Linear Attention for Efficient Bidirectional Sequence Modeling

Arshia Afzal

EPFL

Not all Sequences are Causal

Sometimes all tokens are available in tasks like:

- Image classification
- Masked Language Modeling (MLM)
- Diffusion LLMs
- DNA modeling

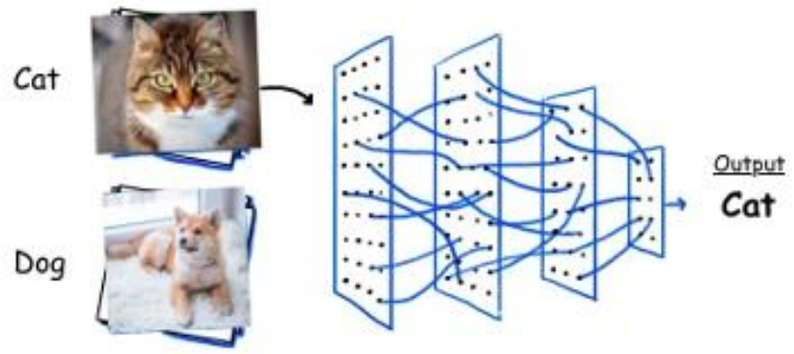
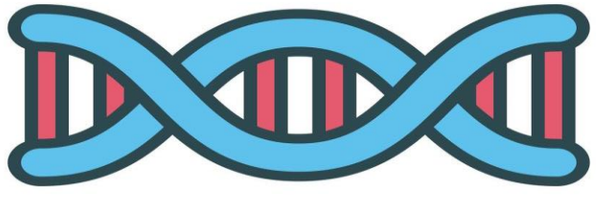
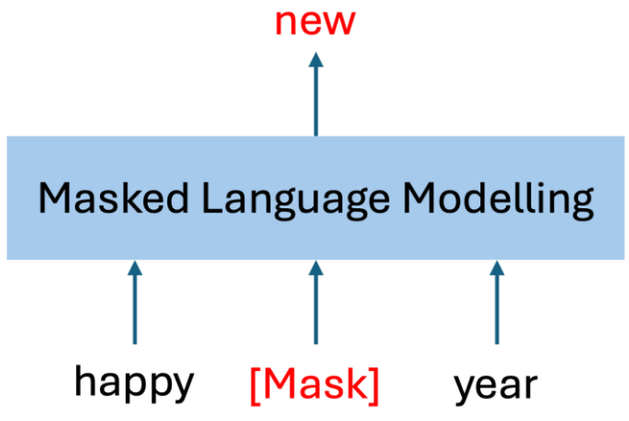


Image classification



DNA Modeling

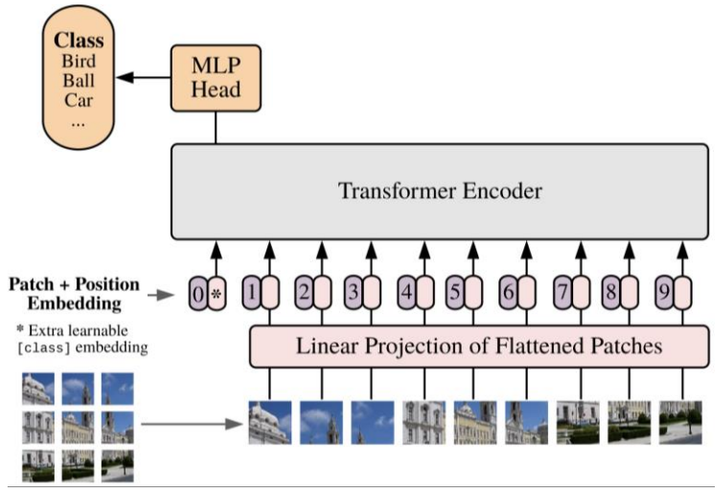


Masked Language Modeling

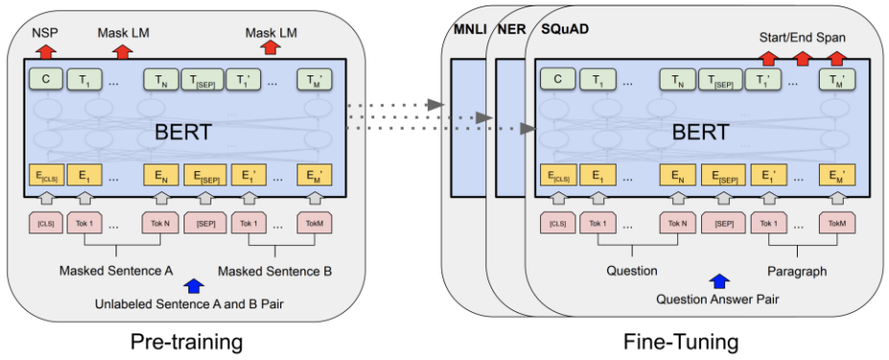
Bi-directional Transformers

Examples of Bi-directional Transformers:

- Vanilla Transformer [Machine Translation]
- Vision Transformers (ViT) [Vision]
- Bidirectional Encoder Representations from Transformers (BERT) [MLM]
- Large Language Diffusion Models (LLaDA)



Vision Transformers (ViT)

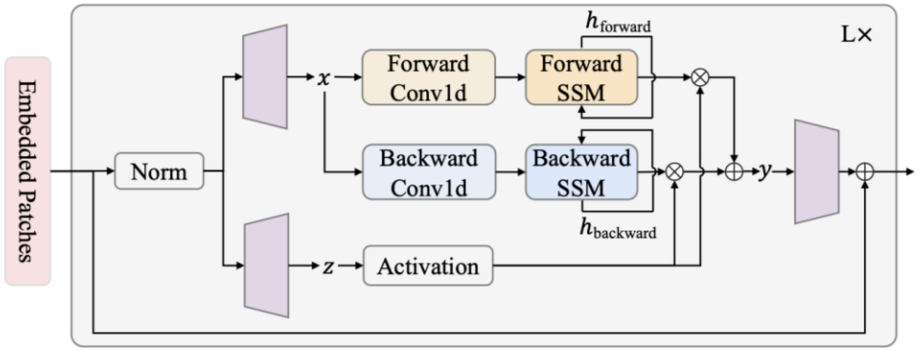


BERT

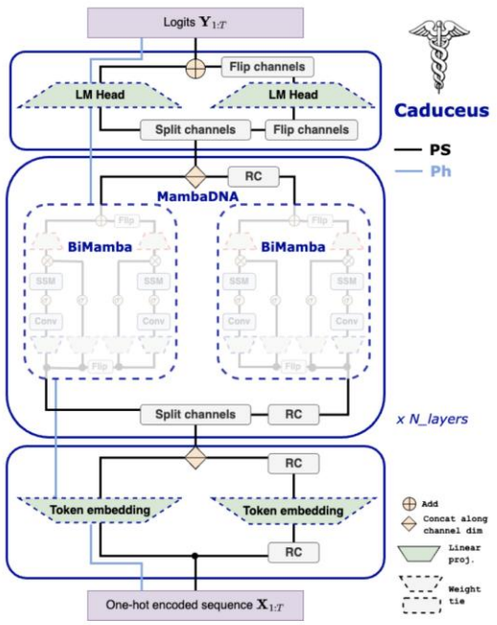
Bi-directional SSMs

Examples of Bi-directional SSMs:

- Vision Mamba (Vim) [Vision]
- Hydra [Vision, MLM]
- Caduceus [DNA]
- Lyra [DNA]



Vision Mamba (Vim)



Caduceus

Hydra

Transformer–SSM Trade-offs (in Bidirectional Form)

Model	Training Speed	Inference Efficiency
Transformer	✅	❌
SSM	❌	✅

Transformer–SSM Trade-offs (in Bidirectional Form)

Model	Training Speed	Inference Efficiency
Transformer	✅	❌
SSM	❌	✅

Model	Top-1 Acc. (%)		Train Time (↓)	
	Small	Base	Small	Base
ViT	72.2	77.9	×1.00	×1.00
DeiT	79.8	81.8	×1.00	×1.00
Hydra	78.6	81.0	×2.50	×2.51
Vim	**80.3**	**81.9**	×14.95	×10.86

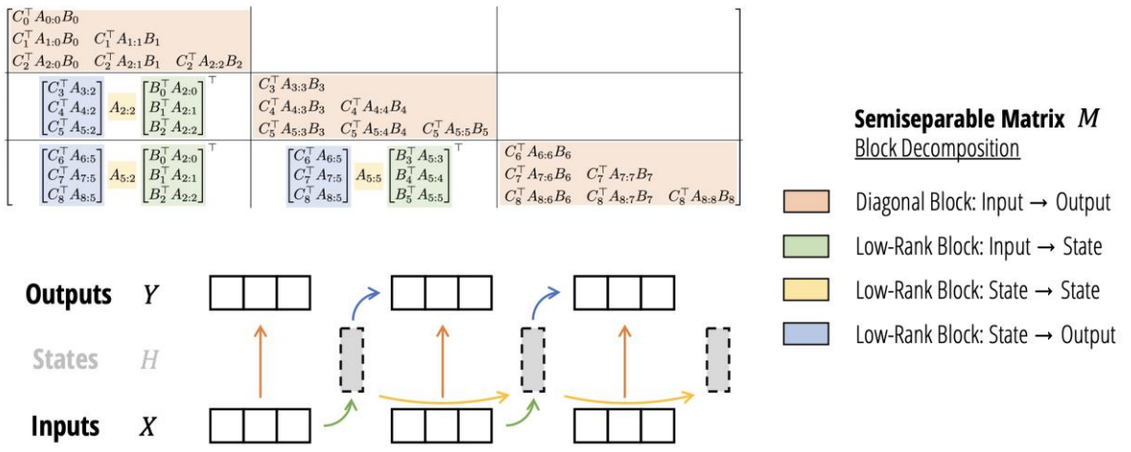
Training Speed on ImageNet-1K Classification

Transformer–SSM Trade-offs (in Bidirectional Form)

Model	Training Speed	Inference Efficiency
Transformer	✅	❌
SSM	❌	✅

Model	Top-1 Acc. (%)		Train Time (↓)	
	Small	Base	Small	Base
ViT	72.2	77.9	×1.00	×1.00
DeiT	79.8	81.8	×1.00	×1.00
Hydra	78.6	81.0	×2.50	×2.51
Vim	**80.3**	**81.9**	×14.95	×10.86

Training Speed on ImageNet-1K Classification



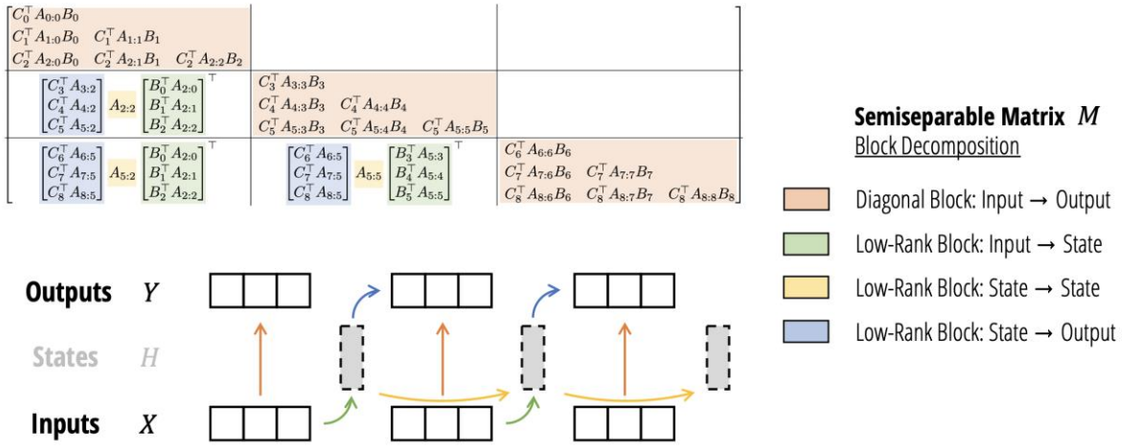
2 SSD Algorithm

Transformer–SSM Trade-offs (in Bidirectional Form)

Model	Training Speed	Inference Efficiency
Transformer	✅	❌
SSM	❌	✅

Model	Top-1 Acc. (%)		Train Time (↓)	
	Small	Base	Small	Base
ViT	72.2	77.9	×1.00	×1.00
DeiT	79.8	81.8	×1.00	×1.00
Hydra	78.6	81.0	×2.50	×2.51
Vim	**80.3**	**81.9**	×14.95	×10.86

Training Speed on ImageNet-1K Classification



2 SSD Algorithm

💥 Flash Linear Attention: Fast Chunkwise Parallel Training

Why Bi-directional Linear Attention?

Can Bidirectional Linear Attention …

- Train as fast as Transformers in fully parallel form?
- Be as efficient as SSMs during inference?
- Support chunkwise processing for balanced memory–speed trade-offs?

Causal Linear Transformers

$$(\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i) = (\mathbf{x}_i \mathbf{W}_\mathbf{q},\ \mathbf{x}_i \mathbf{W}_\mathbf{k},\ \mathbf{x}_i \mathbf{W}_\mathbf{v})$$

$$\mathbf{Y} = \mathrm{softmax}\big(\mathbf{Q}\mathbf{K}^\top + \mathbf{M}^C\big)\,\mathbf{V}, \quad \mathbf{y}_i = \sum_{j=1}^{i} \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_j)}{\sum_{p=1}^{i} \exp(\mathbf{q}_i^\top \mathbf{k}_p)} \mathbf{v}_j, \quad \mathbf{M}^C \in \{-\infty, 0\}^{L \times L}$$

Causal Linear Transformers

$$(\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i) = (\mathbf{x}_i \mathbf{W}_\mathbf{q},\ \mathbf{x}_i \mathbf{W}_\mathbf{k},\ \mathbf{x}_i \mathbf{W}_\mathbf{v})$$

$$\mathbf{Y} = \mathrm{softmax}\big(\mathbf{Q}\mathbf{K}^\top + \mathbf{M}^C\big)\,\mathbf{V}, \quad \mathbf{y}_i = \sum_{j=1}^{i} \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_j)}{\sum_{p=1}^{i} \exp(\mathbf{q}_i^\top \mathbf{k}_p)} \mathbf{v}_j, \quad \mathbf{M}^C \in \{-\infty, 0\}^{L \times L}$$

$$\mathbf{Y} = \textsc{Scale}\left(\mathbf{Q}\mathbf{K}^\top \odot \mathbf{M}^C\right)\mathbf{V}, \quad \mathbf{S}_i = \mathbf{S}_{i-1} + \mathbf{k}_i \mathbf{v}_i^\top, \quad \mathbf{z}_i = \mathbf{z}_{i-1} + \mathbf{k}_i, \quad \mathbf{y}_i = \frac{\mathbf{q}_i^\top \mathbf{S}_i}{\mathbf{q}_i^\top \mathbf{z}_i}, \quad \mathbf{M}^C \in \{1, 0\}^{L \times L}$$

Causal Linear Transformers

$$(\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i) = (\mathbf{x}_i \mathbf{W}_\mathbf{q},\ \mathbf{x}_i \mathbf{W}_\mathbf{k},\ \mathbf{x}_i \mathbf{W}_\mathbf{v})$$

$$\mathbf{Y} = \mathrm{softmax}\big(\mathbf{Q}\mathbf{K}^\top + \mathbf{M}^C\big)\,\mathbf{V}, \quad \mathbf{y}_i = \sum_{j=1}^{i} \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_j)}{\sum_{p=1}^{i} \exp(\mathbf{q}_i^\top \mathbf{k}_p)} \mathbf{v}_j, \quad \mathbf{M}^C \in \{-\infty, 0\}^{L \times L}$$

$$\mathbf{Y} = \textsc{Scale}\left(\mathbf{Q}\mathbf{K}^\top \odot \mathbf{M}^C\right)\mathbf{V}, \quad \mathbf{S}_i = \mathbf{S}_{i-1} + \mathbf{k}_i \mathbf{v}_i^\top, \quad \mathbf{z}_i = \mathbf{z}_{i-1} + \mathbf{k}_i, \quad \mathbf{y}_i = \frac{\mathbf{q}_i^\top \mathbf{S}_i}{\mathbf{q}_i^\top \mathbf{z}_i}, \quad \mathbf{M}^C \in \{1, 0\}^{L \times L}$$

$$\mathbf{Y} = \textsc{Scale}\left(\mathbf{Q}\mathbf{K}^\top \odot \mathbf{M}^C\right)\mathbf{V}, \quad \mathbf{S}_i = \lambda_i \mathbf{S}_{i-1} + \mathbf{k}_i \mathbf{v}_i^\top, \quad \mathbf{M}^C_{ij} = \begin{cases} \Pi_{k=j+1}^{i} \lambda_k, & i \geq j; \\ 0, & i < j. \end{cases}$$

Causal Linear Transformers

Model	Recurrence	Memory read-out
Linear Attention [48, 47]	$\mathbf{S}_t = \mathbf{S}_{t-1} + \boldsymbol{v}_t \boldsymbol{k}_t^\top$	$\boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t$
+ Kernel	$\mathbf{S}_t = \mathbf{S}_{t-1} + \boldsymbol{v}_t \phi(\boldsymbol{k}_t)^\top$	$\boldsymbol{o}_t = \mathbf{S}_t \phi(\boldsymbol{q}_t)$
+ Normalization	$\mathbf{S}_t = \mathbf{S}_{t-1} + \boldsymbol{v}_t \phi(\boldsymbol{k}_t)^\top, \;\; \boldsymbol{z}_t = \boldsymbol{z}_{t-1} + \phi(\boldsymbol{k}_t)$	$\boldsymbol{o}_t = \mathbf{S}_t \phi(\boldsymbol{q}_t) / (\boldsymbol{z}_t^\top \phi(\boldsymbol{q}_t))$
DeltaNet [101]	$\mathbf{S}_t = \mathbf{S}_{t-1}(\mathbf{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\top) + \beta_t \boldsymbol{v}_t \boldsymbol{k}_t^\top$	$\boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t$
Gated RFA [81]	$\mathbf{S}_t = g_t \mathbf{S}_{t-1} + (1 - g_t) \boldsymbol{v}_t \boldsymbol{k}_t^\top, \;\; \boldsymbol{z}_t = g_t \boldsymbol{z}_{t-1} + (1 - g_t) \boldsymbol{k}_t$	$\boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t / (\boldsymbol{z}_t^\top \boldsymbol{q}_t)$
S4 [32, 106]	$\mathbf{S}_t = \mathbf{S}_{t-1} \odot \exp(-(\boldsymbol{\alpha} \mathbf{1}^\top) \odot \exp(\boldsymbol{A})) + \boldsymbol{B} \odot (\boldsymbol{v}_t \mathbf{1}^\top)$	$\boldsymbol{o}_t = (\mathbf{S}_t \odot \boldsymbol{C}) \mathbf{1} + \boldsymbol{d} \odot \boldsymbol{v}_t$
ABC [82]	$\mathbf{S}_t^{\boldsymbol{k}} = \mathbf{S}_{t-1}^{\boldsymbol{k}} + \boldsymbol{k}_t \boldsymbol{\phi}_t^\top, \;\; \mathbf{S}_t^{\boldsymbol{v}} = \mathbf{S}_{t-1}^{\boldsymbol{v}} + \boldsymbol{v}_t \boldsymbol{\phi}_t^\top$	$\boldsymbol{o}_t = \mathbf{S}_t^{\boldsymbol{v}} \operatorname{softmax}\left(\mathbf{S}_t^{\boldsymbol{k}} \boldsymbol{q}_t\right)$
DFW [63]	$\mathbf{S}_t = \mathbf{S}_{t-1} \odot (\boldsymbol{\beta}_t \boldsymbol{\alpha}_t^\top) + \boldsymbol{v}_t \boldsymbol{k}_t^\top$	$\boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t$
RetNet [108]	$\mathbf{S}_t = \gamma \mathbf{S}_{t-1} + \boldsymbol{v}_t \boldsymbol{k}_t^\top$	$\boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t$
Mamba [31]	$\mathbf{S}_t = \mathbf{S}_{t-1} \odot \exp(-(\boldsymbol{\alpha}_t \mathbf{1}^\top) \odot \exp(\boldsymbol{A})) + (\boldsymbol{\alpha}_t \odot \boldsymbol{v}_t) \boldsymbol{k}_t^\top$	$\boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t + \boldsymbol{d} \odot \boldsymbol{v}_t$
GLA [124]	$\mathbf{S}_t = \mathbf{S}_{t-1} \odot (\mathbf{1} \boldsymbol{\alpha}_t^\top) + \boldsymbol{v}_t \boldsymbol{k}_t^\top = \mathbf{S}_{t-1} \mathrm{Diag}(\boldsymbol{\alpha}_t) + \boldsymbol{v}_t \boldsymbol{k}_t^\top$	$\boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t$
RWKV-6 [79]	$\mathbf{S}_t = \mathbf{S}_{t-1} \mathrm{Diag}(\boldsymbol{\alpha}_t) + \boldsymbol{v}_t \boldsymbol{k}_t^\top$	$\boldsymbol{o}_t = (\mathbf{S}_{t-1} + (\boldsymbol{d} \odot \boldsymbol{v}_t) \boldsymbol{k}_t^\top) \boldsymbol{q}_t$
HGRN-2 [92]	$\mathbf{S}_t = \mathbf{S}_{t-1} \mathrm{Diag}(\boldsymbol{\alpha}_t) + \boldsymbol{v}_t (\mathbf{1} - \boldsymbol{\alpha}_t)^\top$	$\boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t$
mLSTM [9]	$\mathbf{S}_t = f_t \mathbf{S}_{t-1} + i_t \boldsymbol{v}_t \boldsymbol{k}_t^\top, \;\; \boldsymbol{z}_t = f_t \boldsymbol{z}_{t-1} + i_t \boldsymbol{k}_t$	$\boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t / \max\{1, \lvert \boldsymbol{z}_t^\top \boldsymbol{q}_t \rvert\}$
Mamba-2 [19]	$\mathbf{S}_t = \gamma_t \mathbf{S}_{t-1} + \boldsymbol{v}_t \boldsymbol{k}_t^\top$	$\boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t$
GSA [131]	$\mathbf{S}_t^{\boldsymbol{k}} = \mathbf{S}_{t-1}^{\boldsymbol{k}} \mathrm{Diag}(\boldsymbol{\alpha}_t) + \boldsymbol{k}_t \boldsymbol{\phi}_t^\top, \;\; \mathbf{S}_t^{\boldsymbol{v}} = \mathbf{S}_{t-1}^{\boldsymbol{v}} \mathrm{Diag}(\boldsymbol{\alpha}_t) + \boldsymbol{v}_t \boldsymbol{\phi}_t^\top$	$\boldsymbol{o}_t = \mathbf{S}_t^{\boldsymbol{v}} \operatorname{softmax}\left(\mathbf{S}_t^{\boldsymbol{k}} \boldsymbol{q}_t\right)$
Gated DeltaNet [125]	$\mathbf{S}_t = \mathbf{S}_{t-1} \left(\alpha_t (\mathbf{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\top)\right) + \beta_t \boldsymbol{v}_t \boldsymbol{k}_t^\top$	$\boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t$

❓Can have bi-directional framework for above linear transformers.

Key Components of Bidirectional Linear Transformers

- Full Linear Attention
- Equal Bi-directional RNN
- Chunkwise Parallel Representation

Key Representation of Bidirectional Linear Transformers

- Full Linear Attention
- Equal Bi-directional RNN
- Chunkwise Parallel Representation

Full Linear Attention

$$\mathbf{Y} = \mathrm{softmax}\left(\mathbf{Q}\mathbf{K}^{\top}\right)\mathbf{V}$$

Full Softmax Attention

Full Linear Attention

$$\mathbf{Y} = \mathrm{softmax}\left(\mathbf{Q}\mathbf{K}^\top\right)\mathbf{V}$$ Full Softmax Attention

$$\mathbf{Y} = \textsc{Scale}\left(\mathbf{Q}\mathbf{K}^\top \odot \mathbf{M}\right)\mathbf{V}$$ Full Linear Attention

Full Linear Attention

$$\mathbf{Y} = \mathrm{softmax}\left(\mathbf{Q}\mathbf{K}^\top\right)\mathbf{V}$$ Full Softmax Attention

$$\mathbf{Y} = \textsc{Scale}\left(\mathbf{Q}\mathbf{K}^\top \odot \mathbf{M}\right)\mathbf{V}$$ Full Linear Attention

$$\mathbf{M}_{ij} = \begin{cases} \Pi_{k=j}^{i-1}\lambda_k, & i > j \\ 1 & i = j \\ \Pi_{k=i+1}^{j}\lambda_k, & i < j. \end{cases}, \qquad \mathbf{M}_{ij} = \lambda^{|i-j|}, \qquad \mathbf{M}_{ij} = 1.$$

LION-Full Linear Attention

$$\mathbf{Y} = \mathrm{softmax}\left(\mathbf{Q}\mathbf{K}^\top\right)\mathbf{V}$$ Full Softmax Attention

$$\mathbf{Y} = \textsc{Scale}\left(\mathbf{Q}\mathbf{K}^\top \odot \mathbf{M}\right)\mathbf{V}$$ Full Linear Attention

$$\mathbf{M}_{ij} = \begin{cases} \Pi_{k=j}^{i-1}\lambda_k, & i > j \\ 1 & i = j \\ \Pi_{k=i+1}^{j}\lambda_k, & i < j. \end{cases}, \quad \mathbf{M}_{ij} = \lambda^{|i-j|}, \quad \mathbf{M}_{ij} = 1.$$

$$\left(\underbrace{\begin{pmatrix} \mathbf{q}_1^\top\mathbf{k}_1 & \mathbf{q}_1^\top\mathbf{k}_2 & \cdots & \mathbf{q}_1^\top\mathbf{k}_L \\ \mathbf{q}_2^\top\mathbf{k}_1 & \mathbf{q}_2^\top\mathbf{k}_2 & \cdots & \mathbf{q}_2^\top\mathbf{k}_L \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_L^\top\mathbf{k}_1 & \mathbf{q}_L^\top\mathbf{k}_2 & \cdots & \mathbf{q}_L^\top\mathbf{k}_L \end{pmatrix}}_{\mathbf{A}=\mathbf{Q}\mathbf{K}^\top} \odot \underbrace{\begin{pmatrix} 1 & \lambda_2 & \lambda_2\lambda_3 & \cdots & \lambda_2\cdots\lambda_L \\ \lambda_1 & 1 & \lambda_3 & \cdots & \lambda_3\cdots\lambda_L \\ \lambda_1\lambda_2 & \lambda_2 & 1 & \cdots & \lambda_4\cdots\lambda_L \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{L-1}\cdots\lambda_1 & \lambda_{L-1}\cdots\lambda_2 & \lambda_{L-1}\cdots\lambda_3 & \cdots & 1 \end{pmatrix}}_{\mathbf{M}}\right)\begin{pmatrix} \mathbf{v}_1^\top \\ \mathbf{v}_2^\top \\ \mathbf{v}_3^\top \\ \vdots \\ \mathbf{v}_L^\top \end{pmatrix}$$

Anti-Causal

Causal

Diagonal

This is full parallel form for training!

Key Representation of Bidirectional Linear Transformers

- Full Linear Attention
- Equal Bi-directional RNN
- Chunkwise Parallel Representation

Equivalent Bi-directional RNN for Linear Attention

$$\underbrace{\begin{pmatrix} \mathbf{q}_1^\top \mathbf{k}_1 & \mathbf{q}_1^\top \mathbf{k}_2 & \cdots & \mathbf{q}_1^\top \mathbf{k}_L \\ \mathbf{q}_2^\top \mathbf{k}_1 & \mathbf{q}_2^\top \mathbf{k}_2 & \cdots & \mathbf{q}_2^\top \mathbf{k}_L \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_L^\top \mathbf{k}_1 & \mathbf{q}_L^\top \mathbf{k}_2 & \cdots & \mathbf{q}_L^\top \mathbf{k}_L \end{pmatrix}}_{\mathbf{A} = \mathbf{Q}\mathbf{K}^\top} = \underbrace{\begin{pmatrix} \frac{1}{2}\mathbf{q}_1^\top \mathbf{k}_1 & & & \\ \mathbf{q}_2^\top \mathbf{k}_1 & \frac{1}{2}\mathbf{q}_2^\top \mathbf{k}_2 & & \\ \vdots & \vdots & \ddots & \\ \mathbf{q}_L^\top \mathbf{k}_1 & \mathbf{q}_L^\top \mathbf{k}_2 & \cdots & \frac{1}{2}\mathbf{q}_L^\top \mathbf{k}_L \end{pmatrix}}_{\mathbf{A}^F} + \underbrace{\begin{pmatrix} \frac{1}{2}\mathbf{q}_1^\top \mathbf{k}_1 & \mathbf{q}_1^\top \mathbf{k}_2 & \cdots & \mathbf{q}_1^\top \mathbf{k}_L \\ & \frac{1}{2}\mathbf{q}_2^\top \mathbf{k}_2 & \cdots & \mathbf{q}_2^\top \mathbf{k}_L \\ & & \ddots & \vdots \\ & & & \frac{1}{2}\mathbf{q}_L^\top \mathbf{k}_L \end{pmatrix}}_{\mathbf{A}^B}$$

Equivalent Bi-directional RNN for Linear Attention

$$\underbrace{\begin{pmatrix} \mathbf{q}_1^\top\mathbf{k}_1 & \mathbf{q}_1^\top\mathbf{k}_2 & \cdots & \mathbf{q}_1^\top\mathbf{k}_L \\ \mathbf{q}_2^\top\mathbf{k}_1 & \mathbf{q}_2^\top\mathbf{k}_2 & \cdots & \mathbf{q}_2^\top\mathbf{k}_L \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_L^\top\mathbf{k}_1 & \mathbf{q}_L^\top\mathbf{k}_2 & \cdots & \mathbf{q}_L^\top\mathbf{k}_L \end{pmatrix}}_{\mathbf{A}=\mathbf{Q}\mathbf{K}^\top} = \underbrace{\begin{pmatrix} \frac{1}{2}\mathbf{q}_1^\top\mathbf{k}_1 & & & \\ \mathbf{q}_2^\top\mathbf{k}_1 & \frac{1}{2}\mathbf{q}_2^\top\mathbf{k}_2 & & \\ \vdots & \vdots & \ddots & \\ \mathbf{q}_L^\top\mathbf{k}_1 & \mathbf{q}_L^\top\mathbf{k}_2 & \cdots & \frac{1}{2}\mathbf{q}_L^\top\mathbf{k}_L \end{pmatrix}}_{\mathbf{A}^F} + \underbrace{\begin{pmatrix} \frac{1}{2}\mathbf{q}_1^\top\mathbf{k}_1 & \mathbf{q}_1^\top\mathbf{k}_2 & \cdots & \mathbf{q}_1^\top\mathbf{k}_L \\ & \frac{1}{2}\mathbf{q}_2^\top\mathbf{k}_2 & \cdots & \mathbf{q}_2^\top\mathbf{k}_L \\ & & \ddots & \vdots \\ & & & \frac{1}{2}\mathbf{q}_L^\top\mathbf{k}_L \end{pmatrix}}_{\mathbf{A}^B}$$

$$\underbrace{\begin{pmatrix} \mathbf{1} & \boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_2\boldsymbol{\lambda}_3 & \cdots & \boldsymbol{\lambda}_2\cdots\boldsymbol{\lambda}_L \\ \boldsymbol{\lambda}_1 & \mathbf{1} & \boldsymbol{\lambda}_3 & \cdots & \boldsymbol{\lambda}_3\cdots\boldsymbol{\lambda}_L \\ \boldsymbol{\lambda}_1\boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_2 & \mathbf{1} & \cdots & \boldsymbol{\lambda}_4\cdots\boldsymbol{\lambda}_L \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\lambda}_{L-1}\cdots\boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_{L-1}\cdots\boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_{L-1}\cdots\boldsymbol{\lambda}_3 & \cdots & \mathbf{1} \end{pmatrix}}_{\mathbf{M}} = \underbrace{\begin{pmatrix} \mathbf{1} & & & & \\ \boldsymbol{\lambda}_1 & \mathbf{1} & & & \\ \boldsymbol{\lambda}_1\boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_2 & \mathbf{1} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \boldsymbol{\lambda}_{L-1}\cdots\boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_{L-1}\cdots\boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_{L-1}\cdots\boldsymbol{\lambda}_3 & \cdots & \mathbf{1} \end{pmatrix}}_{\mathbf{M}^F} + \underbrace{\begin{pmatrix} \mathbf{1} & \boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_2\boldsymbol{\lambda}_3 & \cdots & \boldsymbol{\lambda}_2\cdots\boldsymbol{\lambda}_L \\ & \mathbf{1} & \boldsymbol{\lambda}_3 & \cdots & \boldsymbol{\lambda}_3\cdots\boldsymbol{\lambda}_L \\ & & \mathbf{1} & \cdots & \boldsymbol{\lambda}_4\cdots\boldsymbol{\lambda}_L \\ & & & \ddots & \vdots \\ & & & & \mathbf{1} \end{pmatrix}}_{\mathbf{M}^B} - \mathbf{I}$$

Equivalent Bi-directional RNN for Linear Attention

$$\underbrace{\begin{pmatrix} \mathbf{q}_1^\top\mathbf{k}_1 & \mathbf{q}_1^\top\mathbf{k}_2 & \cdots & \mathbf{q}_1^\top\mathbf{k}_L \\ \mathbf{q}_2^\top\mathbf{k}_1 & \mathbf{q}_2^\top\mathbf{k}_2 & \cdots & \mathbf{q}_2^\top\mathbf{k}_L \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_L^\top\mathbf{k}_1 & \mathbf{q}_L^\top\mathbf{k}_2 & \cdots & \mathbf{q}_L^\top\mathbf{k}_L \end{pmatrix}}_{\mathbf{A}=\mathbf{Q}\mathbf{K}^\top} = \underbrace{\begin{pmatrix} \frac{1}{2}\mathbf{q}_1^\top\mathbf{k}_1 & & & \\ \mathbf{q}_2^\top\mathbf{k}_1 & \frac{1}{2}\mathbf{q}_2^\top\mathbf{k}_2 & & \\ \vdots & \vdots & \ddots & \\ \mathbf{q}_L^\top\mathbf{k}_1 & \mathbf{q}_L^\top\mathbf{k}_2 & \cdots & \frac{1}{2}\mathbf{q}_L^\top\mathbf{k}_L \end{pmatrix}}_{\mathbf{A}^F} + \underbrace{\begin{pmatrix} \frac{1}{2}\mathbf{q}_1^\top\mathbf{k}_1 & \mathbf{q}_1^\top\mathbf{k}_2 & \cdots & \mathbf{q}_1^\top\mathbf{k}_L \\ & \frac{1}{2}\mathbf{q}_2^\top\mathbf{k}_2 & \cdots & \mathbf{q}_2^\top\mathbf{k}_L \\ & & \ddots & \vdots \\ & & & \frac{1}{2}\mathbf{q}_L^\top\mathbf{k}_L \end{pmatrix}}_{\mathbf{A}^B}$$

$$\underbrace{\begin{pmatrix} \mathbf{1} & \boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_2\boldsymbol{\lambda}_3 & \cdots & \boldsymbol{\lambda}_2\cdots\boldsymbol{\lambda}_L \\ \boldsymbol{\lambda}_1 & \mathbf{1} & \boldsymbol{\lambda}_3 & \cdots & \boldsymbol{\lambda}_3\cdots\boldsymbol{\lambda}_L \\ \boldsymbol{\lambda}_1\boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_2 & \mathbf{1} & \cdots & \boldsymbol{\lambda}_4\cdots\boldsymbol{\lambda}_L \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\lambda}_{L-1}\cdots\boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_{L-1}\cdots\boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_{L-1}\cdots\boldsymbol{\lambda}_3 & \cdots & \mathbf{1} \end{pmatrix}}_{\mathbf{M}} = \underbrace{\begin{pmatrix} \mathbf{1} & & & & \\ \boldsymbol{\lambda}_1 & \mathbf{1} & & & \\ \boldsymbol{\lambda}_1\boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_2 & \mathbf{1} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \boldsymbol{\lambda}_{L-1}\cdots\boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_{L-1}\cdots\boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_{L-1}\cdots\boldsymbol{\lambda}_3 & \cdots & \mathbf{1} \end{pmatrix}}_{\mathbf{M}^F} + \underbrace{\begin{pmatrix} \mathbf{1} & \boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_2\boldsymbol{\lambda}_3 & \cdots & \boldsymbol{\lambda}_2\cdots\boldsymbol{\lambda}_L \\ & \mathbf{1} & \boldsymbol{\lambda}_3 & \cdots & \boldsymbol{\lambda}_3\cdots\boldsymbol{\lambda}_L \\ & & \mathbf{1} & \cdots & \boldsymbol{\lambda}_4\cdots\boldsymbol{\lambda}_L \\ & & & \ddots & \vdots \\ & & & & \mathbf{1} \end{pmatrix}}_{\mathbf{M}^B} - \mathbf{I}$$

$$\mathbf{Y} = \big(\textsc{scale}(\mathbf{Q}\mathbf{K}^\top \odot \mathbf{M})\big)\mathbf{V} = (\mathbf{C}^{-1}(\mathbf{Q}\mathbf{K}^\top \odot \mathbf{M}))\mathbf{V}$$

$$\mathbf{C}_i = \underbrace{\mathbf{q}_i^\top \sum_{j=1}^{i} \mathbf{M}_{ij}\mathbf{k}_j - \frac{1}{2}\mathbf{q}_i^\top\mathbf{k}_i}_{\mathbf{C}_i^F} + \underbrace{\mathbf{q}_i^\top \sum_{j=i}^{L} \mathbf{M}_{ij}\mathbf{k}_j - \frac{1}{2}\mathbf{q}_i^\top\mathbf{k}_i}_{\mathbf{C}_i^B}$$

Parallel Form for Diagonal Decay

$$\mathbf{Y} = (\mathbf{C}^F + \mathbf{C}^B)^{-1}\big(\mathbf{A}^F \odot \mathbf{M}^F + \underbrace{\mathbf{A}^F \odot \mathbf{M}^B}_{\mathbf{A}^F \odot \mathbf{I}} + \underbrace{\mathbf{A}^B \odot \mathbf{M}^F}_{\mathbf{A}^B \odot \mathbf{I}} + \mathbf{A}^B \odot \mathbf{M}^B - \mathbf{A}^F \odot \mathbf{I} - \mathbf{A}^B \odot \mathbf{I}\big)\mathbf{V}$$

$$= (\,\mathbf{C}^F + \mathbf{C}^B\,)^{-1}\big(\underbrace{(\mathbf{A}^F \odot \mathbf{M}^F)\mathbf{V}}_{\text{Forward}} + \underbrace{(\mathbf{A}^B \odot \mathbf{M}^B)\mathbf{V}}_{\text{Backward}}\big).$$

✅ The backward (anti-causal) component of full linear attention is equivalent to a recurrent neural network operating in the reverse direction.

$$\underbrace{\begin{pmatrix} \frac{1}{2}\frac{\mathbf{q}_L^\top \mathbf{k}_L}{\mathbf{q}_L^\top \mathbf{z}_L} & & & \\ \frac{\mathbf{q}_{L-1}^\top \mathbf{k}_L}{\mathbf{q}_2^\top \mathbf{z}_L} & \frac{1}{2}\frac{\mathbf{q}_{L-1}^\top \mathbf{k}_{L-1}}{\mathbf{q}_2^\top \mathbf{z}_L} & & \\ \vdots & \vdots & \ddots & \\ \frac{\mathbf{q}_1^\top \mathbf{k}_L}{\mathbf{q}_1^\top \mathbf{z}_L} & \frac{\mathbf{q}_1^\top \mathbf{k}_{L-1}}{\mathbf{q}_1^\top \mathbf{z}_L} & \cdots & \frac{1}{2}\frac{\mathbf{q}_1^\top \mathbf{k}_1}{\mathbf{q}_1^\top \mathbf{z}_L} \end{pmatrix}}_{F(\mathbf{A}^B)} \odot \underbrace{\begin{pmatrix} 1 & & & \\ \boldsymbol{\lambda}_L & 1 & & \\ \boldsymbol{\lambda}_L\boldsymbol{\lambda}_{L-1} & \boldsymbol{\lambda}_{L-1} & 1 & \\ \vdots & \vdots & \vdots & \ddots \\ \boldsymbol{\lambda}_L \cdots \boldsymbol{\lambda}_2 & \boldsymbol{\lambda}_L \cdots \boldsymbol{\lambda}_3 & \boldsymbol{\lambda}_L \cdots \boldsymbol{\lambda}_4 & \cdots \quad 1 \end{pmatrix}}_{F(\mathbf{M}^B)} \begin{pmatrix} \mathbf{v}_L^\top \\ \mathbf{v}_{L-1}^\top \\ \mathbf{v}_{L-2}^\top \\ \vdots \\ \mathbf{v}_1^\top \end{pmatrix}$$

$$\mathbf{Y} = (\mathbf{C}^F + \mathbf{C}^B)^{-1}(\,\mathbf{Y}^F + \mathbf{Y}^B\,), \text{ where}$$

$$\mathbf{Y}^F = (\mathbf{A}^F \odot \mathbf{M}^F)\mathbf{V}, \quad \mathbf{Y}^B = (\mathbf{A}^B \odot \mathbf{M}^B)\mathbf{V} = \textsc{Flip}\Big(\big(F(\mathbf{A}^B) \odot F(\mathbf{M}^B)\big)\textsc{Flip}(\mathbf{V})\Big).$$

LION-RNN

$$\mathbf{S}_i^{F/B} = \lambda_i \mathbf{S}_{i-1}^{F/B} + \mathbf{k}_i \mathbf{v}_i^\top, \tag{19a}$$

$$\mathbf{z}_i^{F/B} = \lambda_i \mathbf{z}_{i-1}^{F/B} + \mathbf{k}_i, \tag{19b}$$

$$c_i^{F/B} = \mathbf{q}_i{}^\top \mathbf{z}_i^{F/B} - \frac{\mathbf{q}_i{}^\top \mathbf{k}_i}{2}, \tag{19c}$$

$$\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top \mathbf{S}_i^{F/B} - \frac{\mathbf{q}_i{}^\top \mathbf{k}_i}{2} \mathbf{v}_i, \tag{20a}$$

$$\text{OUTPUT:} \quad \mathbf{y}_i = \frac{\mathbf{y}_i^F + \mathbf{y}_i^B}{c_i^F + c_i^B} \tag{20b}$$

LION-RNN

$$\mathbf{S}_i^{F/B} = \lambda_i \mathbf{S}_{i-1}^{F/B} + \mathbf{k}_i \mathbf{v}_i^\top, \tag{19a}$$

$$\mathbf{z}_i^{F/B} = \lambda_i \mathbf{z}_{i-1}^{F/B} + \mathbf{k}_i, \tag{19b}$$

$$c_i^{F/B} = \mathbf{q}_i{}^\top \mathbf{z}_i^{F/B} - \frac{\mathbf{q}_i{}^\top \mathbf{k}_i}{2}, \tag{19c}$$

$$\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top \mathbf{S}_i^{F/B} - \frac{\mathbf{q}_i{}^\top \mathbf{k}_i}{2} \mathbf{v}_i, \tag{20a}$$

$$\text{OUTPUT:} \quad \mathbf{y}_i = \frac{\mathbf{y}_i^F + \mathbf{y}_i^B}{c_i^F + c_i^B} \tag{20b}$$

LION-RNN

$$\mathbf{S}_i^{F/B} = \lambda_i \mathbf{S}_{i-1}^{F/B} + \mathbf{k}_i \mathbf{v}_i^\top, \tag{19a}$$

$$\mathbf{z}_i^{F/B} = \lambda_i \mathbf{z}_{i-1}^{F/B} + \mathbf{k}_i, \tag{19b}$$

$$c_i^{F/B} = \mathbf{q}_i{}^\top \mathbf{z}_i^{F/B} - \frac{\mathbf{q}_i{}^\top \mathbf{k}_i}{2}, \tag{19c}$$

$$\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top \mathbf{S}_i^{F/B} - \frac{\mathbf{q}_i{}^\top \mathbf{k}_i}{2} \mathbf{v}_i, \tag{20a}$$

$$\text{OUTPUT:} \quad \mathbf{y}_i = \frac{\mathbf{y}_i^F + \mathbf{y}_i^B}{c_i^F + c_i^B} \tag{20b}$$

- LION-LIT for $\lambda_i = 1$ which is bi-directional form of Vanilla Linear Transformer [25].
- LION-D for $\lambda_i = \lambda$ fixed decay, and bi-directional form of RetNet (with scaling) [44].
- LION-S , where $\lambda_i = \sigma(\mathbf{W}\mathbf{x}_i + b)$ is the bidirectional extension of GRFA [34] (with shifted SiLU activation) and also inspired by the selectivity of Mamba2.

LION-RNN

$$\mathbf{S}_i^{F/B} = \lambda_i \mathbf{S}_{i-1}^{F/B} + \mathbf{k}_i \mathbf{v}_i^\top, \quad (19a)$$

$$\mathbf{z}_i^{F/B} = \lambda_i \mathbf{z}_{i-1}^{F/B} + \mathbf{k}_i, \quad (19b)$$

$$c_i^{F/B} = \mathbf{q}_i{}^\top \mathbf{z}_i^{F/B} - \frac{\mathbf{q}_i{}^\top \mathbf{k}_i}{2}, \quad (19c)$$

$$\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top \mathbf{S}_i^{F/B} - \frac{\mathbf{q}_i{}^\top \mathbf{k}_i}{2} \mathbf{v}_i, \quad (20a)$$

$$\text{OUTPUT:} \quad \mathbf{y}_i = \frac{\mathbf{y}_i^F + \mathbf{y}_i^B}{c_i^F + c_i^B} \quad (20b)$$

- LION-LIT for $\lambda_i = 1$ which is bi-directional form of Vanilla Linear Transformer [25].
- LION-D for $\lambda_i = \lambda$ fixed decay, and bi-directional form of RetNet (with scaling) [44].
- LION-S, where $\lambda_i = \sigma(\mathbf{W}\mathbf{x}_i + b)$ is the bidirectional extension of GRFA [34] (with shifted SiLU activation) and also inspired by the selectivity of Mamba2.

For diagonal forget gate: $\mathbf{S}_i = \mathbf{\Lambda}_i \mathbf{S}_{i-1} + \mathbf{k}_i \mathbf{v}_i^\top$

$$\mathbf{Y} = \mathbf{M}\mathbf{V}$$

$$\mathbf{M}_{ij} = \mathbf{Q}_i^\top \mathbf{\Lambda}^{\times}_{0:i} (\mathbf{\Lambda}^{\times}_{0:j})^{-1} \mathbf{K}_j =$$

$$\mathbf{Q}_i^\top (\boldsymbol{\lambda}_0 \odot \boldsymbol{\lambda}_1 \odot \boldsymbol{\lambda}_2 \ldots \boldsymbol{\lambda}_i)(\boldsymbol{\lambda}_0 \odot \boldsymbol{\lambda}_1 \odot \boldsymbol{\lambda}_2 \ldots \boldsymbol{\lambda}_j)^{-1} \mathbf{K}_j =$$

$$(\mathbf{Q}_i \odot \boldsymbol{\lambda}_0 \odot \boldsymbol{\lambda}_1 \odot \boldsymbol{\lambda}_2 \ldots \boldsymbol{\lambda}_i)^\top (\boldsymbol{\lambda}_0 \odot \boldsymbol{\lambda}_1 \odot \boldsymbol{\lambda}_2 \ldots \boldsymbol{\lambda}_j)^{-1} \odot \mathbf{K}_j$$

$$\mathbf{L}_i = \boldsymbol{\lambda}_0 \odot \boldsymbol{\lambda}_1 \odot \cdots \odot \boldsymbol{\lambda}_i$$

$$\mathbf{U}_i = (\boldsymbol{\lambda}_0 \odot \boldsymbol{\lambda}_1 \odot \cdots \odot \boldsymbol{\lambda}_i)^{-1}$$

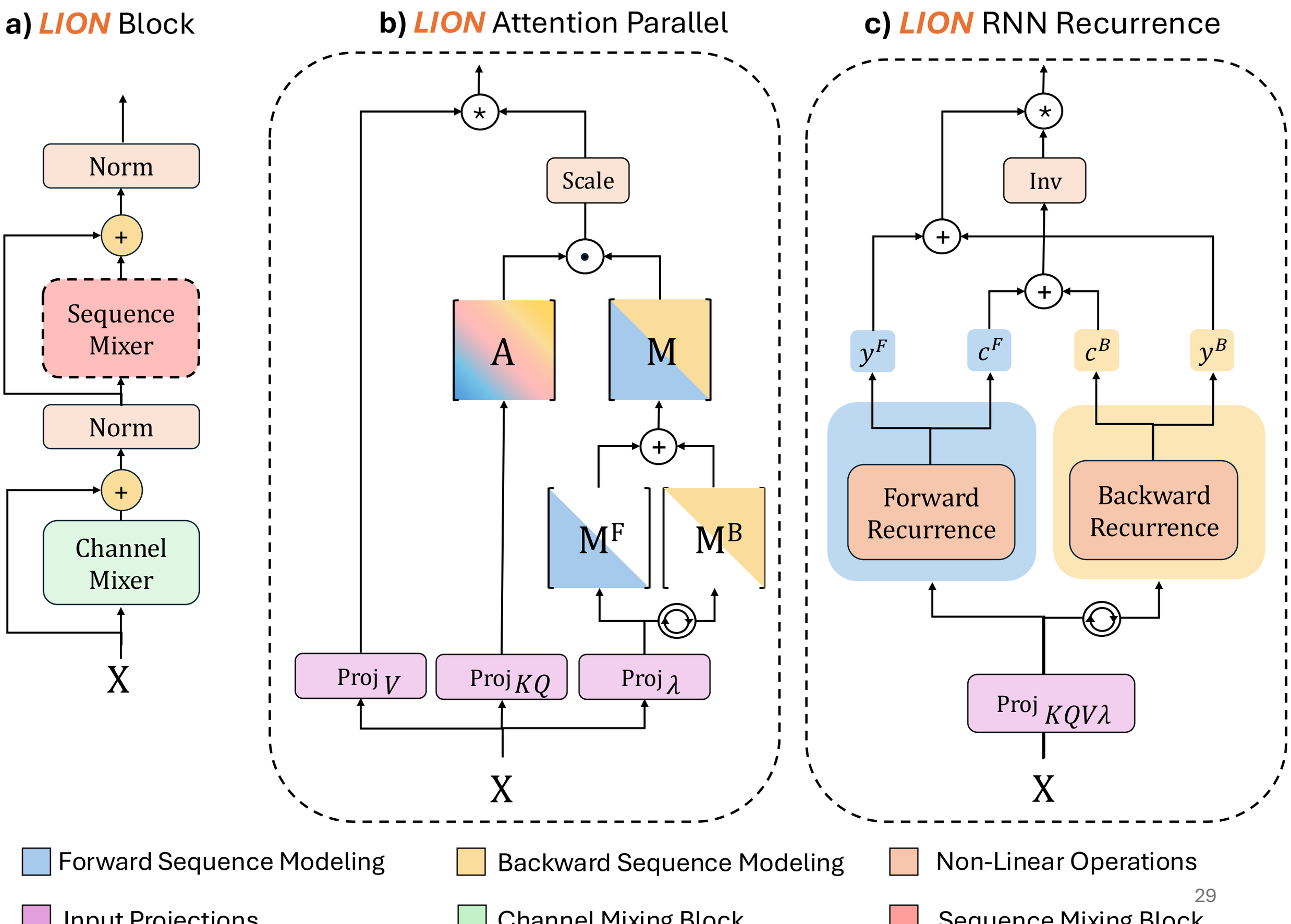
$$\mathbf{Y} = \text{TRIL}\left[\underbrace{(\mathbf{Q} \odot \mathbf{L})}_{\mathbf{Q}^*} \underbrace{(\mathbf{K} \odot \text{Inv}(\mathbf{L}))^\top}_{\mathbf{K}^{*\top}}\right] \mathbf{V}$$

$$\mathbf{L} = \texttt{cumprod}(\mathbf{D}), \quad \text{with } \mathbf{D}_i = \text{DIAG}(\mathbf{\Lambda}_i)$$

$$\mathbf{Y} = (\text{TRIL}\left[(\mathbf{Q} \odot \mathbf{L}^F)\ (\mathbf{K} \odot \text{Inv}(\mathbf{L}^F))^\top\right] + \text{TRIU}\left[(\mathbf{Q} \odot \mathbf{L}^B)\ (\mathbf{K} \odot \text{Inv}(\mathbf{L}^B))^\top\right])\mathbf{V} \quad (81)$$

$$\mathbf{L}^F = \texttt{cumprod}(\mathbf{D}), \quad \textit{with } \mathbf{D}_i = \text{DIAG}(\mathbf{\Lambda}_i) \quad (82)$$

$$\mathbf{L}^B = \texttt{cumprod}(\textit{Flip}(\mathbf{D})), \quad \textit{with } \mathbf{D}_i = \text{DIAG}(\mathbf{\Lambda}_i) \quad (83)$$


a) LION Block
Norm
+
Sequence Mixer
Norm
+
Channel Mixer
X
b) LION Attention Parallel
*
Scale
⊙
A
M
+
M^F
M^B
Proj V
Proj KQ
Proj λ
X
c) LION RNN Recurrence
*
Inv
+
+
y^F
c^F
c^B
y^B
Forward Recurrence
Backward Recurrence
Proj KQVλ
X
Forward Sequence Modeling
Backward Sequence Modeling
Non-Linear Operations
Input Projections
Channel Mixing Block
Sequence Mixing Block

Bi-directional Linear Transformers

Model	Causal Recurrence	LION Bi-directional RNN	LION Full Linear Attention
LinAtt [25]	$\mathbf{S}_i = \mathbf{S}_{i-1} + \mathbf{k}_i \mathbf{v}_i^\top$ $\mathbf{y}_i = \mathbf{q}_i^\top \mathbf{S}_i$	$\mathbf{S}_i^{F/B} = \mathbf{S}_{i-1}^{F/B} + \mathbf{k}_i \mathbf{v}_i^\top$ $\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{S}_i^{F/B} - \frac{1}{2}\mathbf{k}_i\mathbf{v}_i), \quad \mathbf{y}_i = \mathbf{y}_i^F + \mathbf{y}_i^B$	$\mathbf{Y} = \mathbf{Q}\mathbf{K}^\top\mathbf{V}$
+ Scaling	$\mathbf{z}_i = \mathbf{z}_{i-1} + \mathbf{k}_i$ $\mathbf{y}_i = \mathbf{q}_i^\top \mathbf{S}_i / \mathbf{q}_i^\top \mathbf{z}_i$	$\mathbf{z}_i^{F/B} = \mathbf{z}_{i-1}^{F/B} + \mathbf{k}_i, \quad c_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{z}_i^{F/B} - \frac{1}{2}\mathbf{k}_i)$ $\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{S}_i^{F/B} - \frac{1}{2}\mathbf{k}_i\mathbf{v}_i), \quad \mathbf{y}_i = \frac{\mathbf{y}_i^F + \mathbf{y}_i^B}{c_i^F + c_i^B}$ LION-LIT	$\mathbf{Y} = \text{SCALE}(\mathbf{Q}\mathbf{K}^\top)\mathbf{V}$
RetNet [44]	$\mathbf{S}_i = \lambda\mathbf{S}_{i-1} + \mathbf{k}_i \mathbf{v}_i^\top$ $\mathbf{y}_i = \mathbf{q}_i^\top \mathbf{S}_i$	$\mathbf{S}_i^{F/B} = \lambda\mathbf{S}_{i-1}^{F/B} + \mathbf{k}_i\mathbf{v}_i^\top$ $\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{S}_i^{F/B} - \frac{1}{2}\mathbf{k}_i\mathbf{v}_i), \quad \mathbf{y}_i = \mathbf{y}_i^F + \mathbf{y}_i^B$	$\mathbf{Y} = (\mathbf{Q}\mathbf{K}^\top \odot \mathbf{M})\mathbf{V}$, $\mathbf{M}_{ij} = \lambda^{\lvert i-j\rvert}$
+ Scaling	$\mathbf{z}_i = \lambda\mathbf{z}_{i-1} + \mathbf{k}_i$ $\mathbf{y}_i = \mathbf{q}_i^\top \mathbf{S}_i / \mathbf{q}_i^\top \mathbf{z}_i$	$\mathbf{z}_i^{F/B} = \lambda\mathbf{z}_{i-1}^{F/B} + \mathbf{k}_i, \quad c_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{z}_i^{F/B} - \frac{1}{2}\mathbf{k}_i)$ $\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{S}_i^{F/B} - \frac{1}{2}\mathbf{k}_i\mathbf{v}_i), \quad \mathbf{y}_i = \frac{\mathbf{y}_i^F + \mathbf{y}_i^B}{c_i^F + c_i^B}$ LION-D	$\mathbf{Y} = \text{SCALE}(\mathbf{Q}\mathbf{K}^\top \odot \mathbf{M})\mathbf{V}$, $\mathbf{M}_{ij} = \lambda^{\lvert i-j\rvert}$
Gated RFA [34]	$\mathbf{S}_i = \sigma(\mathbf{W}\mathbf{x}_i)\mathbf{S}_{i-1} + \mathbf{k}_i \mathbf{v}_i^\top$ $\mathbf{z}_i = \sigma(\mathbf{W}\mathbf{x}_i)\mathbf{z}_{i-1} + \mathbf{k}_i$ $\mathbf{y}_i = \mathbf{q}_i^\top \mathbf{S}_i / \mathbf{q}_i^\top \mathbf{z}_i$	$\mathbf{S}_i^{F/B} = \sigma(\mathbf{W}\mathbf{x}_i)\mathbf{S}_{i-1}^{F/B} + \mathbf{k}_i \mathbf{v}_i^\top$ $\mathbf{y}_i = \mathbf{q}_i^\top \mathbf{S}_i$ $\mathbf{z}_i^{F/B} = \sigma(\mathbf{W}\mathbf{x}_i)\mathbf{z}_{i-1}^{F/B} + \mathbf{k}_i, \quad c_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{z}_i^{F/B} - \frac{1}{2}\mathbf{k}_i)$ $\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{S}_i^{F/B} - \frac{1}{2}\mathbf{k}_i\mathbf{v}_i), \quad \mathbf{y}_i = \frac{\mathbf{y}_i^F + \mathbf{y}_i^B}{c_i^F + c_i^B}$ LION-S	$\mathbf{Y} = \text{SCALE}(\mathbf{Q}\mathbf{K}^\top \odot \mathbf{M}^*)\mathbf{V}$, $\mathbf{M^*}_{ij} = \begin{cases} \Pi_i^{j+1}\lambda_k & \text{if } i > j, \\ \Pi_{i+1}^{j}\lambda_k & \text{if } i < j, \\ 1 & \text{if } i = j, \end{cases}$
Mamba-2 [6]	$\mathbf{S}_i = \lambda_i\mathbf{S}_{i-1} + \mathbf{k}_i \mathbf{v}_i^\top$ $\mathbf{y}_i = \mathbf{q}_i^\top \mathbf{S}_i$	$\mathbf{S}_i^{F/B} = \lambda_i\mathbf{S}_{i-1}^{F/B} + \mathbf{k}_i\mathbf{v}_i^\top$ $\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{S}_i^{F/B} - \frac{1}{2}\mathbf{k}_i\mathbf{v}_i), \quad \mathbf{y}_i = \mathbf{y}_i^F + \mathbf{y}_i^B$	$\mathbf{Y} = (\mathbf{Q}\mathbf{K}^\top \odot \mathbf{M})\mathbf{V}$ $\mathbf{M} = \mathbf{M}^*$
RWKV-6[33] Mamba [13] GLA [51]	$\mathbf{S}_i = \text{Diag}(\boldsymbol{\lambda}_i)\mathbf{S}_{i-1} + \mathbf{k}_i \mathbf{v}_i^\top$ $\mathbf{y}_i = \mathbf{q}_i^\top \mathbf{S}_i$	$\mathbf{S}_i^{F/B} = \text{Diag}(\boldsymbol{\lambda}_i)\mathbf{S}_{i-1}^{F/B} + \mathbf{k}_i\mathbf{v}_i^\top$ $\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{S}_i^{F/B} - \frac{1}{2}\mathbf{k}_i\mathbf{v}_i), \quad \mathbf{y}_i = \mathbf{y}_i^F + \mathbf{y}_i^B$	$\mathbf{Y^F} = \text{Tril}(\mathbf{Q} \odot \mathbf{L^F})(\mathbf{K} \odot (\mathbf{L^F})^{-1})\mathbf{V}$ $\mathbf{Y^B} = \text{Triu}(\mathbf{Q} \odot \mathbf{L^B})(\mathbf{K} \odot (\mathbf{L^B})^{-1})\mathbf{V}$ $\mathbf{Y} = \mathbf{Y^F} + \mathbf{Y^B}$
HGRN-2 [36]	$\mathbf{S}_i = \text{Diag}(\boldsymbol{\lambda}_i)\mathbf{S}_{i-1} + (1 - \boldsymbol{\lambda}_i)\mathbf{v}_i^\top$ $\mathbf{y}_i = \mathbf{q}_i^\top \mathbf{S}_i$	$\mathbf{S}_i^{F/B} = \text{Diag}(\boldsymbol{\lambda}_i)\mathbf{S}_{i-1}^{F/B} + (1 - \boldsymbol{\lambda}_i)\mathbf{v}_i^\top, \quad \mathbf{k}_i = (1 - \lambda_i)$ $\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{S}_i^{F/B} - \frac{1}{2}\mathbf{k}_i\mathbf{v}_i), \quad \mathbf{y}_i = \mathbf{y}_i^F + \mathbf{y}_i^B$	$\mathbf{Y^F} = \text{Tril}(\mathbf{Q} \odot \mathbf{L^F})(\mathbf{K} \odot (\mathbf{L^F})^{-1})\mathbf{V}$ $\mathbf{Y^B} = \text{Triu}(\mathbf{Q} \odot \mathbf{L^B})(\mathbf{K} \odot (\mathbf{L^B})^{-1})\mathbf{V}$ $\mathbf{Y} = \mathbf{Y^F} + \mathbf{Y^B}$
xLSTM [3]	$\mathbf{S}_i = f_i\mathbf{S}_{i-1} + i_i \mathbf{k}_i \mathbf{v}_i^\top$ $\mathbf{z}_i = f_i\mathbf{z}_{i-1} + i_i \mathbf{k}_i$ $\mathbf{y}_i = \mathbf{q}_i^\top \mathbf{S}_i / \max(\lvert \mathbf{q}_i^\top \mathbf{z}_i \rvert, 1)$	$\mathbf{S}_i^{F/B} = f_i\mathbf{S}_{i-1}^{F/B} + i_i \mathbf{k}_i \mathbf{v}_i^\top,$ $\mathbf{y}_i = \mathbf{q}_i^\top \mathbf{S}_i$ $\mathbf{z}_i^{F/B} = f_i\mathbf{z}_{i-1}^{F/B} + i_i\mathbf{k}_i, \quad c_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{z}_i^{F/B} - \frac{1}{2}\mathbf{k}_i)$ $\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{S}_i^{F/B} - \frac{1}{2}\mathbf{k}_i\mathbf{v}_i), \quad \mathbf{y}_i = \frac{\mathbf{y}_i^F + \mathbf{y}_i^B}{\max(c_i^F + c_i^B, 1)}$	$\mathbf{Y} = \text{SCALE}_{max}(\mathbf{Q}\mathbf{U}^\top) \odot \mathbf{M})\mathbf{V}$, $\mathbf{M}_{ij} = \begin{cases} \Pi_i^{j+1} f_k & \text{if } i > j, \\ \Pi_{i+1}^{j} f_k & \text{if } i < j, \\ 1 & \text{if } i = j, \end{cases}$
DeltaNet [52]	$\mathbf{S}_i = (1 - \beta_i \mathbf{k}_i \mathbf{k}_i^\top)\mathbf{S}_{i-1} + \beta_i \mathbf{k}_i \mathbf{v}_i^\top$ $\mathbf{y}_i = \mathbf{q}_i^\top \mathbf{S}_i$	$\mathbf{S}_i^{F/B} = (1 - \beta_i \mathbf{k}_i \mathbf{k}_i^\top)\mathbf{S}_{i-1}^{F/B} + \beta_i \mathbf{k}_i \mathbf{v}_i^\top$ $\mathbf{y}_i^{F/B} = \mathbf{q}_i{}^\top(\mathbf{S}_i^{F/B} - \frac{1}{2}\mathbf{k}_i\mathbf{v}_i), \quad \mathbf{y}_i = \mathbf{y}_i^F + \mathbf{y}_i^B$	$\mathbf{Y} = (\mathbf{Q}\mathbf{K}^\top)\mathbf{T}\mathbf{V}$, $\mathbf{T} = \frac{1}{2}(\mathbf{T}^F + \mathbf{T}^B)$, $\mathbf{T}^F = (\mathbf{I} + \text{tril}(\text{diag}(\beta_i)\mathbf{K}_i\mathbf{K}_i^\top, -1))^{-1}\text{diag}(\beta_i)$, $\mathbf{T}^B = (\mathbf{I} + \text{triu}(\text{diag}(\beta_i)\mathbf{K}_i\mathbf{K}_i^\top, -1))^{-1}\text{diag}(\beta_i)$

Key Representation of Bidirectional Linear Transformers

- Full Linear Attention
- Equal Bi-directional RNN
- Chunkwise Parallel Representation

Chunkwise Parallel Form of LION

$Q[1]K[3]^T$

$Q[3]K[1]^T$

$q_1^T k_1$	$q_1^T k_2$	$q_1^T k_3$	$q_1^T k_4$	$q_1^T k_5$	$q_1^T k_6$	$q_1^T k_7$	$q_1^T k_8$	$q_1^T k_9$
$q_2^T k_1$	$q_2^T k_2$	$q_2^T k_3$	$q_2^T k_4$	$q_2^T k_5$	$q_2^T k_6$	$q_2^T k_7$	$q_2^T k_8$	$q_2^T k_9$
$q_3^T k_1$	$q_3^T k_2$	$q_3^T k_3$	$q_3^T k_4$	$q_3^T k_5$	$q_3^T k_6$	$q_3^T k_7$	$q_3^T k_8$	$q_3^T k_9$
$q_4^T k_1$	$q_4^T k_2$	$q_4^T k_3$	$q_4^T k_4$	$q_4^T k_5$	$q_4^T k_6$	$q_4^T k_7$	$q_4^T k_8$	$^T_4 k_9$
$q_5^T k_1$	$q_5^T k_2$	$q_5^T k_3$	$q_5^T k_4$	$q_5^T k_5$	$q_5^T k_6$	$q_5^T k_7$	$q_5^T k_8$	$q_5^T k_9$
$q_6^T k_1$	$q_6^T k_2$	$q_6^T k_3$	$q_6^T k_4$	$q_6^T k_5$	$q_6^T k_6$	$q_6^T k_7$	$q_6^T k_8$	$q_6^T k_9$
$q_7^T k_1$	$q_7^T k_2$	$q_7^T k_3$	$q_7^T k_4$	$q_7^T k_5$	$q_7^T k_6$	$q_7^T k_7$	$q_7^T k_8$	$q_7^T k_9$
$q_8^T k_1$	$q_8^T k_2$	$q_8^T k_3$	$q_8^T k_4$	$q_8^T k_5$	$q_8^T k_6$	$q_8^T k_7$	$q_8^T k_8$	$q_8^T k_9$
$q_9^T k_1$	$q_9^T k_2$	$q_9^T k_3$	$q_9^T k_4$	$q_9^T k_5$	$q_9^T k_6$	$q_9^T k_7$	$q_9^T k_8$	$q_9^T k_9$

$$A = QK^T$$

Chunkwise Parallel Form of LION

1	λ	λ^2	λ^3	λ^4	λ^5	λ^6	λ^7	λ^8
λ	1	λ	λ^2	λ^3	λ^4	λ^5	λ^6	λ^7
λ^2	λ	1	λ	λ^2	λ^3	λ^4	λ^5	λ^6
λ^3	λ^2	λ	1	λ	λ^2	λ^3	λ^4	λ^5
λ^4	λ^3	λ^2	λ	1	λ	λ^2	λ^3	λ^4
λ^5	λ^4	λ^3	λ^2	λ	1	λ	λ^2	λ^3
λ^6	λ^5	λ^4	λ^3	λ^2	λ	1	λ	λ^2
λ^7	λ^6	λ^5	λ^4	λ^3	λ^2	λ	1	λ
λ^8	λ^7	λ^6	λ^5	λ^4	λ^3	λ^2	λ	1

$L\left(\frac{1}{L}\right)^T \lambda^{|i-j|}$

$L\left(\frac{1}{L}\right)^T \lambda^{|i-j|}$

$$M = \lambda^{|i-j|}, \qquad L_i = \lambda^i$$

Chunkwise Parallel Form of LION

$$L_F[1]\left(\frac{1}{L_F[3]}\right)^T$$

1	λ_2	$\lambda_2\lambda_3$	$\lambda_2\lambda_3\lambda_4$	$\lambda_2 \dots \lambda_5$	$\lambda_2 \dots \lambda_6$	$\lambda_2 \dots \lambda_7$	$\lambda_2 \dots \lambda_8$	$\lambda_2 \dots \lambda_9$
λ_1	1	λ_3	$\lambda_3\lambda_4$	$\lambda_3\lambda_4\lambda_5$	$\lambda_3 \dots \lambda_6$	$\lambda_3 \dots \lambda_7$	$\lambda_3 \dots \lambda_8$	$\lambda_3 \dots \lambda_9$
$\lambda_1\lambda_2$	λ_2	1	λ_4	$\lambda_4\lambda_5$	$\lambda_4\lambda_5\lambda_6$	$\lambda_4 \dots \lambda_7$	$\lambda_4 \dots \lambda_8$	$\lambda_4 \dots \lambda_9$
$\lambda_1\lambda_2\lambda_3$	$\lambda_2\lambda_3$	λ_3	1	λ_5	$\lambda_5\lambda_6$	$\lambda_5\lambda_6\lambda_7$	$\lambda_5 \dots \lambda_8$	$\lambda_5 \dots \lambda_9$
$\lambda_1 \dots \lambda_4$	$\lambda_2\lambda_3\lambda_4$	$\lambda_3\lambda_4$	λ_4	1	λ_6	$\lambda_6\lambda_7$	$\lambda_6\lambda_7\lambda_8$	$\lambda_6 \dots \lambda_9$
$\lambda_1 \dots \lambda_5$	$\lambda_2 \dots \lambda_5$	$\lambda_3\lambda_4\lambda_5$	$\lambda_4\lambda_5$	λ_5	1	λ_7	$\lambda_7\lambda_8$	$\lambda_7\lambda_8\lambda_9$
$\lambda_1 \dots \lambda_6$	$\lambda_2 \dots \lambda_6$	$\lambda_3 \dots \lambda_6$	$\lambda_4\lambda_5\lambda_6$	$\lambda_5\lambda_6$	λ_6	1	λ_8	$\lambda_8\lambda_9$
$\lambda_1 \dots \lambda_7$	$\lambda_2 \dots \lambda_7$	$\lambda_3 \dots \lambda_7$	$\lambda_4 \dots \lambda_7$	$\lambda_5\lambda_6\lambda_7$	$\lambda_6\lambda_7$	λ_7	1	λ_9
$\lambda_1 \dots \lambda_8$	$\lambda_2 \dots \lambda_8$	$\lambda_3 \dots \lambda_8$	$\lambda_4 \dots \lambda_8$	$\lambda_5 \dots \lambda_8$	$\lambda_6\lambda_7\lambda_8$	$\lambda_7\lambda_8$	λ_8	1

$$L_B[3]\left(\frac{1}{L_B[1]}\right)^T$$

$$\mathbf{M}_{[ij]} = \begin{cases} \mathbf{L}^F_{[i]} \frac{1}{\mathbf{L}^F_{[j]}}^\top & \text{if } i > j, \\ \mathbf{L}^B_{[j]} \frac{1}{\mathbf{L}^B_{[i]}}^\top & \text{if } i < j, \\ \text{Tril}\left(\mathbf{L}^F_{[i]} \frac{1}{\mathbf{L}^F_{[i]}}^\top\right) + \text{Triu}\left(\mathbf{L}^B_{[i]} \frac{1}{\mathbf{L}^B_{[i]}}^\top\right) - \mathbf{I} & \text{if } i = j, \end{cases}$$

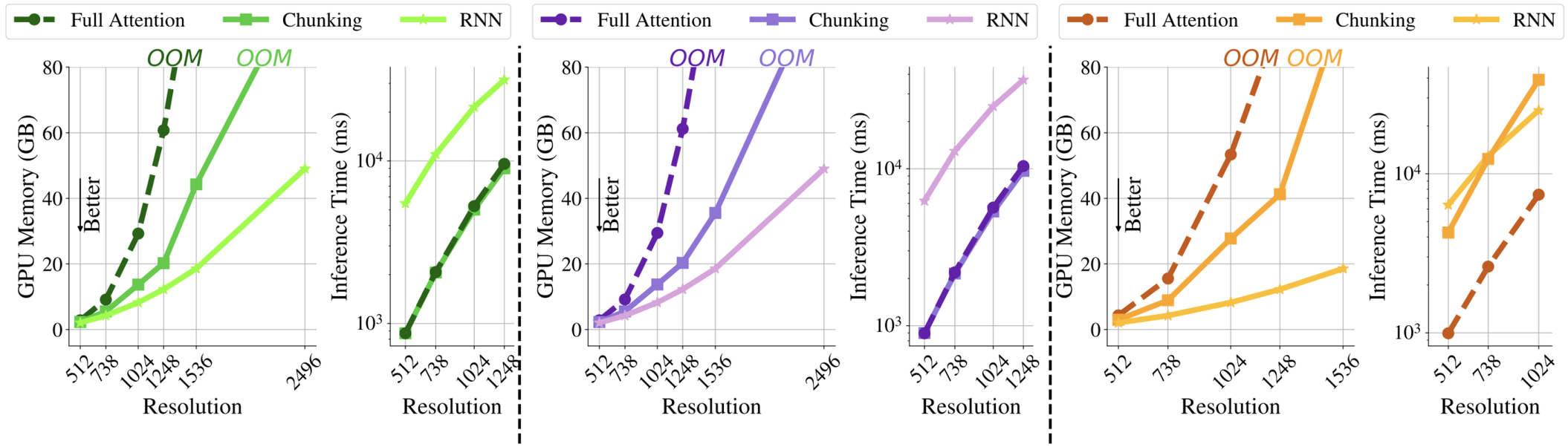
$$\mathbf{L}^F_{[i]} = \texttt{cumprod}(\lambda)_{iC+1:(i+1)C},$$
$$\mathbf{L}^B_{[i]} = \texttt{cumprod}(\text{Flip}(\lambda))_{iC+1:(i+1)C}$$

LION-Chunk

$$\mathbf{A}_{[ij]} = \mathbf{Q}_{[i]}\mathbf{K}_{[j]}^{\top} \odot \mathbf{M}_{[ij]}, \quad \mathbf{C}_{[ij]} = \mathbf{C}_{[ij-1]} + \mathit{Sum}(\mathbf{A}_{[ij]}), \quad \mathbf{S}_{[ij]} = \mathbf{S}_{[ij-1]} + \mathbf{A}_{[ij]}\mathbf{V}_{[j]}, \quad \mathbf{Y}_{[i]} = \frac{\mathbf{S}_{[iN]}}{\mathbf{C}_{[iN]}} \tag{21}$$

LION-Chunk

$$\mathbf{A}_{[ij]} = \mathbf{Q}_{[i]}\mathbf{K}_{[j]}^{\top} \odot \mathbf{M}_{[ij]}, \quad \mathbf{C}_{[ij]} = \mathbf{C}_{[ij-1]} + \mathit{Sum}(\mathbf{A}_{[ij]}), \quad \mathbf{S}_{[ij]} = \mathbf{S}_{[ij-1]} + \mathbf{A}_{[ij]}\mathbf{V}_{[j]}, \quad \mathbf{Y}_{[i]} = \frac{\mathbf{S}_{[iN]}}{\mathbf{C}_{[iN]}} \quad (21)$$

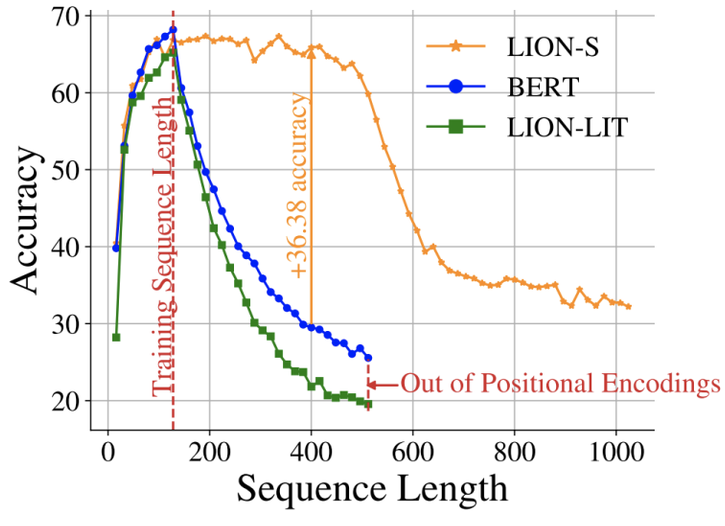


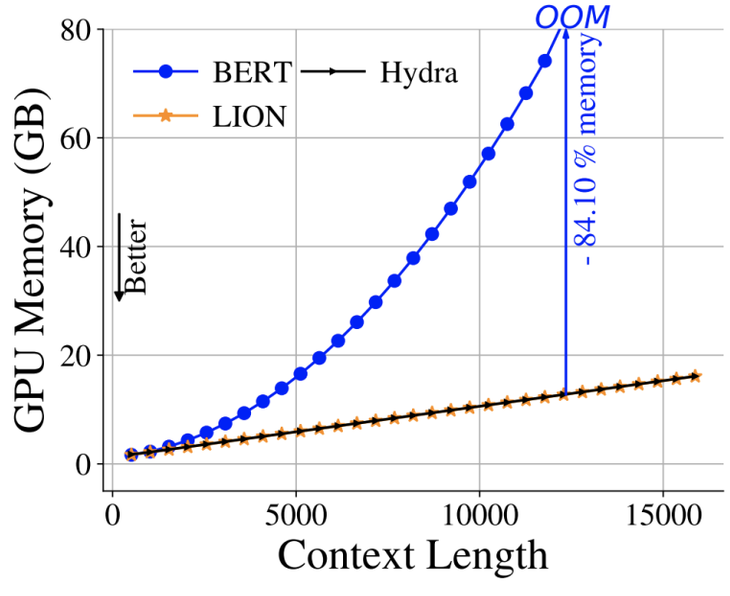
LION-Lit LION-D LION-S

Experiments

- Masked Language Modeling

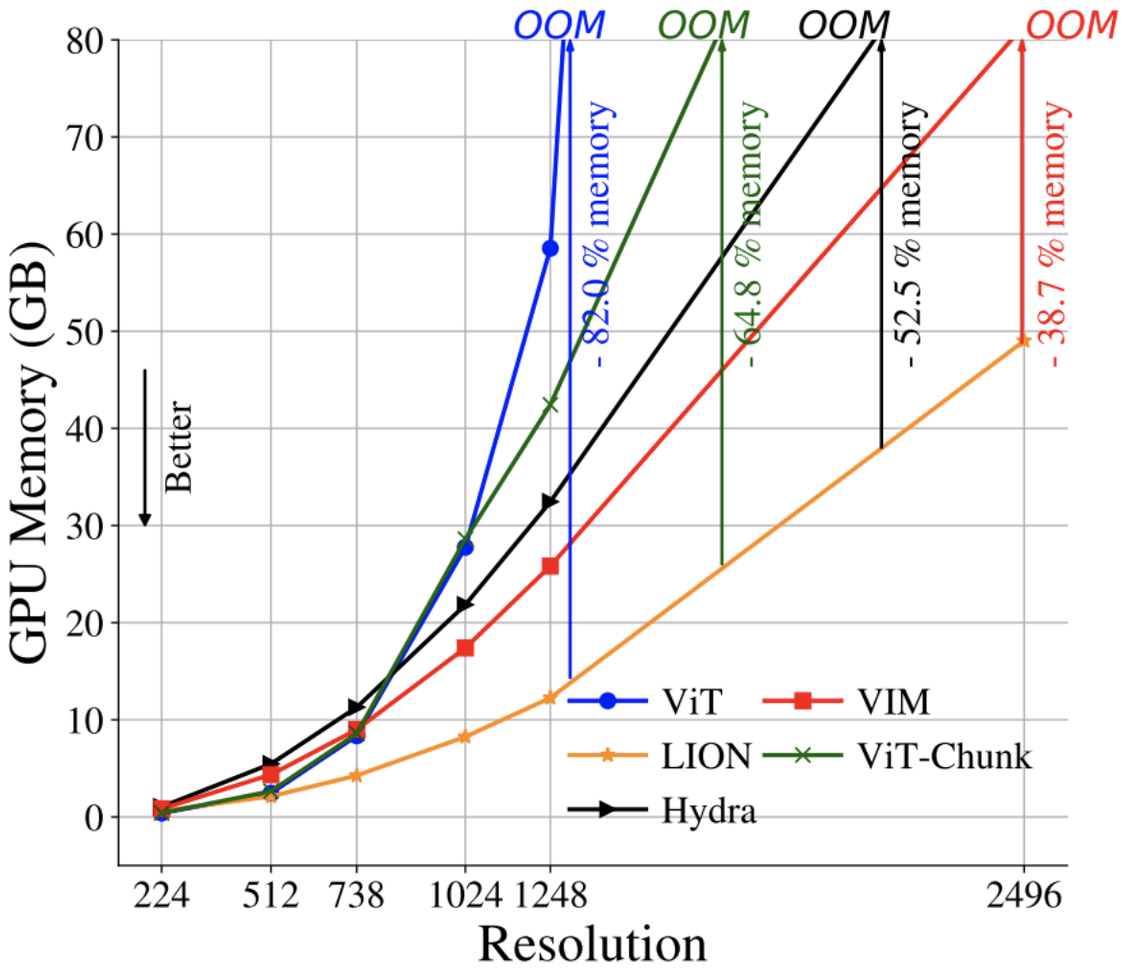
Model	MLM Acc.	GLUE	Train. time
BERT	<u>69.88</u>	**82.95**	×1
Hydra	**71.18**	<u>81.77</u>	×3.13
LION-LIT	67.11	80.76	×**0.95**
LION-D	68.64	81.34	×<u>1.10</u>
LION-S	69.16	81.58	×1.32

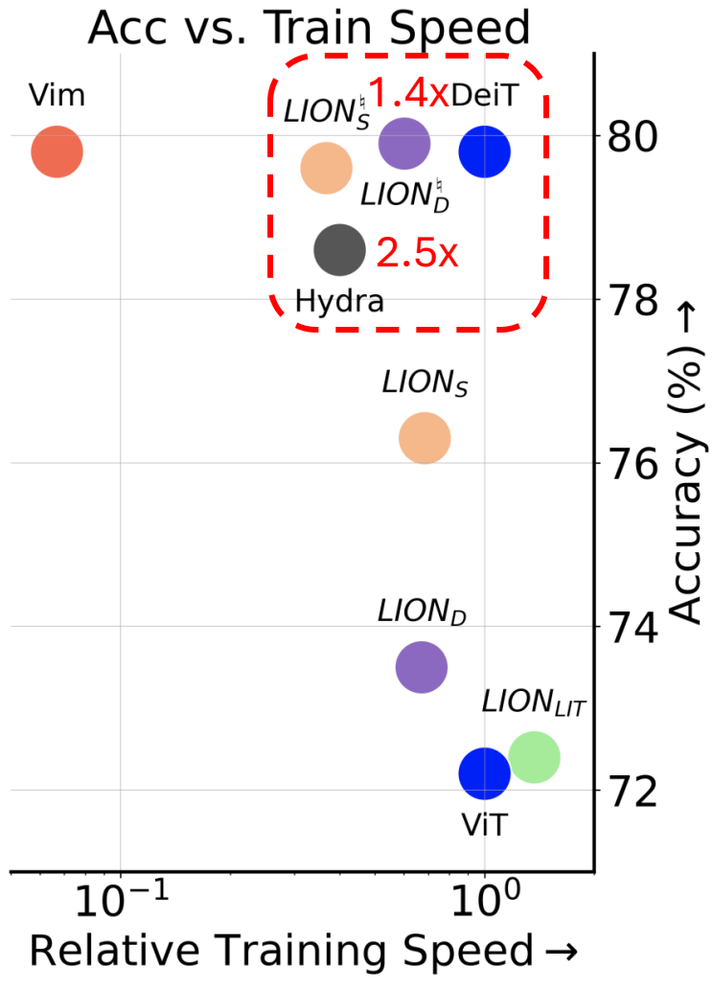




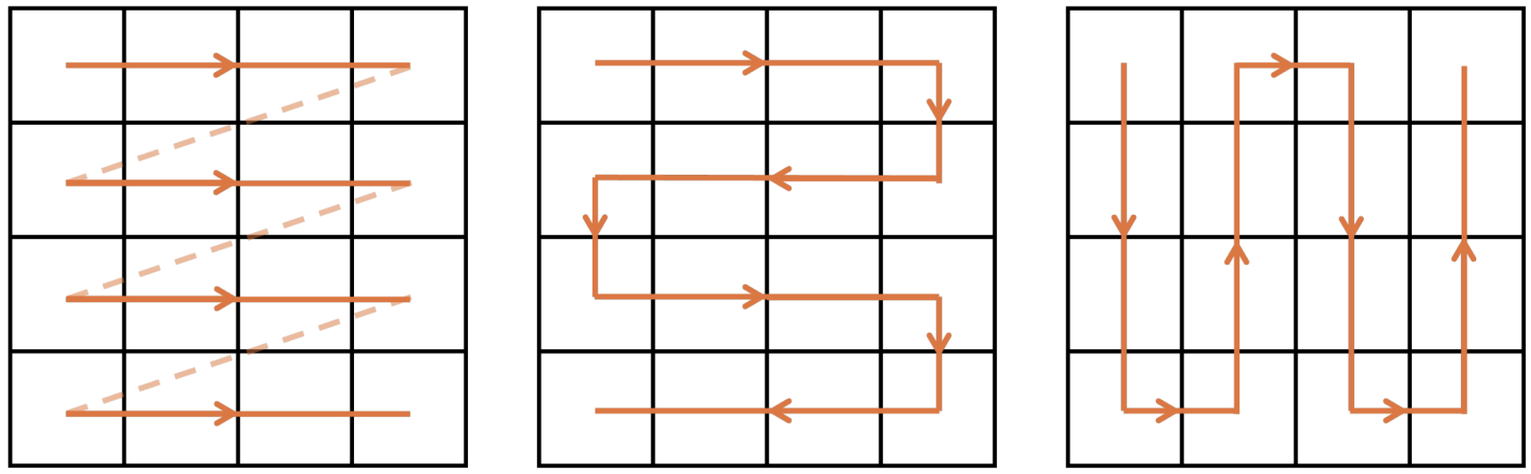
Experiments

- Image Classification





Linear Vision Transformers Require Multiple Scans



LION-LIT	22M	86M	72.4	74.7	×**0.74**	×**0.73**
LION-D	22M	86M	73.5	77.8	×1.49	×1.39
LION-D♮	22M	86M	79.9	80.2	×1.66	×1.48
LION-S	22M	86M	74.0	76.3	×2.03	×1.46
LION-S♮	22M	86M	79.6	79.9	×2.72	×1.68

Future Directions

- Using LION Masks as 2D Positional Embeddings in Softmax Transformers
- Leveraging LION to Accelerate Hydra's Training
- Extending LION to Other Domains: DNA and Diffusion Language Models

Authors and Links

Arshia Afzal

Elias Abad Rocamora

Leyla Naz Candogan

Pol Puigdemont Plana

Yungtao Wu

Francesco Tonin

Mahsa Shoaran

Volkan Cevher

Code Blog Paper