ZeroS: Zero-Sum Linear Attention for Efficient Transformers

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]AWS

Motivation & Problem

- **Quadratic vs. Linear:** Softmax attention is accurate but $O(N^2)$. Linear attention variants achieve $O(N)$ yet often underperform.
- **Two fundamental limitations (affecting linear and even softmax):**
  - *Convex-combination bottleneck:* per-head reads are convex averages, permitting only additive blending; single-layer contrast/differencing is impossible.
  - *Uniform weight bias / dilution:* a roughly uniform $1/N$ component acts like persistent average pooling in long contexts, weakening focus.
- **Idea:** *Remove the zero-order term* behind that bias and reweight residuals to allow signed, zero-sum weights—enabling contrastive operations in one layer while preserving $O(N)$ complexity.

ZeroS in One Slide

Zero-Sum Linear Attention (ZeroS) removes the constant zero-order term and reweights the remaining softmax residuals to obtain *signed*, *zero-sum* attention weights.

$$\bar{s}_t = \frac{1}{t}\sum_{j=1}^{t} s_{t,j}, \quad \delta_{t,i} = s_{t,i} - \bar{s}_t,$$

$$\varepsilon_{t,i} = \frac{e^{s_{t,i}}}{\sum_{r=1}^{t} e^{s_{t,r}}} - \frac{1}{t} - \frac{\delta_{t,i}}{t}, \quad \sum_i \varepsilon_{t,i} = 0,$$

$$w_{t,i} = \underbrace{\sigma_t^1 \frac{\delta_{t,i}}{t}}_{\text{1st order}} + \underbrace{\sigma_t^h \varepsilon_{t,i}}_{\text{2nd+ order}}, \qquad \sum_{i=1}^{t} w_{t,i} = 0.$$

Here $\sigma_t^1, \sigma_t^h \in (0,1)$ are learned gates (per step/channel) controlling linear vs. nonlinear competition components.

Expressivity & Stability (intuition)

- **Beyond convex hull:** subtracting $1/t$ unlocks signed, zero-sum weights. A single layer can *contrast* tokens (e.g., simple differencing), which convex mixing cannot represent in one step.
- **Stable by construction:** with bounded logits and $\|v_i\| \le B$, zero-sum weights satisfy $\max_i |w_{t,i}| = O(1/t)$; the head output is $O(B)$, independent of t (length-stable).
- **Residual alignment:** zero-sum produces centered updates ($\sum_i w_{t,i} = 0$), matching decoder residual stream design.

From Averaging to Contrast

- Keeping $1/t$ in the *first* layer preserves the affine hull; subsequent ZeroS layers (zero-sum) add contrastive directions without shrinking the reachable set.
- Practically, we default to removing $1/t$ in all layers; optionally retain it in the first layer (*minor impact* in our experiments).

Radial–Angular Decoupling (with RoPE)

ZeroS combines a **radial** reweighted zero-sum softmax with an **angular** cosine term:

$$\boldsymbol{o}_t = \sum_{i=1}^{t} \underbrace{r_{t,i}}_{\text{signed, zero-sum}} \cdot \underbrace{\cos\theta_{t,i}}_{\text{direction}} \boldsymbol{v}_i, \quad \cos\theta_{t,i} = \hat{\boldsymbol{q}}_t^\top \hat{\boldsymbol{k}}_i \;\text{ or }\; \hat{\boldsymbol{q}}_t^\top \mathbf{R}_{t-i} \hat{\boldsymbol{k}}_i \;(RoPE)$$

Unlike positive-only kernels, the sign flip of $\cos\theta$ *survives* into the weights via $r_{t,i}$, restoring contrastive, distance-aware behavior at *linear* cost.

Linear-Time Implementation (ZeroS-Lin)

We design $O(N)$ logits and a prefix-sum scan:

$$\boldsymbol{u}_i = \boldsymbol{x}_i \mathbf{W}_u, \quad \bar{\boldsymbol{u}}_i = \frac{e^\tau \boldsymbol{\mu} + \sum_{j=1}^{i} \boldsymbol{u}_j}{e^\tau + i}, \quad s_i = -\frac{1}{\sqrt{d}} \boldsymbol{u}_i \bar{\boldsymbol{u}}_i^\top$$

Maintain prefix sums (per head/channel):

$$E_t = \sum_{i\le t} e^{s_i},\; P_t = \sum_{i\le t} s_i,\; \mathbf{F}_t = \sum_{i\le t} e^{s_i}\hat{\boldsymbol{k}}_i^\top \boldsymbol{v}_i,\; \mathbf{G}_t = \sum_{i\le t} s_i \hat{\boldsymbol{k}}_i^\top \boldsymbol{v}_i,\; \mathbf{H}_t = \sum_{i\le t} \hat{\boldsymbol{k}}_i^\top \boldsymbol{v}_i.$$

Compute scalars and output:

$$\alpha_t = \frac{\sigma_t^h}{E_t}, \quad \beta_t = \frac{\sigma_t^1 - \sigma_t^h}{t}, \quad \gamma_t = -\frac{\sigma_t^h + \beta_t P_t}{t}, \qquad \boldsymbol{o}_t = \hat{\boldsymbol{q}}_t(\alpha_t \mathbf{F}_t + \beta_t \mathbf{G}_t + \gamma_t \mathbf{H}_t).$$

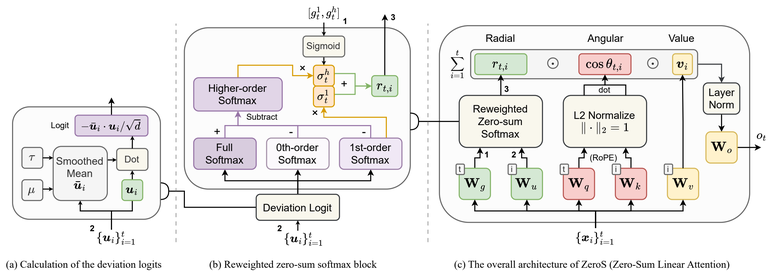
Complexity: $O(d^2)$ per step, $O(Nd^2)$ total; same hidden-state size as linear attention.

ZeroS for Softmax Attention (ZeroS-SM)

Apply the same zero-order removal and residual reweighting to the *quadratic* softmax matrix:

$$\mathbf{S} = \frac{1}{\sqrt{d}}\mathbf{Q}\mathbf{K}^\top + \mathbf{M}, \quad \mathbf{A} = \mathrm{softmax}(\mathbf{S}), \quad \mathbf{W} = (\mathbf{g}^1\mathbf{1}^\top) \odot \mathbf{\Delta} + (\mathbf{g}^h\mathbf{1}^\top) \odot \boldsymbol{\varepsilon}, \; \mathbf{O} = \mathbf{W}\mathbf{V}.$$

Empirically, ZeroS-SM *exceeds* vanilla softmax at matched budgets.



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

Figure 1. ZeroS block: deviation logits $\{s_i\}$, reweighted zero-sum softmax ($r_{t,i}$), angular modulation (RoPE), linear-time scan.

Empirical Highlights (matched parameter budgets)

Benchmarks. In-context learning (MAD, RegBench, MQAR), language modeling (WikiText-103, OWT2), ImageNet-1K (DeiT-Tiny), and multivariate TSF (ETT, Weather, Solar).

Model	Compress	Fuzzy	In-Ctx	Memorize	Noisy	Select	Avg
Transformer	51.6	29.8	94.1	85.2	86.8	99.6	74.5
GLA	38.8	6.9	80.8	63.3	81.6	88.6	60.0
Mamba	52.7	6.7	90.4	89.5	90.1	86.3	69.3
ZeroS	44.0	14.9	99.9	88.1	96.1	97.8	73.5
ZeroS-SM	45.2	28.0	**100**	84.3	96.6	98.5	**75.4**

Ablations & Analysis

- **Zero-sum helps:** reintroducing the 0^{th}-order term hurts In-Context Recall and WikiText PPL.
- **Residual reweighting matters:** replacing the reweighted zero-sum branch with vanilla softmax degrades performance, confirming the expressive gap vs. convex mixing.
- **Gating:** removing gates (σ^1, σ^h) causes consistent but moderate drops—gates learn task-adaptive entropy/competition.
- **Normalization:** removing LayerNorm particularly harms algorithmic tasks (e.g., In-Context Recall), highlighting variance control in linear attention.

Conclusion & Outlook

ZeroS removes the uniform zero-order term and reweights softmax residuals to produce *signed*, *zero-sum* linear attention:

- Enables single-layer *contrastive* operations and restores angular sign-flips at $O(N)$ complexity.
- Achieves accuracy on par with or better than softmax and surpasses prior linear methods across NLP, vision, and time series.
- Maintains numerical stability (length-independent bounds) and integrates seamlessly with RoPE and existing linear-attention accelerators.