ZeroS: Zero-Sum Linear Attention for Efficient Transformers

NeurIPS 2025 Spotlight

Jiecheng Lu[1], Xu Han[2], Yan Sun[1], Viresh Pati[1],
Yubin Kim[1], Siddhartha Somani[1], Shihao Yang[1]

[1]Georgia Institute of Technology [2]Amazon Web Services

2025

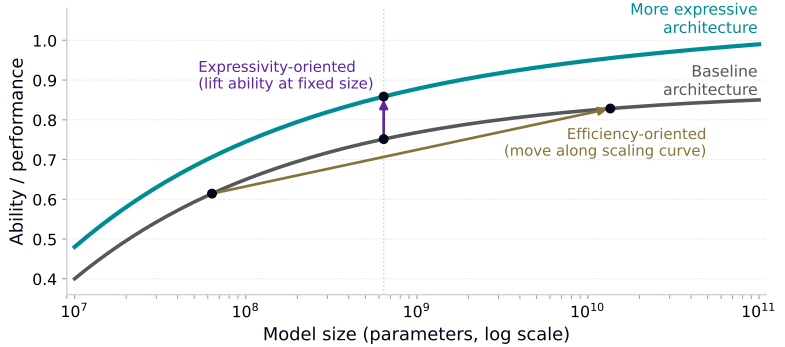
Outline

Transformers and Attention Complexity

- The Transformer architecture is ubiquitous in NLP, vision, speech and RL.
- Standard self-attention mixes all pairs of tokens with $O(N^2)$ time and memory complexity.
- Long contexts are challenging: memory footprint and runtime scale quadratically with sequence length.
- Linear-time variants approximate or replace softmax attention, aiming for $O(N)$ complexity.
- A persistent gap remains: linear methods often underperform the softmax baseline.

What Limits (Linear) Attention?

- Existing linear attentions approximate softmax with kernels or factorisations with non-negative feature maps to ensure stability.
- Strictly positive weights form convex combinations of values: only additive blending is possible. A uniform weight bias arises from the constant term in the softmax expansion, diluting focus on long sequences.
- Negative weights and subtractive operations cannot be expressed in a single layer.



Taylor Expansion of Softmax

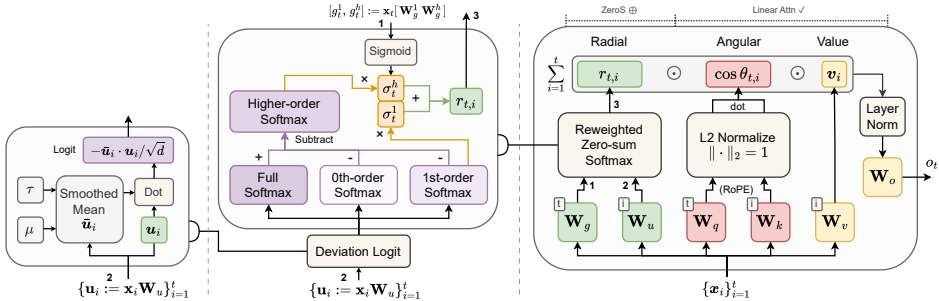
- For logits $\{s_i\}$ at a given timestep, the softmax has a series expansion:

$$\mathrm{softmax}(s_i) \approx \frac{1}{t} + \frac{1}{t}\delta_i + \frac{1}{2t}(\delta_i^2 - \frac{1}{t}\sum_j \delta_j^2) + O(\|\delta\|^3)$$

  where $\delta_i = s_i - \bar{s}$ and $\bar{s} = \frac{1}{t}\sum_j s_j$.
- The zero-order term $\frac{1}{t}$ ensures weights sum to one but contributes a uniform averaging bias.
- Higher-order terms capture competitive interactions among tokens but remain within a convex hull.

Zero-Sum Intuition

- Remove the constant $\frac{1}{t}$ baseline from softmax weights — obtain zero-mean residuals.
- Allow weights to be positive or negative while preserving their sum equal to zero.
- Contrastive operations become possible: a single layer can now perform differencing and suppression.
- To maintain stability, reweight the residuals using learned gates σ_t^1 and σ_t^h.
- The reweighted zero-sum mechanism underlies our proposed ZeroS attention.



(a) Calculation of the deviation logits (b) Reweighted zero-sum softmax block (c) The overall architecture of ZeroS (Zero-Sum Linear Attention)

Expressive Power: Convex vs. Zero-Sum

Proposition

Let $\{\mathbf{v}_i\}$ be value vectors and $v_{\text{avg}} = \frac{1}{t}\sum_i \mathbf{v}_i$. Softmax produces convex combinations of $\mathbf{v}_i$, i.e. points in

$$\mathcal{C} = \Big\{ \sum_i \alpha_i \mathbf{v}_i : \alpha_i \geq 0,\ \sum_i \alpha_i = 1 \Big\}.$$

Removing the zero-order term yields zero-sum weights w_i with $\sum_i w_i = 0$, representing points in

$$\mathcal{Z} = \Big\{ \sum_i w_i (\mathbf{v}_i - v_{\text{avg}}) : \sum_i w_i = 0 \Big\}.$$

Then $\mathcal{C} \subsetneq v_{\text{avg}} + \mathcal{Z}$. Zero-sum attention enlarges the space of attainable head outputs.

- Negative weights enrich the span beyond convex mixtures; only the average direction is removed.
- Subsequent layers or multiple heads can recover the lost average direction.

Reweighted Zero-Sum Softmax

- Given logits $s_{t,i}$ at step t, compute the mean $\bar{s}_t$ and deviations $\delta_{t,i} = s_{t,i} - \bar{s}_t$.
- Define the residual weights

$$w_{t,i} = \sigma_t^1 \frac{\delta_{t,i}}{t} + \sigma_t^h \varepsilon_{t,i},$$

  where $\varepsilon_{t,i}$ is the second- and higher-order softmax residual after subtracting $\frac{1}{t}$ and $\frac{\delta_{t,i}}{t}$.
- σ_t^1, σ_t^h are learnable gates (sigmoid activations) depending on the current query.
- In the first layer, an optional gate σ_t^0 can reintroduce the $1/t$ term for completeness.
- Summation of weights satisfies $\sum_i w_{t,i} = 0$, ensuring zero-sum outputs.

Stability and Lipschitz Properties

- ZeroS maintains numerical stability even with negative weights.
- The update norm is bounded by $Bt \max_i |w_{t,i}|$, and bounded logits guarantee $\max_i |w_{t,i}| = O(1/t)$.
- A ℓ_2 Lipschitz bound holds with a $1/\sqrt{t}$ decay factor — gradients remain stable for long sequences.
- LayerNorm further controls variance and obviates the need for explicit scaling.

Radial-Angular Decoupling

- Standard softmax attention couples magnitude and direction: $\exp(\|q\| \, \|k\| \cos\theta)$.
- Linear attentions often ignore angular sign flips due to non-negative feature maps.
- We separate magnitude (radial) and direction (angular) components:

$$o_t = \sum_{i=1}^{t} r_{t,i} \, \cos\theta_{t,i} \, v_i,$$

  where $r_{t,i}$ is the reweighted zero-sum radial term and $\cos\theta_{t,i} = \hat{q}_t^\top \hat{k}_i$ or with positional rotation (RoPE).
- This enables contrastive contributions: positive and negative modulation from angular alignment.

Linear-Time Implementation

- Replace $s_{t,i}$ with logits depending only on the key step i: for example

$$s_i = -\frac{1}{\sqrt{d}}\, u_i\, \bar{u}_i^\top,$$

  where $u_i = x_i W_u$ and $\bar{u}_i$ is a running average with trainable smoothness.
- Compute prefix sums for $E_t = \sum_i e^{s_i}$, $P_t = \sum_i s_i$, and three state matrices $\mathbf{F}_t, \mathbf{G}_t, \mathbf{H}_t$.
- The output can be written as a linear combination

$$o_t = \hat{q}_t(\alpha_t \mathbf{F}_t + \beta_t \mathbf{G}_t + \gamma_t \mathbf{H}_t),$$

  where the coefficients $\alpha_t, \beta_t, \gamma_t$ depend on the gates and prefix sums.
- Overall complexity is $O(Nd^2)$ time and $O(d^2)$ memory — on par with other linear attention implementations.

In-Context Learning: MAD Benchmark

- The Mechanistic Architecture Design (MAD) suite probes algorithmic capabilities: recall, selective copying, compression and more.
- We replace the attention module in a baseline model with ZeroS and evaluate on six tasks.
- ZeroS outperforms other linear-time methods (Hyena, Mamba, DeltaNet, Gated Linear Attention) and approaches the standard Transformer.

Model	Compress	Fuzzy Recall	InCtx Recall	Memorize	Noisy Recall	Selective Copy
Hyena	45.2	7.9	81.7	89.5	78.8	93.1
Mamba	52.7	6.7	90.4	89.5	90.1	86.3
LinAttn	31.1	8.2	91.0	74.9	75.6	93.1
Transformer	51.6	29.8	94.1	85.2	86.8	99.6
ZeroS	44.0	14.9	99.9	88.1	96.1	97.8
ZeroS-SM	45.2	28.0	100	84.3	96.6	98.5

Table: MAD results (higher is better).

Language Modeling: WikiText-103

- We train models on WikiText-103 at modest scale.
- ZeroS improves perplexity over the Transformer baseline; ZeroS-SM achieves further gains.

Model	PPL (val)	PPL (test)	Params (M)
Transformer	24.40	24.78	44.65
gMLP	28.08	29.13	47.83
Mamba	22.58	23.19	44.99
ZeroS	23.91	24.61	46.31
ZeroS-SM	**23.62**	**24.17**	44.69

Table: WikiText-103 perplexities (lower is better).

OpenWebText2 and Other Tasks

- On the larger OpenWebText2 dataset, ZeroS tracks close to softmax attention and surpasses alternative linear architectures.
- We also replace attention in DeiT-Tiny for ImageNet classification, achieving 75.5% top-1 accuracy versus 72.2% for the baseline.
- On time-series forecasting, ZeroS consistently improves both MSE and MAE compared to GLA and AFT.

Task	Improvement	Notes
OWT2 (PPL)	**Up to 7%**	12-layer GPT-2 model
ImageNet-1k	+3.3 pt	DeiT-Tiny architecture
Time Series	Lower MSE/MAE	Six forecasting datasets

Ablation Studies

- Adding back the zero-order term reduces performance in tasks demanding in-context reasoning.
- Replacing the reweighted zero-sum softmax with standard softmax degrades MAD scores.
- Omitting the gating mechanism lowers flexibility across tasks.
- Layer Normalisation is crucial for stable training; without it, performance drops on challenging recall tasks.

Variant	MAD Avg %	WikiText PPL
ZeroS	73.5	24.61
w/ zero-order term	69.4	24.74
w/o reweighted softmax	67.6	24.97
w/o gating	70.8	24.65

Table: Ablation results highlighting key design elements.

Efficiency and Practicality

- ZeroS shares the $O(Nd^2)$ computational profile of other linear attentions.
- Implementation can reuse existing linear attention kernels — we run three prefix scans per head.
- Benchmarks on a 768-hidden 12-layer GPT-2 model (seq. length 1024) show inference latency within the range of LinAttn and close to FlashAttention.
- Memory usage during training is comparable to the fastest linear methods.

Conclusion

- We identified fundamental limitations of existing linear attentions: convex combinations and uniform weight bias constrain expressivity.
- ZeroS removes the constant term in softmax, enabling zero-sum weights with both positive and negative contributions.
- Reweighted zero-sum softmax and radial-angular decoupling yield richer interactions while maintaining $O(N)$ complexity.
- Extensive experiments show ZeroS matches or exceeds standard softmax attention across a wide range of benchmarks, closing the linear-quadratic performance gap.

Limitations and Outlook

- Our focus is on algorithmic expressivity; we do not introduce specialised GPU kernels like FlashAttention or Mamba.
- Large-scale language models were not explored due to resource constraints; scaling behaviour remains to be assessed.
- Future work: optimised implementations, combining ZeroS with local windows, exploring non-causal tasks, and extending to multi-modal architectures.

Thank You!

Questions?

Code: `https://github.com/LJC-FVNR/SequenceLab`

Email: `jlu414@gatech.edu`