pLSTM: parallelizable Linear Source Transition Mark networks

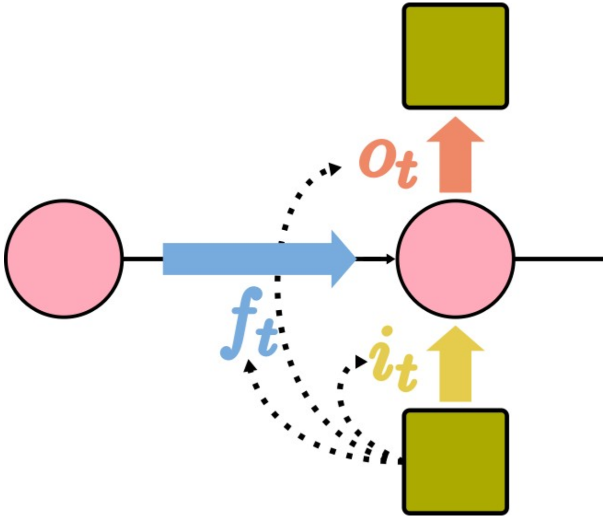
Korbinian Pöppel, ASAP Seminar
13.08.2025

JOHANNES KEPLER UNIVERSITY LINZ

co-authors: Richard Freinschlag, Thomas Schmied, Wei Lin, Sepp Hochreiter

Introduction: Linear RNNs

qLSTM, DeltaNet, LRU, RWKV-4/5/6, GLA, Mamba, Griffin, mLSTM, Gated DeltaNet, TTT, DeltaProduct..



No previous (hidden) state dependence

→ Parallelization by unrolling the sum

Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model

Lianghui Zhu [1*] **Bencheng Liao** [2 1*] **Qian Zhang** [3] **Xinlong Wang** [4] **Wenyu Liu** [1] **Xinggang Wang** [1]

VISION-RWKV: EFFICIENT AND SCALABLE VISUAL PERCEPTION WITH RWKV-LIKE ARCHITECTURES

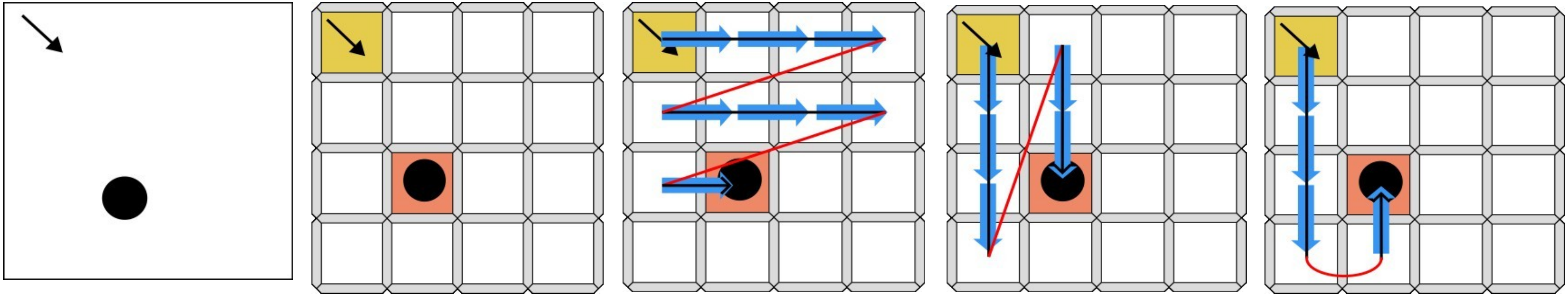
Yuchen Duan[1,2*], **Weiyun Wang**[3,2*], **Zhe Chen**[4,2*], **Xizhou Zhu**[5,2,6], **Lewei Lu**[6], **Tong Lu**[4], **Yu Qiao**[2], **Hongsheng Li**[1], **Jifeng Dai**[5,2], **Wenhai Wang**[1,2✉]
[1]The Chinese University of Hong Kong, [2]Shanghai AI Laboratory, [3]Fudan University, [4]Nanjing University, [5]Tsinghua University, [6]SenseTime Research

Vision-LSTM: xLSTM as Generic Vision Backbone

Benedikt Alkin [1,2] **Maximilian Beck** [1,3] **Korbinian Pöppel** [1,3]
Sepp Hochreiter [1,2,3] **Johannes Brandstetter** [1,2]

JOHANNES KEPLER
UNIVERSITY LINZ

Multi-Dimensional Data
Arrow Pointing Task



Multi-Dimensional RNNs

Multi-Dimensional RNNs / LSTMs (Graves et al. 2007)

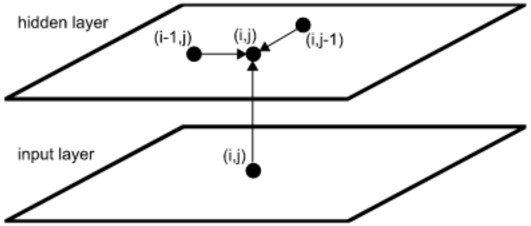


Figure 1: 2D RNN Forward pass.

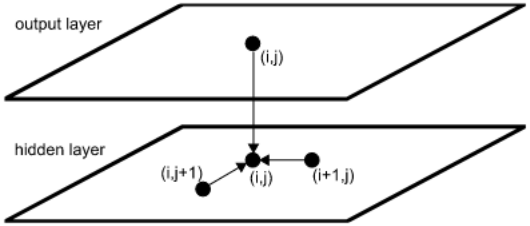
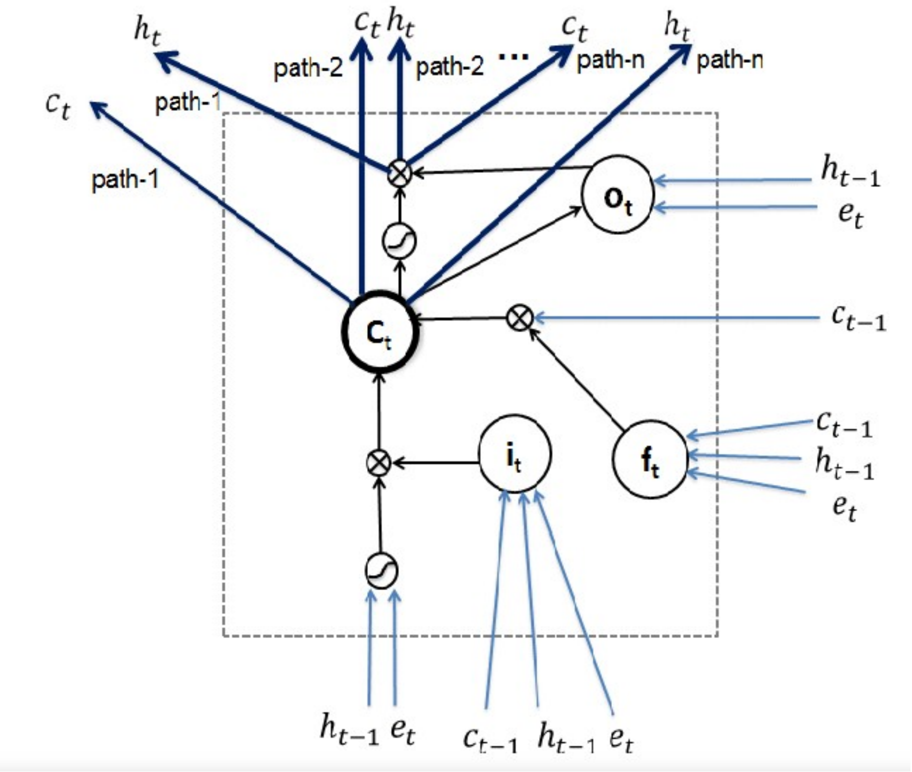


Figure 2: 2D RNN Backward pass.

DAG-LSTM

DAG-Structured Long
Short-Term Memory
for Semantic Compositionality
(Zhu et al. 2016)

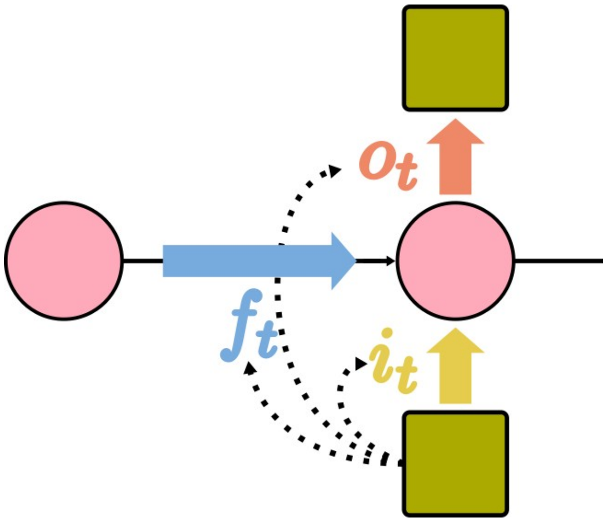


Can we combine linear RNNs with MD-RNNs?

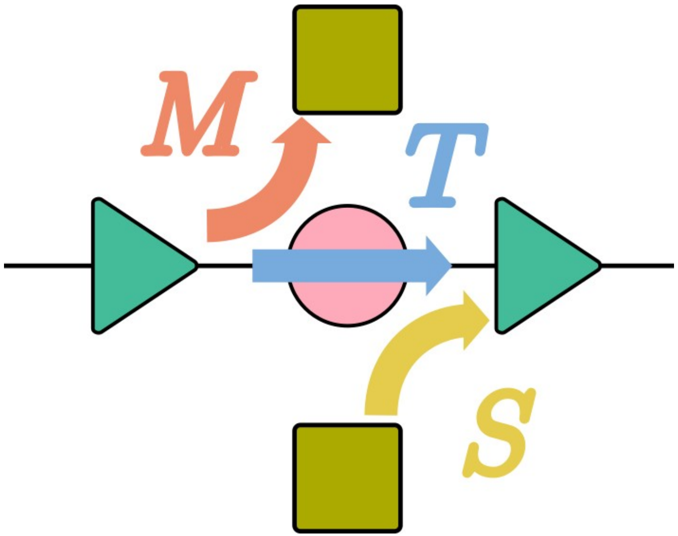
- with parallelization ☑
- with long-range stability ☑
- with directional propagation ☑

From Linear RNNs to pLSTMs

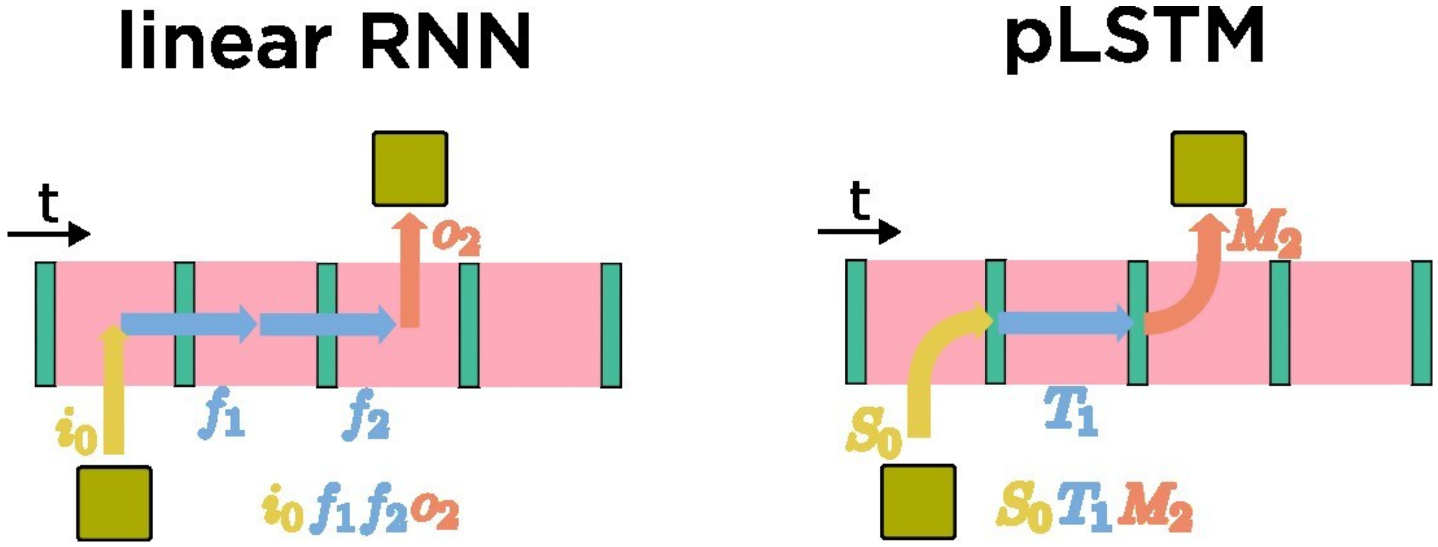
- Input, forget, output gate



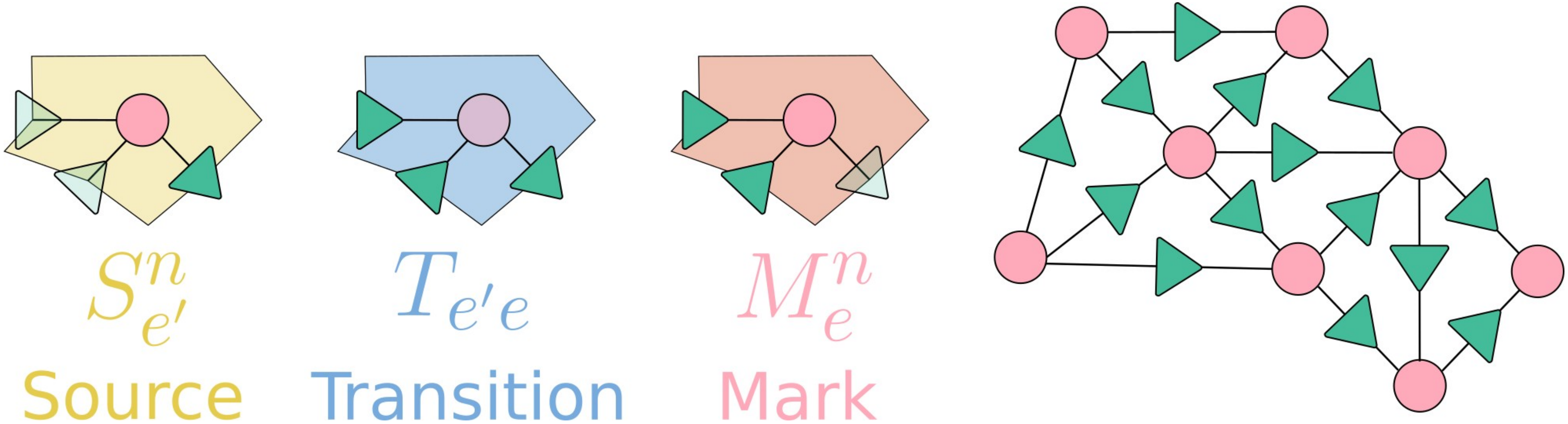
- Source, Transition, Mark



Linear RNN vs pLSTM on a sequence



pLSTM on a DAG



Recurrent Computation: Topological Order

- At current node:
  - Add source contributions to out-going edge states
  - Add transitions from in-coming to out-going edge states
  - Compute Mark from in-coming edge states for network output at current node
- Iterate to next node in the topological order sequence

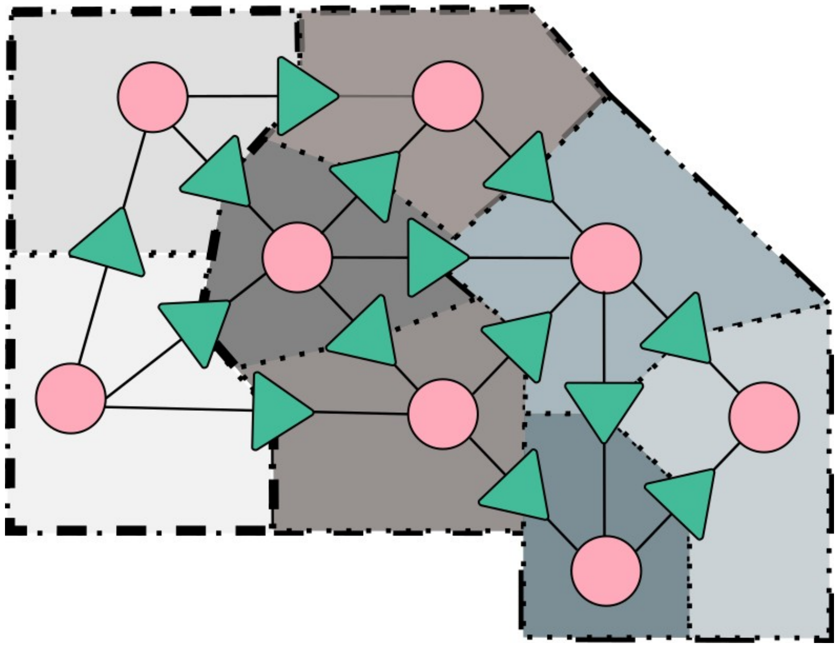
Naive Parallel Computation: Sum All Paths

- Everything is linear
- In between every pair of nodes:
  - Sum all paths in between

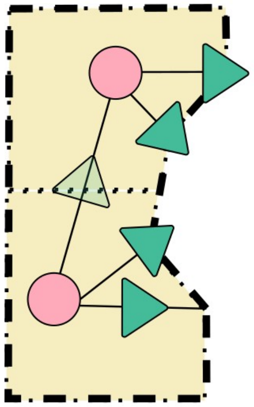
Problem: exponentially many paths (for multi-dim.)

Loosely self-similar DAG

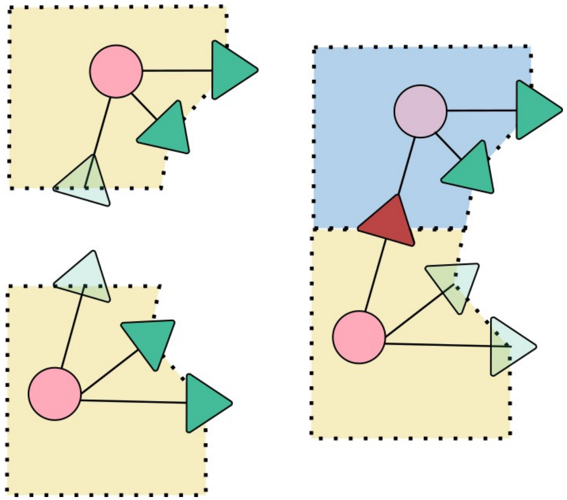
- Low in- and out-degree
- Can be partitioned recursively
- Similar structure for all sub-partitions

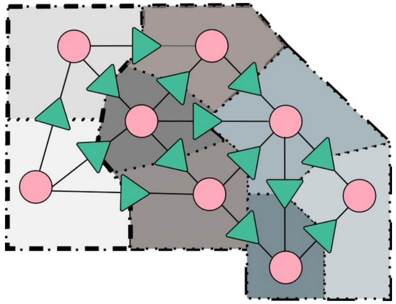


Combining Source and Transition



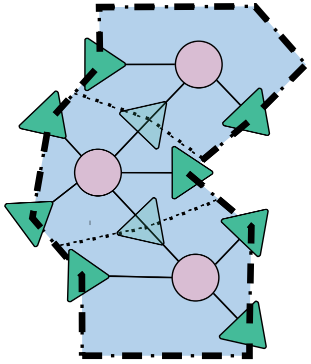
$= \sum$



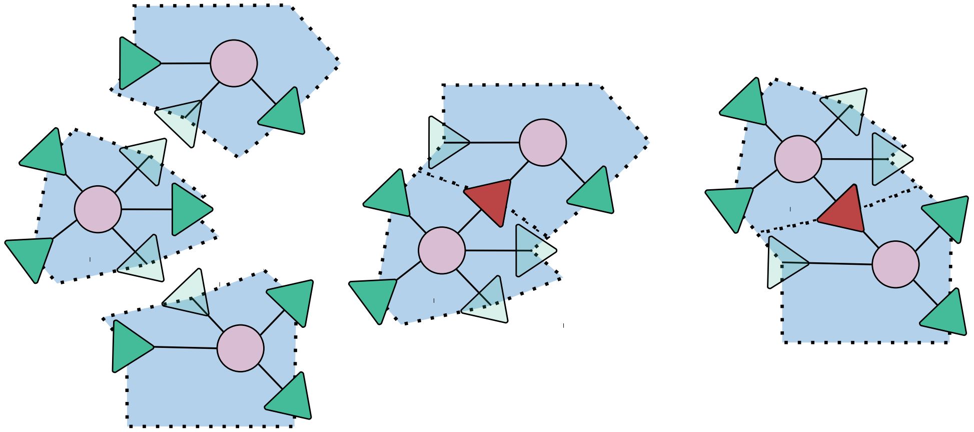


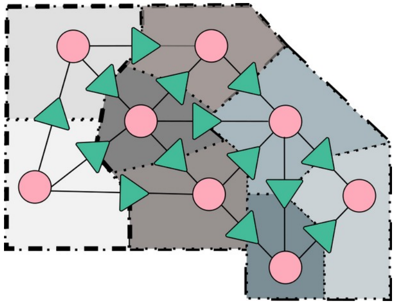
Source (Level n) = Combination of Source (Level n-1) + Transitions (Level n-1)

Combining Transitions



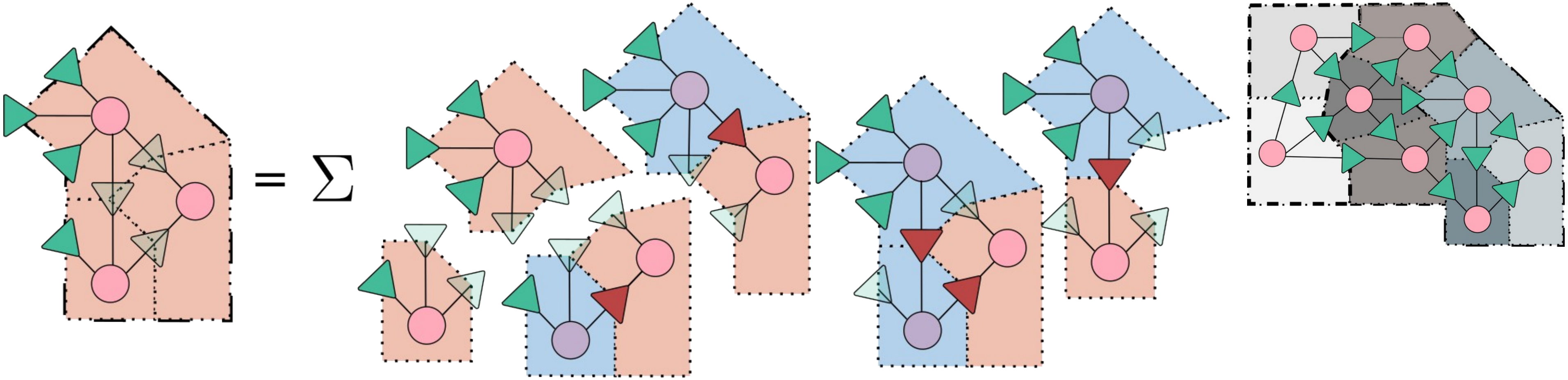
$= \sum$





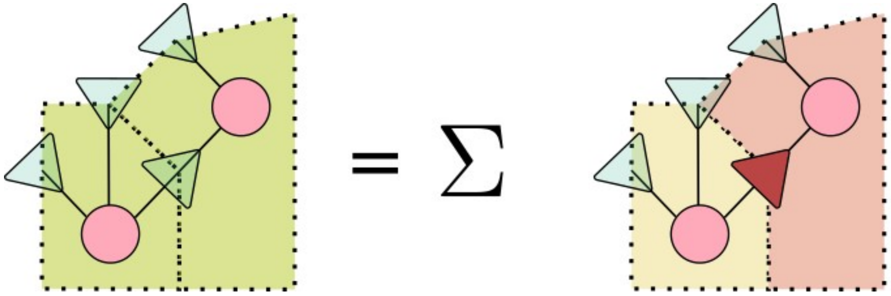
Transition (Level n) = Combination of Transitions (Level n-1)

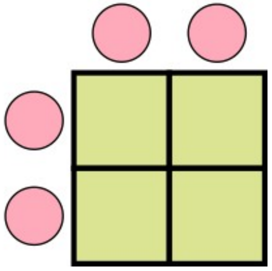
Combining Source and Transition

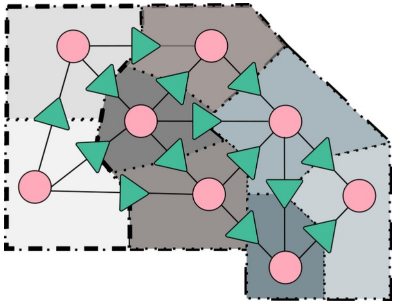


Mark (Level n) = Combination of Transitions (Level n-1) and Mark (Level n-1)

Source and Mark: Gating Matrix



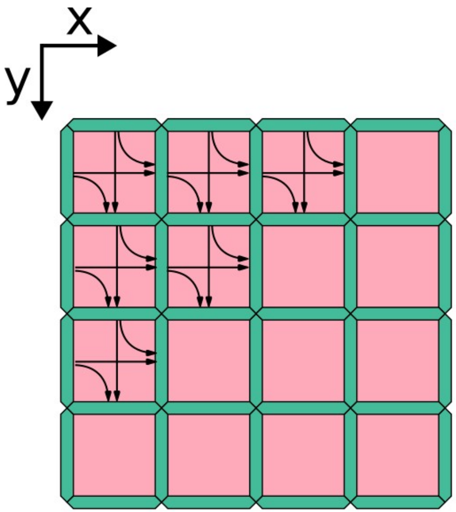




Gating Matrix Entries (like an input dependent “Attention Mask”)

pLSTM on Regular Grids: Images

- Create a DAG: Go Top-Left to Bottom-Right
  + 3 other combinations

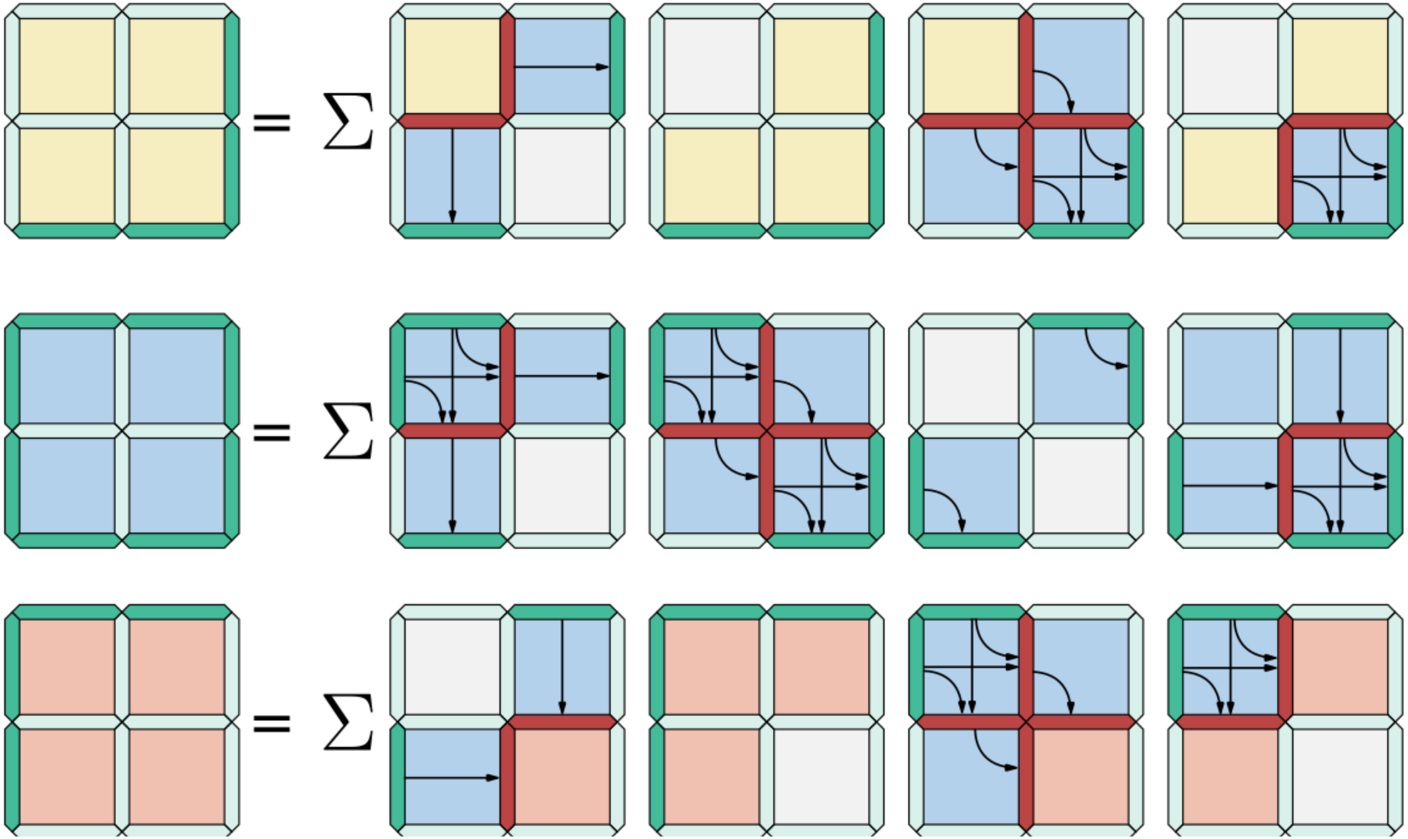


$$S = \begin{bmatrix} S_x & S_y \end{bmatrix}$$

$$T = \begin{bmatrix} T_{xx} & T_{xy} \\ T_{yx} & T_{yy} \end{bmatrix}$$

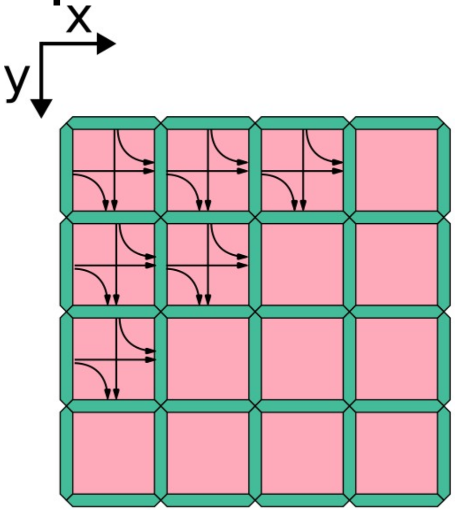
$$M = \begin{bmatrix} M_x \\ M_y \end{bmatrix}$$

Parallelization 2D



Long-Range Stability

- Exponentially many paths
  - from every edge two options
  - Activation blowup



$$S = \begin{bmatrix} S_x & S_y \end{bmatrix}$$

$$T = \begin{bmatrix} T_{xx} & T_{xy} \\ T_{yx} & T_{yy} \end{bmatrix}$$

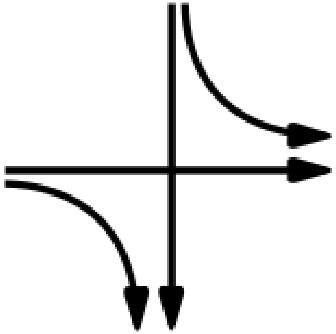
$$M = \begin{bmatrix} M_x \\ M_y \end{bmatrix}$$

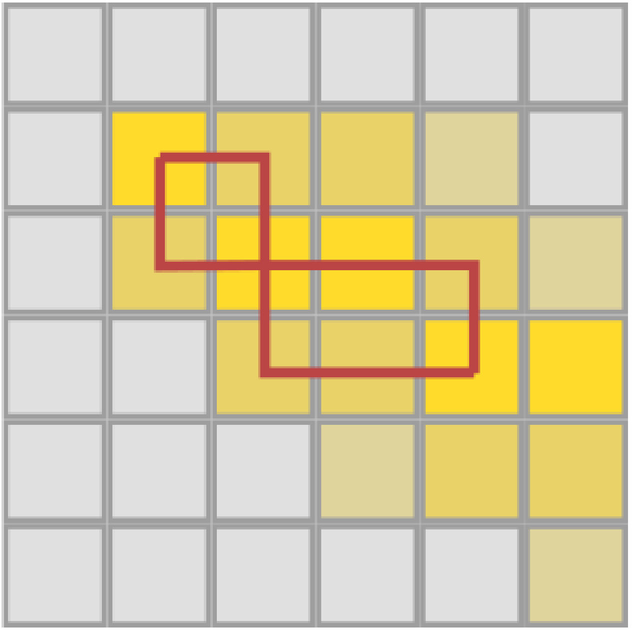
P-mode: L_1 Norm on Transitions

$$|T_{xx}| + |T_{yx}| = 1$$

$$|T_{xy}| + |T_{yy}| = 1$$

$$T = \begin{bmatrix} \alpha & \alpha \\ 1-\alpha & 1-\alpha \end{bmatrix}$$



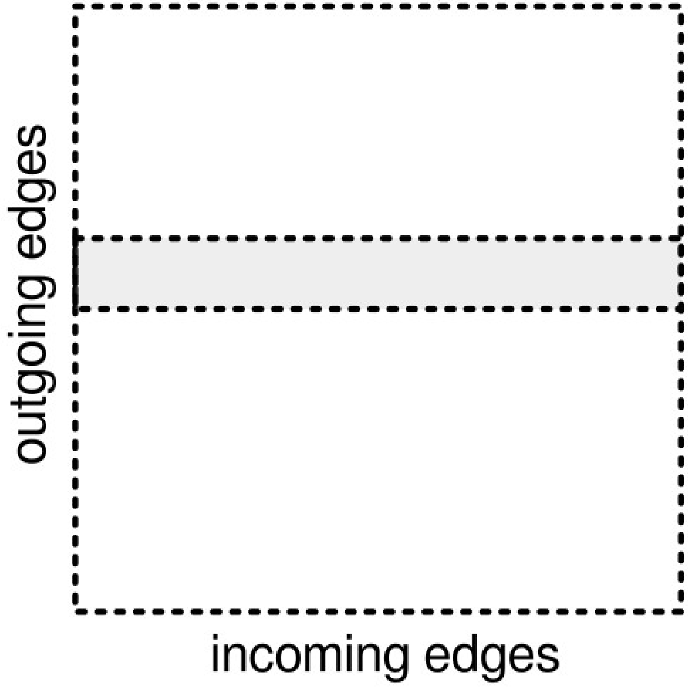


→ Directional Propagation

similar: Mamba2D
(Baty et. al. 2024)
but there fixed 45° angle

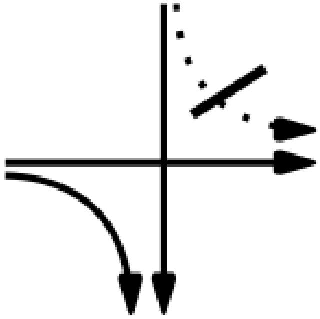
P-mode on a DAG

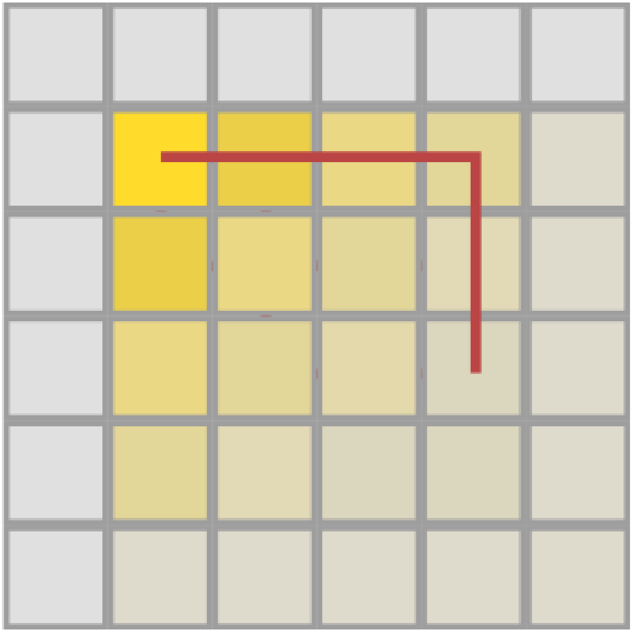
- L1 norm of the line-graph adjacency matrix composed of transition entries



D-mode: Single Connecting Path

$$T = \begin{bmatrix} \gamma_x & 1 \\ 0 & \gamma_y \end{bmatrix}$$



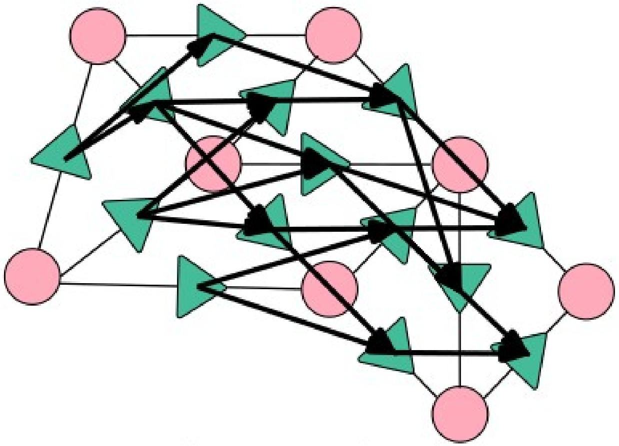


→ Diffusive Distribution

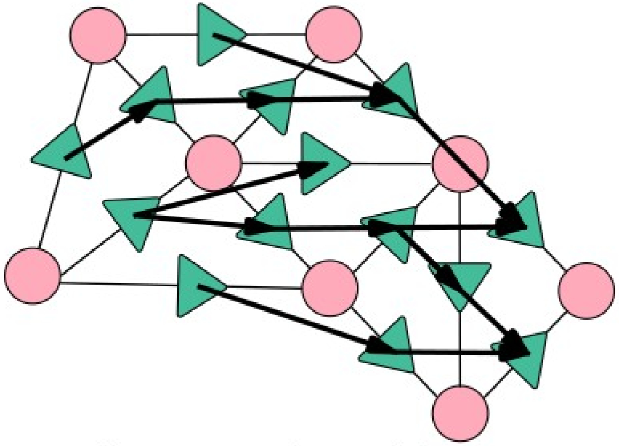
similar: 2DMamba
(Zhang et. al. 2024)

D-mode on a DAG

- Line graph is pruned to a multi-tree



line graph

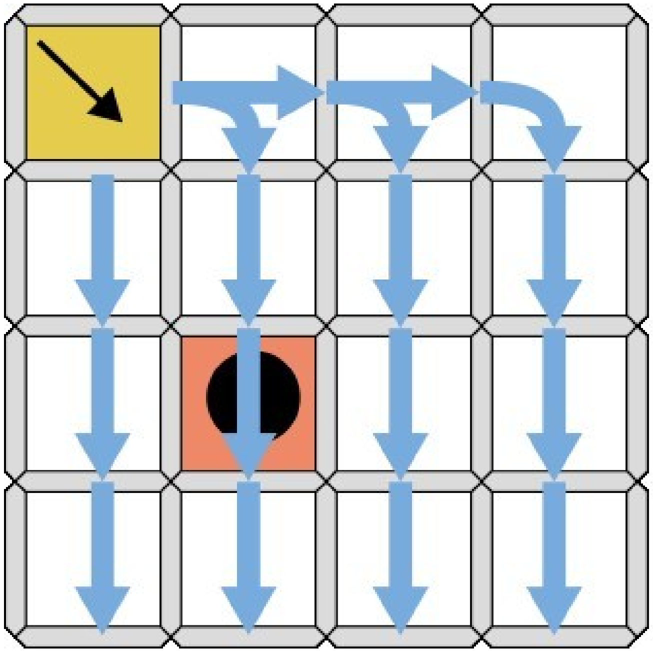


line graph multitree

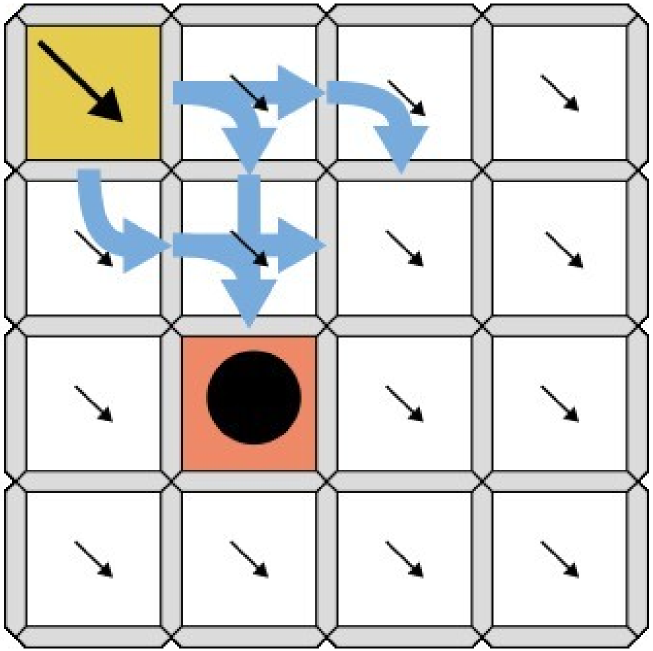
General Computational Graphs

- every finite computational graph can be unrolled to a DAG
- Back-Propagation uses gradients = linear approximations of the graph
- pLSTM analysis holds for all neural networks

pLSTM for Arrow Pointing

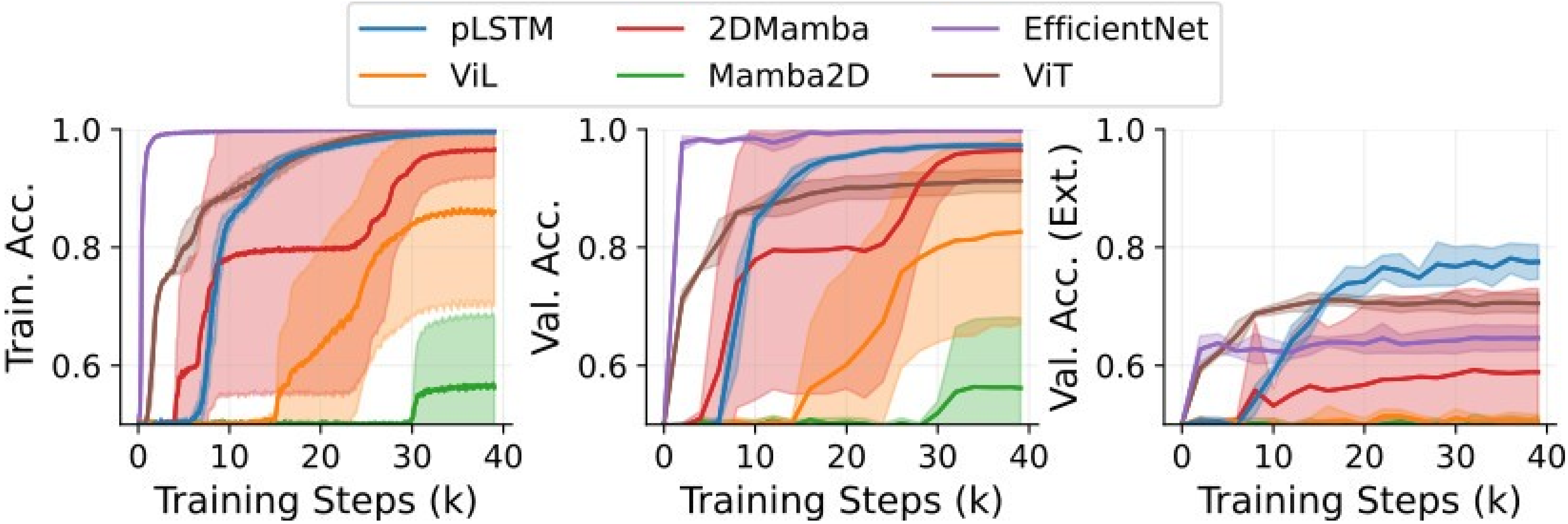


D-mode



P-mode

Arrow Pointing Extrapolation Results



ImageNet-1K

- pLSTM roughly on par with existing methods
- room for improvement by optimizing training schedule + backbone

Model	Epochs	#Params	FLOPS	IN-1K
EfficientNet-B0 [Tan and Le, 2019]	?	5M	0.39G	77.1
DeiT-T [Touvron et al., 2021]	300	6M	1.3G	72.2
DeiT-III-T (reimpl.) [Touvron et al., 2022]	800+20	6M	1.1G	75.4
VRWKV-T [Duan et al., 2024]	300	6M	1.2G	75.1
Vim-T [Zhu et al., 2024]	300	7M	1.5G	76.1
ViL-T [Alkin et al., 2025]	800+20	6M	1.3G	**78.3**
pLSTM-Vis-T	800+20	6M	1.4G	75.2
EfficientNet-B4 [Tan and Le, 2019]	?	19M	1.8G	**82.9**
DeiT-S [Touvron et al., 2021]	300	22M	4.6G	79.8
DeiT-III-S (reimpl.) [Touvron et al., 2022]	400+20	22M	4.6G	80.3
ConvNeXt-S (*iso.*) [Liu et al., 2022]	300	22M	4.3G	79.7
VRWKV-S [Duan et al., 2024]	300	24M	4.6G	80.1
Vim-S [Zhu et al., 2024]	300	26M	5.3G	80.5
Mamba2D-T [Baty et al., 2024]	300	27M	-	82.4
2DMamba-T [Zhang et al., 2024a]	?	30M	4.9G	82.8
ViL-S [Alkin et al., 2025]	400+20	23M	4.7G	81.5
pLSTM-Vis-S	400+20	23M	4.9G	80.7
EfficientNet-B6 [Tan and Le, 2019]	?	43M	19G	**84.0**
DeiT-B [Touvron et al., 2021]	300	86M	17.6G	81.8
DeiT-III-B (reimpl.) [Touvron et al., 2022]	400+20	87M	16.8G	83.5
ConvNeXt-B (*iso.*) [Liu et al., 2022]	300	87M	16.9G	82.0
VRWKV-B [Duan et al., 2024]	300	94M	18.2G	82.0
2DMamba-S [Zhang et al., 2024a]	?	50M	8.8G	83.8
ViL-B [Alkin et al., 2025]	400+5	89M	17.9G	82.4
pLSTM-Vis-B	400+20	89M	18.2 G	82.5

JOHANNES KEPLER UNIVERSITY LINZ

pLSTM on molecule graphs

model	MUTAG	NCI1	PROTEINS	PTC_FM	AVG
GAT [Veličković et al., 2018]	0.7822 ± 0.09	0.7968 ± 0.03	0.7215 ± 0.03	0.6105 ± 0.05	0.7277 ± 0.06
GCN [Kipf and Welling, 2017]	0.7234 ± 0.08	0.7852 ± 0.02	0.7395 ± 0.03	**0.6162 ± 0.04**	0.7161 ± 0.05
GIN [Xu et al., 2018]	0.8251 ± 0.10	**0.8175 ± 0.02**	0.7350 ± 0.04	0.6097 ± 0.10	**0.7468 ± 0.08**
LSTM GNN [Liang et al., 2016]	0.7450 ± 0.11	0.7951 ± 0.02	**0.7503 ± 0.04**	0.6076 ± 0.04	0.7245 ± 0.06
MPNN [Gilmer et al., 2017]	0.7450 ± 0.09	0.8012 ± 0.02	0.7350 ± 0.04	0.5786 ± 0.07	0.7149 ± 0.06
PLSTM	**0.8512 ± 0.06**	0.7324 ± 0.03	0.7502 ± 0.05	0.6133 ± 0.08	0.7368 ± 0.06

- pLSTM roughly on par with existing methods
- room for improvement by including more inductive biases

Conclusion

- Contributions:
  - translating linear RNNs to multiple dimensions + general DAGs, including a parallelization scheme
  - Long-range stabilization scheme for linear networks on DAGs with P- and D-mode
  - Arrow-Pointing Extrapolation (APE) task as synthetic benchmark for long-range directional capabilities
- Limitations / Future work:
  - Mechanistically investigate solution modes of pLSTM (and others) on APE
  - Improve results on common benchmark, by integrating further inductive biases
  - Harder real-world benchmarks that need long-range multi-dimensional relationships (e.g. maps?)

Bibliography

- Graves, Alex, Santiago Fernandez, and Juergen Schmidhuber. “Multi-Dimensional Recurrent Neural Networks.” arXiv, May 14, 2007. http://arxiv.org/abs/0705.2011.
- Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A Large-Scale Hierarchical Image Database.” In 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–55. Ieee, 2009.
- Zhu, Xiao-Dan, Parinaz Sobhani, and Hongyu Guo. “DAG-Structured Long Short-Term Memory for Semantic Compositionality.” In North American Chapter of the Association for Computational Linguistics, 2016. https://api.semanticscholar.org/CorpusID:2016937.
- Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.” In International Conference on Learning Representations, 2021. https://openreview.net/forum?id=YicbFdNTTy.
- Touvron, Hugo, Matthieu Cord, and Hervé Jégou. “DeiT III: Revenge of the ViT.” arXiv, April 14, 2022. https://doi.org/10.48550/arXiv.2204.07118.
- Zhu, Lianghui, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. “Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model.” In Forty-First International Conference on Machine Learning, 2024. https://openreview.net/forum?id=YbHCqn4qF4.
- Duan, Yuchen, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. “Vision-RWKV: Efficient and Scalable Visual Perception with RWKV-Like Architectures.” arXiv, March 31, 2025. https://doi.org/10.48550/arXiv.2403.02308.
- Alkin, Benedikt, Maximilian Beck, Korbinian Pöppel, Sepp Hochreiter, and Johannes Brandstetter. “Vision-LSTM: xLSTM as Generic Vision Backbone.” In The Thirteenth International Conference on Learning Representations, 2025. https://openreview.net/forum?id=SiH7DwNKZZ.
- Baty, Enis, Alejandro Hernández Díaz, Chris Bridges, Rebecca Davidson, Steve Eckersley, and Simon Hadfield. “Mamba2D: A Natively Multi-Dimensional State-Space Model for Vision Tasks.” arXiv, December 20, 2024. https://doi.org/10.48550/arXiv.2412.16146.
- Zhang, Jingwei, Anh Tien Nguyen, Xi Han, Vincent Quoc-Huy Trinh, Hong Qin, Dimitris Samaras, and Mahdi S. Hosseini. “2DMamba: Efficient State Space Model for Image Representation with Applications on Giga-Pixel Whole Slide Image Classification.” arXiv, December 1, 2024. https://doi.org/10.48550/arXiv.2412.00678.