Theory, Analysis, and Best Practices for Sigmoid Self-Attention

Jason Ramapuram*, Federico Danieli*, Eeshan Dhekane*, Floris Weers*, Dan Busbridge*, Pierre Ablin*, Tatiana Likhomanenko*, Jagrit Digani, Zijin Gu, Amitis Shidani, Russ Webb
ICLR 2025 · Apple

Abstract

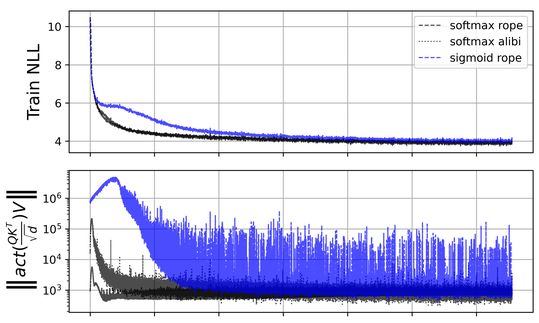
Attention is typically obtained as the softmax of dot product between k and q, a non-local operation requiring noticeable overhead. Building on prior work, we revisit a **purely local**, **more efficient alternative**: sigmoid attention. Our work unifies prior art and establishes best practices for sigmoid attention as a drop-in softmax replacement in transformers.

Key Contributions

- Extension of FlashAttention2 with the sigmoid kernel, **increasing kernel throughput by 15-18%.**
- Prove SigmoidAttn is a **universal function approximator** on seq-to-seq tasks.
- **Analysis of SigmoidAttn's regularity** and provide its worst-case Jacobian bound.
- **SigmoidAttn matches SoftmaxAttn performance** in various tasks and domains.

Main Problem

- **Large initial attention norm** for SigmoidAttn leads to worse final loss.
- Norm of SigmoidAttn is **unbounded** as we **increase sequence length**.

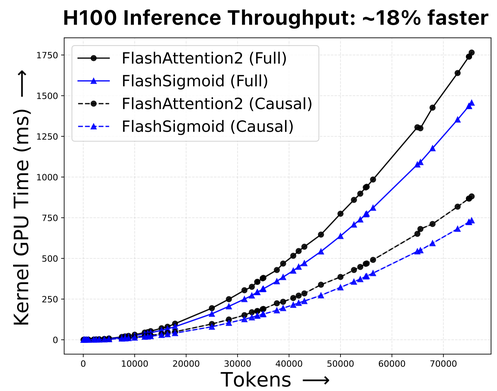


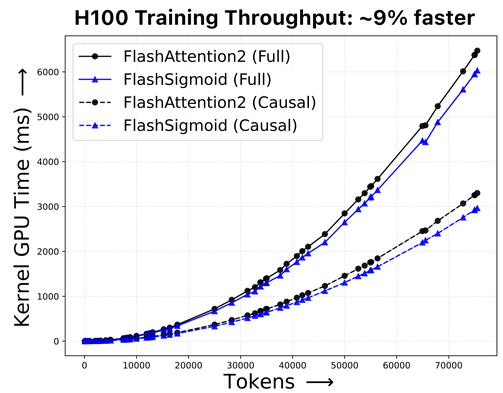
Sigmoid attention in one simple equation

$$\mathrm{SigmoidAttn}(\boldsymbol{X}) = \sigma(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d_{qk}})\boldsymbol{V},$$
$$\text{with } \sigma : u \mapsto \mathrm{sigmoid}(u+b) := (1+e^{-(u+b)})^{-1}.$$

In most scenarios, setting $b = -\ln(\max n)$ works well. If this still fails, we suggest hybrid norm.

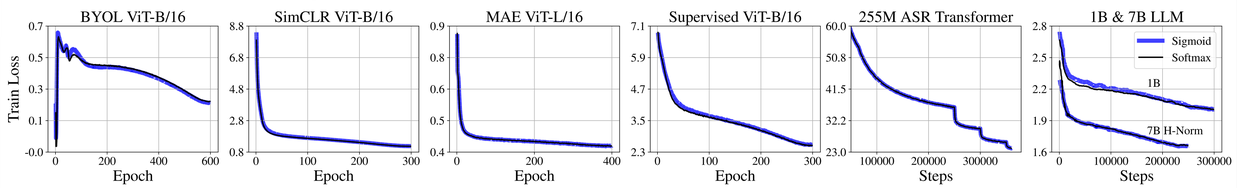
Sigmoid matches softmax performance and increases throughput





Throughput benefits are at **all** token counts.

Sigmoid attention works in multiple tasks and domains.



	Seq. Len	ARC Easy (%)	ARC (%)	Hellaswag (%)	Piqa (%)	Sciq (%)	Winogrande (%)	Lambada (%)	TriviaQA (1-shot) (%)	WebQS (1-shot) (%)	AVG (%)	Step time (s)
Softmax 1B(ALiBi)	2k	62.2	26.8	42.4	59.0	**72.3**	88.1	58.4	**19.9**	**15.4**	**49.4**	0.38
Sigmoid 1B(ALiBi)	2k	**62.8**	**28.8**	**42.5**	**59.7**	70.3	**88.6**	**59.7**	19.1	13.8	**49.5**	**0.34**
Softmax 7B(ALiBi) H-Norm	4k	71.2	39.9	53.2	65.5	75.6	91.8	**67.2**	37.7	**21.8**	59.0	3.85
Sigmoid 7B(ALiBi) H-Norm	4k	**72.7**	**40.5**	**53.5**	**66.2**	**76.0**	**92.5**	66.5	**39.5**	**21.8**	**59.7**	3.4

Universal Function Approximation

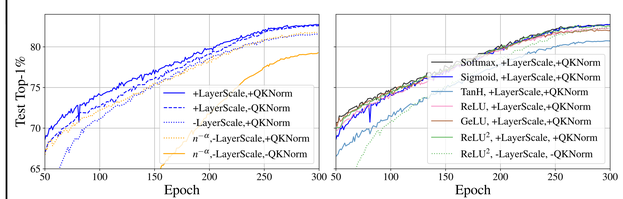
$\mathcal{T}_{\sigma}^{h,d_v,r}$ is the class of transformer networks obtainable by combining an arbitrary number of SigmoidAttn layers (each of h heads of dimension d_v) followed by FFN layers of hidden dimension r. For any given continuous, permutation-equivariant function $f : \Omega \subset \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ with compact support Ω, and for **any arbitrarily small error** ε, there exists a sigmoid attention transformer $g \in \mathcal{T}_{\sigma}^{4,1,4}$ s.t:

$$\left(\int_{\Omega} \|f(X) - g(X)\|_p^p dX\right) \leq \varepsilon, \qquad \text{for} \quad 1 \leq p < \infty\,.$$

Sequence length generalization

255M ASR Model	test-clean ↓	test-other ↓	Ted 00-10s ↓	Ted 10-20s↓	Ted 20-30s ↓	Ted 30s+ ↓
Softmax (RoPE)	**2.2**	5.5	12.7	10.6	12.8	9.5
Sigmoid (RoPE)	2.3	**5.4**	**12.3**	**10.1**	**12.3**	**8.6**
Softmax (ALiBi)	**2.2**	5.4	**12.3**	10.7	**12.1**	**8.6**
Sigmoid (ALiBi)	2.3	**5.1**	**12.3**	**10.5**	12.6	9.1

Vision Ablations



- LayerScale important in vision(not ASR, LM).
- Other activations perform well in vision only.

TM and © 2025 Apple Inc. All rights reserved.