On the Power of Convolution Augmented Transformer

Mingchen Li[1], **Xuechen Zhang**[1], **Yixiao Huang**[2], **Samet Oymak**[1],

[1]University of Michigan
[2]UC Berkeley
milii@umich.edu, zxuechen@umich.edu, yixiaoh@berkeley.edu, oymak@umich.edu

Abstract

The transformer architecture has catalyzed revolutionary advances in language modeling. However, recent architectural recipes, such as state-space models, have bridged the performance gap. Motivated by this, we examine the benefits of Convolution-Augmented Transformer (CAT) for recall, copying, and length generalization tasks. CAT incorporates convolutional filters in the K/Q/V embeddings of an attention layer. Through CAT, we show that the locality of the convolution synergizes with the global view of the attention. Unlike comparable architectures, such as Mamba or transformer, CAT can provably solve the associative recall (AR) and copying tasks using a single layer while also enjoying guaranteed length generalization. We also establish computational tradeoffs between convolution and attention by characterizing how convolution can mitigate the need for full attention by summarizing the context window and creating salient summary tokens to attend. Evaluations on real datasets corroborate our findings and demonstrate that CAT and its variations indeed enhance the language modeling performance.

1 Introduction

The attention mechanism is the central component of the transformer architecture (Vaswani et al. 2017) which empowers modern large language models. Through the self-attention layer, all pairs of tokens get to interact with each other which equips the model with a global view of the context window. On the other hand, without positional-encoding (PE), self-attention lacks *locality*. For instance, without PE or causal masking, self-attention layer is permutation-equivariant and does not distinguish between nearby vs distant tokens. In contrast, convolution operator is a well-established primitive that facilitates local feature aggregation based on relative positions and provides a natural alternative to PE. While convolution has enjoyed major success in vision during the last three decades, its explicit use in language modeling is relatively recent (Dauphin et al. 2017). On the other hand, there is a growing recent interest in using convolution-based blocks in language models: For instance, state-space models (SSM) (Gu, Goel, and Ré 2021) and linear RNNs (Orvieto et al. 2023) are efficient parameterizations of long convolutional filters. These models have enjoyed significant success in long-range sequence modeling as they provide fast inference and parallelizable training. However, purely convolutional architectures are known to suffer from recall capability as they lack the global view of the context window (Arora et al. 2024). These insights motivated a recent push toward hybrid architectures (De et al. 2024; Arora et al. 2024; Park et al. 2024; Arora et al. 2023) that combine the strengths of both attention and convolution-like approaches, including short convolutional filters, SSMs, linear RNNs, or Mamba.

In this work, we explore the synergy between attention and convolution which reveals new theoretical principles that inform hybrid architecture design. Specifically, we introduce an intuitive hybrid architecture called *Convolution-Augmented Transformer* (CAT)[1]. CAT incorporates convolutional filters to the K/Q/V embeddings of the attention layer as depicted on the left hand side of Figure 2. We explore the capabilities of the CAT layer through mechanistic tasks including associative recall (AR), selective copying (Gu and Dao 2023; Jing et al. 2019), and length generalization. For instance, AR is a fundamental task motivated from the associative memory in cognitive science (Ba et al. 2016). This task underpins critical applications such as bigram retrieval, where a specific sequence, such as 'Rings' following 'The Lord of the', must be correctly retrieved. It is also a generalization of the induction head task (Olsson et al. 2022) and recent research has shown the ability of LLMs to solve mechanistic tasks is highly correlated with their real performance in NLP tasks (Olsson et al. 2022; Fu et al. 2022; Arora et al. 2024; Nichani, Damian, and Lee 2024; Poli et al. 2024).

We theoretically and empirically show that, within the CAT layer, attention and convolution exhibit strong synergy and complementarity to solve these mechanistic tasks while enjoying length generalization benefits. As a concrete example, the left side of Figure 1 displays the AR performance for various test-time sequence lengths. As the sequence length grows, we observe two distinct failure modes: Mamba's accuracy degrades due to its finite state dimension whereas attention-only models degrade due to the length extension bottlenecks of PE. In contrast, CAT maintains perfect accuracy and length

[1]The transformer architecture consists of attention and MLP layers. For theoretical analysis and synthetic experiments, we will entirely focus on the *Convolution Augmented Attention* layer described in Fig. 2. For this reason, we will use the CAT acronym to refer to both Convolution-Augmented Transformer and Attention.

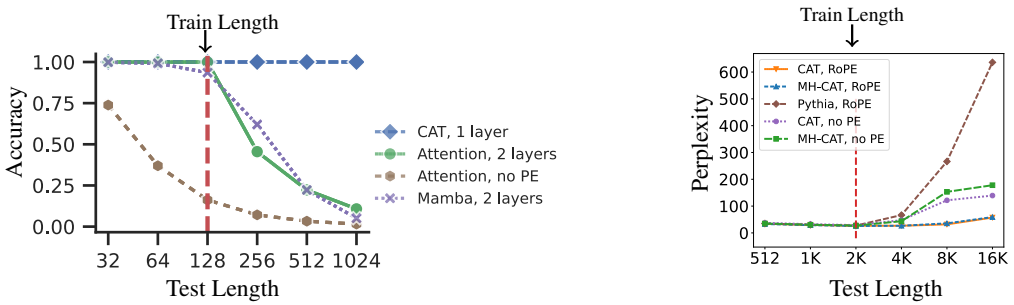

Figure 1: Evaluations on synthetic and real data. The models are trained on 128 and 2,048 context length (vertical dashed lines) and tested on varying context lengths respectively. **Left figure:** We conduct synthetic experiments on the Associative Recall task and contrast 1-layer CAT with 2-layers of alternative architectures. The embedding dimension is 128. We find that CAT is the only model that solves AR with length generalization in line with our theory (also see Fig. 6). **Right figure:** Evaluations on language modeling where we pretrain CAT models by equipping Pythia with short convolutions (window size 21) on SlimPajama dataset. Convolution allows the model to pretrain without positional encoding and further improves perplexity when combined with RoPE. Importantly, it also generalizes to longer context lengths more robustly with or without RoPE. For length generalization, we used YaRN (Peng et al. 2023) which incorporates position interpolation (Chen et al. 2023) (for RoPE only) and temperature scaling (see Sec. 6.2).

generalization because attention and convolution patch these failure modes in a complementary fashion. Overall, we make the following contributions:

- We propose the convolution-augmented attention layer and prove that it can solve the N-gram AR (NAR) and Selective Copying tasks using a single layer (Theorems 1 and 4). Comparison to alternatives (Mamba, Based, attention, linear attention) reveals that CAT can uniquely solve NAR with length generalization.

- To explain this, we establish a length generalization result on the loss landscape (Theorem 2): Under mild assumptions, all CAT models that solve AR for a particular length provably generalize to all other context lengths.

- We evaluate CAT on real data and demonstrate that even 1-dimensional short convolutions noticeably aids language modeling: In line with theory, convolution enables the model to train stably without PE and improves length generalization. We also develop a multihead version of CAT which yields further improvements (see Table 2).

- We show that long convolutions, such as SSMs, bring the benefit of *context summarization* and mitigates the need for dense attention: We describe the Landmark CAT model (following Landmark Attention (Mohtashami and Jaggi 2023)) which first creates landmark/summary tokens through convolution and then attends on these landmarks to efficiently locate the most relevant subsets of the input sequence (Sec. 5). Focusing on the AR problem, we characterize fundamental tradeoffs between the embedding dimension, amount of summarization, and the sparsity of attention. Through these, we show that the use of long convolutions can provably enable the success of sparse/cheaper attention.

2 Related Works

Convolution-like sequence models. Gated-convolutions (Dauphin et al. 2017) and state-space models, such as S4 (Gu, Goel, and Ré 2021), utilize long convolutions to reduce the computational demands associated with attention mechanisms. Performance enhancements have also been achieved through novel filter parametrization (Gupta, Gu, and Berant 2022; Gu et al. 2022). Despite these innovations, challenges in Multi-query Associative Recall (MQAR) prompted the development of input-dependent convolution techniques. Notable developments in this area include, Liquid S4 (Hasani et al. 2022), Mamba (Gu and Dao 2023; Dao and Gu 2024) and (Yang et al. 2019; Kosma, Nikolentzos, and Vazirgiannis 2023) where convolution filters are directly parametrized by inputs and include correlation terms between input tokens to enhance state mixing. (Li et al. 2022) empirically explores the reason underlying the success of convolutional models.

Expressivity, recall, length generalization. Recent works (Arora et al. 2024; Jelassi et al. 2024; Arora et al. 2023; Fu et al. 2022) explore the limitations of purely convolutional models, including Mamba, and demonstrate that, these models inherently lack the capability to solve recall problems unless they have large state dimensions (i.e. memory). (Jelassi et al. 2024) provides a construction for 2-layer self-attention to solve AR with length generalization. Interestingly, this construction uses Hard Alibi, which is a variation of Alibi PE (Press, Smith, and Lewis 2021) that utilize explicit linear biases in attention. Their Hard Alibi restricts the attention layer to focus on and aggregate only the recent N tokens. In this regard, this construction is related to our short convolution. However, while this work is constructive, we also prove that CAT has good loss landscape and CAT solutions to AR provably length generalize. It has been observed that PE can hurt length generalization and reasoning. In fact, (Kazemnejad et al. 2024) has found NoPE to be viable. On the other hand, in our real data evaluations, we have found pure NoPE to

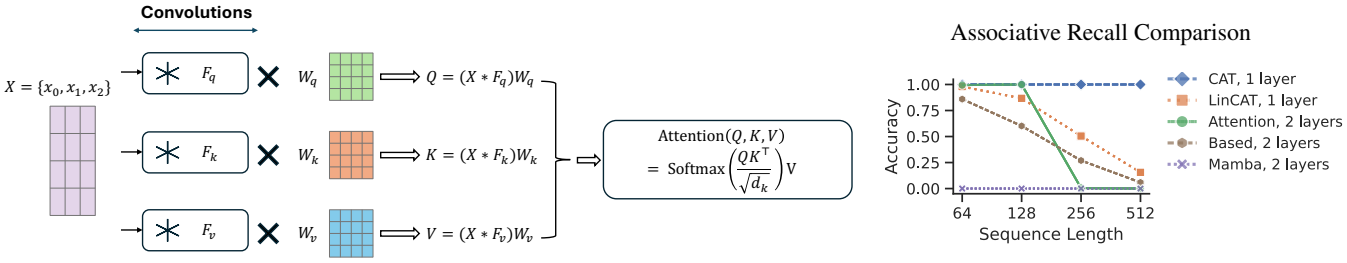

Figure 2: **Left figure**: Illustration of the Convolution-Augmentated Attention (CAT) block, where separate filters are applied to the K/Q/V embeddings, before self-attention (see Sec. 3.1 for details). **Right figure**: Performance of 1-layer CAT models trained on multi-query AR (MQAR, see Sec. 3.2 for details) tasks with model embedding dimension 64 and varying sequence length. The LinCAT replaces the standard attention in CAT with linear attention. We observe that the CAT model outperforms the baseline models across all sequence lengths with only 1 layer compared to 2 layers baselines.

be highly brittle as it either fails to converge or optimization is unreasonably slow. Our AR experiments also corroborate that NoPE by itself is indeed not a viable strategy.

Hybrid architectures and mechanistic design. There is a growing interest in integrating different language modeling primitives to obtain best-of-all-world designs. To this end, mechanistic tasks such as AR, copying, and in-context learning have been important to demystify the functionalities of language models (Olsson et al. 2022; Park et al. 2024). Recent research has shown that performance on mechanistic tasks highly correlates with the model's ability to generalize to real-world tasks, thus have been utilized to guide architecture design in LLMs (Arora et al. 2023; Poli et al. 2024). Gating mechanisms have been integrated within convolutional frameworks to enhance the model's selectivity (Zhang et al. 2024). Models employing gating functions, have shown substantial improvements in AR tasks (Fu et al. 2022; Poli et al. 2023). Additionally, recent innovations on hybrid architecture, such as BaseConv (Arora et al. 2023, 2024), GLA (Yang et al. 2023), MambaFormer (Park et al. 2024), and (Ma et al. 2024, 2022; Ren et al. 2024) have provided more effective solutions to AR tasks. This comprehensive foundation of hybrid architectures informs our exploration into the convolution-attention synergy.

3 Problem Setup

3.1 Convolutional-Augmented Attention

Let us first introduce helpful notation. $\boldsymbol{I}_d$ is the identity matrix of size d. $\boldsymbol{D}_i$ denotes the causal delay filter that shifts a signal $\boldsymbol{x}$ i-timesteps forward i.e. $(\boldsymbol{x} * \boldsymbol{D}_i)_j = \boldsymbol{x}_{j-i}$. For an integer $n \geq 1$, we denote the set $\{0, \ldots, n-1\}$ by $[n]$. We use lowercase and upper-case bold letters (e.g., $\boldsymbol{m}, \boldsymbol{M}$) to represent vectors and matrices. m_i denotes the i-th entry of a vector $\boldsymbol{m}$.

Below, we introduce the Convolution-Augmented Attention layer, which incorporates learnable filters into the K/Q/V embeddings. Let $\boldsymbol{X} = [\boldsymbol{x}_0 \ \ldots \ \boldsymbol{x}_{L-1}]^\top \in \mathbb{R}^{L \times d}$ denote the input to the layer containing L tokens with embedding dimension d. Let $\boldsymbol{F} \in \mathbb{R}^W$ denote the convolutional filter with temporal length W. We examine two convolution types which handle multi-head attention in different ways:

• **1D per-head convolution:** For each attention head, we have a distinct 1D filter $\boldsymbol{F} \in \mathbb{R}^W$. $\boldsymbol{F}$ is applied temporally to each of the d embedding dimensions. This results in $\boldsymbol{F} * \boldsymbol{X}$ where $(\boldsymbol{F} * \boldsymbol{X})_i = \sum_{j \in [W]} F_j \boldsymbol{x}_{i-j}$, with F_j being the j-th entry of $\boldsymbol{F}$.

• **Multi-head convolution:** Suppose we have H sequences $\bar{\boldsymbol{X}} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_H] \in \mathbb{R}^{L \times d \times H}$ each corresponding to one of the H attention heads. We use a filter $\bar{\boldsymbol{F}} = [\boldsymbol{F}_1, \ldots, \boldsymbol{F}_H] \in \mathbb{R}^{W \times H \times H}$. Each $\boldsymbol{F}_i$ is convolved with $\bar{\boldsymbol{X}}$ to obtain the i-th head's output of size $L \times d$.

Observe that both convolution types are identical when there is a single attention head. However, multi-head convolution is more expressive because it mixes the attention heads. In Table 2, we also examine multi-head convolution where we mix the attention maps rather than embeddings. The architecture of CAT is illustrated in Fig. 2 and is formally defined as follows:

Definition 1 (Convolution-Augmented Attention (CAT))**.** *A CAT layer incorporates learnable convolutional filters to the key/query/value embeddings. For a single-head CAT, the key embeddings are given by* $\boldsymbol{K} = (\boldsymbol{X} * \boldsymbol{F}_k)\boldsymbol{W}_k$ *with weights* $\boldsymbol{F}_k, \boldsymbol{W}_k$ *(same for query and value embeddings).*

		Input	Query	Output
Single Query	AR	a 2 c 1	a	2
	NAR	(a b) 2 (b a) q (a a) 4	b a	q
	SC	a [n] [n] c [n] k	$\perp$	a c k
Multi Query	AR	a 2 c 1	c a	1 2
	NAR	(a b) 2 (b a) q (a a) 4	(b a) (a a)	q 4

Table 1: Illustrative examples of synthetic tasks. In all AR-based tasks, keys and queries are highlighted in red and the values in green. For NAR tasks, parentheses denote N-gram queries; note that the parentheses are not part of the input. In SC tasks, signal tokens are in green and noise tokens in gray, and the model begins output when $\perp$ appears in the sequence.

3.2 Mechanistic Tasks for Language Modeling

To proceed, we describe the Associative Recall and Selective Copying tasks that will help us mechanistically study CAT. Table 1 provides an illustration of these tasks which are

adapted from the sequence modeling literature (Gu and Dao 2023; Arora et al. 2023; Poli et al. 2024; Olsson et al. 2022).

Definition 2 (Associative Recall Problem)**.** *Consider a discrete input sequence $X = [x_0, x_1, \ldots, x_{L-1}]$, with tokens drawn from a vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$. The AR problem is defined as follows: Suppose that there is a unique index i ($0 \le i < L-1$) such that $x_i = x_{L-1}$. A model f successfully solves the AR problem if $f(X) = x_{i+1}$ for all inputs X. In this problem, x_i becomes the key, x_{i+1} is the associated value, and the last token x_{L-1} is the query.*

Building on the AR problem, we introduce its N-gram variation: The model needs to identify the copy of the last N tokens in the context window and return the associated value.

Definition 3 (N-gram AR Problem)**.** *Consider a discrete input sequence $X = [x_0, x_1, \ldots, x_{L-1}]$, with tokens drawn from a vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$. Let $X_{\{i,j\}} = [x_i, x_{i+1}, \ldots, x_j]$ denote the subsequence of X from index i to j. The N-gram associative recall (NAR) problem is formulated as follows: for $X_{\{L-N,L-1\}}$ (which are the last N tokens), there exists a unique index i ($0 \le i < L-N$) such that $X_{\{i,i+N-1\}} = X_{\{L-N,L-1\}}$. A model f solves NAR if $f(X) = x_{i+N}$ for all inputs X.*

Selective copying (SC) task is originally introduced by (Jing et al. 2019) and it is utilized by the recent Mamba (Gu and Dao 2023) and Griffin (De et al. 2024) papers to assess their model's approximation capabilities. In SC, given an input sequence X containing noisy tokens, the model should denoise X and return the *signal tokens* within.

Definition 4 (Selective Copying)**.** *Consider a vocabulary $\mathcal{V}$ composed of a set of signal tokens $\mathcal{S}$, a set of noise tokens $\mathcal{N}$, and special token $\perp$ i.e. $\mathcal{V} = \mathcal{S} \cup \mathcal{N} \cup \{\perp\}$. Let X be a sequence whose tokens are drawn from $\mathcal{S} \cup \mathcal{N}$ and let $X_{\mathcal{S}}$ be the sub-sequence of X that includes all signal tokens in order. f solves* selective copying *over $\mathcal{S}$ if it autoregressively outputs $X_{\mathcal{S}}$ following the prompt $[X \perp]$ for all inputs X. f solves* unique selective copying *if it outputs all unique tokens of $X_{\mathcal{S}}$ in order for all X.*

Multi-Query Associative Recall We also introduce the multi-query versions of the AR and NAR tasks, abbreviated as MQAR and MQNAR, respectively. In these tasks, a model receives multiple queries in a single input and must generate corresponding outputs in a single forward pass, at varying positions in the sequence. This approach was first introduced in (Arora et al. 2023), which demonstrated that while the Mamba model successfully addresses AR tasks, it struggles with MQAR when operating with a limited model dimension. This highlights the increased complexity of MQAR tasks where models need to memorize more sequence information and recall queries at different positions.

Definition 5 (Multi-Query Associative Recall (MQAR))**.** *Consider a discrete input sequence $X = [x_0, x_1, \ldots, x_{L-1}]$ with tokens drawn from a vocabulary $\mathcal{V}$. Let $X_{\{i,j\}} = [x_i, \ldots, x_j]$ denote a subsequence of X from index i to j. The **multi-query N-gram associative recall (MQNAR)** problem is defined as follows: for every N-gram query $Q_k = X_{k-N+1\ldots k}$, $N \le k < L$, determine if there exists a $N \le j < k$ such that $X_{\{j-N+1,j\}} = Q_k$. If so, output the value x_{j+1} as the result, else output a special token to indicate no match is found. A model f solves MQNAR if it outputs the correct values for all N-gram queries and all inputs X. The standard **MQAR** problem is a special instance of MQNAR by setting $N = 1$.*

Table 1 provides examples of the synthetic tasks we consider in this work. Specifically, we conduct AR and NAR experiment on their multi-queiry variants to evaluate the model's ability to recall multiple queries. For the selective copying task, we generate the output auto-regressively by predicting the signal tokens in the input sequence after the special token $\perp$.

4 Provable Benefits of Convolution-Augmented Attention

Before diving into the theoretical results, we make a few clarifying remarks. We assume that all token embeddings have unit ℓ_2 norm. Secondly, a CAT layer maps each query to a vector-valued output $f(\boldsymbol{X}) \in \mathbb{R}^d$. To sample the discrete output token, we will simply return the nearest neighbor in the vocabulary of token embeddings. For associative recall problems, we will use a single head attention layer with weights $\boldsymbol{W}_q, \boldsymbol{W}_k$ are chosen as suitably scaled identity matrices. With this choice, attention essentially implements a *nearest neighbor retrieval*. It suffices for the theory thanks to the simple nature of the AR problem where we wish to identify the replica of a query within the context window. In general, we can easily contrive natural generalizations of AR and Selective Copy problems that necessitate a more sophisticated attention mechanism (see (Poli et al. 2024)). One such generalization is, given query q, we wish to retrieve a general key k (possibly $k \neq q$) and return the value associated with k.

N-gram AR. Our first result shows that a single CAT layer can solve the NAR problem under fairly general conditions.

Theorem 1 (Solving NAR)**.** *Let $\boldsymbol{F} \in \mathbb{R}^N$ be a causal 1-D convolutional filter of length N and $\texttt{norm}(\boldsymbol{X})$ normalize the rows of a matrix to unit ℓ_2 norm. Consider a single CAT layer $f(\boldsymbol{X}) = (\boldsymbol{X}_v \boldsymbol{W}_v)^\top \mathbb{S}(\boldsymbol{X}_k \boldsymbol{W}_k \boldsymbol{W}_q^\top \boldsymbol{q})$ where $\boldsymbol{q}$ is the final token of $\boldsymbol{X}_q$ and $\boldsymbol{X}_q = \texttt{norm}(\boldsymbol{X} * \boldsymbol{F}_q) \in \mathbb{R}^{L \times d}$ (same for $\boldsymbol{X}_k, \boldsymbol{X}_v$). Set $\boldsymbol{F}_q = \boldsymbol{F}$ and $\boldsymbol{W}_k = \boldsymbol{W}_q = \sqrt{c}\boldsymbol{I}_d$. Use either*

- ***Value delay:** $\boldsymbol{F}_k = \boldsymbol{F}_q$, $\boldsymbol{F}_v = \boldsymbol{D}_{-1}$ and $\boldsymbol{W}_v = 2\boldsymbol{I}_d$ or,*
- ***Key delay:** $\boldsymbol{F}_k = \boldsymbol{D}_1 * \boldsymbol{F}_q$, $\boldsymbol{F}_v = \boldsymbol{D}_0$ and $\boldsymbol{W}_v = \boldsymbol{I}_d$*

Let $\varepsilon > 0$ be the minimum ℓ_2 distance between two distinct tokens embeddings. For almost all choices of $\boldsymbol{F}$, there is a scalar $c_0 > 0$ depending on $\boldsymbol{F}$ such that, setting $c = c_0 \log(4L/\varepsilon)$, CAT layer solves the NAR problem of Def. 3 for all input sequences up to length L.

As a corollary, using a simple 1-D convolutional filter on the key embeddings solves the AR problem.

Corollary 1 (1-D CAT solves AR)**.** *Consider a CAT layer employing 1-D convolution on key embeddings with the delay filter $\boldsymbol{F}_k = \boldsymbol{D}_1 = [0\ 1\ 0\ \ldots\ 0]$ and $\boldsymbol{F}_q = \boldsymbol{F}_v = \boldsymbol{D}_0$. This model solves AR.*

Length generalization. Our next result shows that the global minima of CAT provably exhibit length generalization, thereby shedding light on the empirical benefits of CAT in Figure 1. Concretely, even if we train CAT to solve AR for a

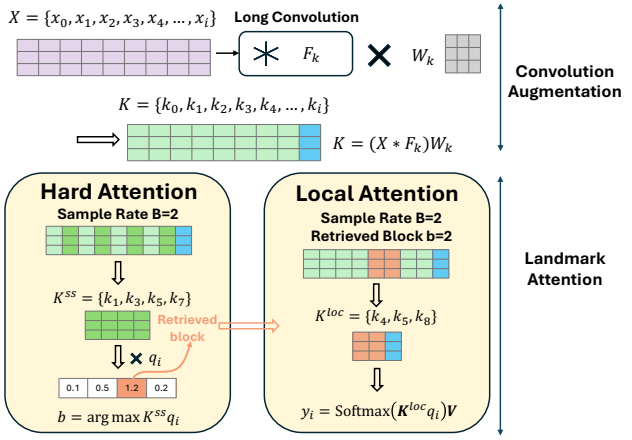

Figure 3: Illustration of the Landmark CAT. We first apply long convolution on the input sequence and subsample it to obtain landmark tokens representing individual blocks. Hard Attention computes the similarity between the query and landmarks to retrieve the most relevant block. Local Attention concatenates the retrieved block with the final block containing the query and computes the output token.

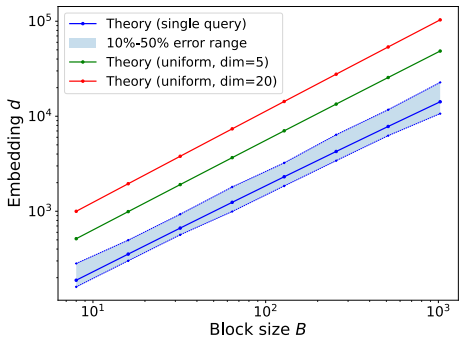

Figure 4: Behavior of the embedding dimension as a function of block size for context length $L = 2^{20} \approx 1$ million (noise level $\sigma^2 = 1$). Shaded region highlights te range of d that exhibits 10%-50% empirical success. Proposition 1 accurately captures the empirical behavior. For the success of uniform AR, we need larger d as the dimension of the query space S grows.

fixed context length, the AR capability will generalize to all other context lengths. This result is distinct from Theorem 1 because **it establishes length generalization for all CAT models** that approximately solve the AR problem for a context length, rather than constructing one such solution. The proof is provided in Section C.2.

Theorem 2 (Length generalization)**.** *Let $\boldsymbol{F}_v \in \mathbb{R}_+^{2W+1}$ be a convolutional filter from time $t = -W$ to $t = W$ where $W \le L - 1$. Consider a CAT layer of the form $f(\boldsymbol{X}) = \boldsymbol{X}_v^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_{L-1})$ where $\boldsymbol{X} \in \mathbb{R}^{L\times d}, \boldsymbol{X}_v = \boldsymbol{X} * \boldsymbol{F}_v \in \mathbb{R}^{L\times d}$ and $\boldsymbol{x}_{L-1}$ is the last token of $\boldsymbol{X}$ and $\boldsymbol{W} = \boldsymbol{W}_k \boldsymbol{W}_q^\top$. Suppose that token embeddings have unit norm. Consider any model $f = (\boldsymbol{W}, \boldsymbol{F}_v)$ that can solve the AR problem defined in Def. 2 up to ε-accuracy on all sequences of length $L \ge 3$. That is, for all $(\boldsymbol{X}, \boldsymbol{y})$ where query $\boldsymbol{x}_{L-1}$ repeats twice and $\boldsymbol{y}$ being the associated value token, we have $\|\boldsymbol{y} - f(\boldsymbol{X})\|_{\ell_2} \le \varepsilon$. Define the minimum embedding distance within vocabulary $\mathcal{V}$ as $\Delta = (1 - \max_{\boldsymbol{a}\neq\boldsymbol{b}\in\mathcal{V}} (\boldsymbol{a}^\top \boldsymbol{b})^2)^{1/2}$ and assume that $\Delta > 0$. There are absolute constants $R_0, R > 0$ such that, if $\varepsilon_0 := \varepsilon/\Delta \le R_0/L$, we have that*

- *The filter obeys $\|\boldsymbol{F}_v - \boldsymbol{D}_{-1}\|_{\ell_1} \le L\varepsilon_0$, which is in line with Theorem 1.*
- *Let $\boldsymbol{X}$ be an input sequence of length L' following Def. 2. Let $\boldsymbol{s}_\star(\boldsymbol{X}) \in \mathbb{R}^{L'}$ be the "golden attention map" with entries equal to $1/2$ at the positions of the query $\boldsymbol{x}_{L'-1}$ and 0 otherwise. For all such $\boldsymbol{X}$, the attention map of f obeys $\|\mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_{L'-1}) - \boldsymbol{s}_\star(\boldsymbol{X})\|_{\ell_1} \le L'\varepsilon_0$.*
- *For all $\boldsymbol{X}$ of length L' following Def. 2, we have that $\|\boldsymbol{y} - f(\boldsymbol{X})\|_{\ell_2} \le RL'\varepsilon_0$.*

Here it worths noting that all CAT models that approximately solve the AR problem ends up learning convolution and attention weights that are consistent with the constructive result of Theorem 1. This simple "universal solution" is in contrast to attention-only models where length generalization not only requires standard positional encoding but also additional adjustments to extend the context window of PE (Peng et al. 2023; Chen et al. 2023).

Additionally, in Appendix C.3, we generalize the length generalization result to the N-gram AR problem under slightly stronger assumptions, which is specified in Assumption 1. The reader is referred to Proposition 3. Besides showcasing the value of convolution-attention hybrids, these results also motivate future research into the optimization landscape: Under what conditions gradient methods provably converge to generalizable CAT models, namely those described in Theorem 2? Answers to such questions could build on the recent optimization theoretic results on the transformer/attention models (Tian et al. 2023; Tarzanagh et al. 2023; Deora et al. 2023; Oymak et al. 2023; Li et al. 2024; Nichani, Damian, and Lee 2024; Ildiz et al. 2024; Ataee Tarzanagh et al. 2023; Makkuva et al. 2024; Collins et al. 2024) and extend them to hybrid designs.

Selective Copy. Our next result shows that, 1-layer CAT model can solve the *unique selective copy* problem. That is, it can provably generate all signal tokens in the correct order as long as the input contains each distinct signal token at most once. Corroborating this, our experiments demonstrate that 1-layer CAT performs on par with or better than alternative architectural choices. The proof is deferred to Section C.4.

Theorem 3 (Selective Copy)**.** *Consider the setting of Def. 4. There is a 1-layer CAT using exponential-decay query-convolution (i.e. $F_{q,i} = \rho^i$) and $d = |\mathcal{S}| + 3$ dimensional token embeddings such that, it outputs all signal tokens in order for all inputs where signal tokens appear uniquely.*

Selective Copy problem is distinct from AR in the sense that, it requires a global view of the token positions as the model has to distinguish the order of the distinct signal tokens within the context window. In Theorem 4, we actually describe two ways to achieve this (see appendix for the details): The first option is using an infinitely long convolution $F_{q,i} = \rho^i$ which admits a simple parameterization as a state-

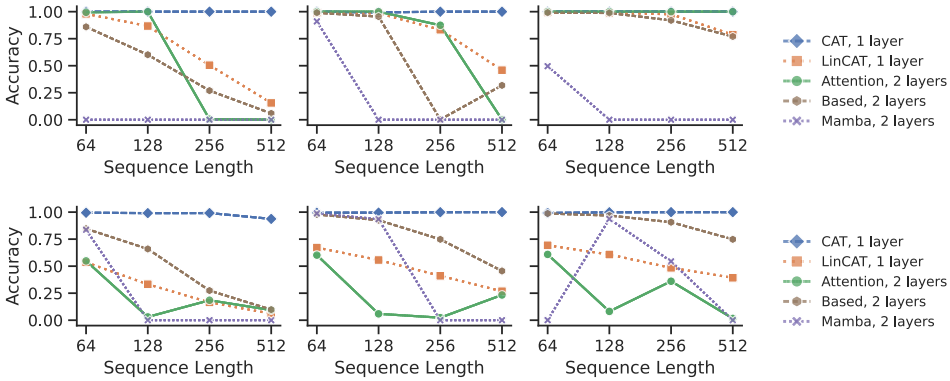

Figure 5: Evaluation of models on MQAR and MQNAR tasks with varying model dimensions and sequence lengths. Model dimensions are **32, 64, 128** for each column of the figures, from left to right. **Top:** Models trained on the MQAR setup. **Bottom:** Models trained on the MQNAR setup. Note that CAT models employ a single-layer architecture, whereas all other models utilize two layers. Refer to Section 6.1 for detailed setup descriptions.

space model (Gu, Goel, and Ré 2021). We show that this convolution choice can aggregate all signal tokens in the query embedding while distinguishing their order. This also partly explains how Mamba/SSMs are equally effective in solving Selective Copying. An alternative construction is using a short convolution together with a simple positional encoding. Here, convolution equips the query with local context (specifically the summary of the signal tokens generated so far) and PE provides the global context on the locations of remaining signal tokens. This synergy of PE and short convolution is in line with our real language modeling experiments where CAT with PE outperforms CAT without PE in terms of perplexity as well as length generalization. We defer the Selective Copy experiment to Appendix B.1.

5 Benefits of Long Convolution for Enabling Sparse-Attention

So far we have discussed the benefits of short convolutions to equip transformer with local context to solve AR and its variations. During this discussion, we have used dense attention which has exact recall capabilities thanks to its ability to scan the full context window. In this section, we ask the following: Can convolution also help mitigate the need for dense attention? Intuitively, we should be able to tradeoff the accuracy of attention computation with computation. Here, we describe how long convolutions can enable this by effectively summarizing the context window so that we can identify where to attend in (extremely) long-context settings.

Specifically, we will prove that, long convolutions (such as SSMs) allow us to utilize sparse attention while retaining (high-probability) recall guarantees. These findings complement the recent research that establish the recall limitations of purely recurrent models (Arora et al. 2024, 2023). Our theory will also shed light on the mechanics of landmark attention (Mohtashami and Jaggi 2023). While (Mohtashami and Jaggi 2023) does not rely on convolution, we will describe how convolution can generate *landmark tokens* by summarizing/hashing the chunks of the context window, and attention can efficiently solve recall by attending only to these summary tokens.

Landmark Convolutional Attention (LCAT): Figure 3 describes the LCAT block that apply on input sequence $\boldsymbol{X}$. Let $\boldsymbol{F}_k \in \mathbb{R}^L$ be the convolutional filter on keys, B be the sampling rate, and $\bar{L} = \lceil L/B \rceil$. Setting $\boldsymbol{K} = (\boldsymbol{X} * \boldsymbol{F}_k)\boldsymbol{W}_k \in \mathbb{R}^{L\times d}$, we obtain $\boldsymbol{K}^{\text{ss}} \in \mathbb{R}^{\bar{L}\times d}$ by sampling $\boldsymbol{K}$ at every B tokens. Additionally, define $\boldsymbol{X}_i$ to be the ith block of $\boldsymbol{X}$ of size B spanning tokens $(i-1)B+1$ to iB. Let $\boldsymbol{V} = (\boldsymbol{F}_v * \boldsymbol{X})\boldsymbol{W}_v$ denote the value embeddings. For a query $\boldsymbol{q}_i$ for $i \in [L]$, the LCAT layer outputs:

$$\text{(1) Hard Attention:} \quad b = \arg\max_{j \neq \lceil i/B \rceil} \boldsymbol{K}^{\text{ss}}\boldsymbol{q}_i \qquad \text{(LCAT)}$$

$$\text{(2) Local Attention:} \quad \boldsymbol{y} = \mathbb{S}(\boldsymbol{K}^{\text{loc}}\boldsymbol{q}_i)\boldsymbol{V}^l \qquad (1)$$

$$\text{where} \quad \boldsymbol{K}^{\text{loc}} = \text{concat}(\boldsymbol{K}_{\lceil i/B \rceil}, \boldsymbol{K}_b).$$

Above, *hard attention* phase aims to retrieve the correct block associated to the query. This block is merged with the local block $\lceil i/B \rceil$ that contains the query itself similar to sliding window attention. We then apply *dense local attention* on the concatenated blocks $\boldsymbol{K}^{\text{loc}}$.

Computational complexity of LCAT: For a fixed query, (LCAT) requires $O(d(L/B + B))$ computations. This is in contrast to $O(dL)$ of vanilla attention. Choosing a suitable block size (e.g. $B = O(\sqrt{L})$), this model should save up to $\times\sqrt{L}$ in computational savings. Importantly, our theory will highlight the interplay between the embedding dimension d and the allowable acceleration by characterizing the exact performance of (LCAT) under a random context model.

Definition 6 (Random Context Model)**.** *The query token $\boldsymbol{x}_L$ occurs twice in the sequence and has unit ℓ_2 norm. All other tokens of $\boldsymbol{X}$ are IID and drawn with IID $\mathcal{N}(0, \sigma^2/d)$ entries.*

The following proposition shows that, (LCAT) will solve AR if and only if $\frac{d}{2B\log\bar{L}} \geq 1 + o(1)$.

Proposition 1. *Recall $\bar{L} = \lceil L/B \rceil$ is the number of blocks. Let $\boldsymbol{W}_v = 2\boldsymbol{I}_d$, $\boldsymbol{F}_v = \boldsymbol{D}_{-1}$, and $\boldsymbol{W}_k = \boldsymbol{W}_q = \sqrt{c}\cdot\boldsymbol{I}_d$ with*

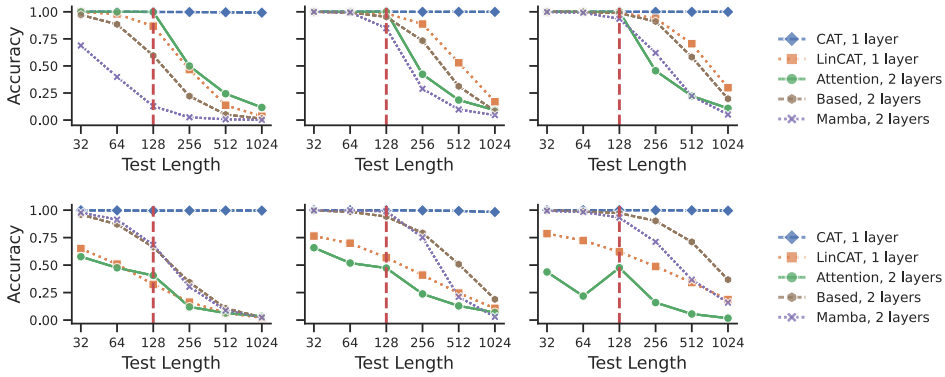

Figure 6: Evaluation of models on length generalization. Model dimensions are **32, 64, 128** for each column of the figures, from left to right. The models are trained with sequence length 128 (vertical red dashed lines) and tested on varying test length. **Top:** Models trained on the MQAR. **Bottom:** Models trained on the MQNAR. Note that CAT models establish length generalization aligned with Theorem 2 .

$c \to \infty$. *Set key convolution as* $F_{k,i} = 1$ *for* $0 \le i < B$ *and zero otherwise.*
(A) *If* $d \ge 2\sigma^2 B(\sqrt{\log \bar{L}} + t)^2$, *then* (LCAT) *solves AR for fixed* $\boldsymbol{x}_L$ *with probability at least* $1 - 3e^{-t^2/4}$.
(B) *Conversely, for any* $\varepsilon > 0$ *there is* $C_\varepsilon > 0$ *as follows: If* $\bar{L} \ge C_\varepsilon$ *and* $d \le 2\sigma^2 B(\sqrt{(1-\varepsilon)\log \bar{L}} - t)^2$, *then* (LCAT) *fails to solve AR with the same probability.*
(C) *Finally, suppose we wish to solve AR* ***uniformly*** *for all queries* $\boldsymbol{x}_L$ *over a subspace* S. *This succeeds with the same probability whenever* $d \ge 2\sigma^2 B(\sqrt{\log \bar{L}} + \sqrt{\texttt{dim}(S)} + t)^2$.

Figure 4 corroborates the predictive accuracy of Proposition 1: As the block size increases, the embedding dimension to maintain success of AR grows approximately linearly. One can expand on this proposition in two directions. Firstly, a fundamental bottleneck in (LCAT) is the requirement $d \gtrsim B \log \bar{L}$. This arises from a *memory-recall tradeoff* (Arora et al. 2024; Jelassi et al. 2024) as we are summarizing the information of block $\boldsymbol{X}_i$ of length B through its landmark token. However, once this requirement is satisfied, the model can identify the correct block in $O(\bar{L})$ cost. To avoid paying the additional $O(B)$ cost of local attention, we could apply the LCAT approach hierarchically within the selected block to reduce the compute cost to $d(\bar{L} + \log B)$ per token. The dominant term $d\bar{L}$ captures the recall capacity of the LCAT model: Consistent with our theorem and lower bounds of (Arora et al. 2024), for AR to succeed, we need

$$\text{recall\_capacity} = d\bar{L} \ge L = \text{required\_memory}$$

Secondly, Proposition (1) chooses a particular long convolution where landmarks become the mean of the input tokens within the block. In practice, we can use a state-space model (Gu et al. 2022) to parameterize convolution efficiently. A particular SSM choice of state dimension 1 is simply using exponential smoothing. This yields the following SSM variant of Proposition 1.

Proposition 2. *Consider the setting of Proposition 1 with the exponential smoothing filter* $\boldsymbol{F}_i = \rho^i$ *for* $i \ge 0$. *Set* $\rho = e^{-1/B}$ *so that* $\rho^B = e^{-1}$. *Suppose* $d \ge 50B(\sqrt{\log \bar{L}} + t)^2$. *Then,* (LCAT) *solves AR with probability at least* $1 - 3e^{-t^2/4}$.

Above, we fixed the decay rate ρ for exposition purposes. More generally, any ρ choice with an effective context size of $O(B)$ would result in similar guarantee.

6 Experiments

6.1 Model Evaluation on N-gram AR and Length Generalization Capability

For the experiments on N-AR problems, we employ the CAT architecture as detailed in Section 3.1. We utilize convolution kernels with a width of $W = 3$ and explore model embedding sizes of $d = 32$, 64, and 128 across MQAR and MQNAR problems to assess the impact of model dimension on performance. In addition to the standard attention mechanism, we introduce a perturbation strategy by implementing linear attention on the convoluted Q, K, and V embeddings, referred to as LinCAT. We adhere strictly to the parameters set by (Arora et al. 2023). More detailed information on the training setup can be found in Section A including the data generation and hyperparameters. For reporting results, we conduct each experiment three times and present the maximum accuracy achieved across these runs, aligning with the methodologies of (Arora et al. 2023) and (Arora et al. 2024).

As illustrated in Fig. 5, the CAT model consistently outperforms all baseline models across a range of sequence lengths and model dimensions. Notably, both Mamba and Based models exhibit improved performance as the model dimension increases, particularly with shorter sequence lengths. This improvement is due to the memory-recall tradeoff (Arora et al. 2024) where models store and recall sequence information more as their dimensionalities expand. In contrast, thanks to the short convolution, the *single-layer* CAT model maintains 100% accuracy across all experimental settings, aligned with our theorem 1. Interestingly, aside from CAT, Mamba is the only model demonstrating the potential to effectively address the MQAR task within a single-layer network architecture. We will discuss this observation in further detail in Section B.

Model	Wikitext ppl↓	Lambada std ppl↓	Lambada openai ppl↓	Lambada std ppl↓	Lambada openai ppl↓	Piqa acc↑	Hella acc_norm↑	Winogrande acc↑	Arc-E acc↑	Arc-C acc_norm↑	Avg Acc↑
Pythia	27.410	74.663	34.023	0.281	0.343	0.651	0.355	0.529	0.443	0.235	0.405
CAT, no PE	29.216	86.318	42.260	0.266	0.321	0.640	0.339	0.515	0.436	0.237	0.393
CAT, RoPE	26.776	65.423	38.557	0.288	0.341	0.654	0.362	0.507	0.461	0.239	0.407
MH-CAT, no PE	27.417	58.959	32.822	0.296	0.355	0.644	0.352	**0.531**	0.460	0.240	0.411
MH-CAT, RoPE	**25.858**	**47.593**	**28.273**	**0.330**	**0.377**	**0.662**	**0.376**	0.512	**0.466**	0.231	**0.422**
TF++ (Touvron et al. 2023)*	28.390	NA	42.690	NA	0.310	0.633	0.340	0.504	0.445	**0.242**	NA
Mamba (Gu and Dao 2023)*	28.390	NA	39.660	NA	0.306	0.650	0.354	0.501	0.463	0.236	NA
GLA (Yang et al. 2023)*	28.650	NA	43.350	NA	0.303	0.648	0.345	0.514	0.451	0.227	NA

Table 2: Experiment results for model pretraining. * are results from (Yang et al. 2023), which uses a same dataset and training procedure as ours. We use the same hyperparameters as (Yang et al. 2023) for fair comparison. For perplexity, lower is better, and for accuracy, higher is better. The average accuracy in last column is calculated by averaging the accuracy across all tasks but excluding the perplexity tasks. The best and second best results are highlighted in boldface and underline, respectively.

Evaluation of Length Generalization. In Fig. 6, we train models with 128 sequence length (the vertical red dashed line) and evaluate their performance on varying sequence lengths from 32 to 1,024. Fig. 6 shows the results of length generalization, which is aligned with our Theorem 2: CAT models maintain 100% accuracy while all other models exhibit a sharp decline in performance as the sequence length increases. This decrease is due to the increased demand of recall which requires the model to store and retrieve more information as the sequence length grows. The CAT model, however, is able to maintain its performance by leveraging the convolutional filters to shift the context and retrieve the necessary information. Remarkably, in Fig. 5, we observe non-monotonic accuracy behavior for Mamba and Attention-only models as a function of sequence length. This is due to the fact that these models are more sensitive and harder to optimize in AR problems. In Fig 6, we used a denser hyperparameter grid and more trials to ensure smoother curves with better reproducibility.

6.2 Evaluations on Language Modeling

After mechanics tasks evaluation, we further explore the efficacy of the CAT model in real-world NLP tasks by integrating a 1D CAT structure into the Pythia (Biderman et al. 2023) framework. We pretrain the modified 370M-parameter model on the SlimPajama (Soboleva et al. 2023) dataset, involving 15B tokens. We then assess the model on a variety of downstream zero-shot tasks, including Wikitext, Lambada, Piqa, Hella, Winogrande, Arc-E, and Arc-C, a methodology commonly used in the field to evaluate generalization capabilities across diverse tasks (Gu and Dao 2023; Arora et al. 2023, 2024). The results are in Table 2, where CAT outperforms both Pythia and state-of-the-art convolutional models, align with its superior performance in mechanics tasks.

In this series of experiments, the CAT model is trained in two variants: one incorporating rotary positional embedding (Su et al. 2024) (PE) and another without positional embedding (noPE). We observe that the CAT model with PE not only consistently outperforms the Pythia model but also achieves performance better than state-of-the-art models, including Mamba (Gu and Dao 2023), TF++ (Touvron et al. 2023), and GLA (Yang et al. 2023). Notably, the CAT model secures a superior perplexity gain compared to the standard model while maintaining a similar level of parameters.

Regarding the noPE variant, training a Pythia model without positional encoding leads directly to divergence and extremely large losses during training, affirming the critical role of positional encoding in enabling standard transformer models to learn and converge. Intriguingly, despite the absence of positional encoding, the CAT model still performs competitively with the leading models. This suggests that the convolutional structure in the CAT model effectively captures positional information within the data. We conjecture that the short convolutions provide positional information for neighboring tokens, while the deep multi-layer network structure hierarchically aggregates this information to establish long-range positional information.

This observation aligns with our synthetic experiments, where the CAT model demonstrated the capability to handle the AR task without positional encoding. These insights indicate that the convolutional structure could potentially replace positional encoding, which might benefit length extrapolation and generalization in the model. This offers a promising direction for further model design and optimization.

Length Generalization Figure 1 presents the results from a length generalization experiment with the CAT model, in which we also train the model on SlimPajama with 2,048 context length, and evaluate its zero-shot performance on Wikitext across varying test lengths. We also implemented YaRN (Peng et al. 2023) which incorporates position interpolation (PI) (Chen et al. 2023) and tempreture scaling on all three RoPE models to facilitate length generalization. The results indicate that among the three RoPE models, the CAT model consistently demonstrates excellent performance across all test sequence lengths. In contrast, the Pythia model exhibits a sharp decline in performance as the length increases. We suggest that is due to the additional positional embeddings introduced by PI that was absent during the training phase. Despite this, CAT models proficiently manage the relative positioning of tokens (especially overcome the new positional embeddings by leveraging convolution information), which significantly boosts its ability for length generalization. Additionally, the CAT model without PE is superior to the Pythia model with RoPE, suggesting the effectiveness of the convolutional structure within the CAT model in capturing essential positional data in length extrapolation.

Acknowledgements

This work was supported in part by the National Science Foundation grants CCF-2046816, CCF-2403075, CCF-2008020, the Office of Naval Research award N000142412289, and a gift by Google Research.

References

Arora, S.; Eyuboglu, S.; Timalsina, A.; Johnson, I.; Poli, M.; Zou, J.; Rudra, A.; and Ré, C. 2023. Zoology: Measuring and Improving Recall in Efficient Language Models. *arXiv:2312.04927*.

Arora, S.; Eyuboglu, S.; Zhang, M.; Timalsina, A.; Alberti, S.; Zinsley, D.; Zou, J.; Rudra, A.; and Ré, C. 2024. Simple linear attention language models balance the recall-throughput tradeoff. *arXiv preprint arXiv:2402.18668*.

Ataee Tarzanagh, D.; Li, Y.; Zhang, X.; and Oymak, S. 2023. Max-margin token selection in attention mechanism. *Advances in Neural Information Processing Systems*, 36: 48314–48362.

Ba, J.; Hinton, G. E.; Mnih, V.; Leibo, J. Z.; and Ionescu, C. 2016. Using fast weights to attend to the recent past. *Advances in neural information processing systems*, 29.

Biderman, S.; Schoelkopf, H.; Anthony, Q. G.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2397–2430. PMLR.

Candes, E. J. 2008. The restricted isometry property and its implications for compressed sensing. *Comptes rendus. Mathematique*, 346(9-10): 589–592.

Candes, E. J.; and Tao, T. 2006. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12): 5406–5425.

Chen, S.; Wong, S.; Chen, L.; and Tian, Y. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Collins, L.; Parulekar, A.; Mokhtari, A.; Sanghavi, S.; and Shakkottai, S. 2024. In-context learning with transformers: Softmax attention adapts to function lipschitzness. *arXiv preprint arXiv:2402.11639*.

Dao, T.; and Gu, A. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*.

Dauphin, Y. N.; Fan, A.; Auli, M.; and Grangier, D. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, 933–941. PMLR.

De, S.; Smith, S. L.; Fernando, A.; Botev, A.; Cristian-Muraru, G.; Gu, A.; Haroun, R.; Berrada, L.; Chen, Y.; Srinivasan, S.; et al. 2024. Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models. *arXiv preprint arXiv:2402.19427*.

Deora, P.; Ghaderi, R.; Taheri, H.; and Thrampoulidis, C. 2023. On the optimization and generalization of multi-head attention. *arXiv preprint arXiv:2310.12680*.

Fu, D. Y.; Dao, T.; Saab, K. K.; Thomas, A. W.; Rudra, A.; and Ré, C. 2022. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; Gupta, A.; and Ré, C. 2022. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35: 35971–35983.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Gupta, A.; Gu, A.; and Berant, J. 2022. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, 35: 22982–22994.

Hasani, R.; Lechner, M.; Wang, T.-H.; Chahine, M.; Amini, A.; and Rus, D. 2022. Liquid structural state-space models. *arXiv preprint arXiv:2209.12951*.

Ildiz, M. E.; Huang, Y.; Li, Y.; Rawat, A. S.; and Oymak, S. 2024. From Self-Attention to Markov Models: Unveiling the Dynamics of Generative Transformers. *arXiv preprint arXiv:2402.13512*.

Jelassi, S.; Brandfonbrener, D.; Kakade, S. M.; and Malach, E. 2024. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*.

Jing, L.; Gulcehre, C.; Peurifoy, J.; Shen, Y.; Tegmark, M.; Soljacic, M.; and Bengio, Y. 2019. Gated orthogonal recurrent units: On learning to forget. *Neural computation*, 31(4): 765–783.

Kazemnejad, A.; Padhi, I.; Natesan Ramamurthy, K.; Das, P.; and Reddy, S. 2024. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36.

Kosma, C.; Nikolentzos, G.; and Vazirgiannis, M. 2023. Time-Parameterized Convolutional Neural Networks for Irregularly Sampled Time Series. *arXiv preprint arXiv:2308.03210*.

Li, Y.; Cai, T.; Zhang, Y.; Chen, D.; and Dey, D. 2022. What makes convolutional models great on long sequence modeling? *arXiv preprint arXiv:2210.09298*.

Li, Y.; Huang, Y.; Ildiz, M. E.; Rawat, A. S.; and Oymak, S. 2024. Mechanics of next token prediction with self-attention. In *International Conference on Artificial Intelligence and Statistics*, 685–693. PMLR.

Ma, X.; Yang, X.; Xiong, W.; Chen, B.; Yu, L.; Zhang, H.; May, J.; Zettlemoyer, L.; Levy, O.; and Zhou, C. 2024. Megalodon: Efficient LLM Pretraining and Inference with Unlimited Context Length. *arXiv preprint arXiv:2404.08801*.

Ma, X.; Zhou, C.; Kong, X.; He, J.; Gui, L.; Neubig, G.; May, J.; and Zettlemoyer, L. 2022. Mega: moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*.

Makkuva, A. V.; Bondaschi, M.; Girish, A.; Nagle, A.; Jaggi, M.; Kim, H.; and Gastpar, M. 2024. Attention with markov: A framework for principled analysis of transformers via markov chains. *arXiv preprint arXiv:2402.04161*.

Mohtashami, A.; and Jaggi, M. 2023. Landmark attention: Random-access infinite context length for transformers. *NeurIPS*.

Nichani, E.; Damian, A.; and Lee, J. D. 2024. How Transformers Learn Causal Structure with Gradient Descent. *arXiv preprint arXiv:2402.14735*.

Olsson, C.; Elhage, N.; Nanda, N.; Joseph, N.; DasSarma, N.; Henighan, T.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Orvieto, A.; Smith, S. L.; Gu, A.; Fernando, A.; Gulcehre, C.; Pascanu, R.; and De, S. 2023. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning*, 26670–26698. PMLR.

Oymak, S.; Rawat, A. S.; Soltanolkotabi, M.; and Thrampoulidis, C. 2023. On the role of attention in prompt-tuning. In *International Conference on Machine Learning*, 26724–26768. PMLR.

Park, J.; Park, J.; Xiong, Z.; Lee, N.; Cho, J.; Oymak, S.; Lee, K.; and Papailiopoulos, D. 2024. Can Mamba Learn How to Learn? A Comparative Study on In-Context Learning Tasks. *arXiv preprint arXiv:2402.04248*.

Peng, B.; Quesnelle, J.; Fan, H.; and Shippole, E. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Poli, M.; Massaroli, S.; Nguyen, E.; Fu, D. Y.; Dao, T.; Baccus, S.; Bengio, Y.; Ermon, S.; and Ré, C. 2023. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, 28043–28078. PMLR.

Poli, M.; Thomas, A. W.; Nguyen, E.; Ponnusamy, P.; Deiseroth, B.; Kersting, K.; Suzuki, T.; Hie, B.; Ermon, S.; Ré, C.; et al. 2024. Mechanistic Design and Scaling of Hybrid Architectures. *arXiv preprint arXiv:2403.17844*.

Press, O.; Smith, N. A.; and Lewis, M. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Ren, L.; Liu, Y.; Lu, Y.; Shen, Y.; Liang, C.; and Chen, W. 2024. Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling. *arXiv preprint arXiv:2406.07522*.

Slepian, D. 1962. The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41(2): 463–501.

Soboleva, D.; Al-Khateeb, F.; Myers, R.; Steeves, J. R.; Hestness, J.; and Dey, N. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Tarzanagh, D. A.; Li, Y.; Thrampoulidis, C.; and Oymak, S. 2023. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*.

Tian, Y.; Wang, Y.; Zhang, Z.; Chen, B.; and Du, S. 2023. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. *arXiv preprint arXiv:2310.00535*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yang, B.; Bender, G.; Le, Q. V.; and Ngiam, J. 2019. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in neural information processing systems*, 32.

Yang, S.; Wang, B.; Shen, Y.; Panda, R.; and Kim, Y. 2023. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*.

Zhang, X.; Chang, X.; Li, M.; Roy-Chowdhury, A.; Chen, J.; and Oymak, S. 2024. Selective Attention: Enhancing Transformer through Principled Context Control. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 11061–11086. Curran Associates, Inc.