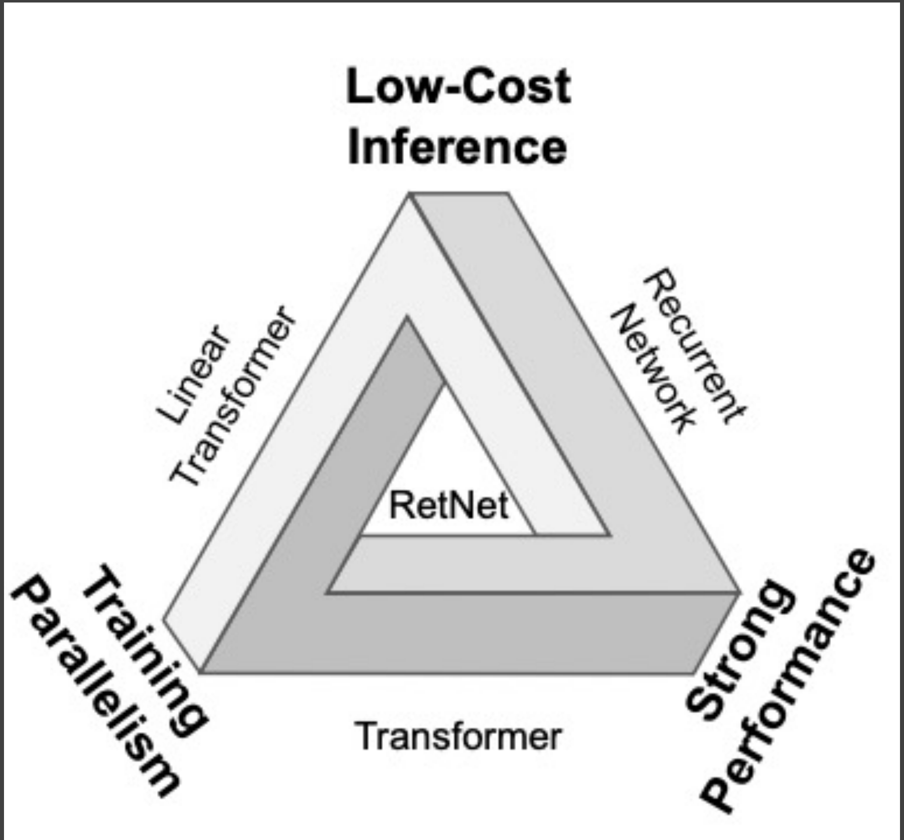
The Alan Turing Institute

Retentive Networks

Edward Gunn

DARe, EME

Sun, Y., “Retentive Network: A Successor to Transformer for Large Language Models”, <i>arXiv e-prints</i>, 2023. doi:10.48550/arXiv.2307.08621.


Low-Cost Inference
Linear Transformer
Recurrent Network
RetNet
Training Parallelism
Strong Performance
Transformer

Architectures	Training Parallelization	Inference Cost	Long-Sequence Memory Complexity	Performance
Transformer	✔	$O(N)$	$O(N^2)$	✔✔
Linear Transformer	✔	$O(1)$	$O(N)$	✘
Recurrent NN	✘	$O(1)$	$O(N)$	✘
RWKV	✘	$O(1)$	$O(N)$	✔
H3/S4	✔	$O(1)$	$O(N \log N)$	✔
Hyena	✔	$O(N)$	$O(N \log N)$	✔
RetNet	✔	$O(1)$	$O(N)$	✔✔

Retention

$$Q = XW_Q$$

$$K = XW_K$$

$$A = \Lambda \mathrm{diag}\left(\gamma \odot e^{i\theta}\right)\Lambda^{-1}$$

$$A^{n-m} = \Lambda \mathrm{diag}\left(\gamma \odot e^{i\theta}\right)^{n-m}\Lambda^{-1}$$

$$s_n = As_{n-1} + K_n^\top v_n$$

$$o_n = Q_n s_n$$

$$= \textstyle\sum_{m=1}^{n} Q_n A^{n-m} K_m^\top v_m$$

$$= \textstyle\sum_{m=1}^{n} Q_n \Lambda \,\mathrm{diag}\left(\gamma \odot e^{i\theta}\right)^{n-m} \Lambda^{-1} K_m^\top v_m$$

$$= \textstyle\sum_{m=1}^{n} Q_n \,\mathrm{diag}\left(\gamma \odot e^{i\theta}\right)^{n-m} K_m^\top v_m$$

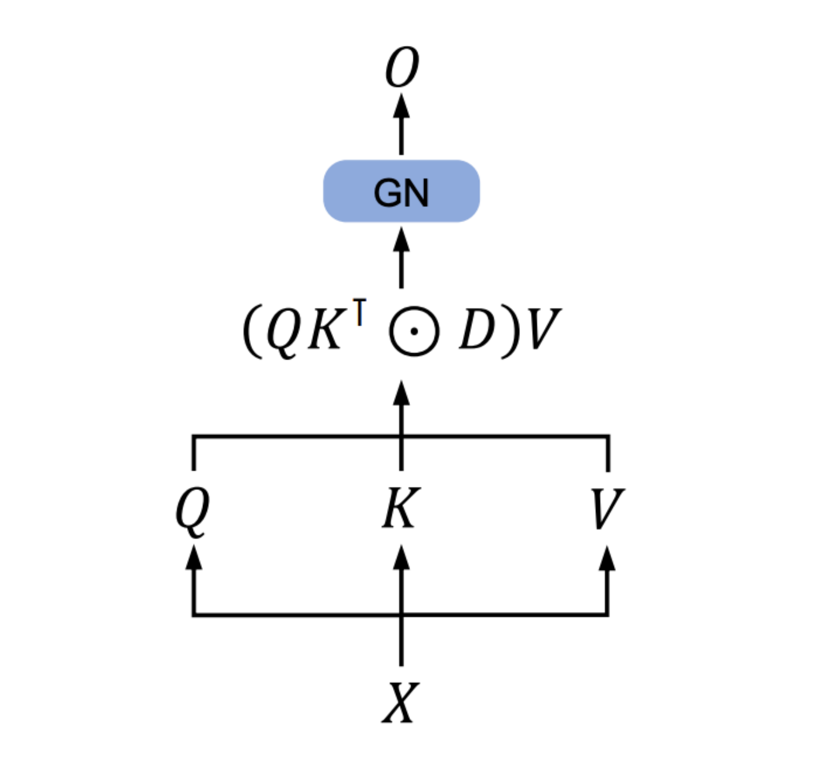
$$= \textstyle\sum_{m=1}^{n} Q_n \mathrm{diag}\left(\gamma \odot e^{i\theta}\right)^{n} \mathrm{diag}\left(\gamma \odot e^{i\theta}\right)^{-m} K_m^\top v_m$$

$$= \textstyle\sum_{m=1}^{n} Q_n \mathrm{diag}\left(\gamma \odot e^{i\theta}\right)^{n} \left(K_m \mathrm{diag}\left(\gamma \odot e^{i\theta}\right)^{-m}\right)^\top v_m$$

$$o_n = \textstyle\sum_{m=1}^{n} \gamma^{n-m} \left(Q_n \mathrm{diag}\left(e^{in\theta}\right)\right) \left(K_m \mathrm{diag}\left(e^{im\theta}\right)\right)^\dagger v_m$$

Parallel Representation

- $Q = (XW_Q) \odot \Theta$
- $K = (XW_K) \odot \overline{\Theta}$
- $V = XW_V$
- $\Theta_n = e^{in\theta}$
- $D_{nm} = \begin{cases} \gamma^{n-m}, n \geq m \\ 0, n < m \end{cases}$
- $Retention(X) = (QK^{\top} \odot D)V$
- GN is GroupNorm



$$D_{nm} = \begin{cases} \gamma^{n-m}, n \geq m \\ 0, n < m \end{cases}$$

Comparison to attention

Attention

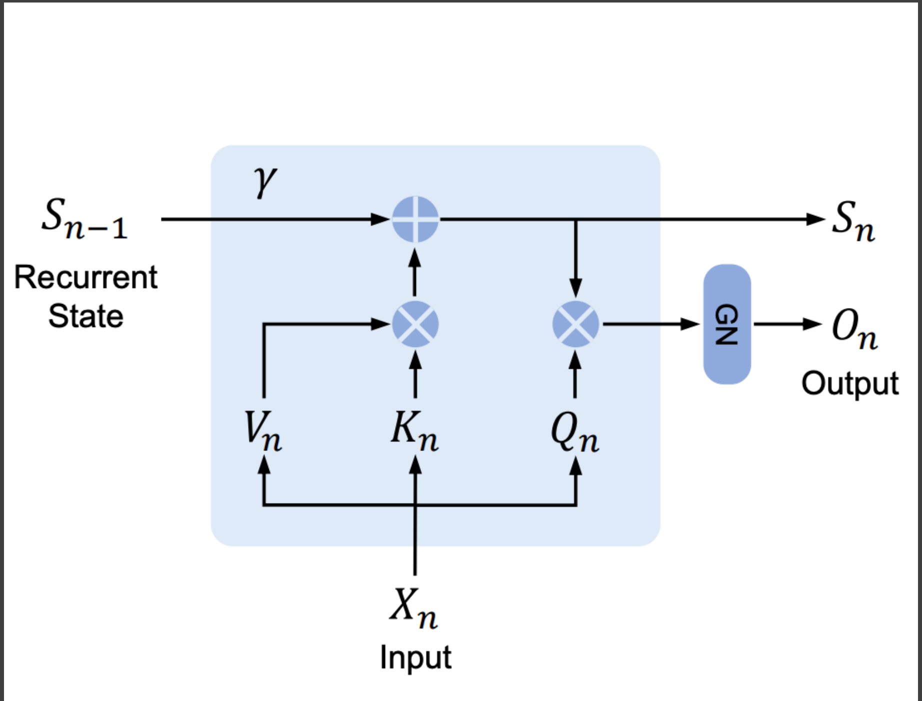
$$softmax(QK^\top)V$$

Retention

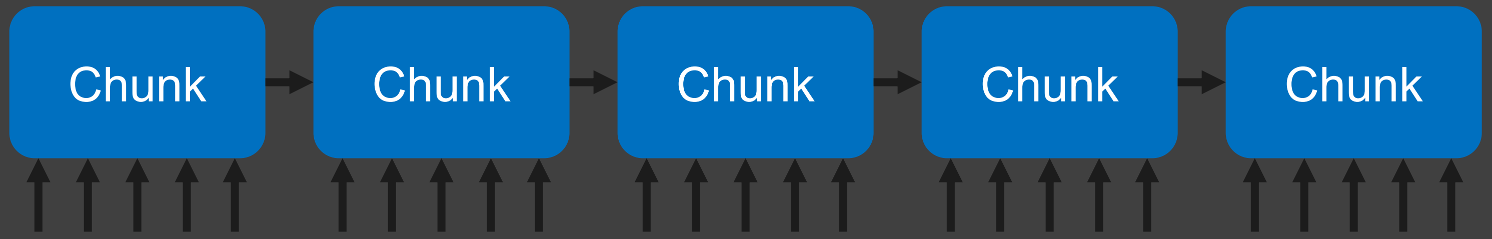
$$(\mathrm{QK}^\top \odot D)V$$

Recurrent Representation

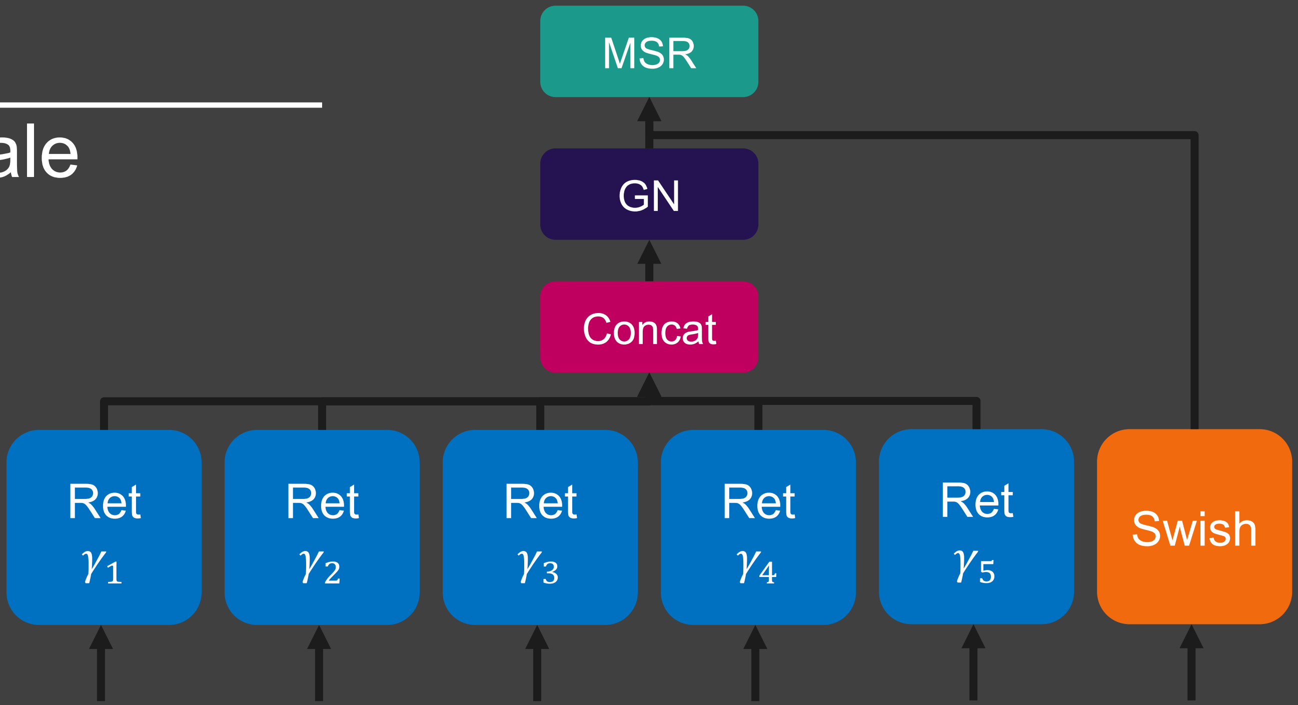
- $S_n = \gamma S_{n-1} + K_n^{\top} V_n$
- $Retention(X_n) = Q_n S_n$
- GN is GroupNorm



Chunkwise Recurrent Representation



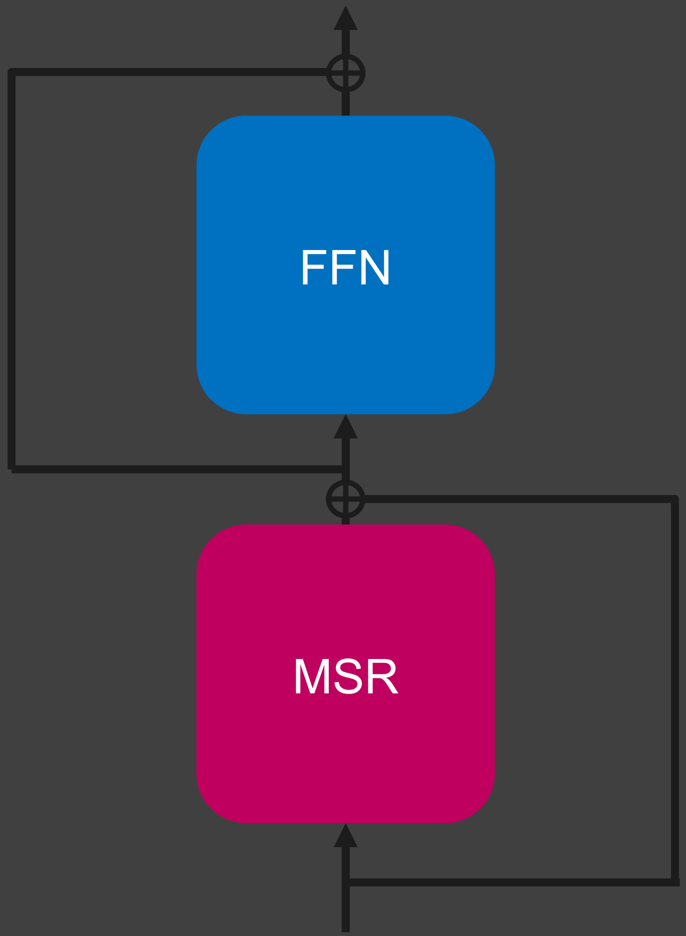
Gated Multi-scale Retention



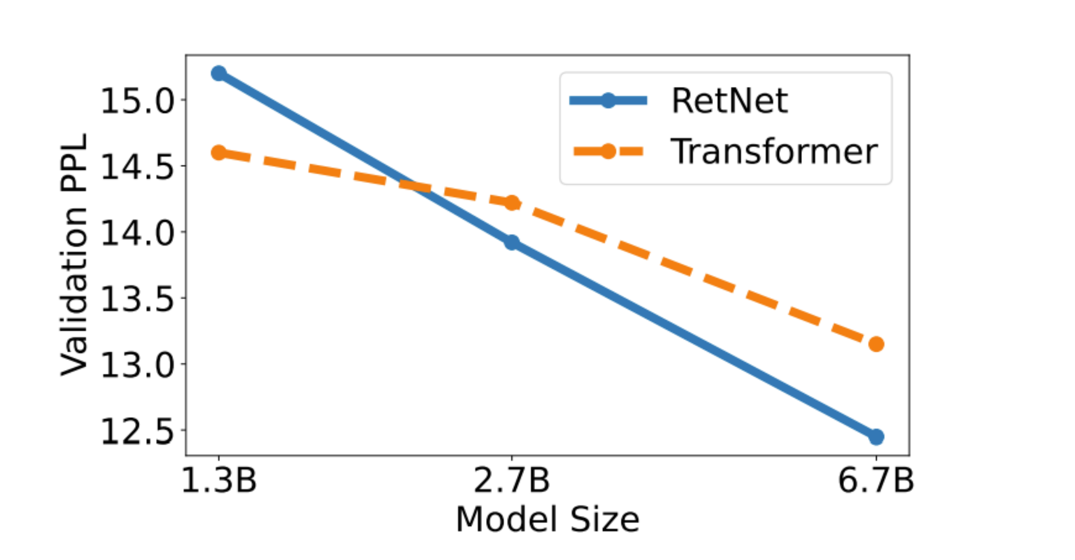
- $\gamma_i = 1 - 2^{-5-i}$
- $head_i = Retention(X, \gamma_i)$
- $Y = GroupNorm(Concat(head_1, \dots, head_h))$
- $MSR(X) = (swish(XW_G) \odot Y)W_O$

RetNet

- $Y_l = MSR\left(LN(X^l)\right) + X^l$
- $X^{l+1} = FFN\left(LN(Y^l)\right) + Y^l$
- LN is LayerNorm

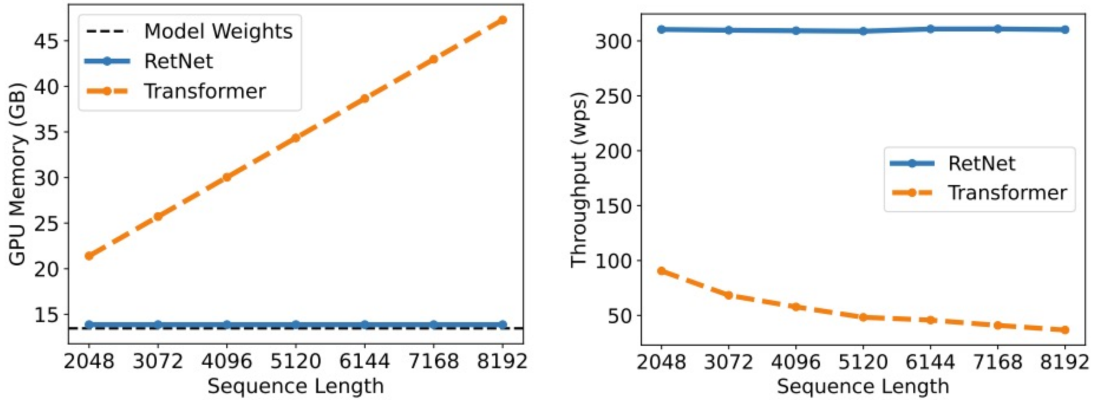


Performance



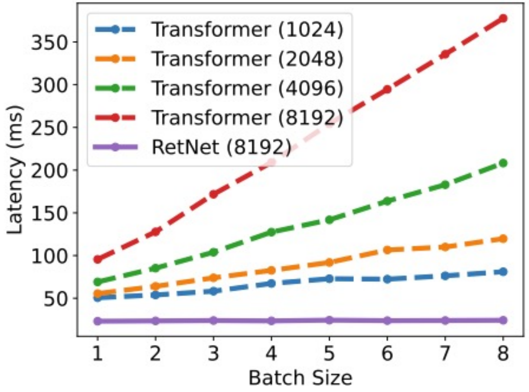
	HS	BoolQ	COPA	PIQA	Winograd	Winogrande	SC	Avg
Zero-Shot								
Transformer	55.9	62.0	69.0	74.6	69.5	56.5	75.0	66.07
RetNet	**60.7**	**62.2**	**77.0**	**75.4**	**77.2**	**58.1**	**76.0**	**69.51**
4-Shot								
Transformer	55.8	58.7	71.0	75.0	71.9	57.3	75.4	66.44
RetNet	**60.5**	**60.1**	**78.0**	**76.0**	**77.9**	**59.9**	**75.9**	**69.76**

Inference Cost



(a) GPU memory cost of Transformer and RetNet.

(b) Throughput of Transformer and RetNet.



(c) Inference latency with different batch sizes.

Ablations

Method	In-Domain	PG22	QMSum	GovReport	SummScreen
RetNet	**26.05**	**45.27**	**21.33**	**16.52**	**22.48**
− swish gate	27.84	49.44	22.52	17.45	23.72
− GroupNorm	27.54	46.95	22.61	17.59	23.73
− γ decay	27.86	47.85	21.99	17.49	23.70
− multi-scale decay	27.02	47.18	22.08	17.17	23.38
Reduce head dimension	27.68	47.72	23.09	17.46	23.41

Questions?