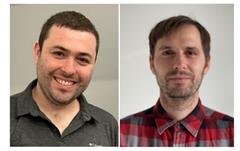
The Low-Rank Bottleneck in Attention

Noah Amsel, Gilad Yehudai, Joan Bruna



Multi-Head Attention Layer

- d Input dimension
- r Rank (query/key dimension)
- H Number of heads

Standard Self-Attention Head:

$\mathbf{X} = [\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_N]$ inputs in $\mathbb{R}^d$

$\mathbf{Q} = \mathbf{W}_Q\mathbf{X}, \quad \mathbf{K} = \mathbf{W}_K\mathbf{X}, \quad \mathbf{V} = \mathbf{W}_V\mathbf{X}$ queries/keys/values in $\mathbb{R}^r$

$\mathbf{W}_O^\top \mathbf{V} \operatorname{softmax}\big(\underbrace{\mathbf{K}^\top\mathbf{Q}}_{=\mathbf{X}^\top\left(\mathbf{W}_K^\top\mathbf{W}_Q\right)\mathbf{X}}\big)$ outputs in $\mathbb{R}^d$

Generalized Multi-Head Cross-Attention:

$$A_{H,r}(\mathbf{X},\mathbf{y}) = \sum_{h=1}^{H} \mathbf{M}_h\mathbf{X}\ \phi_h\left(\mathbf{W}_h\mathbf{X},\ \mathbf{y}\right)$$

- $\mathbf{W}_h$ maps $\mathbb{R}^d \to \mathbb{R}^r$ (like $\mathbf{W}_K$ above)
- $\mathbf{M}_h$ maps $\mathbb{R}^d \to \mathbb{R}^d$ (like $\mathbf{W}_O^\top\mathbf{W}_V$ above)
- ϕ_h is any function that outputs a probability distribution over $[N]$
- "Cross-attention": query comes from $\mathbf{y}$, keys come from $\mathbf{X}$

Standard Hyperparameter Setting

Large d Low rank $r \ll d$ $H = d/r$

Research Question: Is this always a wise choice?

	Attn Is All You Need (2017)	GPT-3 (2020)	Llama-2 (2023)	
# layers	6	96	80	← Massive scaling up
d	512	12,288	8,192	
r	64	128	128	← Insignificant change
H	d/r	d/r	d/r	

Analogous Work: Depth Separation for MLPs

- Why do deep neural nets beat shallow ones?
- Theorem: Some functions on $\mathbb{R}^d$ can be represented by a depth-3 MLP with $\operatorname{poly}(d)$ width, but cannot be approximated by depth-2 MLP unless width is $\exp(d)$
- "Pay a bit for depth, save *many* parameters overall."

Multi-head attention with small query/key dimension is ubiquitous, but 1-head attention with big queries/keys can be more powerful.

Main Theorem

The "nearest neighbor" function on the d-sphere

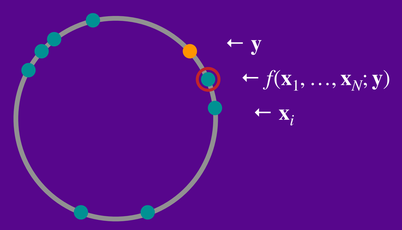
$$f(\mathbf{x}_1,\ldots,\mathbf{x}_N;\mathbf{y}) := \operatorname*{arg\,min}_{\mathbf{x}\in\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}} \|\mathbf{x}-\mathbf{y}\|_2$$

can be represented by one full-rank attention head ($r = d, H = 1$),

$$f(\mathbf{x}_1,\ldots,\mathbf{x}_N;\mathbf{y}) = \mathbf{I}\,\mathbf{X}\,\mathrm{hm}(\mathbf{X}^\top\mathbf{I}\,\mathbf{y})$$

but cannot be ε-approximated by low-rank attention ($r < d$) unless $H \gtrsim \left(d/r\right)^{1/\varepsilon}$:

$$\mathop{\mathbb{E}}_{\mathbf{x}_1\perp\mathbf{x}_2,\mathbf{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})} \left\| f(\mathbf{x}_1,\mathbf{x}_2;\mathbf{y}) - A_{H,r}(\mathbf{x}_1,\mathbf{x}_2;\mathbf{y}) \right\|^2 > \varepsilon$$



Alternative Rank-Separation Theorem

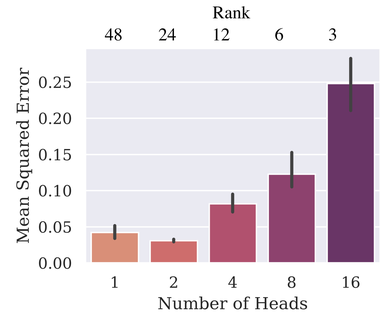
A certain linear combination of biased nearest neighbor functions can be represented by full-rank biased attention with $H = d^2$, but not approximated by low-rank attention unless $H \gtrsim C^{d-r} \,/\, \|V_h\|^2$

Experiments

Task: In-context learning of linear functions in 20 dimensions, as in Garg et. al (2022): $(\mathbf{x}_1, \mathbf{w}^\top\mathbf{x}_1, \ldots, \mathbf{x}_{40}, \mathbf{w}^\top\mathbf{x}_{40}, \mathbf{x}_{41}) \mapsto \mathbf{w}^\top\mathbf{x}_{41}$

Model: Linear embedding + 12-layer transformer with $d = 48$, $H = d/r$.

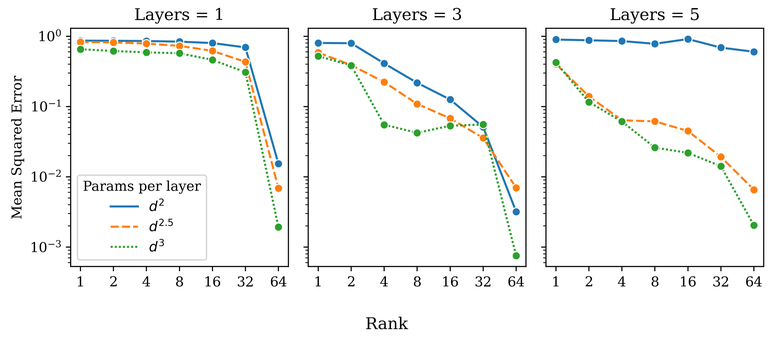
Result: High-rank attention far outperforms low-rank attention with the same number of parameters.



Task: Nearest neighbor function with $d = 64$, $N = 16$, iid Gaussian data

Model: Transformers with $H = d/r$, $d^{1.5}r$, or d^2/r.

Result: 1-layer transformer succeeds iff using full rank attn ($r = d$). Full-rank is always better, even vs. low-rank model with more parameters.



Role of Depth

- Our main theorems concern 1-layer of attention.
- Is large rank still needed for multi-layer transformers?
- Theorem: We can approximate the nearest neighbor function on $N = 2$ points with 2 layers of rank-1 attention *if* we add extra dimensions for positional encodings.
- Conjecture: We cannot do this for general N.

Conclusion

Try increasing r and H more than the standard setting, especially for non-textual data. Theorists shouldn't just assume full-rank attention.