NuerIPS2025

On the Role of **H**idden **S**tates of Modern Hopfie**l**d Network in Transformer

Tsubasa Masumura, Masato Taki

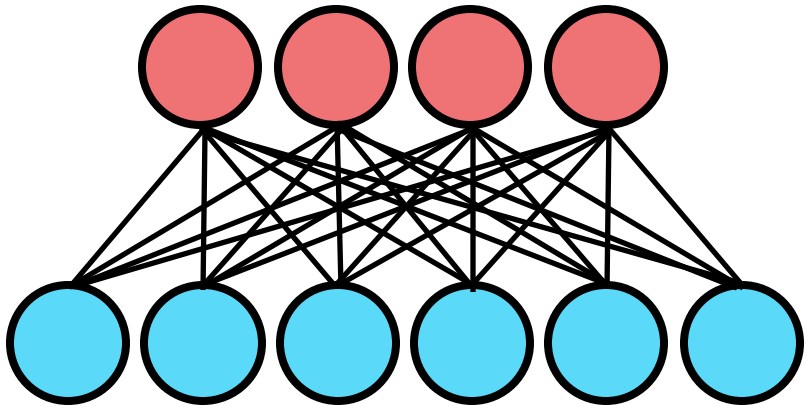
Rikkyo University, Japan

Background

NEURAL INFORMATION PROCESSING SYSTEMS

MHN⇔Transformer[Ramsauer et al, 2020], [Krotov&Hopfield, 2020]

hidden neurons $\boldsymbol{h}$



visible neurons $\boldsymbol{x}$

Update Rule

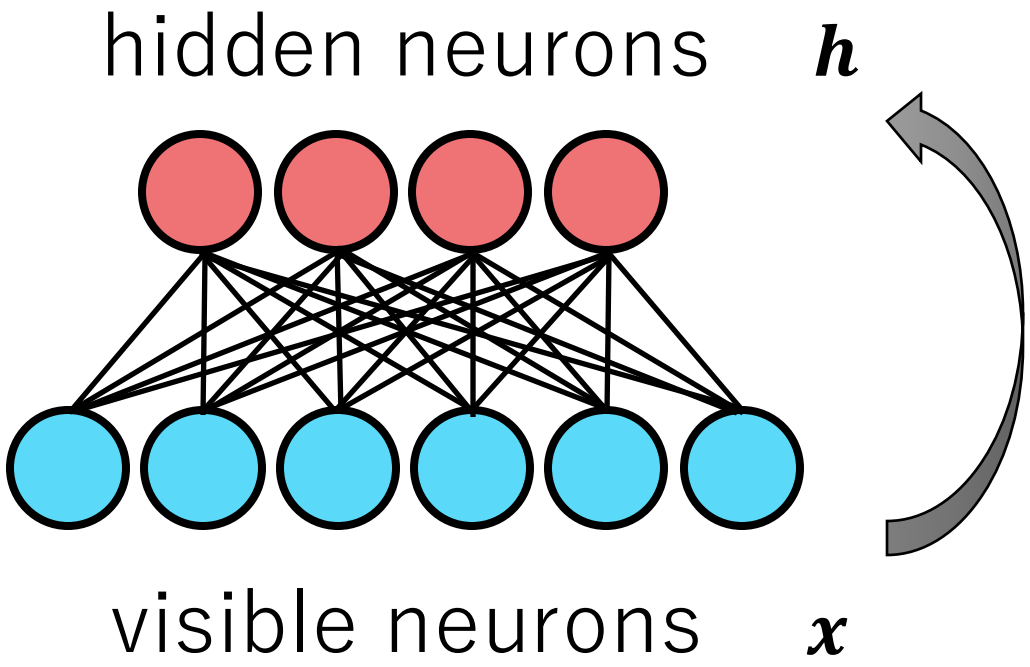
$$\tau_v \frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}\left(\boldsymbol{h}\right) \boldsymbol{W}_1^\top - \boldsymbol{x},$$

$$\tau_h \frac{d\boldsymbol{h}}{dt} = \boldsymbol{g}\left(\boldsymbol{x}\right) \boldsymbol{W}_2 - \boldsymbol{h}.$$

Background

NEURAL INFORMATION PROCESSING SYSTEMS

MHN⇔Transformer[Ramsauer et al, 2020], [Krotov&Hopfield, 2020]



Update Rule

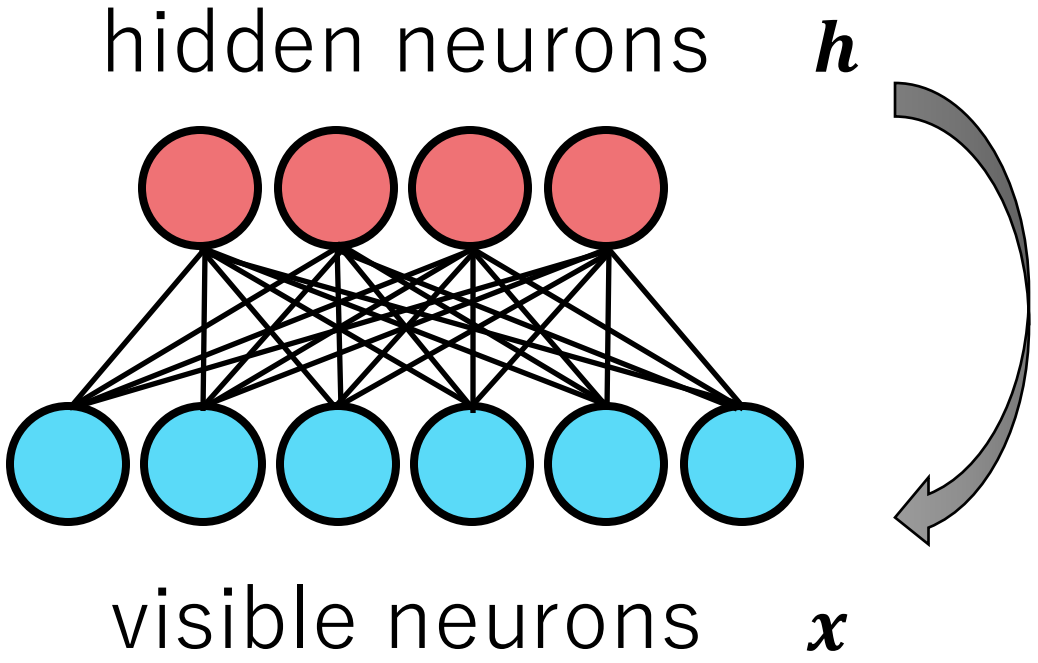
$$\tau_v \frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{h})\boldsymbol{W}_1^\top - \boldsymbol{x},$$

$$\tau_h \frac{d\boldsymbol{h}}{dt} = \boldsymbol{g}(\boldsymbol{x})\boldsymbol{W}_2 - \boldsymbol{h}.$$

Background

NEURAL INFORMATION PROCESSING SYSTEMS

MHN⇔Transformer[Ramsauer et al, 2020], [Krotov&Hopfield, 2020]



Update Rule

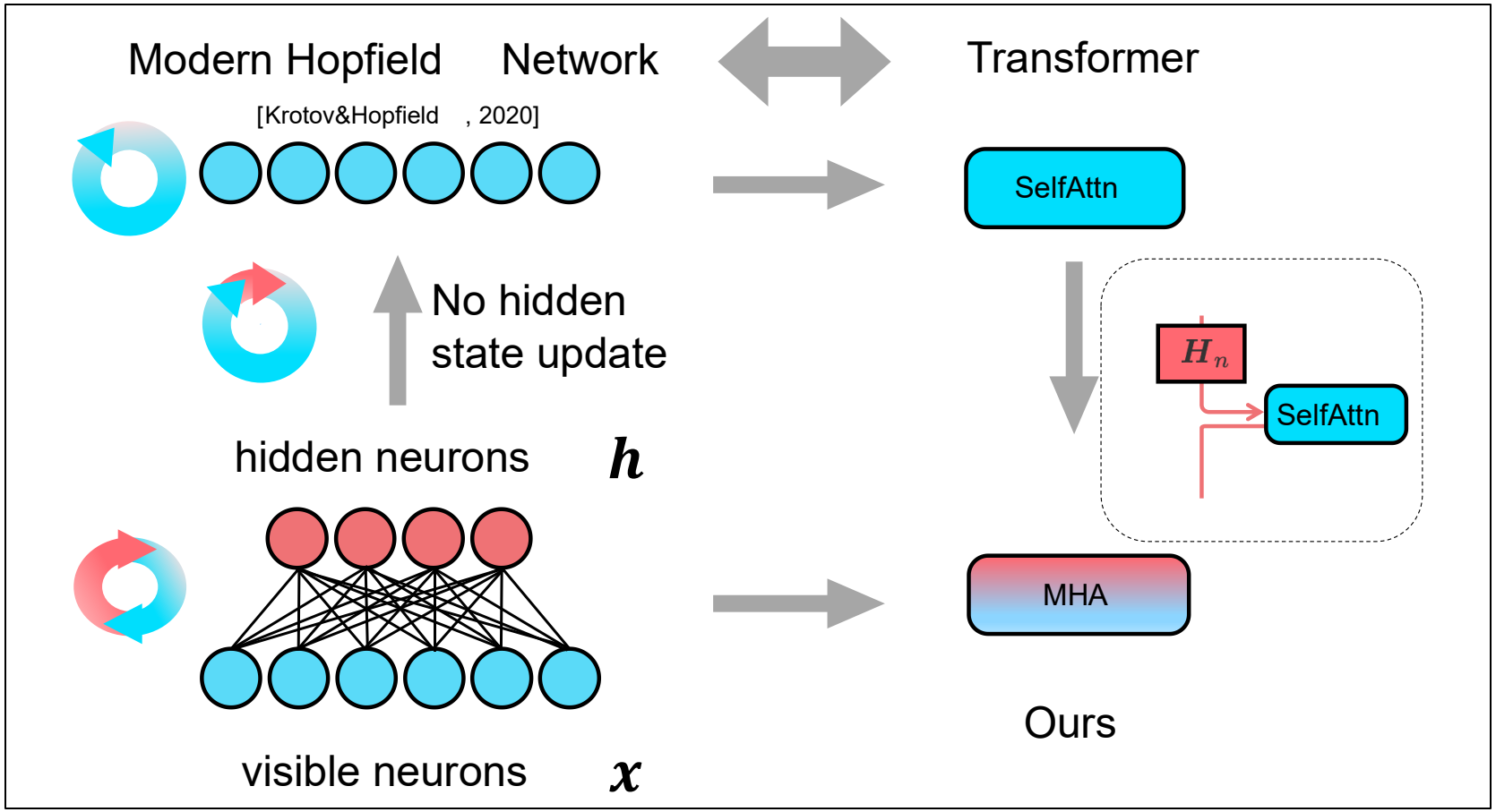
$$\tau_v \frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{h})\boldsymbol{W}_1^\top - \boldsymbol{x},$$

$$\tau_h \frac{d\boldsymbol{h}}{dt} = \boldsymbol{g}(\boldsymbol{x})\boldsymbol{W}_2 - \boldsymbol{h}.$$

Background

NEURAL INFORMATION PROCESSING SYSTEMS

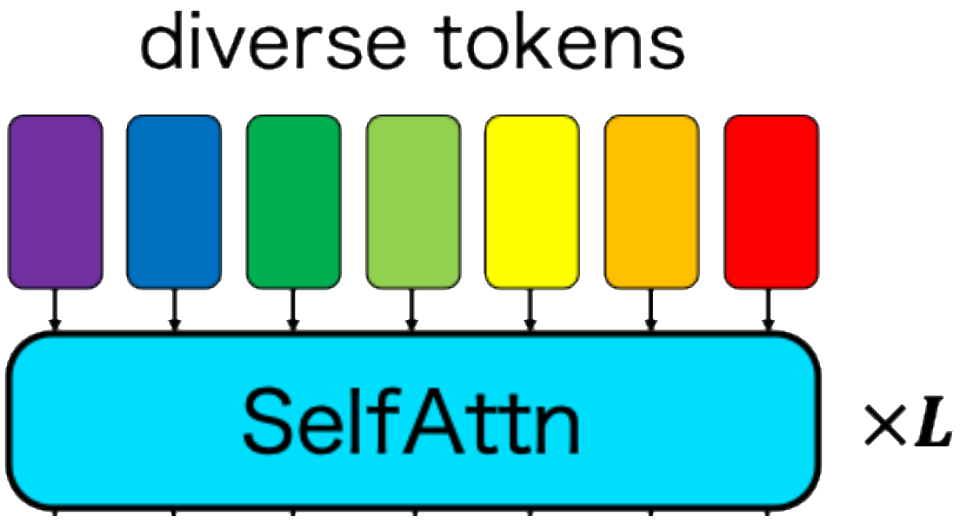
MHN ⇔ Transformer [Ramsauer et al, 2020], [Krotov&Hopfield, 2020]



Background

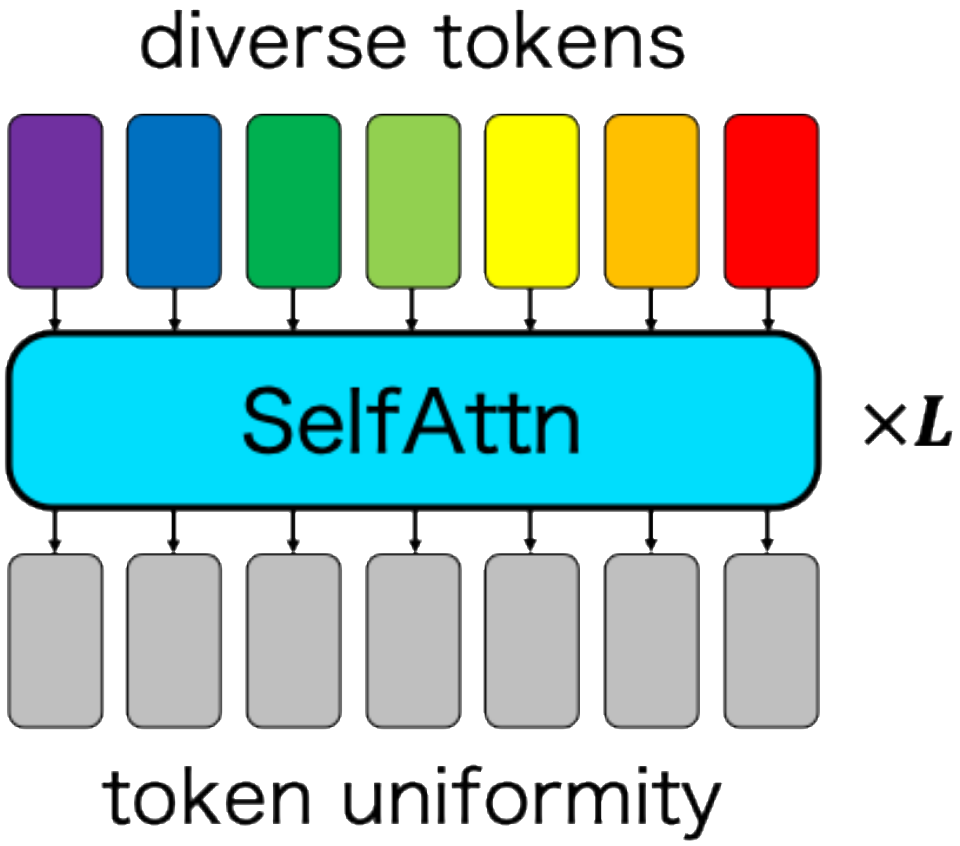
NEURAL INFORMATION PROCESSING SYSTEMS

Token Uniformity/Rank collapse



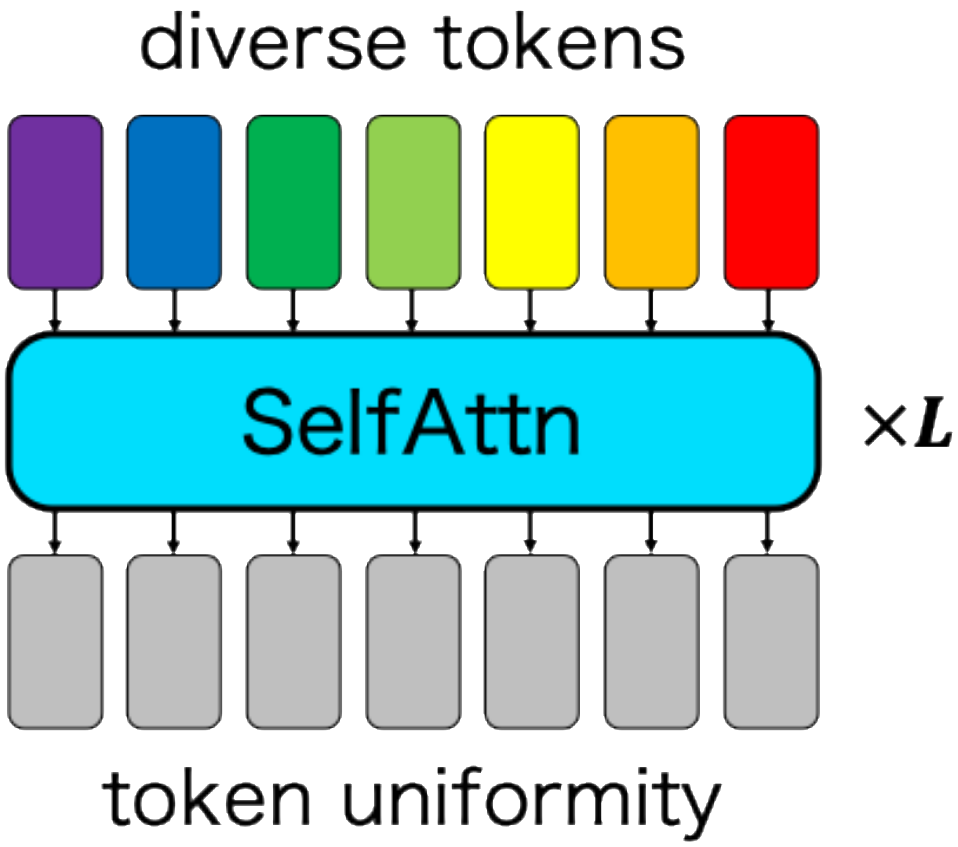
Background

Token Uniformity/Rank collapse



Background

Token Uniformity/Rank collapse



double-exponential decay

Theorem [Dong et al., 2021]

$$\|\mathrm{Res}\left(\mathrm{AttnNet}(\boldsymbol{X})\right)\|_{1,\infty} \leq (rC)^{\frac{3^L-1}{2}} \|\mathrm{Res}\left(\boldsymbol{X}\right)\|_{1,\infty}^{3^L},$$

$$\mathrm{Res}(\boldsymbol{X}) = \boldsymbol{X} - \mathbf{1}\boldsymbol{x}^\top \text{ for } \boldsymbol{x} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{X} - \mathbf{1}\boldsymbol{x}^\top\|.$$

Architecture

NEURAL INFORMATION PROCESSING SYSTEMS

Modern Hopfield Attention(MHA)

MHA operates through the following steps:

1. Compute attention logit

$$\boldsymbol{A}_n = \boldsymbol{Q}_n \boldsymbol{K}_n^\top$$

2. Add hidden state to the attention logit

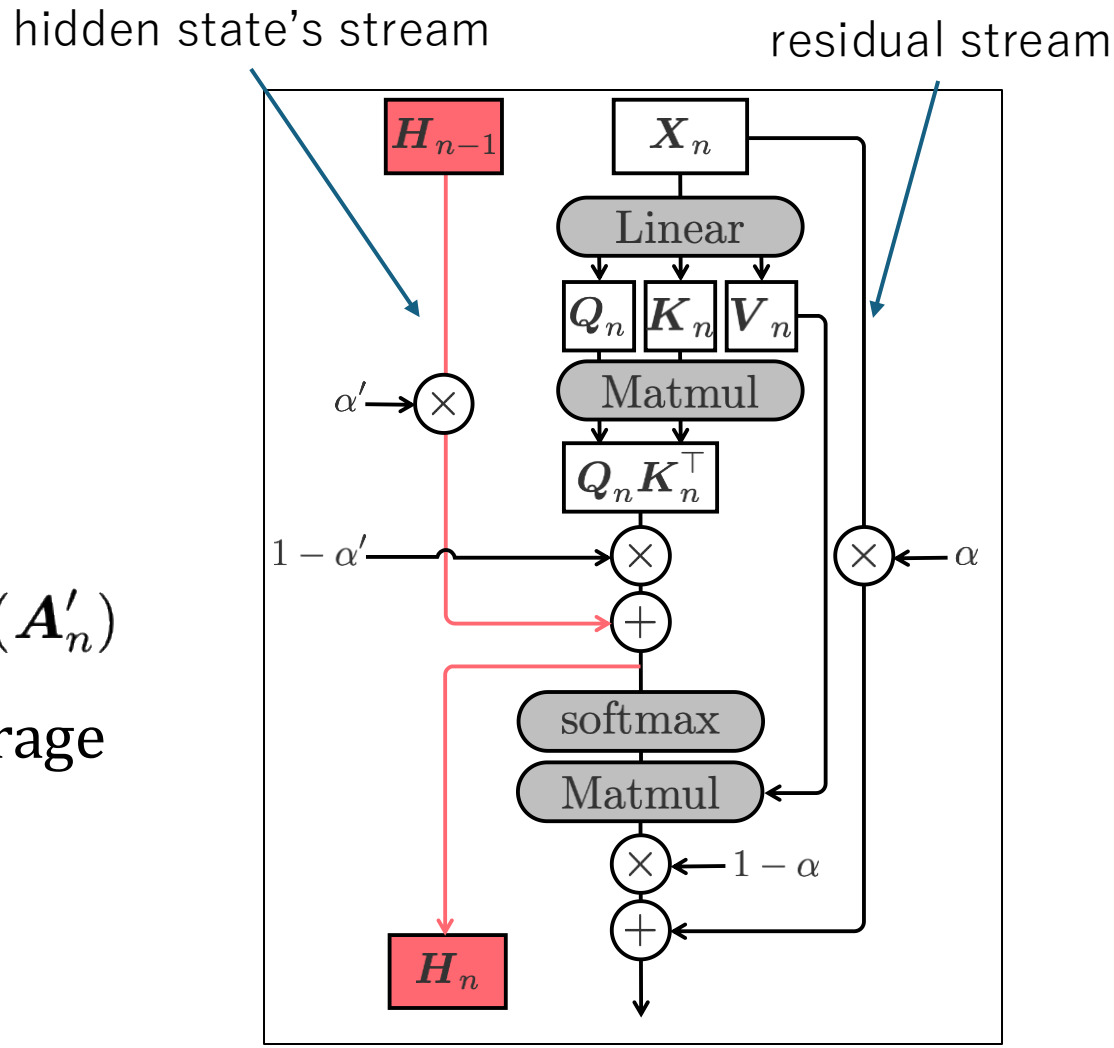
$$\boldsymbol{A}'_n = \alpha' \boldsymbol{H}_{n-1} + (1-\alpha')\boldsymbol{A}_n, \quad \boldsymbol{S}_n = \mathrm{softmax}(\boldsymbol{A}'_n)$$

3. Update hidden state with the moving average

$$\boldsymbol{H}_n = \boldsymbol{A}'_n$$

4. Apply the weighted residual connection

$$\boldsymbol{X}_{n+1} = (1-\alpha)\boldsymbol{S}_n \boldsymbol{V}_n + \alpha \boldsymbol{X}_n$$



Architecture

NEURAL INFORMATION PROCESSING SYSTEMS

Modern Hopfield Attention(MHA)

MHA operates through the following steps:

1. Compute attention logit

$$\boldsymbol{A}_n = \boldsymbol{Q}_n \boldsymbol{K}_n^\top$$

2. Add hidden state to the attention logit

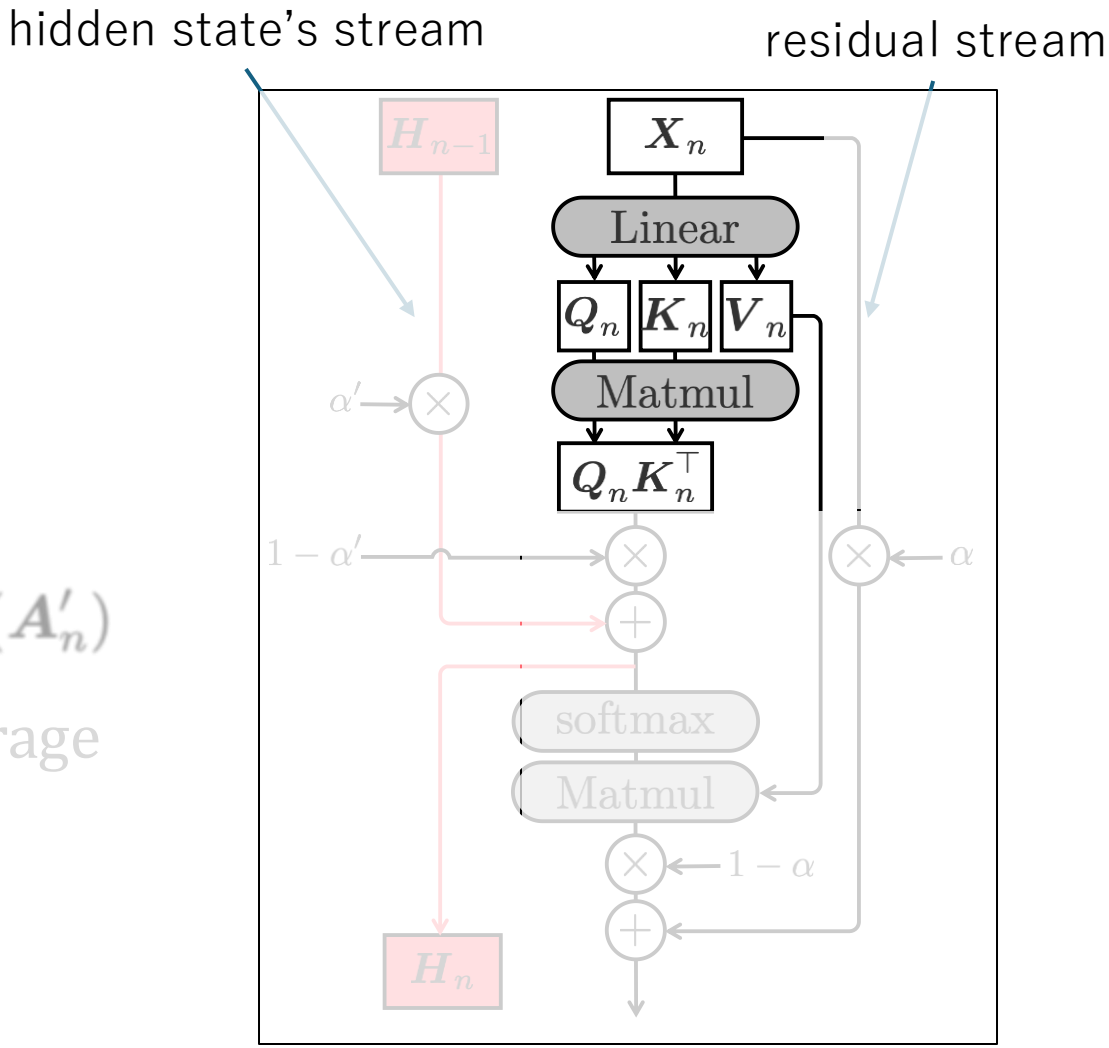
$$\boldsymbol{A}'_n = \alpha' \boldsymbol{H}_{n-1} + (1-\alpha')\boldsymbol{A}_n, \quad \boldsymbol{S}_n = \mathrm{softmax}(\boldsymbol{A}'_n)$$

3. Update hidden state with the moving average

$$\boldsymbol{H}_n = \boldsymbol{A}'_n$$

4. Apply the weighted residual connection

$$\boldsymbol{X}_{n+1} = (1-\alpha)\boldsymbol{S}_n \boldsymbol{V}_n + \alpha \boldsymbol{X}_n$$



Architecture

NEURAL INFORMATION PROCESSING SYSTEMS

Modern Hopfield Attention(MHA)

MHA operates through the following steps:

1. Compute attention logit

$$\boldsymbol{A}_n = \boldsymbol{Q}_n \boldsymbol{K}_n^\top$$

2. Add hidden state to the attention logit

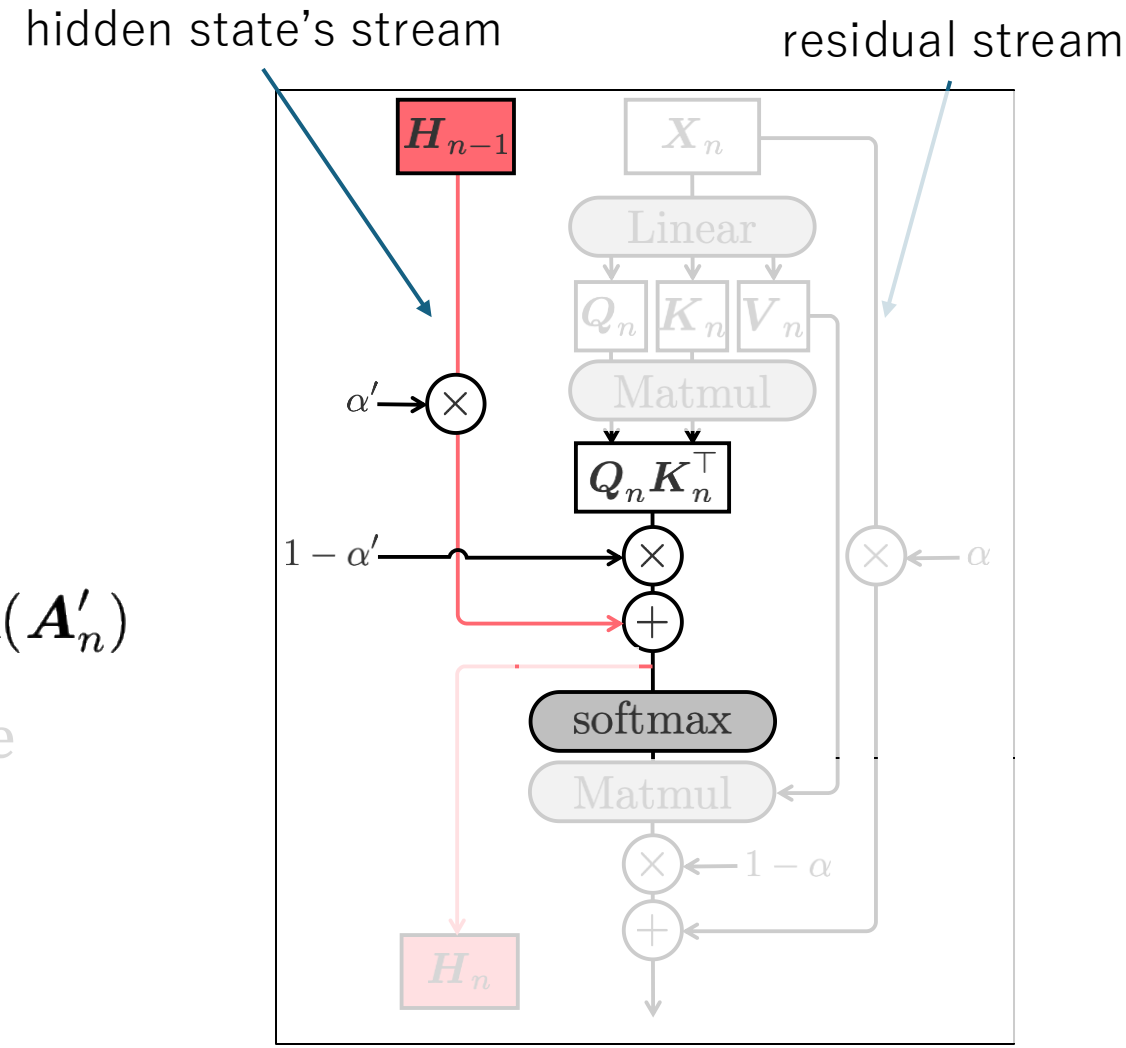
$$\boldsymbol{A}'_n = \alpha' \boldsymbol{H}_{n-1} + (1-\alpha')\boldsymbol{A}_n, \quad \boldsymbol{S}_n = \mathrm{softmax}(\boldsymbol{A}'_n)$$

3. Update hidden state with moving average

$$\boldsymbol{H}_n = \boldsymbol{A}'_n$$

4. Apply the weighted residual connection

$$\boldsymbol{X}_{n+1} = (1-\alpha)\boldsymbol{S}_n \boldsymbol{V}_n + \alpha \boldsymbol{X}_n$$



Architecture

NEURAL INFORMATION PROCESSING SYSTEMS

Modern Hopfield Attention(MHA)

MHA operates through the following steps:

1. Compute attention logit

$$\boldsymbol{A}_n = \boldsymbol{Q}_n \boldsymbol{K}_n^\top$$

2. Add hidden state to the attention logit

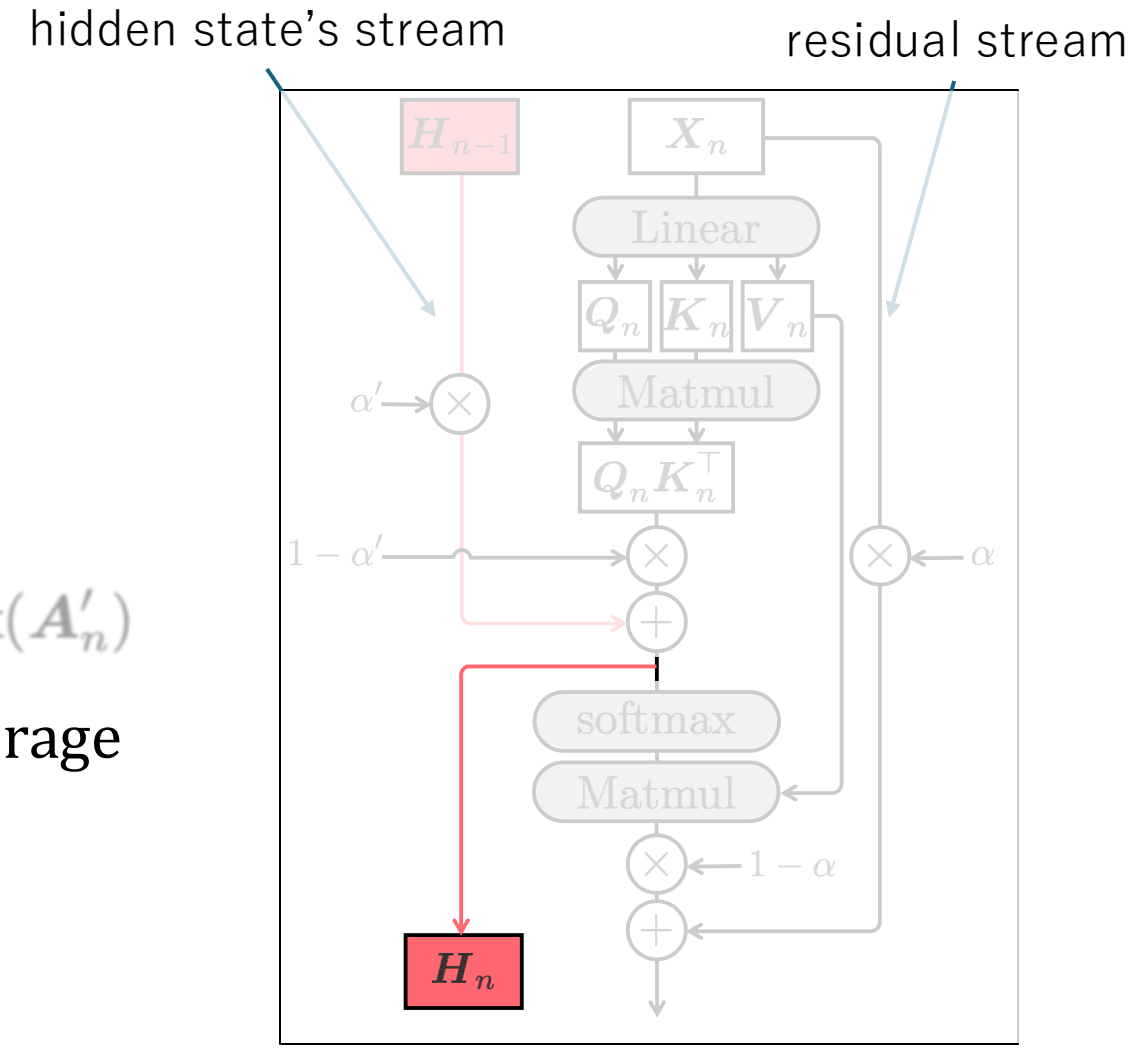
$$\boldsymbol{A}'_n = \alpha' \boldsymbol{H}_{n-1} + (1-\alpha')\boldsymbol{A}_n, \quad \boldsymbol{S}_n = \mathrm{softmax}(\boldsymbol{A}'_n)$$

3. Update hidden state with the moving average

$$\boldsymbol{H}_n = \boldsymbol{A}'_n$$

4. Apply the weighted residual connection

$$\boldsymbol{X}_{n+1} = (1-\alpha)\boldsymbol{S}_n \boldsymbol{V}_n + \alpha \boldsymbol{X}_n$$



Architecture

NEURAL INFORMATION PROCESSING SYSTEMS

Modern Hopfield Attention(MHA)

MHA operates through the following steps:

1. Compute attention logit

$$\boldsymbol{A}_n = \boldsymbol{Q}_n \boldsymbol{K}_n^\top$$

2. Add hidden state to the attention logit

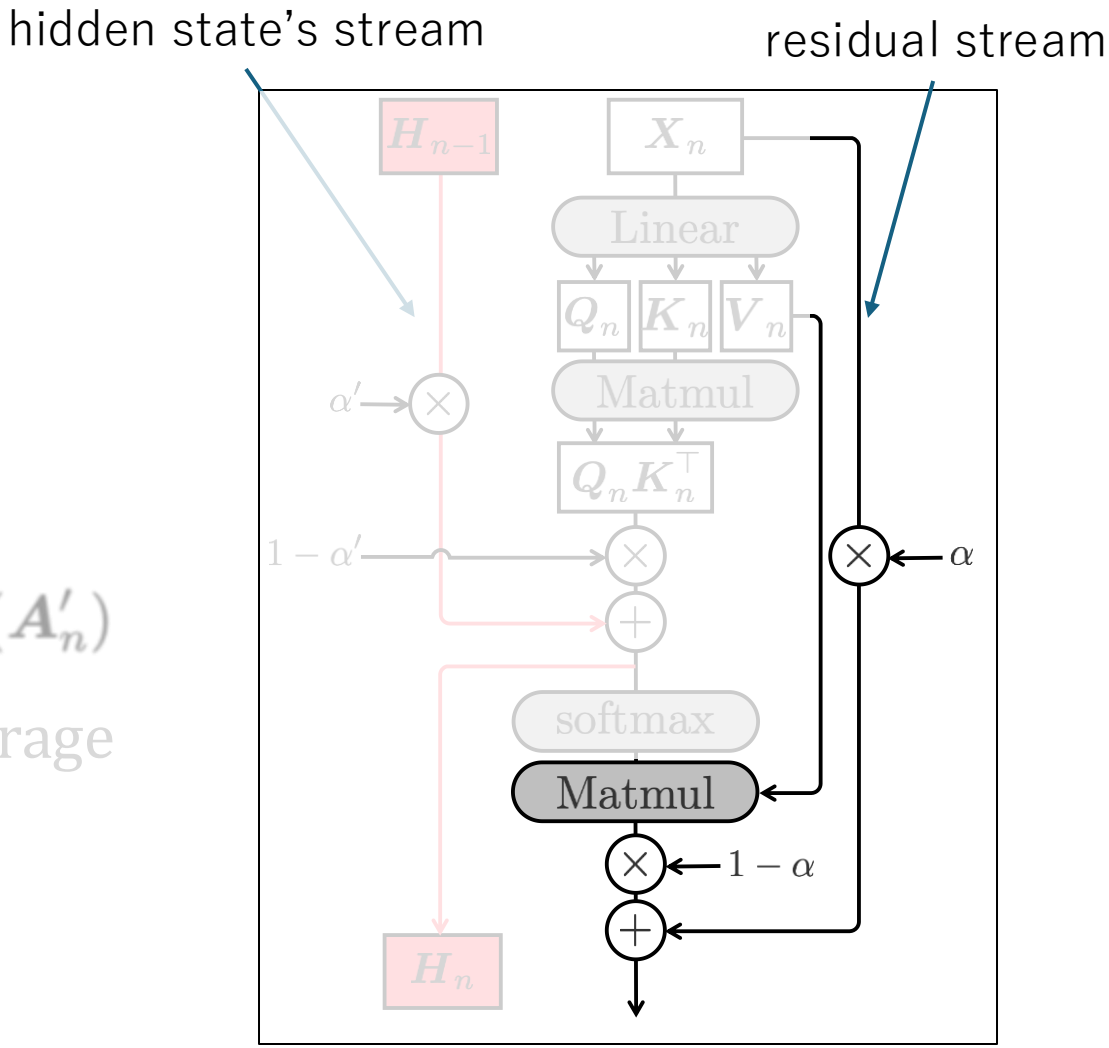
$$\boldsymbol{A}'_n = \alpha' \boldsymbol{H}_{n-1} + (1-\alpha')\boldsymbol{A}_n, \quad \boldsymbol{S}_n = \mathrm{softmax}(\boldsymbol{A}'_n)$$

3. Update hidden state with the moving average

$$\boldsymbol{H}_n = \boldsymbol{A}'_n$$

4. Apply the weighted residual connection

$$\boldsymbol{X}_{n+1} = (1-\alpha)\boldsymbol{S}_n \boldsymbol{V}_n + \alpha \boldsymbol{X}_n$$



Experimental Result

NEURAL INFORMATION PROCESSING SYSTEMS

Pre-train

Vision Tasks(acc)				
model size	model type	CIFAR10	CIFAR100	ImageNet-1k
ViT-Base(86M)	self-attention	96.190	75.360	76.074
	MHA($\alpha = 0.5$)	96.175	**76.215**	76.434
	MHA($\alpha = 0.7$)	**96.490**	75.590	**77.058**

Language Tasks(ppl)				
model size	model type	WikiText-103	DailyMail	BoocCorpus
MiniLlama(43M)	self-attention	14.49	19.36	23.76
	MHA($\alpha = 0.5$)	**14.29**	**18.97**	**23.50**

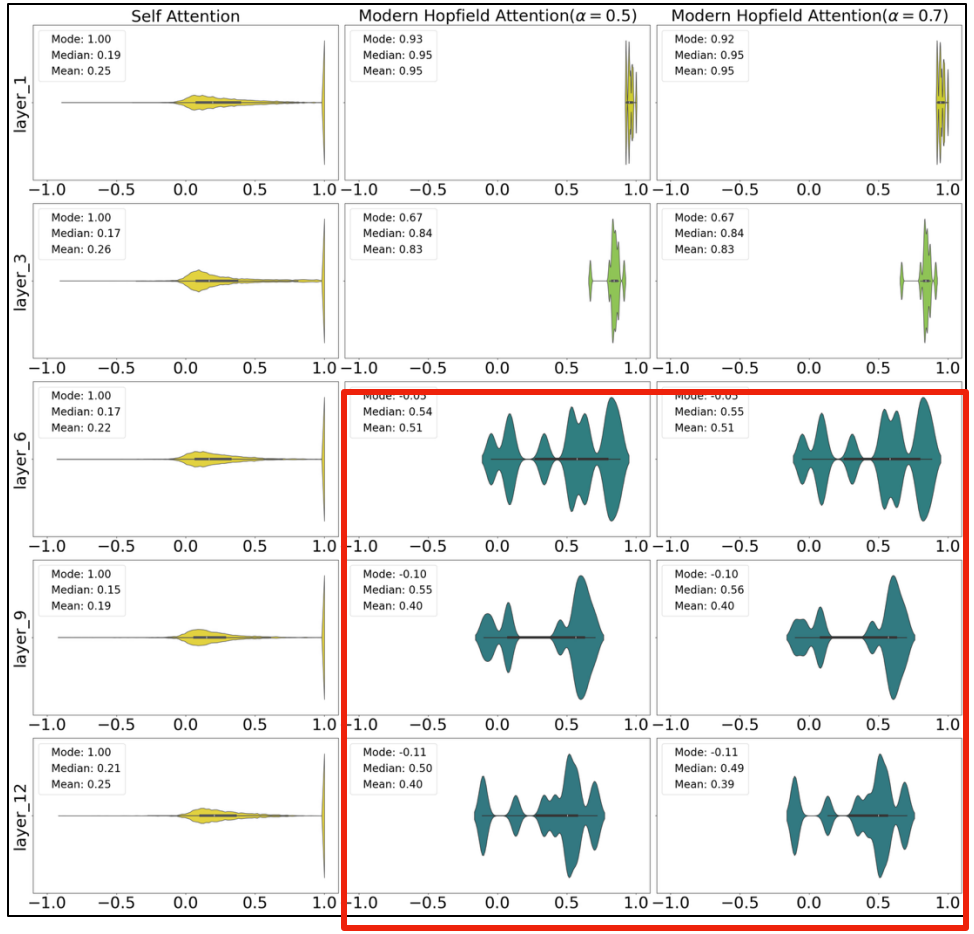
Experimental Result

NEURAL INFORMATION PROCESSING SYSTEMS

Downstream

Vision Tasks(acc)					
model size	model type	Flower102	Food101	stanford dogs	Stanford cars
ViT-Base(86M)	self-attention	81.15	74.51	**95.00**	51.54
	MHA($\alpha = 0.7$)	**93.85**	**87.99**	83.64	**87.54**

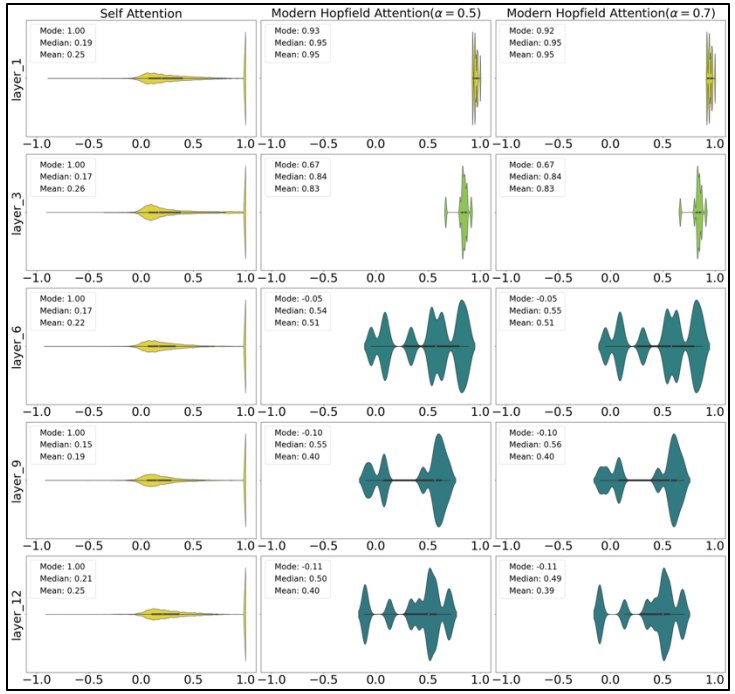
Rank collapse Analysis



diverse distribution

↑token-similarity for ViT-B training with CIFR100

Rank collapse Analysis



Theorem: upper bound of MHA

$$\|\mathrm{Res}\big(\mathrm{AttnNet}(\boldsymbol{X})\big)\|_{1,\infty} \leq \max_{m=0}^{L} \left(r(1-\alpha')C_1\right)^{\frac{3^m+1}{2}} \left(r\alpha' C_2\right)^{3^m(L-m)} \|\mathrm{Res}\big(\boldsymbol{X}\big)\|_{1,\infty}^{3^m}.$$

$m = 0$ term mitigates collapse