NYU

Customizing the Inductive Biases of Softmax Attention using Structured Matrices

Yilun Kuang · Noah Amsel · Sanae Lotfi · Shikai Qiu · Andres Potapczynski · Andrew Gordon Wilson

ICML 2025

In what ways are the inductive biases imposed by softmax attention insufficient?

- The attention mechanism transforms the inputs into **low-dimensional queries and keys**, which causes **information loss** for certain tasks that have **intrinsically high-dimensional inputs.**

- Attention spends **the same FLOPs budget for all input pairs**, without imposing a **distance-dependent compute bias** for neighboring tokens in the sequence.

In-Context Regression

Standard attention down-projects input due to the multi-head design, which is harmful for tasks like in-context regression.

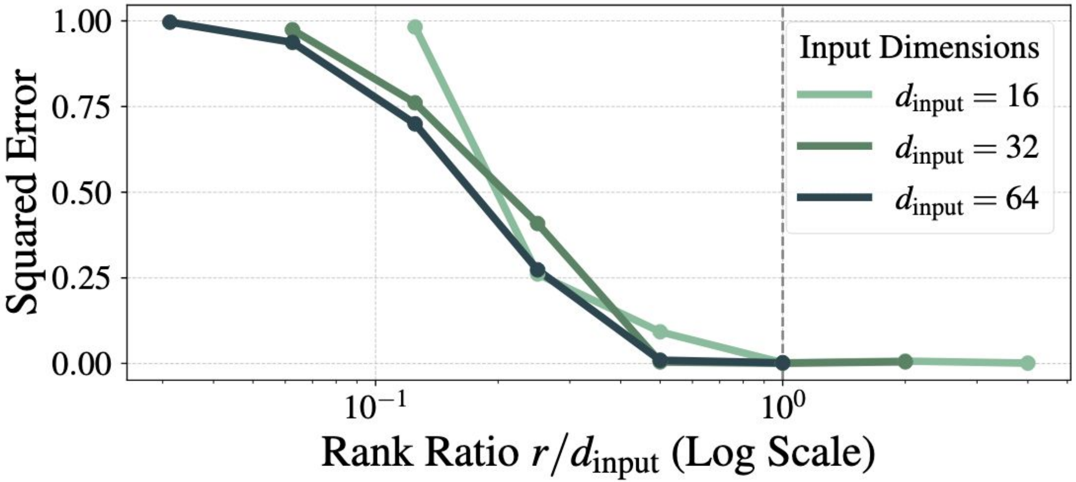
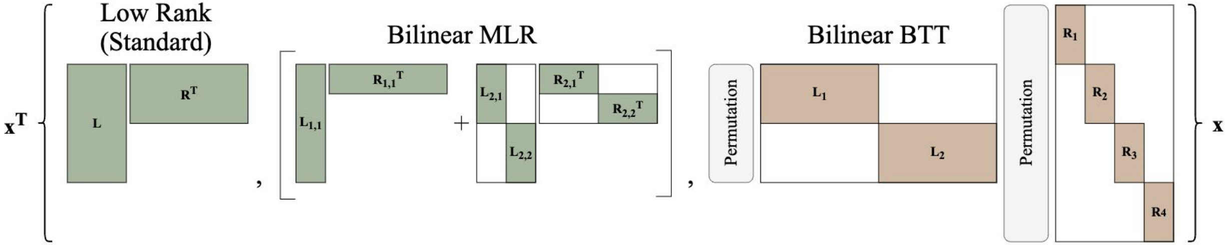


Figure 2. **Multi-head attention cannot solve in-context regression unless the head dimension r is close to the input dimension d_{input}.** We train 6-layer transformers with 8 heads and varying embedding and head dimensions to perform in-context regression: given the prompt $\mathbf{x}_1, \mathbf{w}^\top \mathbf{x}_1, \ldots, \mathbf{x}_{N-1}, \mathbf{w}^\top \mathbf{x}_{N-1}, \mathbf{x}_N$ for $\mathbf{x}_i \in \mathbb{R}^{d_{\text{input}}}$ and unknown $\mathbf{w}$, predict $\mathbf{w}^\top \mathbf{x}_N$. Plot shows error at $N = 2d_{\text{input}}$. For details see Section 4.

NYU

Structured Matrices

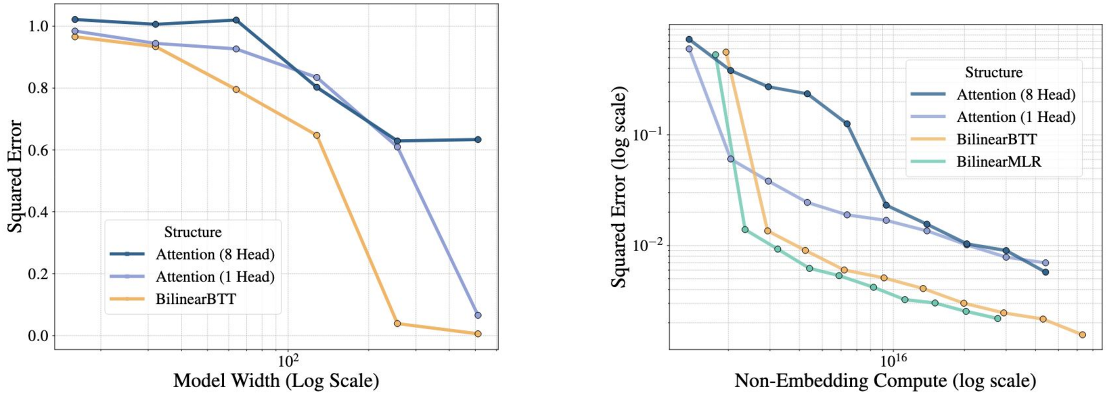


(a) Bilinear Forms with Structured Matrices

Table 1. **Properties of Structured Matrix Families**. For all listed structures, the number of FLOPs needed to compute the bilinear form $\mathbf{x}^\top\mathbf{M}\mathbf{y}$ is $O(\text{\#params})$. For BTT, we assume $a = b = c = d = \sqrt{D}$. MLR and BTT both attain high rank relative to their parameter counts.

Structure	Definition	Parameters	Rank
Dense	$\mathbf{W}$	D^2	D
Low Rank	$\mathbf{L}\mathbf{R}^\top$	$2Dr$	r
Multi-Level Low Rank (MLR)	$\sum_{l=1}^{L}\bigoplus_{k=1}^{p_l}\mathbf{L}_{l,k}\mathbf{R}_{l,k}^\top$	$2D\sum_l r_l$	$\sum_l r_l p_l$
Block Tensor Train (BTT)	$\mathbf{P}_L(\bigoplus_{k'=1}^{b}\mathbf{L}_{k'})\mathbf{P}_R(\bigoplus_{k=1}^{b}\mathbf{R}_k^\top)$	$2D^{\frac{3}{2}}s$	D

In-Context Regression

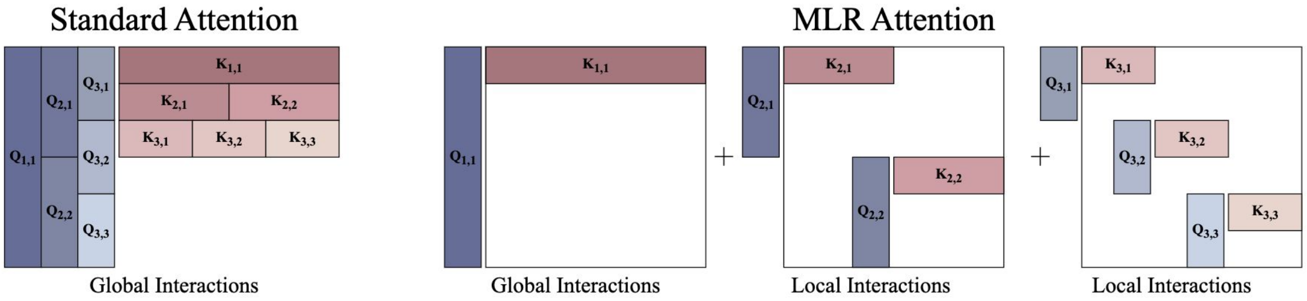


(a) Input Dimension (d_{input}) = 128.

(b) Input Dimension (d_{input}) = 64.

Figure 3. **Both Bilinear BTT and Bilinear MLR outperform standard attention for in-context regression tasks**. (a) Bilinear BTT and 1-head attention are full rank, so they can learn the task using a smaller model width (x-axis) than 8-head attention. Here model width refers to the embedding dimension D. All points on this figure correspond to models that was trained for the same number of steps. (b) For a fixed model width ($D = 256$), both Bilinear MLR and Bilinear BTT have smaller regression error compared to attention with 1 or 8 heads when controlling for compute. Here our bilinear models use 8 heads.

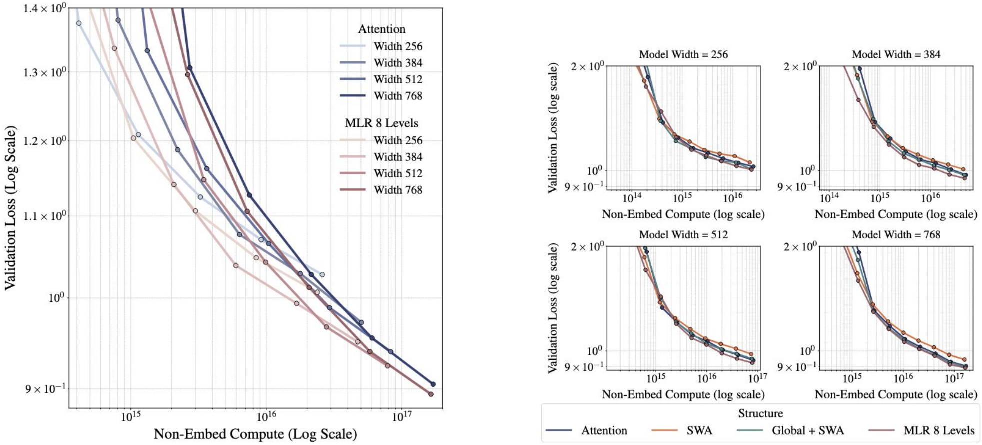
Structured Matrices



(b) Encoding Distance-Dependent Compute Bias with Multi-Level Low Rank (MLR) Attention

Figure 1. **Overview of two ways to customize the inductive biases of softmax attention with structured matrices.** (a) Standard attention computes the dot product between a query and a key via a low-rank bilinear transformation. In this work, we replace the low rank product with other structured matrices such as Multi-Level Low Rank (MLR) and Block Tensor Train (BTT) as introduced in Section 3.2. (b) Standard attention captures pair-wise token interactions in the global context without imposing a distance-dependent compute bias. We introduce an inductive bias for more flexible interactions between nearby tokens by changing the attention score matrix from low rank to Multi-Level Low Rank (MLR), as in Section 3.4.

Language Modeling and Time Series Forecasting



(a) Scaling Law on OpenWebText

(b) Different Forms of Distance-Dependent Compute

Figure 4. **MLR Attention achieves lower validation loss on OpenWebText compared to standard multi-head attention and variants of sliding window attention when controlling for compute.** (a) We plot the validation loss of both MLR attention with 8 levels and standard attention against compute. We vary over 4 values of model width $D \in \{256, 384, 512, 768\}$. MLR attention outperforms standard attention across model widths. (b) We compare MLR attention with variants of sliding window attention (SWA) that also encodes distance-dependent compute bias. Given the training sequence length $T = 1024$ and a 6-layers transformer model, SWA refers to 6 layers of sliding window attention with window size $T' = 128$. Global + SWA refers to a combination of standard attention and sliding window attention with $T' = 128$. The first and fourth layers of the model are standard attention and the rest are sliding window attention. We observe that across the model width D, MLR attention outperforms standard attention, SWA, and Global + SWA when controlling for compute.

Language Modeling and Time Series Forecasting

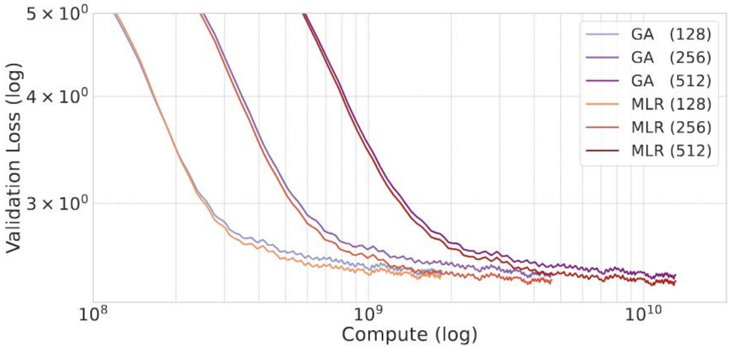


Figure 15. **MLR attention reduces computational cost when compared to standard attention.** We plot the validation loss of three distinct model sizes when substituting the T5 attention mechanism in Chronos (Ansari et al., 2024) from standard attention to MLR attention. As shown in the figure, MLR attention achieves the same validation loss compared to standard attention with smaller FLOPs.

References

Customizing the Inductive Biases of Softmax Attention using Structured Matrices. Yilun Kuang · Noah Amsel · Sanae Lotfi · Shikai Qiu · Andres Potapczynski · Andrew Gordon Wilson.