WAVE: Weighted Autoregressive Varying Gate Attention for Time-Series Forecasting

Jiecheng Lu [1] Xu Han [2] Yan Sun [1] Shihao Yang [1]

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

Motivation

- **Decoder-only Transformers** (GPT-style) are powerful for sequential modelling but rarely used *end-to-end* in long-horizon time-series forecasting (TSF).
- Classic **ARMA** models decouple long-term trends (AR) from short-term fluctuations (MA).
- Existing efficient attentions add exponential decay (EMA) for locality, but this can suppress stable seasonal patterns.

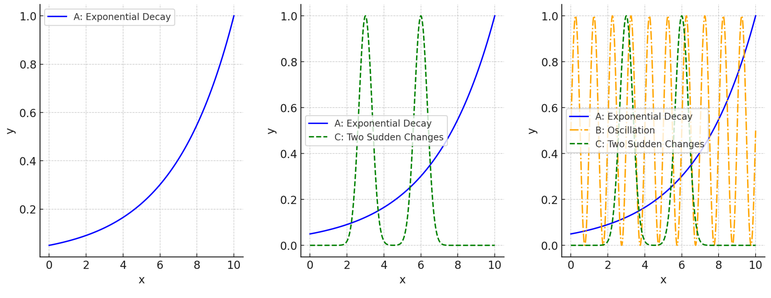

Figure 1. Visualization of different effects with exponential decay strategies and their challenges in gated linear attention. (Left: a) Pure exponential decay strategy in gated linear attention; (Mid: b) Exponential decay facing challenges in capturing long-term dependencies; (Right: c) Exponential decay facing challenges in capturing periodic dependencies

Key Contributions

1. Show that a well-tokenised *AR Transformer* already reaches SOTA TSF performance.
2. Propose **WAVE attention**: adds an MA path to *any* AR attention (softmax, linear, gated linear, ...).
3. Novel *indirect MA weight generation* keeps $\mathcal{O}(N)$ time and no extra parameters (weights are shared).
4. Achieve new state-of-the-art results on 12 public TSF benchmarks.

One-Step Patch Tokenization

Each multivariate series is split into non-overlapping *patch tokens* of size L_P. A single decoder step predicts the next patch, preventing error accumulation.

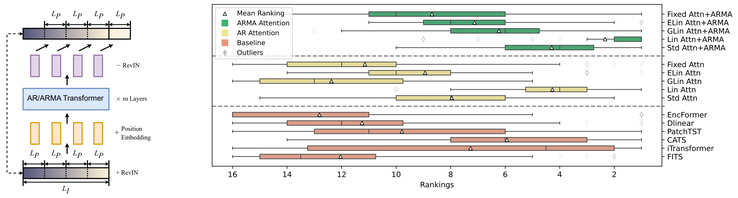

Figure 2. (Left: a) Overall architecture of our decoder Transformer for TSF. (Right: b) Box plots of performance rankings from 48 sub-experiments across 12 datasets. Green represents WAVE Transformers, yellow AR Transformers, and red the baselines, with triangles indicating mean rankings. AR Transformers perform comparably to baselines, while WAVE Transformers significantly outperform their AR counterparts. See Table and for more details.

WAVE Attention Mechanism

Given query $\mathbf{q}_t$, keys $\mathbf{k}_{1:t}$, values $\mathbf{v}_{1:t}$:

$$\underbrace{\mathbf{o}_t^{\mathrm{AR}} = \sum_{i=1}^{t} w_{t,i}\,\mathbf{v}_i}_{\text{AR (long-term)}} \quad + \quad \underbrace{\mathbf{o}_t^{\mathrm{MA}} = \sum_{j=1}^{t-1} \theta_{t-1,j}\,\boldsymbol{\epsilon}_j}_{\text{MA (short-term)}}$$

$$\boldsymbol{\epsilon}_j \;=\; \mathbf{v}_{j+1} \;-\; \mathbf{o}_j^{\mathrm{AR}}$$

The MA weights $\theta_{t-1,j}$ are produced *indirectly*: $\theta = B(I-B)^{-1}$, where $B_{t-1,j} = \phi_q(\mathbf{q}_{t-1}^{\mathrm{MA}})\,\phi_k(\mathbf{k}_j^{\mathrm{MA}})$.

- **AR path:** standard causal attention (softmax / linear / gated).
- **MA path:** lightweight linear attention over residuals $\boldsymbol{\epsilon}$.
- **Weight-sharing:** $\mathbf{W}_q$ reused; $\mathbf{W}_v$ fixed to identity.

Indirect MA Weight Generation

(1) Avoids explicit $N \times N$ MA matrix. (2) Preserves $\mathcal{O}(N)$ runtime of efficient attentions. (3) Activation choice: $\phi_k = \sigma(\alpha k)$, $\phi_q = -\mathrm{LeakyReLU}(-q)$ $\rightarrow$ yields near-diagonal Θ emphasising local context.

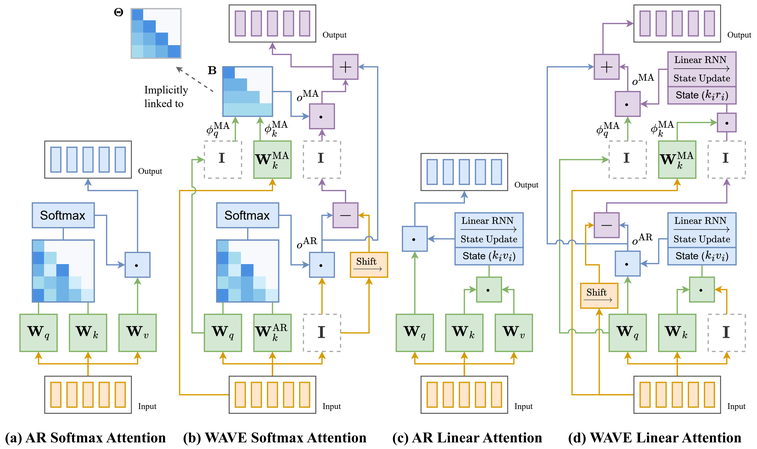

Figure 3. WAVE attention structure with the indirect MA weight generation method applied to softmax and linear attention. See Table 1 for more calculation details.

Table 1. Summary of WAVE attention for various attention mechanisms, detailing the calculation methods for AR output and MA output, where $\boldsymbol{r}_j = \boldsymbol{v}_{j+1} - \boldsymbol{o}_j^{\mathrm{AR}}$.

Model	AR term output $\boldsymbol{o}_t^{\mathrm{AR}}$	Indirect MA term output $\boldsymbol{o}_t^{\mathrm{MA}}$
Standard Softmax Attention (Std Attn)	$\frac{\sum_{i=1}^{t}\exp(\boldsymbol{q}_t(\boldsymbol{k}_i^{\mathrm{AR}})^\top)\boldsymbol{v}_i}{\sum_{i=1}^{t}\exp(\boldsymbol{q}_t(\boldsymbol{k}_i^{\mathrm{AR}})^\top)}$	$\sum_{j=1}^{t-1}\phi_q^{\mathrm{MA}}(\boldsymbol{q}_{t-1})\phi_k^{\mathrm{MA}}(\boldsymbol{k}_j^{\mathrm{MA}})^\top\boldsymbol{r}_j$
Linear Attention (Lin Attn)	$\boldsymbol{q}_t\sum_{i=1}^{t}(\boldsymbol{k}_i^{\mathrm{AR}})^\top\boldsymbol{v}_i$	$\phi_q^{\mathrm{MA}}(\boldsymbol{q}_{t-1})\sum_{j=1}^{t-1}\phi_k^{\mathrm{MA}}(\boldsymbol{k}_j^{\mathrm{MA}})^\top\boldsymbol{r}_j$
Element-wise Linear Attention (ELin Attn)	$\sigma(\boldsymbol{q}_t)\odot\frac{\sum_{i=1}^{t}\exp(\boldsymbol{k}_i^{\mathrm{AR}})\odot\boldsymbol{v}_i}{\sum_{i=1}^{t}\exp(\boldsymbol{k}_i^{\mathrm{AR}})}$	$\phi_q^{\mathrm{MA}}(\boldsymbol{q}_{t-1})\odot\sum_{j=1}^{t-1}\phi_k^{\mathrm{MA}}(\boldsymbol{k}_j^{\mathrm{MA}})\odot\boldsymbol{r}_j$
Gated Linear Attention (GLin Attn)	$\boldsymbol{q}_t\sum_{i=1}^{t}\mathbf{G}_i\odot(\boldsymbol{k}_i^{\mathrm{AR}})^\top\boldsymbol{v}_i$	$\phi_q^{\mathrm{MA}}(\boldsymbol{q}_{t-1})\sum_{j=1}^{t-1}\phi_k^{\mathrm{MA}}(\boldsymbol{k}_j^{\mathrm{MA}})^\top\boldsymbol{r}_j$
Fixed Attention (Fixed Attn)	$\sum_{i=1}^{t}w_{t,i}^{\mathrm{AR}}\boldsymbol{v}_i$	$\phi_q^{\mathrm{MA}}(\boldsymbol{w}_{t-1}^{\mathrm{MA,q}})\sum_{j=1}^{t-1}\phi_k^{\mathrm{MA}}(\boldsymbol{w}_j^{\mathrm{MA,k}})^\top\boldsymbol{r}_j$

Main Results (12 Datasets)

- **Average ranking over 48 TSF tasks** (4 horizons × 12 data sets): △ AR Transformer rank = 4.3 ⋆ **WAVE–Linear rank = 2.3 (best)**
- WAVE improves *every* attention variant: e.g. on "Weather", MSE **4 %** vs AR; on "ETTm1", MSE **8 %**.
- Outperforms previous models like PatchTST, FITS, iTransformer, CATS, DLinear.

Table 2. Summary of main TSF results with forecasting horizons $L_P \in \{12, 24, 48, 96\}$ and $L_I = 512$. Average rankings (AvgRank) of each model on test set MSE, along with the count of first-place rankings (#Top1), are included.

Models	Std Attn	Std Attn +ARMA	Lin Attn	Lin Attn +ARMA	GLin Attn	GLin Attn +ARMA	ELin Attn	ELin Attn +ARMA	Fixed Attn	Fixed Attn +ARMA
AvgRank	7.958	4.292	4.271	2.333	12.375	6.229	8.938	7.125	11.146	8.688
#Top1	0	4	4	25	0	1	1	3	1	3

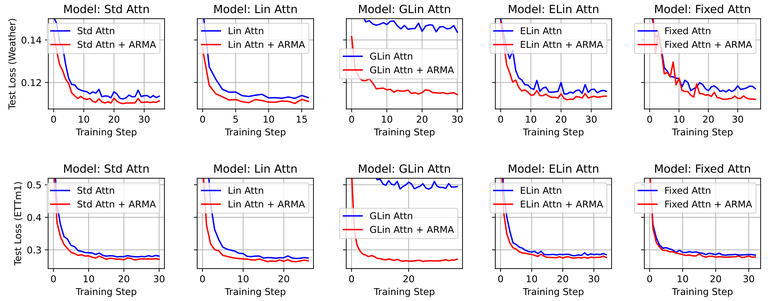

Figure 4. Visualization of test loss curves. We show the testing performance of five attention mechanisms using WAVE structures on the Weather and ETTm1 datasets ($L_I = 512$, $L_P = 48$).

Scalability & Efficiency

- Longer look-back (L_I=4 096) → WAVE MSE improves; many baselines degrade.
- WAVE with 3 layers beats AR with 8 layers → gains come from *structure*, not size.
- Same parameter count, linear memory/time; negligible extra FLOPs.

Conclusion

WAVE bridges classic ARMA ideas and modern Transformers. By explicitly modeling residuals via an efficient MA branch, it decouples short- and long-term dynamics, giving state-of-the-art forecasting accuracy without extra complexity.

jlu414@gatech.edu | shihao.yang@isye.gatech.edu