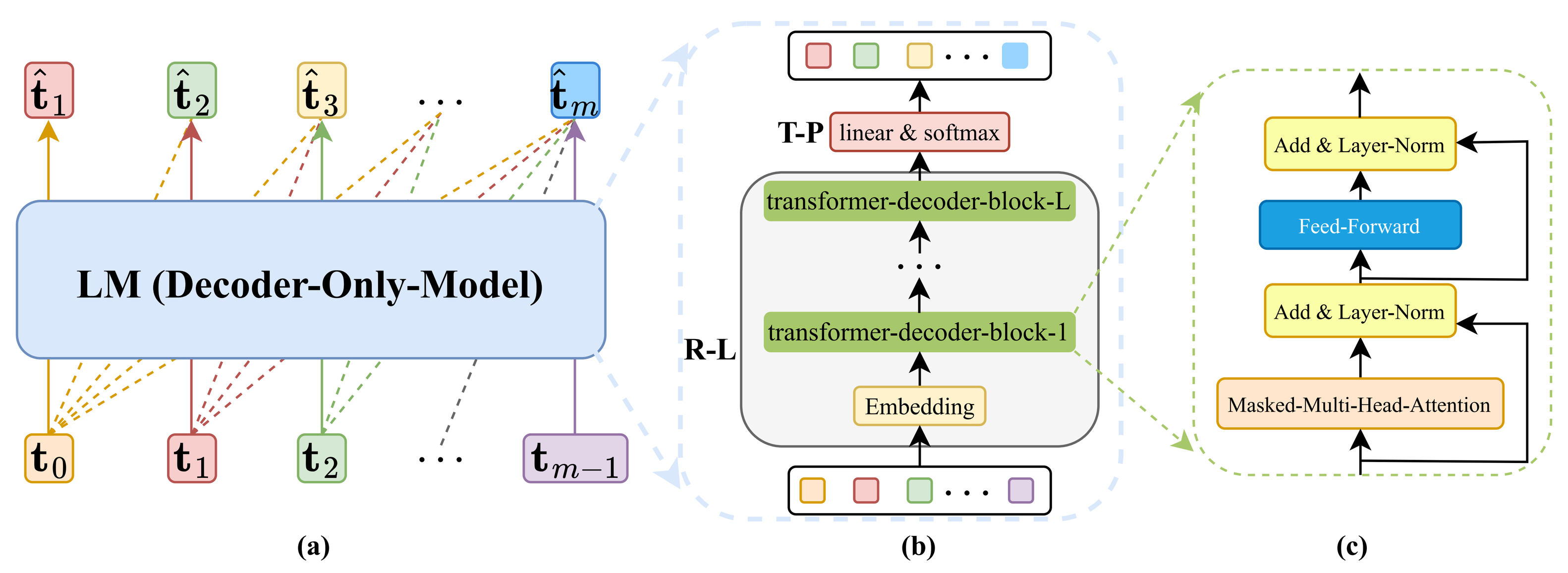

Figure 1. How NTP works utilizing decoder-only model (DOM).

Decoder-only large language models (LLMs) like GPT-3, Llama, and Qwen universally employ Next-Token Prediction (NTP) during pre-training. Despite empirical successes, a rigorous theoretical understanding of NTP's generalization mechanisms remains unexplored. Prior studies demonstrate NTP's empirical efficacy but lack rigorous analysis:

- Shlegeris et al. (2022) [1] show LLMs outperform humans on NTP tasks
- Malach et al. (2023) [2] prove linear predictors fit Chain-of-Thought data
- Bachmann et al. (2024) [3] reveal NTP's limitations in planning tasks

No theoretical framework explains how model parameters enable generalization.

Contributions

- **A novel Rademacher complexity decomposition method:** We consider the dependence between tokens and provide a theoretical framework for NTP pre-training. We establish the Rademacher complexity upper bounds of excess risk by a novel Rademacher complexity decomposition method.
- **A refined covering number for multi-layer, multi-head transformer-decoder models:** We establish bounds for the covering number of a function space derived from a multi-layer, multi-head transformer-decoder model based on masked-self-attention.
- **A generalization bound for DOMs-based NTP pre-training:** We use the Rademacher complexity upper bound and covering number to establish the generalization theory of DOMs-based NTP pre-training.

Preliminaries

Let $\mathcal{T}$ be a token set with vocabulary size $n_v = |\mathcal{T}|$. Given a pre-training dataset $D = \{\mathbf{X}_i\}_{i=1}^N \subseteq \mathcal{X}$ of sequences sampled i.i.d. from $\mathcal{D}$, each sequence $\mathbf{X}_i = \{\mathbf{t}_1^i, \cdots, \mathbf{t}_m^i\} \subseteq \mathcal{T}$ has fixed length m after preprocessing. The context for $\mathbf{t}_j^i$ is $\mathbf{T}_j^i = \{\mathbf{t}_0^i, \cdots, \mathbf{t}_{j-1}^i\}$ with $\mathbf{t}_0^i$ as a fixed begin token.

Next-Token-Prediction The model $\mathbf{LM} : \mathcal{X} \times \mathcal{T} \to \mathcal{T}$ maps context $\mathbf{T}_{j-1}$ and token $\mathbf{t}_{j-1}$ to prediction $\hat{\mathbf{t}}_j = \mathbf{LM}(\mathbf{T}_{j-1}, \mathbf{t}_{j-1})$. This decoder-only model decomposes as $\mathbf{LM} = g \circ h$ where: - $h : \mathcal{T} \to \mathcal{I}$ (Representation-Learner) - $g : \mathcal{I} \to \mathcal{T}$ (Token-Predictor). The empirical risk for $\mathbf{X}_i$ is:

$$\hat{\mathcal{L}}_{\mathbf{X}_i}(g \circ h) = \frac{1}{m}\sum_{j=1}^{m} \ell\left(g(h(\mathbf{T}_{j-1}^i, \mathbf{t}_{j-1}^i)), \mathbf{t}_j^i\right) \quad (1)$$

with cross-entropy loss ℓ. Pre-training minimizes:

$$\min_{g,h} \hat{\mathcal{L}}_D(g \circ h) = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathcal{L}}_{\mathbf{X}_i}(g \circ h) \quad (2)$$

The excess risk is $\mathcal{E}_{\mathcal{D}}(\hat{g}, \hat{h}) = \mathcal{L}_{\mathcal{D}}(\hat{g} \circ \hat{h}) - \min_{g,h} \mathcal{L}_{\mathcal{D}}(g \circ h)$ for optimal $\hat{g}, \hat{h}$.

Decoder-only Models For layer l with weights $\mathcal{W}^l$: $\mathbf{Z}^l = \Pi_{\text{norm}}\left(\sigma(\mathbf{Y}^l \mathbf{W}_{F1}^l)\mathbf{W}_{F2}^l + \mathbf{Y}^l\right)$ where $\mathbf{Y}^l = \Pi_{\text{norm}}\left(\sum_{h=1}^{H} \mathbf{A}_h^l \mathbf{W}_{O_h}^l + \mathbf{Z}^{l-1}\right)$ and $\mathbf{A}_h^l = \text{softmax}\left(\frac{\mathbf{Q}_h^l(\mathbf{K}_h^l)^\top + \mathbf{M}}{\sqrt{d_k}}\right)\mathbf{V}_h^l$ with $\mathbf{M}_{ij} = 0$ if $j \le i$ else $-\infty$. The T-P is: $g(h(\mathbf{Z})) = \text{softmax}\left(h(\mathbf{Z})\mathbf{W}^P\right)$.

Assumptions

Assumption 1 Assume that $\mathbf{X}_i = \{\mathbf{t}_1^i, \cdots, \mathbf{t}_m^i\}$ is generated by a φ-mixing distribution ϕ_i for all i, and there exists an unknown distribution $\mathcal{U}$ such that $U = \{\phi_i\}_{i=1}^N \sim \mathcal{U}$.

Assumption 2 There exists a constant $B_\ell \in \mathbb{R}^+$ satisfying $|\ell(\hat{\mathbf{t}}, \mathbf{t})| \le B_\ell$ for any $\hat{\mathbf{t}}, \mathbf{t} \in \mathcal{T}$, and ℓ is G_ℓ-Lipschitz w.r.t. $\hat{\mathbf{t}}$.

Assumption 3 (1) Π_{norm} is G_π-Lipschitz with the ℓ_2-norm, i.e., $\forall \mathbf{t}_1, \mathbf{t}_2 \in \mathbb{R}^d$, $\|\Pi_{\text{norm}}(\mathbf{t}_1) - \Pi_{\text{norm}}(\mathbf{t}_2)\|_{\ell_2} \le G_\pi \|\mathbf{t}_1 - \mathbf{t}_2\|_{\ell_2}$. (2) $\forall l \in [L]$ and $h \in [H]$, there exists constants C_l such that $\|\mathbf{Q}_h^l(\mathbf{K}_h^l)^\top/\sqrt{d_k}\|_{\ell_\infty} \le C_l$. (3) $\forall l \in [L]$, $\mathbf{W}^l \in \mathcal{W}^l$, there exists constants B_l satisfying $\|\mathbf{W}^l\|_F \le B_l$.

Main Results

Proposition 1 Let $\mathcal{F} : \mathcal{Z} \to \mathbb{R}$ be a composite function satisfying $\mathcal{F} = \ell \circ \mathcal{G} \circ \mathcal{H}$, where ℓ is a loss function and $\mathcal{H}, \mathcal{G}$ are function classes. Given a sample set $S = \{z_1, ..., z_n\} \subseteq \mathcal{Z}$, for any $g \in \mathcal{G}$ satisfying G_g-Lipschitz w.r.t. $h \in \mathcal{H}$ and ℓ satisfying G_ℓ-Lipschitz w.r.t. $g \circ h \in \mathcal{G} \circ \mathcal{H}$, we have

$$\hat{\mathfrak{R}}_S(\ell \circ \mathcal{G} \circ \mathcal{H}) \le G_\ell G_g \hat{\mathfrak{R}}_S(\mathcal{H}) + G_\ell \hat{\mathfrak{R}}_S(\mathcal{G} \circ \hat{h}),$$

where $\hat{h}$ is any given function in $\mathcal{H}$.

Theorem 1 Given a pre-training dataset D containing N token sequences $\{\mathbf{X}_i\}_{i=1}^N \subseteq \mathcal{X}$, satisfying the distribution conditions in **Assumption 1**. Denote $\hat{g}$ and $\hat{h}$ as the optimal R-L and T-P derived by solving eqn-2, respectively. Then, under **Assumption 2**, for some $\varphi_0 > 0$, $\varphi_1 > 0$ and $r > 0$, there holds

$$\mathcal{E}_{\mathcal{D}}(\hat{g}, \hat{h}) \le \underbrace{6\hat{\mathfrak{R}}_D(\ell \circ \mathcal{G} \circ \mathcal{H}) + B_\ell\sqrt{\frac{8\ln\frac{4}{\delta}}{N}}}_{\text{I}} + \underbrace{B_\ell\sqrt{\frac{\|\Delta_m\|_\infty^2 \log\frac{2}{\delta}}{2m}}}_{\text{II}} + 4B_\ell\,\text{disc}(U),$$

with probability at least $1-\delta$, where $\|\Delta_m\|_\infty \le 1 + 2\sum_{k=1}^{m}\varphi(k)$ and $\varphi(k) \le \varphi_0 \exp(-\varphi_1 k^r)$.

Main Results

Theorem 2 Let $D = \{\mathbf{X}_i\}_{i=1}^N$ be a a dataset containing N token sequences and let $\mathbf{Z}_{[N]} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_N] \in \mathbb{R}^{Nm \times n_v}$ be the input matrix generated from D, and denote $\mathbf{Z}_{[N]}^0 \in \mathbb{R}^{Nm \times d}$ as the embedded matrix. The function class of the R-L can be defined as

$$\mathcal{H} := \left\{\mathbf{Z} \mapsto h(\mathbf{Z}) : \|\mathbf{W}^l\|_F \le B_l, \mathbf{W}^l \in \mathcal{W}^l, \forall l \in [L]\right\}.$$

Then, denote $\Theta \approx 12Ld^2$ as the number of model parameters and under **Assumption 3**:

$$\ln \mathcal{N}(\mathcal{H}, \epsilon, \|\cdot\|_F) \le \frac{\Theta H}{L}\sum_{l=1}^{L} \ln\left(1 + \frac{LB_l^2 s_L \|\mathbf{Z}_{[N]}^0\|_F}{\epsilon}\right).$$

Theorem 3 Let $\mathbf{Z}_{[N]} \in \mathbb{R}^{Nm \times n_v}$ be the input sequences generated from dataset D. Then, there exists a constant $C_{\varphi,r} > 0$ such that the following inequality holds with probability at least $1-\delta$:

$$\mathcal{E}_{\mathcal{D}}(\hat{f}, \hat{h}) \lesssim \mathcal{O}\left(\sqrt{\frac{\Theta d H \tau_1}{Nm}}\right) + G_\ell\sqrt{\frac{d n_v}{Nm}} + B_\ell\left(\sqrt{\frac{8\ln\frac{4}{\delta}}{N}} + \sqrt{\frac{C_{\varphi,r}\log\frac{2}{\delta}}{2m}} + 4\,\text{disc}(U)\right),$$

Experiments

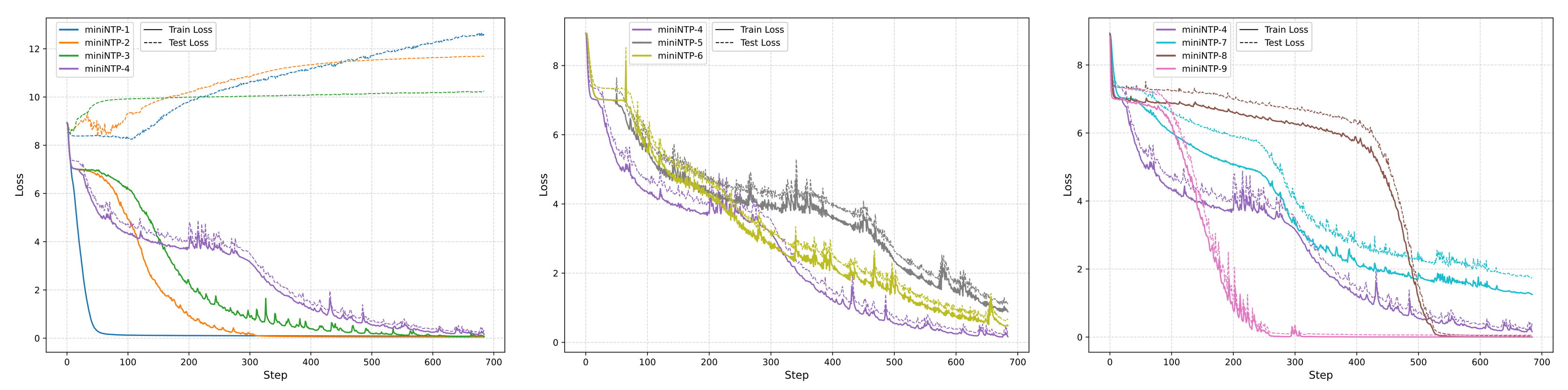

Figure 2. Experiments on MiniMind and DAMO_NLP datasets.

Table 1. Model architectures, training data specifications, hyperparameter configurations, and test PPL ($m = 512$).

Model	Θ	L	H	d	m	$N\%$	Batch Size	Learning Rate	PPL
miniNTP-1	0.029B	8	8	512	64	100	0.5M	5.0e-4	316024.25
miniNTP-2	0.029B	8	8	512	128	100	0.5M	5.0e-4	130613.71
miniNTP-3	0.029B	8	8	512	256	100	0.5M	5.0e-4	24343.04
miniNTP-4	0.029B	8	8	512	512	100	0.5M	5.0e-4	1.49
miniNTP-5	0.029B	8	8	512	512	50	0.5M	5.0e-4	3.17
miniNTP-6	0.029B	8	8	512	512	75	0.5M	5.0e-4	1.95
miniNTP-7	0.002B	6	4	128	512	100	0.5M	1.0e-3	5.76
miniNTP-8	0.09B	12	12	768	512	100	0.5M	6.0e-4	1.13
miniNTP-9	0.31B	24	16	1024	512	100	0.5M	3.0e-4	1.05

References

[1] B. Shlegeris, F. Roger, L. Chan, and E. McLean, "Language models are better than humans at next-token prediction," *arXiv preprint arXiv: 2212.11281*, 2022.

[2] E. Malach, "Auto-regressive next-token predictors are universal learners," *arXiv preprint arXiv: 2309.06979*, 2023.

[3] G. Bachmann and V. Nagarajan, "The pitfalls of next-token prediction," *arXiv preprint arXiv: 2403.06963*, 2024.