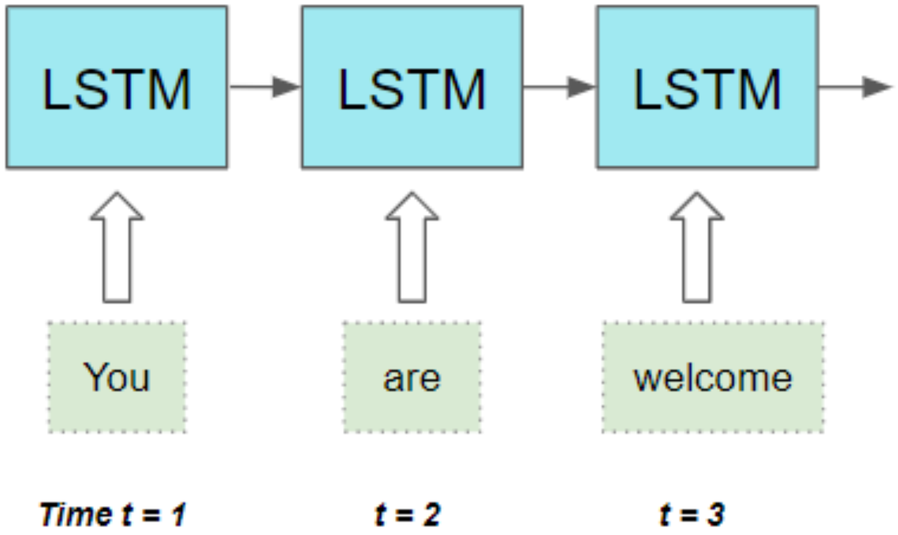
物理に着想を得たモデルによる、長期的かつコヒーレントな再帰システムの探究

宮戸 岳 (University of Tübingen)

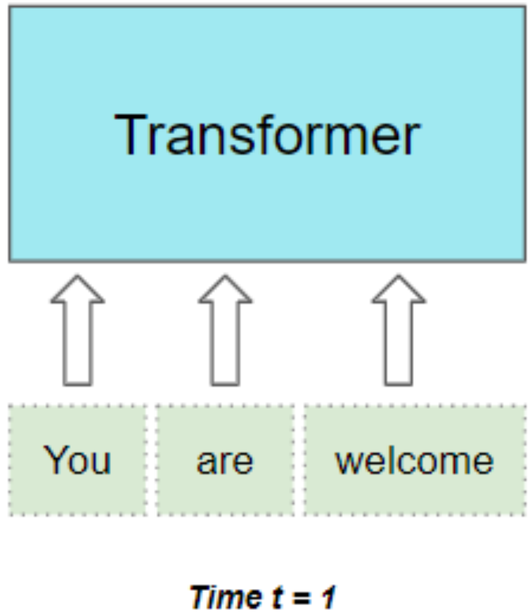
ニューラル(特に言語処理)モデル

before 2017



逐次処理

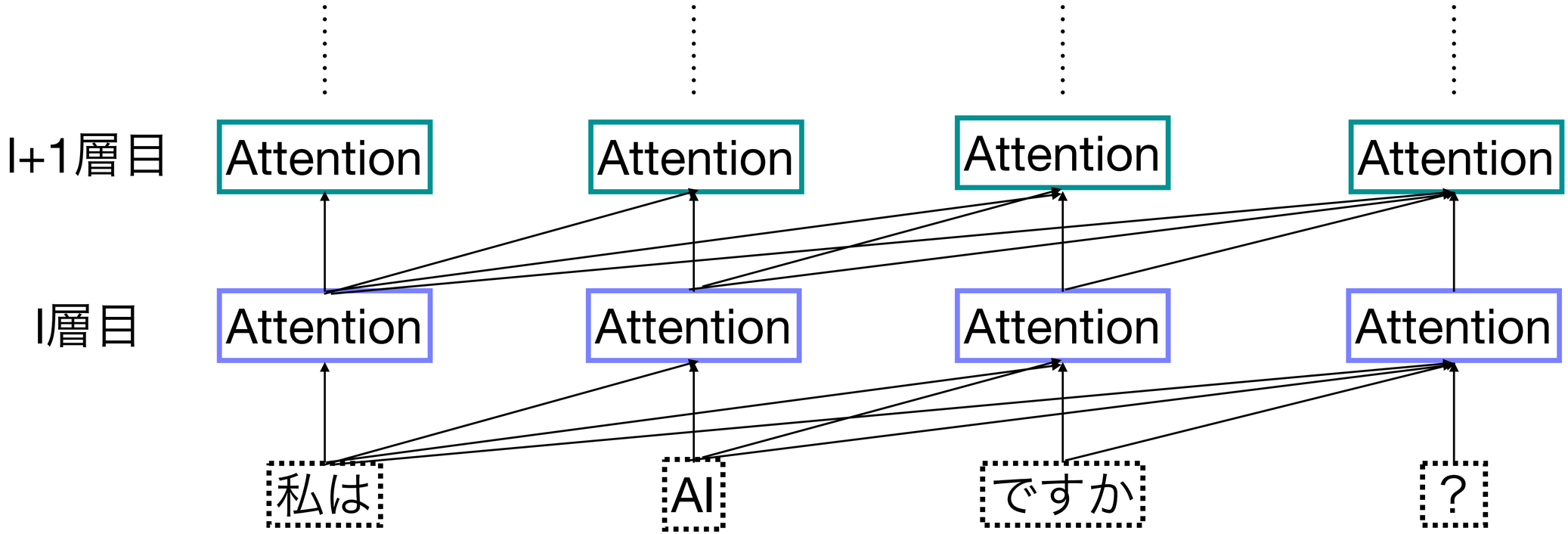
after 2017



並列処理 (GPU フレンドリー)

https://towardsdatascience.com/transformers-explained-visually-part-1-overview-of-functionality-95a6dd460452/

トランスフォーマー (2017)



トランスフォーマー は 端的に言えば**再帰の無い**モデル

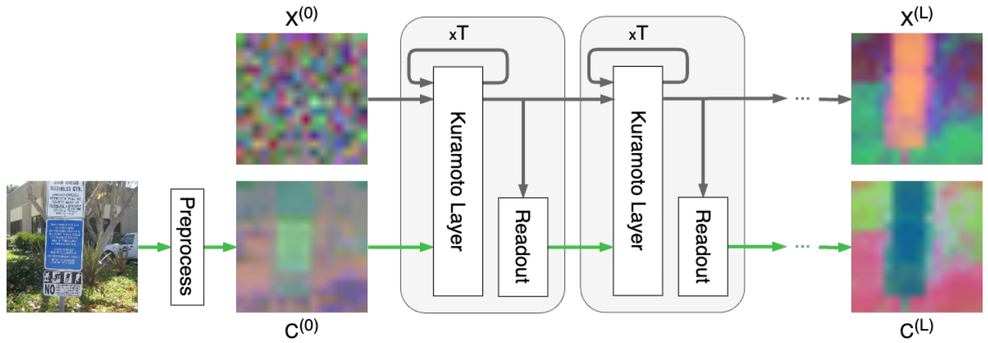
- **各層は別々のパラメーターを持つ**
- **l+1 層目は l層のみの値に依存して計算される**

現在 (2026)

再び、再帰をモデルに取り入れる試みが注目されている

- **層方向の再帰:**
  各層が共通のパラメータを持つ:

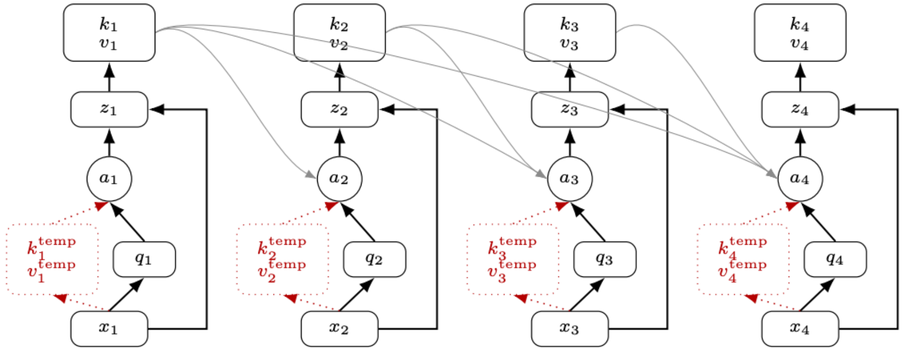
例) TRM, HRM, AKOrN, Looped Trnsfmrs



- 学習分布外への汎化
- テスト時の”長考”を可能にする

- **時間方向の再帰:**
  前時刻の状態に依存したモデル

例) Recurrent transformer



https://arxiv.org/pdf/2604.21215v1

-スケーリングの改善

自分たちの試み

物理現象やモデルにinspireされたarchitectureや解析

- **層方向の再帰:**

  AKOrN (2025)

- **時間方向の再帰:**

  SSB for Deep information propagation (2026)

  Fast-slow recurrence for sequential processing (2026)

Artificial Kuramoto Oscillatory Neurons

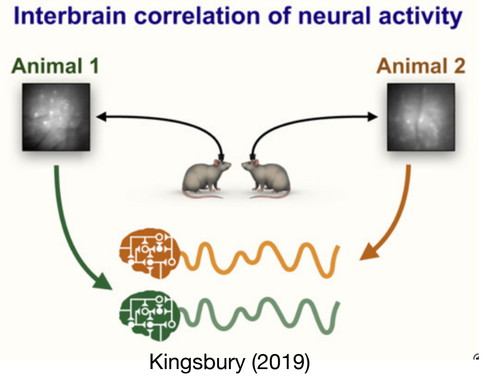
ICLR2025 (Oral)

Takeru Miyato[1] Sindy Löwe[2] Andreas Geiger[1] Max Welling[2]

1: University of Tübingen, 2: University of Amsterdam

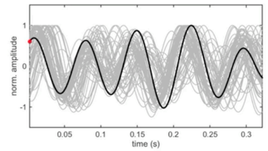
同期現象

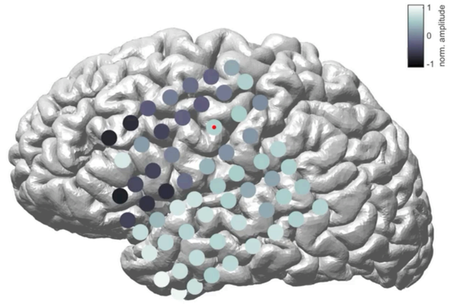
- 同期現象は我々の身の回り至るとこで見られる



Kingsbury (2019)

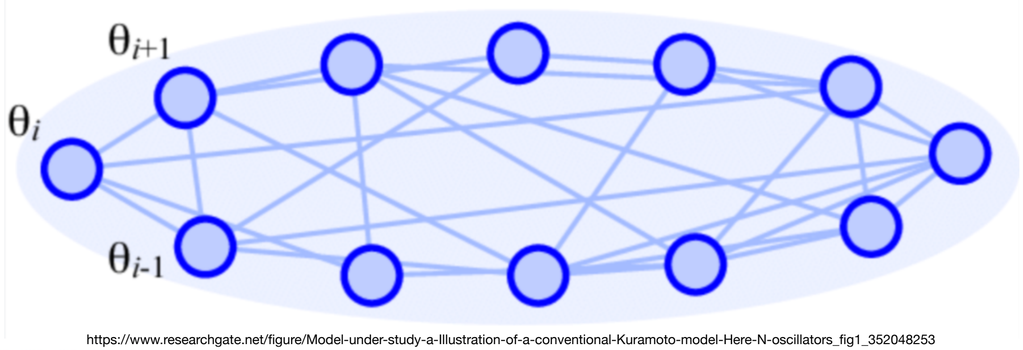
- また同期はいくつかのニューロン機構において重要な土台である
  - W**inner-take-all** competition
  - F**eature binding**
  - I**nformation compression**





蔵本モデル (1975)

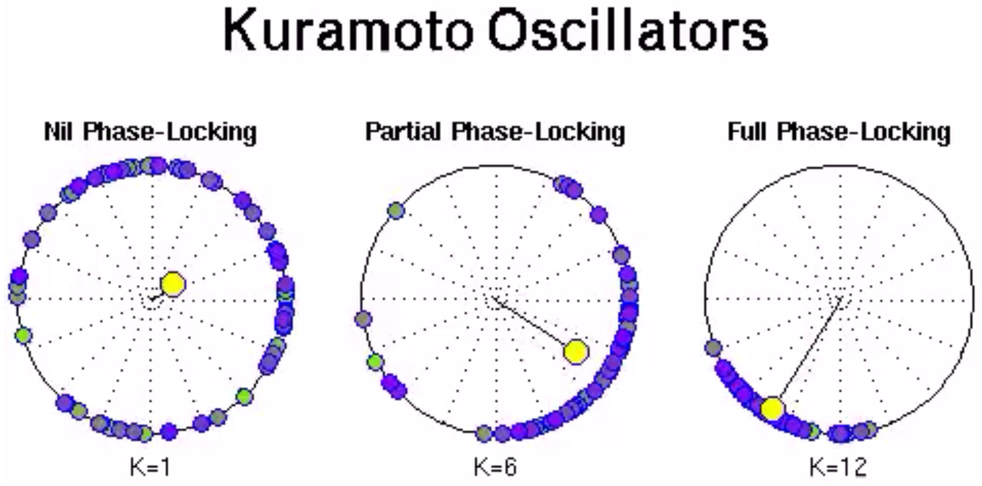
a model to describe synchronization



https://www.researchgate.net/figure/Model-under-study-a-Illustration-of-a-conventional-Kuramoto-model-Here-N-oscillators_fig1_352048253

$$\dot{\theta}_i = \omega_i + \sum_j J_{ij} \sin(\theta_j - \theta_i)$$

- $\theta_i \in [0, 2\pi)$ は各振動子の位相
- $J_{ij} \in R$ はj番目からi番目振動子への重み
- $\omega_i \in R$ は i番目の振動子の自然周波数
- J_{ij} と ω_i の強さのバランスが同期の強さを決める

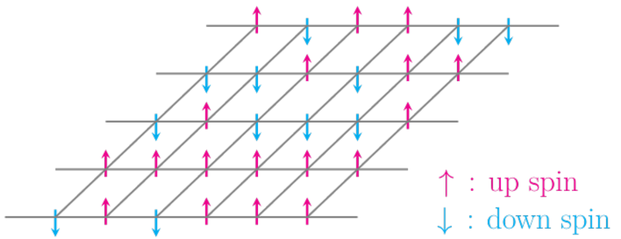


Nil, partial and full phase-locking in an all-to-all network of Kuramoto oscillators. Phase-locking is governed by the coupling strength K and the distribution of intrinisic frequencies ω. Here, the intrinsic frequencies were drawn from a normal distribution (M=0.5Hz, SD=0.5Hz). The yellow disk marks the phase centroid. Its radius is a measure of coherence.

https://en.wikipedia.org/wiki/Kuramoto_model

ニューラルネットの層としての蔵本モデル

- 蔵本モデルをニューラルネットワークの反復層として組み込んだらどうなるだろうか？
  - 同期ダイナミクスによる情報圧縮
    - モデルが圧縮された特徴表現を持つように誘導する自然な方法
  - 蔵本モデルはイジングモデルの連続的アナログとして捉えることができる。組合せ的・状態探索的な計算を連続ダイナミクスとして実行する枠組みを与える。



Artificial Kuramoto Oscillatory Neurons

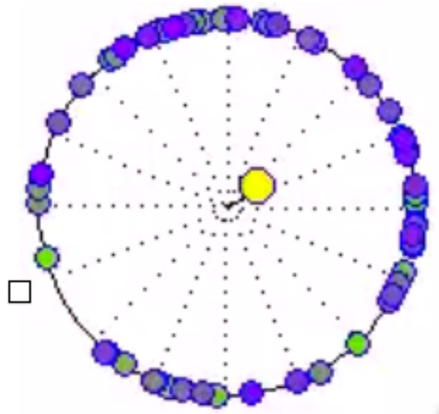
as dynamical alternative to threshold neurons

- 蔵本モデルを超球面に一般化したニューロンモデルを考える (AKOrN)
- AKOrN のニューロン機構がネットワークの振る舞いを大きく変えることが分かった。
  - 標準モデルに匹敵する性能で、物体に結び付いた良い特徴を学習可能
  - パズル推論タスクで特に難しいタスクに汎化
  - 様々な摂動（ランダム・敵対的・自然な）に頑健で、良い較正

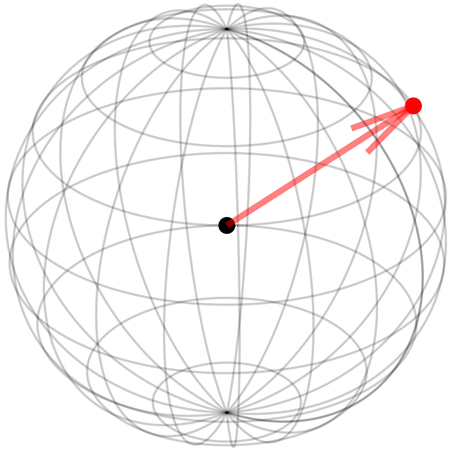
AKOrNモデル

AKOrN
超球面上を動くニューロンモデル

オリジナルの蔵本モデル: S^1上を動くニューロン

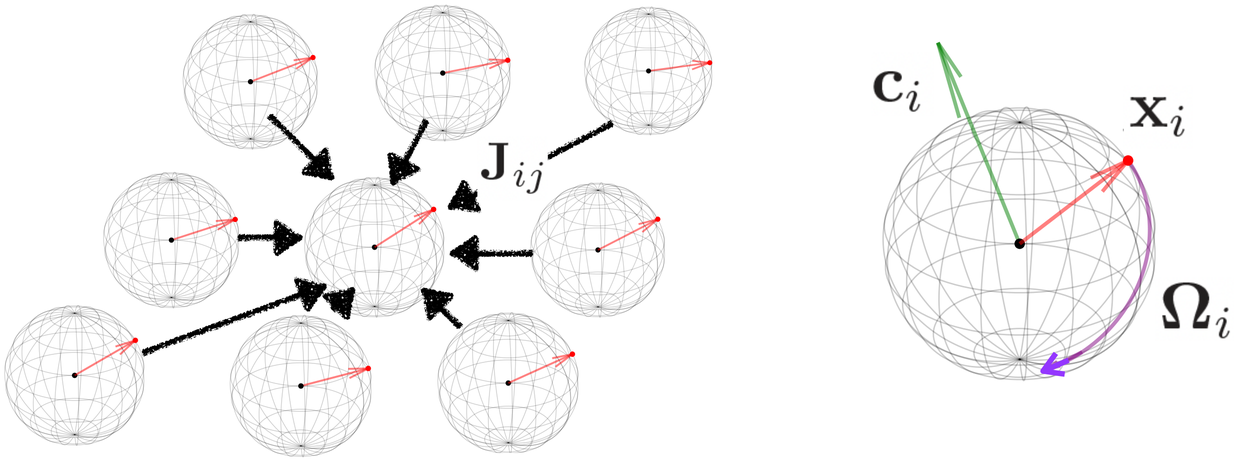


AKOrNモデル: S^{N-1} 上のニューロン



AKOrN
超球面上を動くニューロンモデル

AKOrNモデル: S^{N-1} 上のニューロン



AKOrNダイナミクス: $\dot{\mathbf{x}}_i = \mathbf{\Omega}_i \mathbf{x}_i + \mathrm{Proj}_{\mathbf{x}_i}\left(\mathbf{c}_i + \sum_j \mathbf{J}_{ij}\mathbf{x}_j\right)$ $(\mathbf{x}_i \in \mathbb{R}^N, \|\mathbf{x}_i\| = 1)$

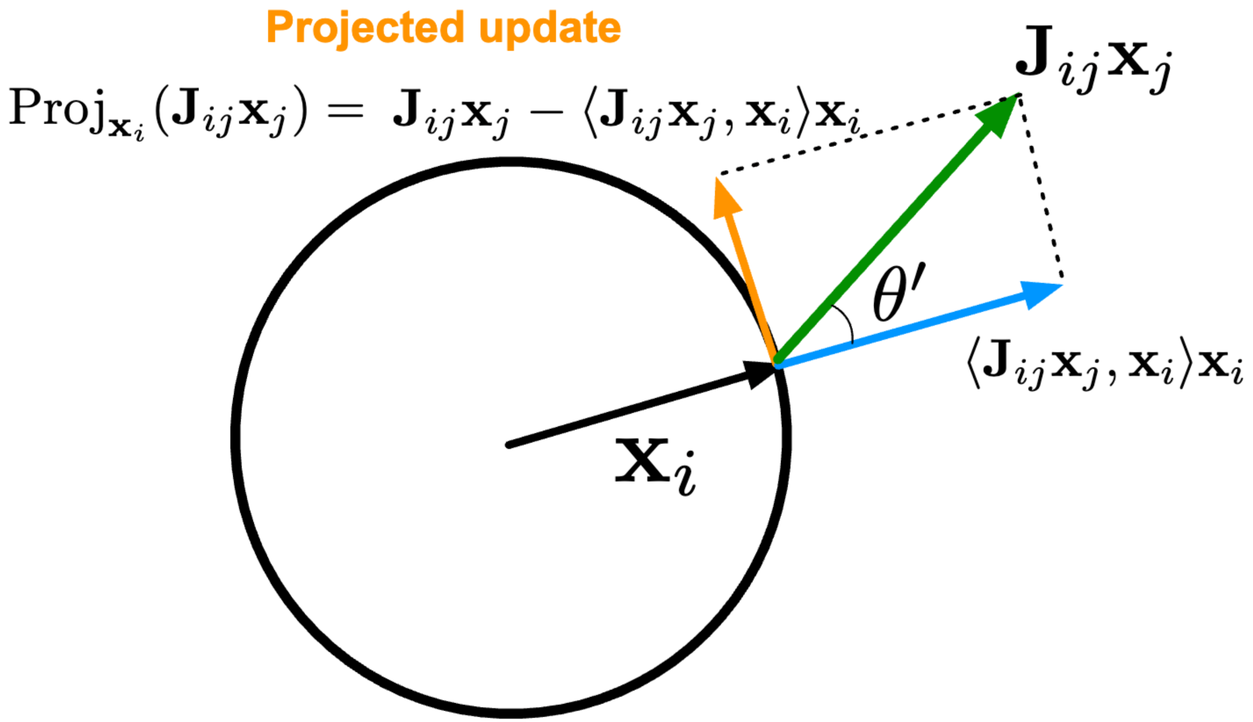
natural freq.

coupling

conditional bias

Proj: 射影オペレーター

$$\mathrm{Proj}_{\mathbf{x}_i}\left(\mathbf{c}_i + \sum_j \mathbf{J}_{ij}\mathbf{x}_j\right) \text{ where } \mathrm{Proj}_{\mathbf{x}_i}(\mathbf{y}_i) = \mathbf{y}_i - \langle \mathbf{y}_i, \mathbf{x}_i \rangle \mathbf{x}_i$$

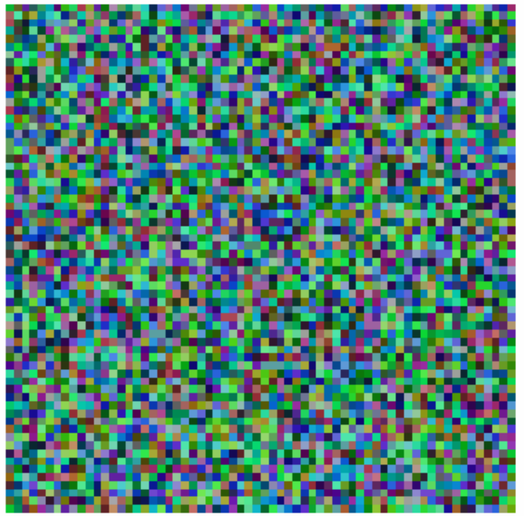


Example: Fish silhouette

64x64 grid example

C

X

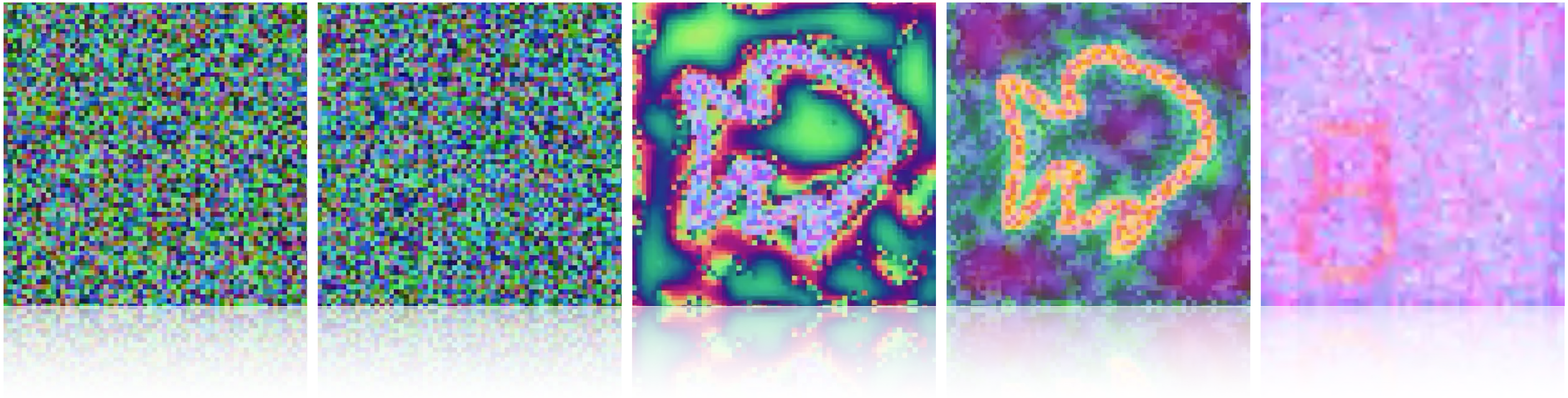


シルエット外の**c** は **1**

シルエット内の**c** は = **0**

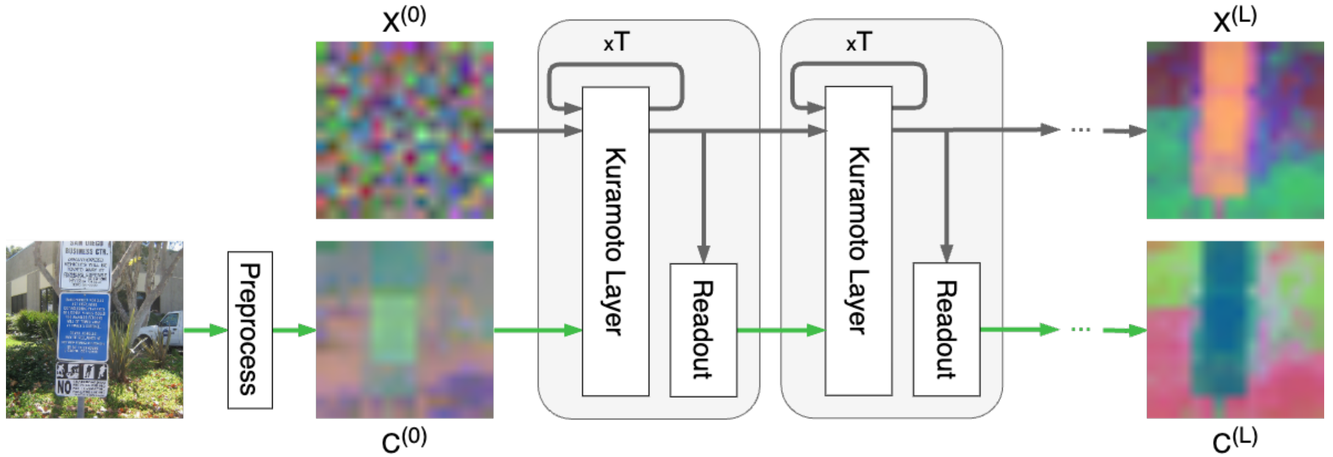
重みJはランダムな局所重み

さまざまな波状の振動子パターン



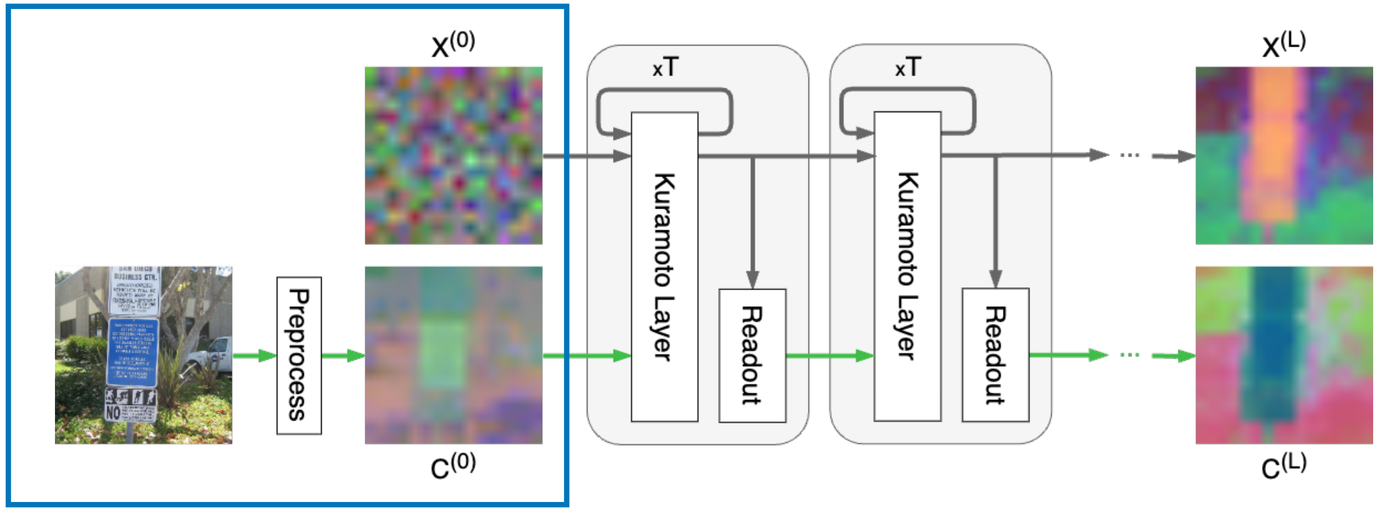
AKOrNを用いたネットワーク

- AKOrN ダイナミクスをネットワークの再帰的な層として用いる
- 通常のニューラルネットとは異なり、振動子ベクトル$\mathbf{x}$ と条件付けバイアスベクトル$\mathbf{c}$のペアを処理する



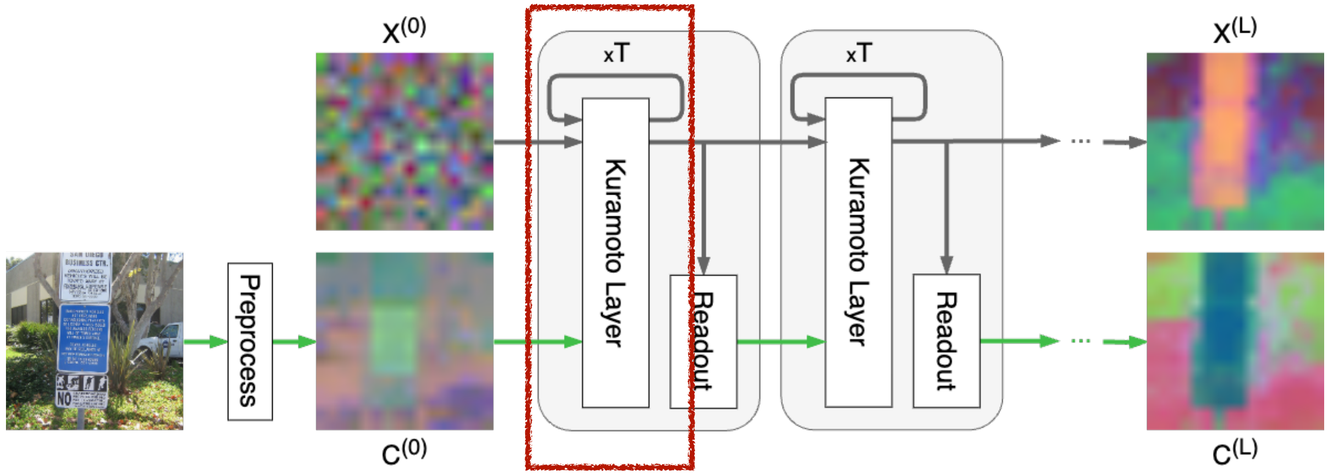
1. XとCの初期化

- I球面上の一様分布からサンプルした振動子$\mathbf{X}^{(0)}$ を初期化する
- 観測入力を処理し、線形層または1層の畳み込み層を用いて $\mathbf{C}^{(0)}$ を作る。
  - 例えばパッチ化など：



2. Xの更新

- 振動子 **X** をバイアス**C**のもと更新する
  - 以下の離散化されたAKOrNのODEをT回適用する

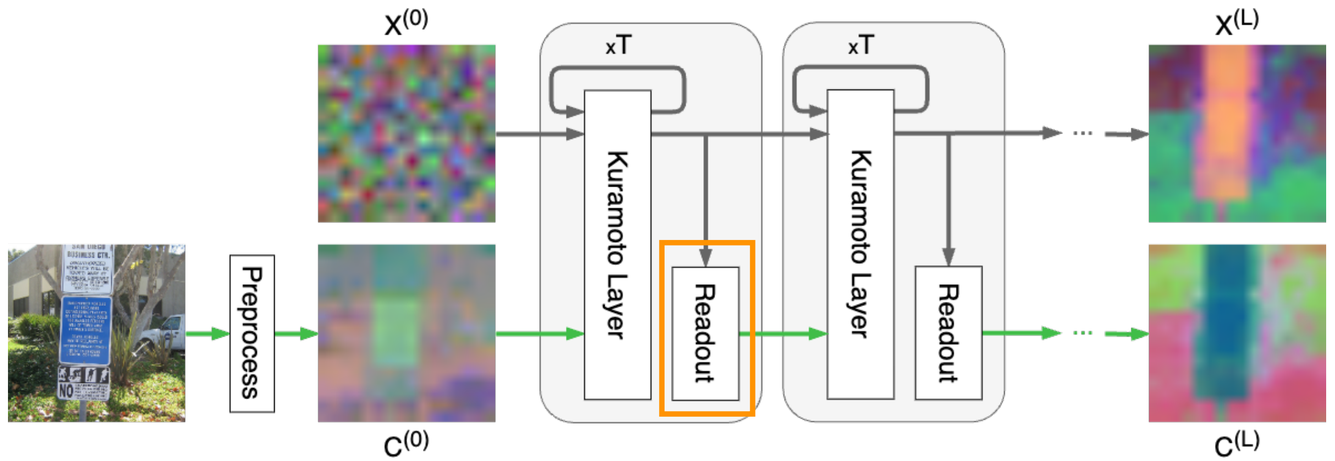


$$\Delta \mathbf{x}_i^{(l,t)} = \mathbf{\Omega}_i^{(l)} \mathbf{x}_i^{(l,t)} + \mathrm{Proj}_{\mathbf{x}_i^{(l,t)}}(\mathbf{c}_i^{(l-1)} + \sum_j \mathbf{J}_{ij}^{(l)} \mathbf{x}_j^{(l,t)})$$

$$\mathbf{x}_i^{(l,t+1)} = \Pi \left[\mathbf{x}_i^{(l,t)} + \gamma \Delta \mathbf{x}_i^{(l,t)} \right],$$

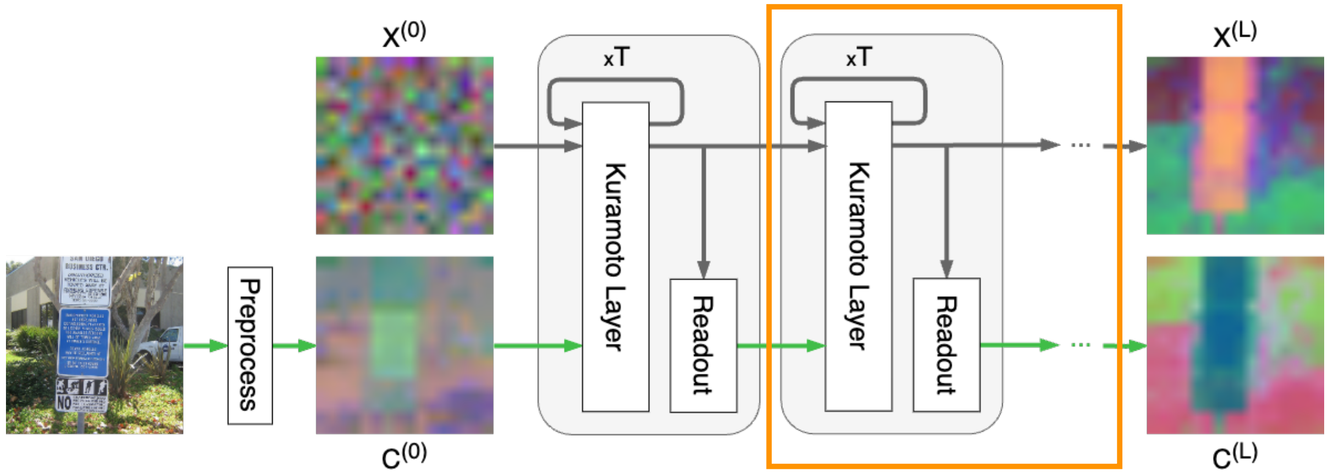
3. $X^{(l,T)}$ から新しいCを作る

- 読み出しブロックが$X^{(l,T)}$から新しい条件信号 $C^{(l)}$ を作る
  - 読み出しブロックは普通のMLPなど
- ペア $\{X^{(l,T)}, C^{(l)}\}$ を次のブロックにわたす



4. ネットワーク出力の計算

- 「KuramotoLayer & Readout」ブロックを重ねる
- $C^{(L)}$ を用いてタスクの最終予測を行う
  - 例: 画像分類では、C(L) をさらにソフトマックス分類器で処理する
- すべてのパラメータ (**J**,**Ω**,γ,Readout) は誤差逆伝播で最適化される。



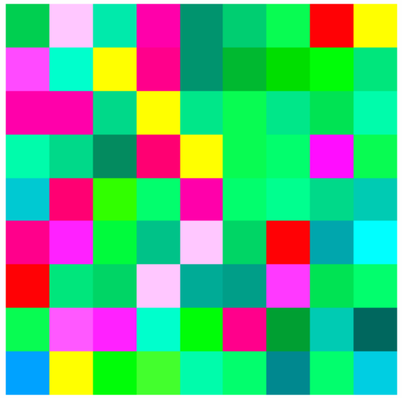
重みについて Local vs. Global

- We tested two types of connectivities: **convolution** and **self-attention**
  - **Convolution**: **Local** connectivity
  - **Attention**: **Global + Modulation**
    - Basically **Attention** works better in general in our experiments

- By constraining J to be symmetric, we can prove there is a Lyapunov function for the dynamics.
  - But, we **don’t** impose such symmetric constraints
    - Asymmetric connections are found to be important

Experiments

Unsupervised object discovery



6	5	7	9	9	3	9	2	6
8	8	6	3	4	7	5	5	7
3	9	7	6	5	5	7	8	6
6	5	3	4	6	4	9	7	2
1	4	2	7	9	5	7	3	6
1	7	1	8	5	7	2	4	6
2	3	3	5	5	2	8	2	8
4	8	7	5	3	1	9	5	5
4	6	9	3	1	1	2	7	5

Sudoku Solving

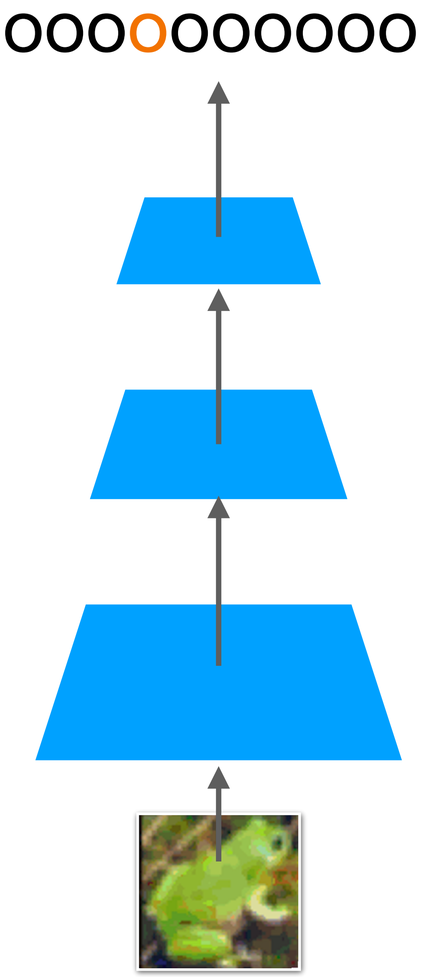
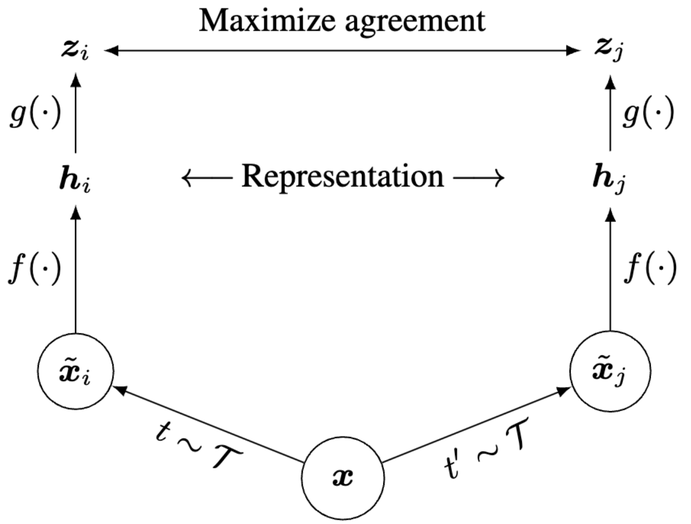


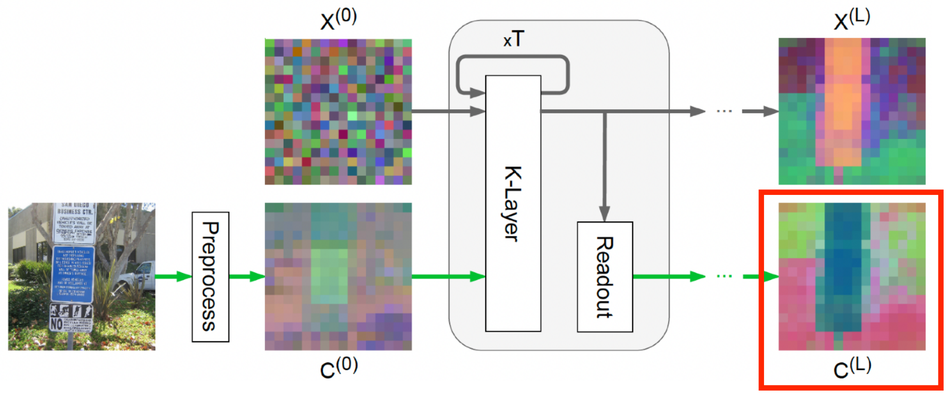
Image classification

Task1. Unsupervised object discovery

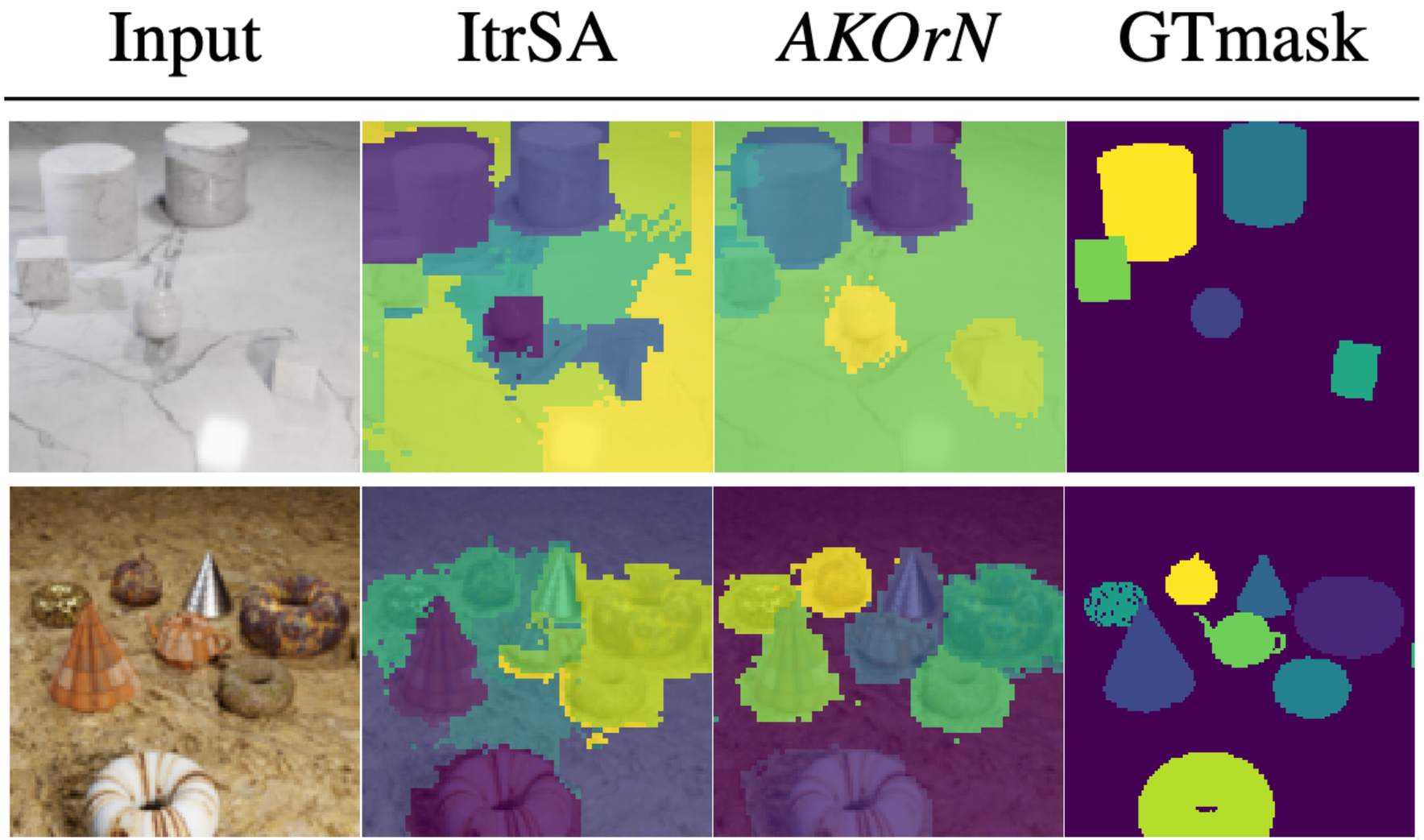
- Training: self-supervised learning using the SimCLR method.
- Evaluation: alignments between learned features and GT object masks.
    1. Apply clustering to the features of $\mathbf{C}^{(\mathrm{L})}$
    2. Compute alignment between the obtained clusters and the ground truth labels



SimCLR (Chen et al. 2020)



AKOrNs Learn object-binding features



スロットベースモデルに匹敵する性能

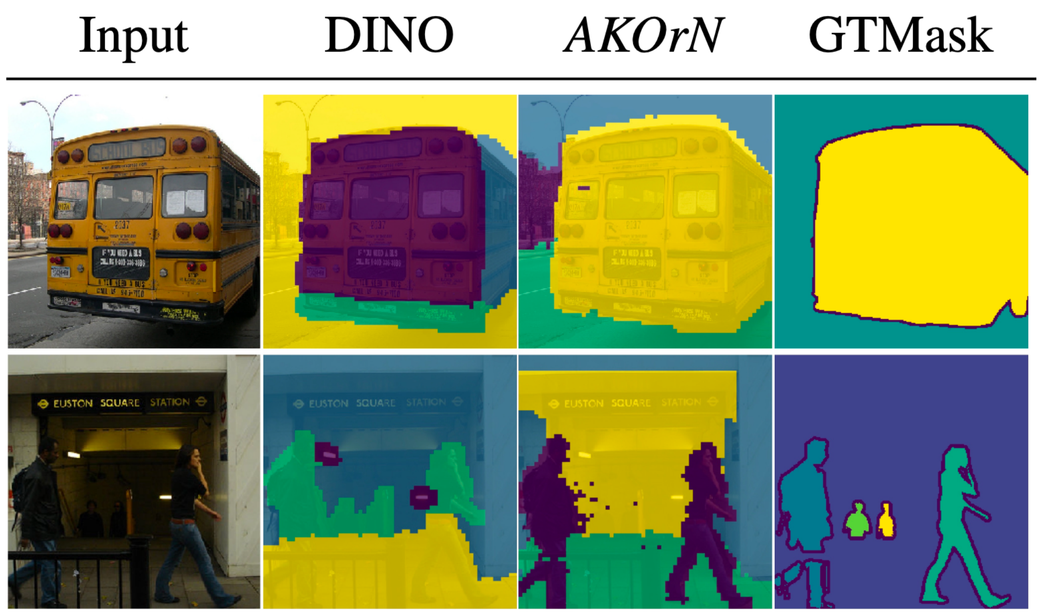
トランスフォーマー（再帰なし）

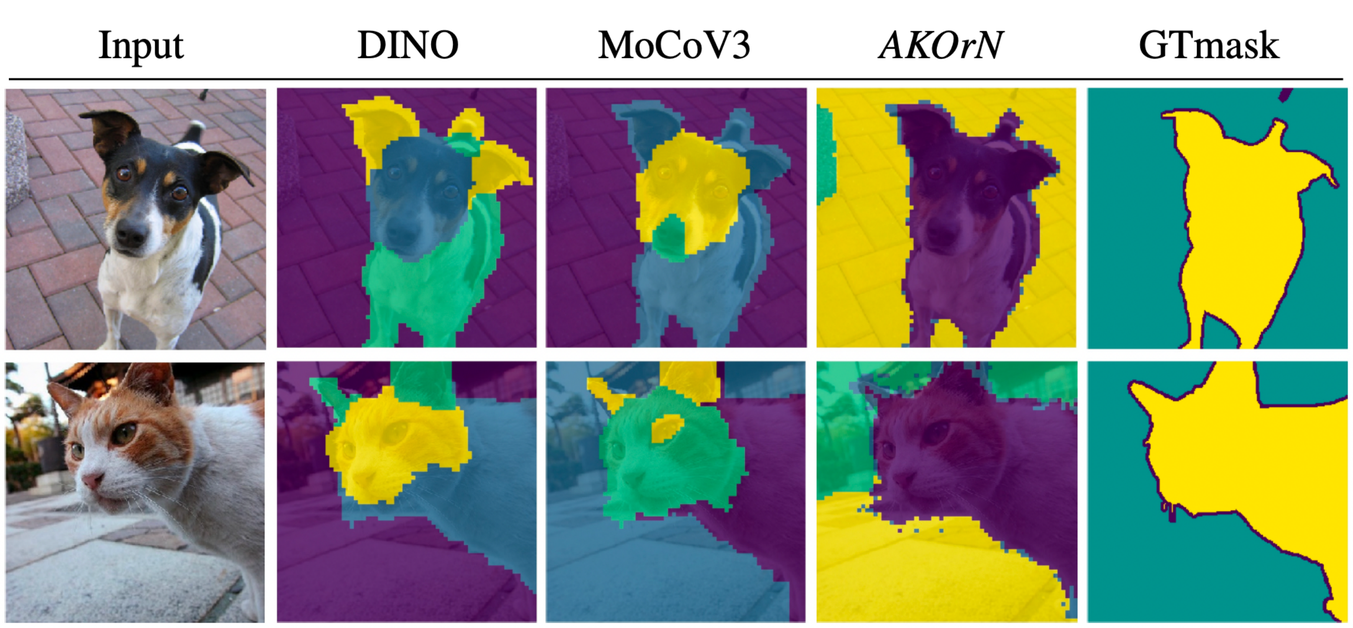
スロットベースモデル

Model	CLEVRTex FG-ARI	CLEVRTex MBO	OOD FG-ARI	OOD MBO	CAMO FG-ARI	CAMO MBO
ViT ($L = 8, T = 1$)	46.4±0.6	25.1±0.7	44.1±0.5	27.2±0.5	32.5±0.6	16.1±1.1
SLATE (Singh et al., 2022)	44.2±NA	50.9±NA	-	-	-	-
*Slot-Attetion (Locatello et al., 2020)	62.4±2.3	-	58.5±1.9	-	57.5±1.0	-
Slot-diffusion (Wu et al., 2023)	69.7±NA	61.9±NA	-	-	-	-
Slot-diffusion+BO (Wu et al., 2023)	78.5±NA	68.7±NA	-	-	-	-
*DTI (Monnier et al., 2021)	79.9±1.4	-	73.7±1.0	-	72.9±1.9	-
*I-SA (Chang et al., 2022)	79.0±3.9	-	83.7±0.9	-	57.2±13.3	-
BO-SA (Jia et al., 2023)	80.5±2.5	-	86.5±0.2	-	63.7±6.1	-
ISA-TS (Biza et al., 2023)	92.9±0.4	-	84.4±0.8	-	86.2±0.8	-
AKOrN$^{\text{attn}}$	88.5±0.9	59.7±0.9	87.7±0.3	60.8±0.6	77.0±0.5	53.4±0.7

Scale to natural images

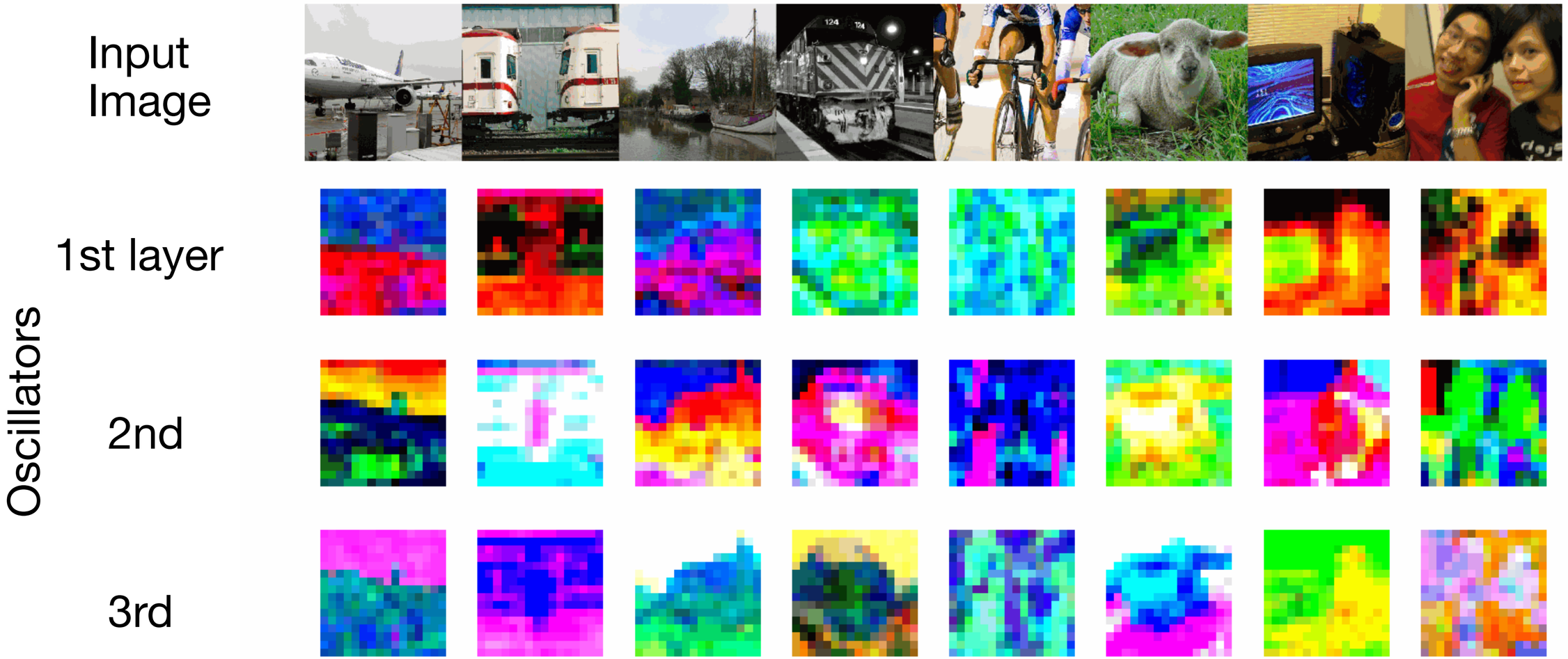
- AKOrN works well beyond synthetic data





Visualization of the oscillators

Shallow layers learn high-frequency and local features, while deep layers generate low-frequency waves.



Sudoku Solving

as a combinatorial problem

- Task: fill blanks so that vertical, horizontal lines, and each 3x3 sub-block have all digits [1-9]
  - Models are trained to classify the digit on each blank square.

- A model has to learn both the rule and exploration strategy

9		1	5		3		6	
3	6		2				1	8
2		7		4	6			9
		4		7	2		5	
1		9	3			8	4	
	7		8			9		
6	5							4
4		8			9	3	7	
				5	1	2		

Evaluation

Board accuracy (perfect match to the ground truth)

Two test sets used for evaluation

- ID set: dist. of #(given_digits) is same as the training set
- OOD set: fewer digits than the training set.

ID set

9		1	5		3		6	
3	6		2				1	8
2		7		4	6			9
		4		7	2		5	
1		9	3			8	4	
	7		8			9		
6	5							4
4		8			9	3	7	
				5	1	2		

31-42 given digits

OOD set

				5			1	
	6		9			4		3
		7			2			
5	8						9	
			7			6		
			3					
1				9		2		
9								
								7

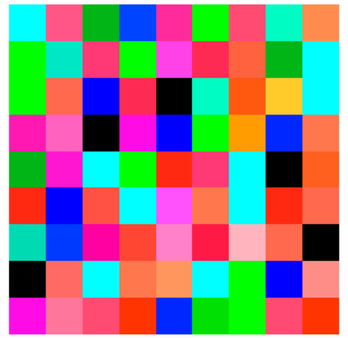
17-34 given digits

Results

- Achieved almost **100%** accuracy on ID boards
- 18% acc on OOD boards, but it can be bump up to **~90%** by increasing test-time compute. Two modes are available in AKOrN
  - 1. Extending Kuramoto updates
  - 2. Energy-based voting.

ID set

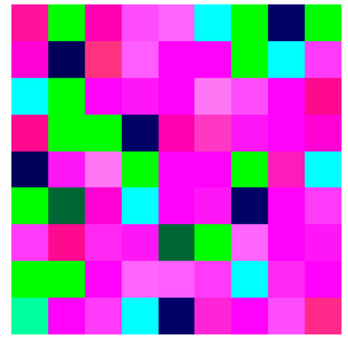
9		1	5		3		6	
3	6		2				1	8
2		7		4	6			9
		4		7	2		5	
1		9	3			8	4	
	7		8			9		
6	5							4
4		8			9	3	7	
				5	1	2		



9	5	1	5	3	3	9	6	1
3	6	5	2	5	1	4	1	8
2	7	7	3	4	6	2	8	9
7	3	4	3	7	2	1	5	5
1	5	9	3	3	2	8	4	6
6	7	9	8	5	2	9	7	6
6	5	1	6	4	1	5	8	4
4	2	8	7	6	9	3	7	5
3	6	1	4	5	1	2	4	2

OOD set

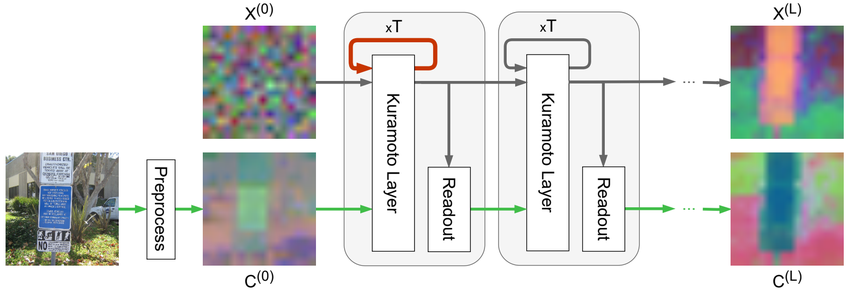
	9				1	8	4	6
	4					5	3	
1	5							
	6	9	4					
4			5			9		3
8	7		3			4		
				7	6			
5	8					1		
2			1	4				



1	9	3	2	2	1	8	4	6
2	4	6	1	5	5	5	3	1
1	5	2	3	7	1	2	6	6
4	6	9	4	7	3	3	7	5
4	9	2	5	8	5	9	5	3
8	7	3	3	2	5	4	4	2
6	5	5	8	7	6	8	4	2
5	8	4	4	6	2	1	3	4
2	4	1	1	4	9	3	2	4

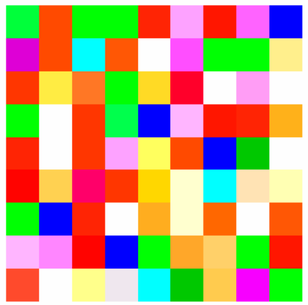
Test-time inference

1. Extending the Kuramoto updates



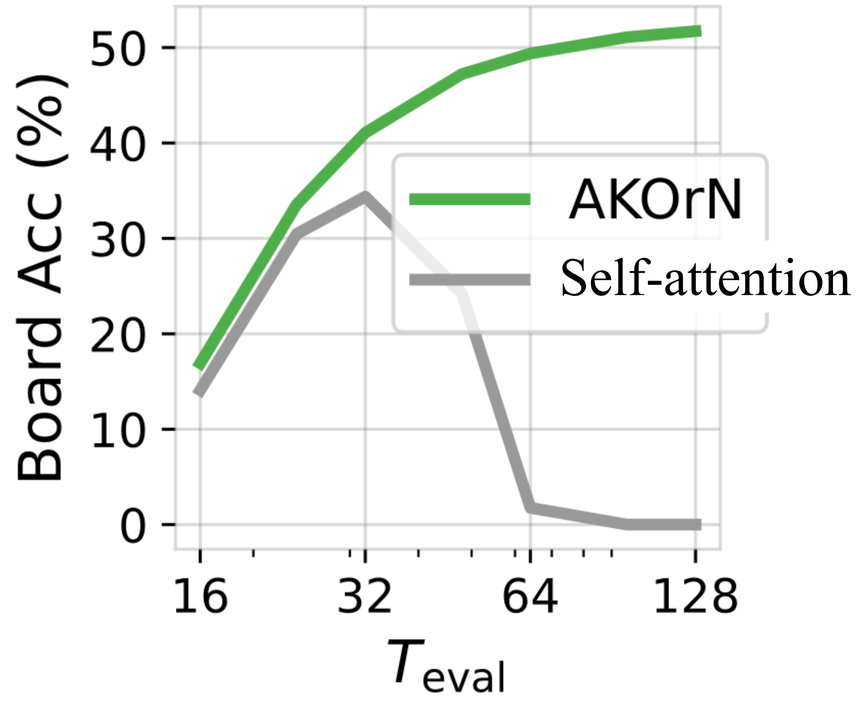
- Increase the number of updates in the AKOrN layer.
- **18%->52%** board accuracy
- The conventional self-attention collapse past 32 step.

2		1	9					4
		3				6	9	
			5					
1			2	4				
						4	7	
						3		
9	4							
			4	6			1	
				3	7			6



t=16

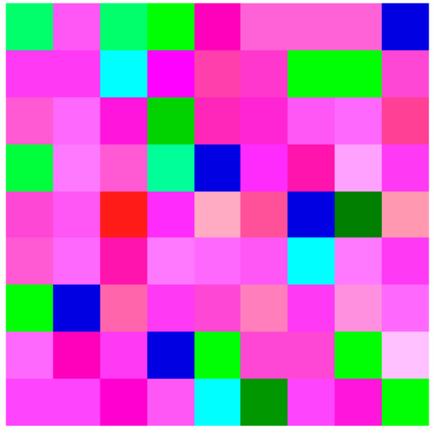
2	6	1	9	8	3	7	5	4
4	5	3	7	1	1	6	9	2
6	9	9	5	2	4	1	8	3
1	3	7	2	4	8	5	6	9
6	2	5	3	9	5	4	7	1
7	9	4	6	9	8	3	2	1
9	4	6	1	5	2	8	3	7
3	7	8	4	6	9	2	1	5
5	1	2	8	3	7	9	4	6



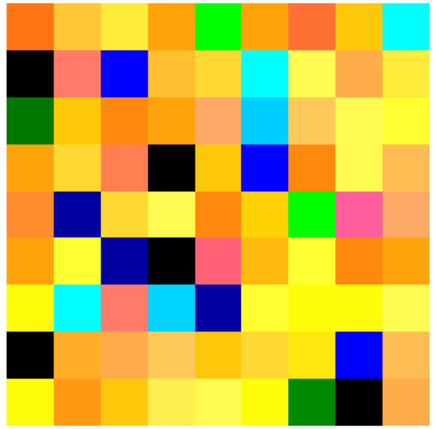
- (T_{train} is set to 16 during training)

2		1	9					4
		3				6	9	
			5					
1			2	4				
						4	7	
						3		
9	4							
			4	6			1	
				3	7			6

7								3
			5	1				
9	3		8					
		2		4	7			
	1	7	2			5	6	
					1	8		2
5	8		1	6				
4	7						2	
		6	7	3		9		



t=0



There still many boards not solved with test-time extension of Kuramoto steps

2	4	1	9	6	9	3	4	4
4	9	3	8	2	2	6	9	1
7	5	4	5	2	3	3	9	5
1	3	7	2	4	4	5	4	3
4	8	3	5	1	4	4	7	2
1	4	4	8	9	5	3	7	3
9	4	2	6	4	3	7	5	8
9	6	2	4	6	8	3	1	1
1	5	9	8	3	7	3	7	6

8	2	8	3	3	8	9	9	1
5	1	9	4	7	1	3	3	3
4	3	3	2	9	2	8	3	6
2	8	1	5	9	9	2	2	8
9	8	3	1	2	7	3	6	6
7	5	6	7	3	3	9	1	5
5	1	3	2	6	7	9	4	2
7	5	9	4	9	1	3	9	2
1	4	3	5	4	9	4	5	4

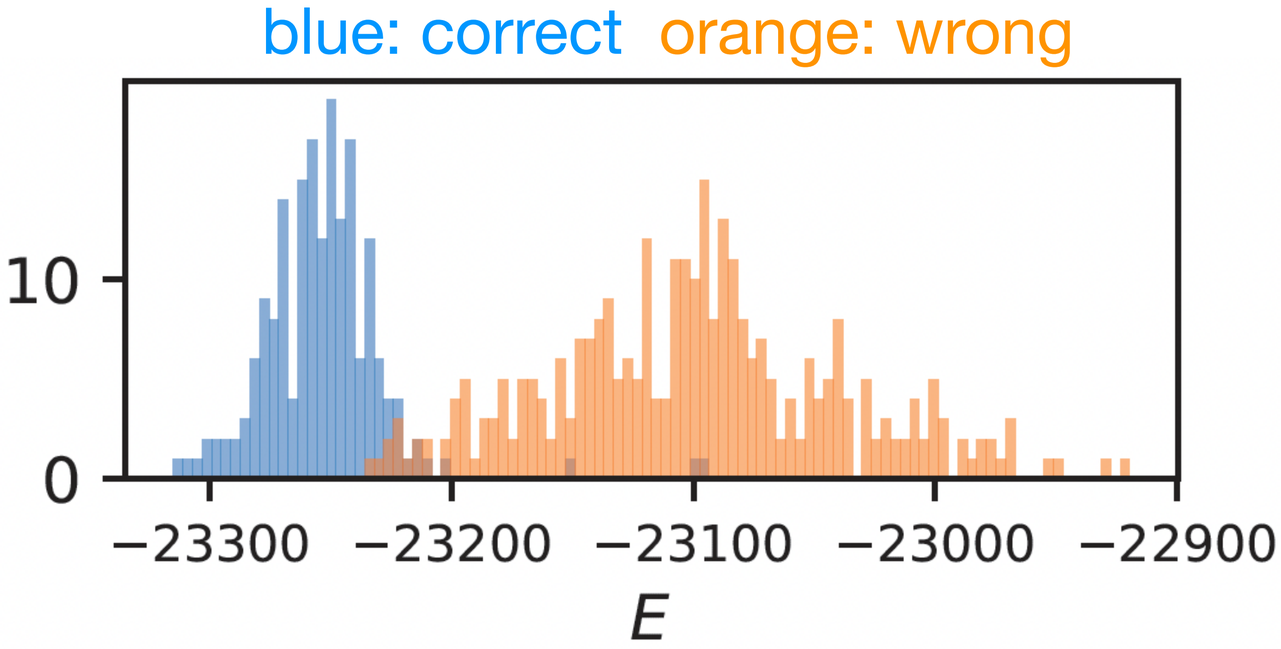
2. Energy-based voting

The model is aware of the prediction correctness

$$E = -\frac{1}{2}\sum_{i,j}\mathbf{x}_i^{\mathrm{T}}\mathbf{J}_{ij}\mathbf{x}_j - \sum_i \mathbf{c}_i^{\mathrm{T}}\mathbf{x}_i$$

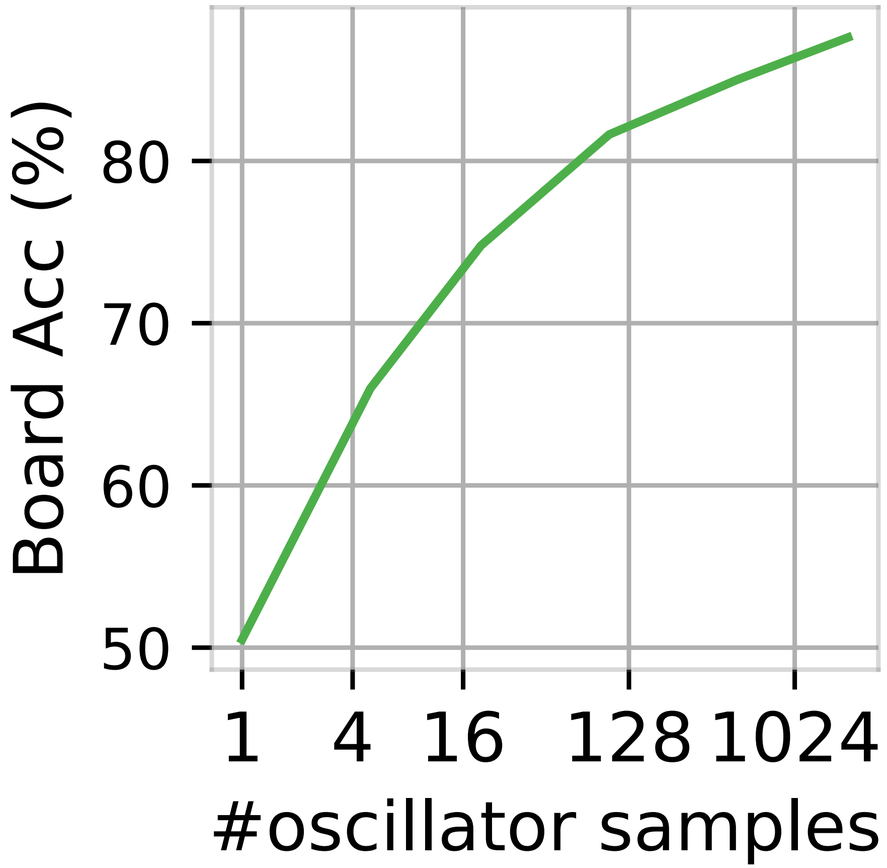
- エネルギー低い状態->予測の正解率が高い
  - **AKOrNが暗にエネルギーベースモデルを学習している**

Histogram of energy values of different oscillators



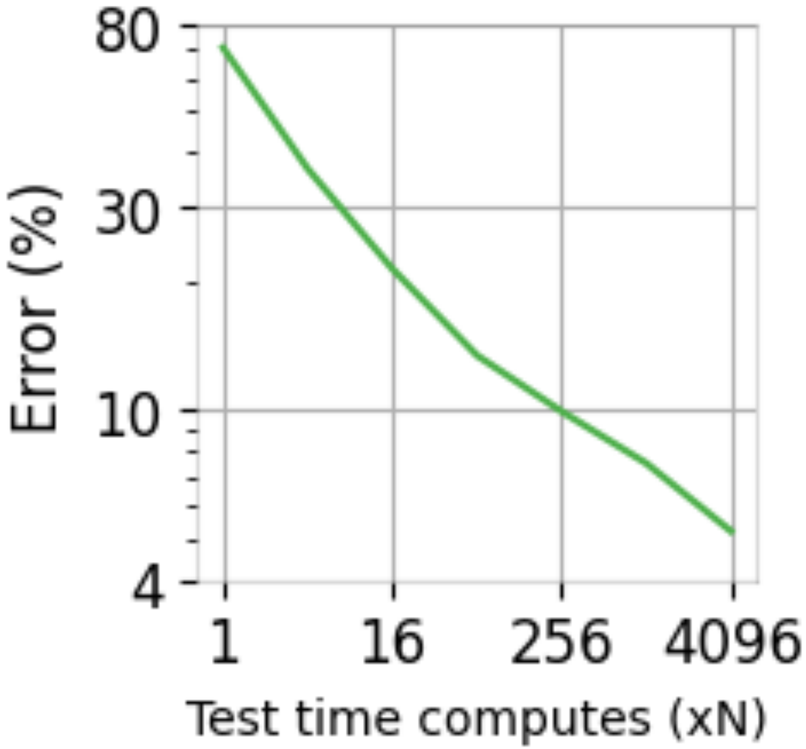
2. Energy-based voting

- Energy-based vote (E-vote):
  Sample multiple initial oscillators and choose **the lowest-energy** oscillators as model's prediction
- E-vote is found to be quite effective: the accuracy goes from 52 to **90**



Scaling on test-time-compute

- Near log-log linear scaling
  - Error →0 when Computes→∞
- But, the digit placement strategy seems not good yet.



Overall performance

- AKOrN gets best board accuracy on both scenarios.
  - Better than diffusion, transformers, and diffusion+energy-based models (IRED).

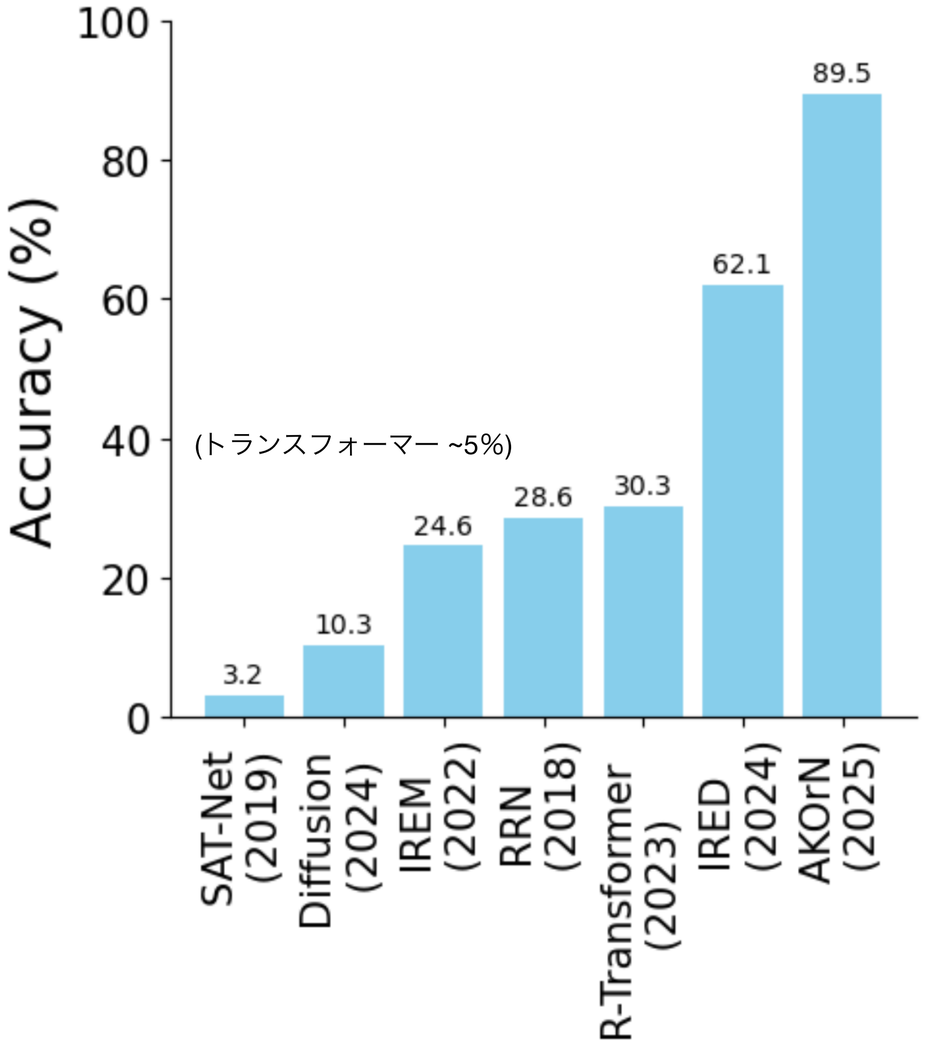
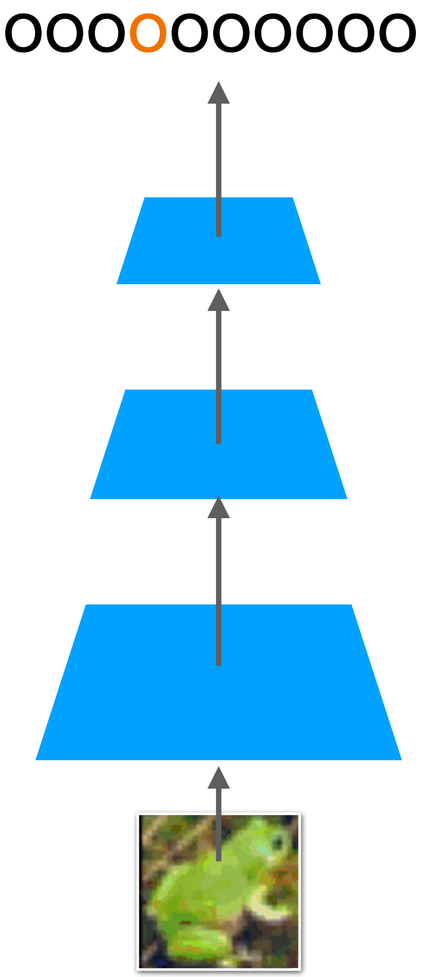


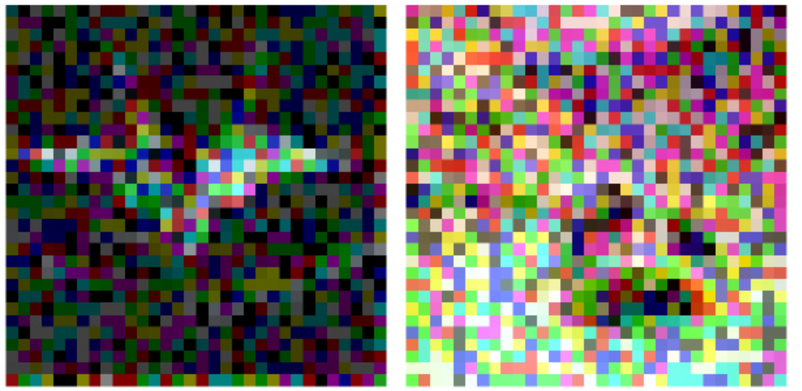
Image classification

- Trained the AKOrN model on CIFAR10 datasets
  - We stack 3 AKOrN iterative layers
- Evaluate the trained models on clean, adversarial, and naturally corrupted CIFAR10 datasets.

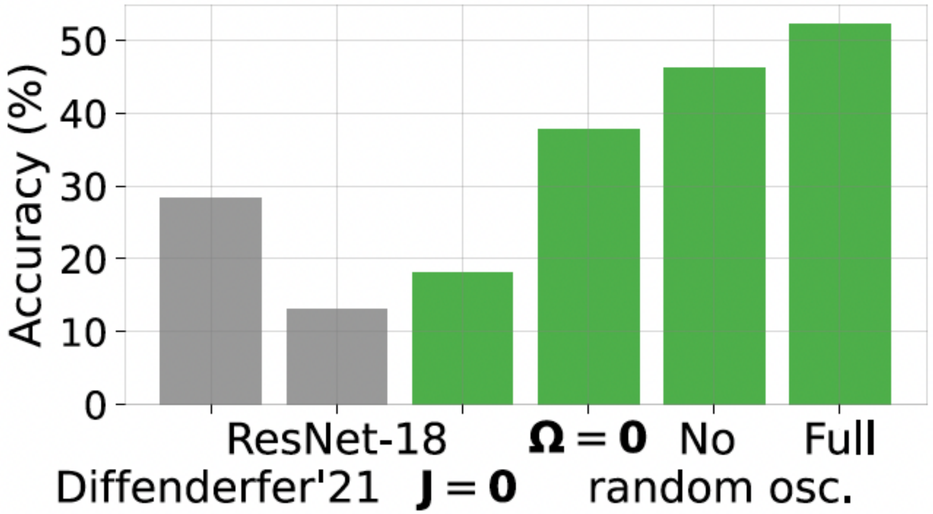


Robustness to strong random noises

- We find that the AKOrN models are highly robust to the random noises.

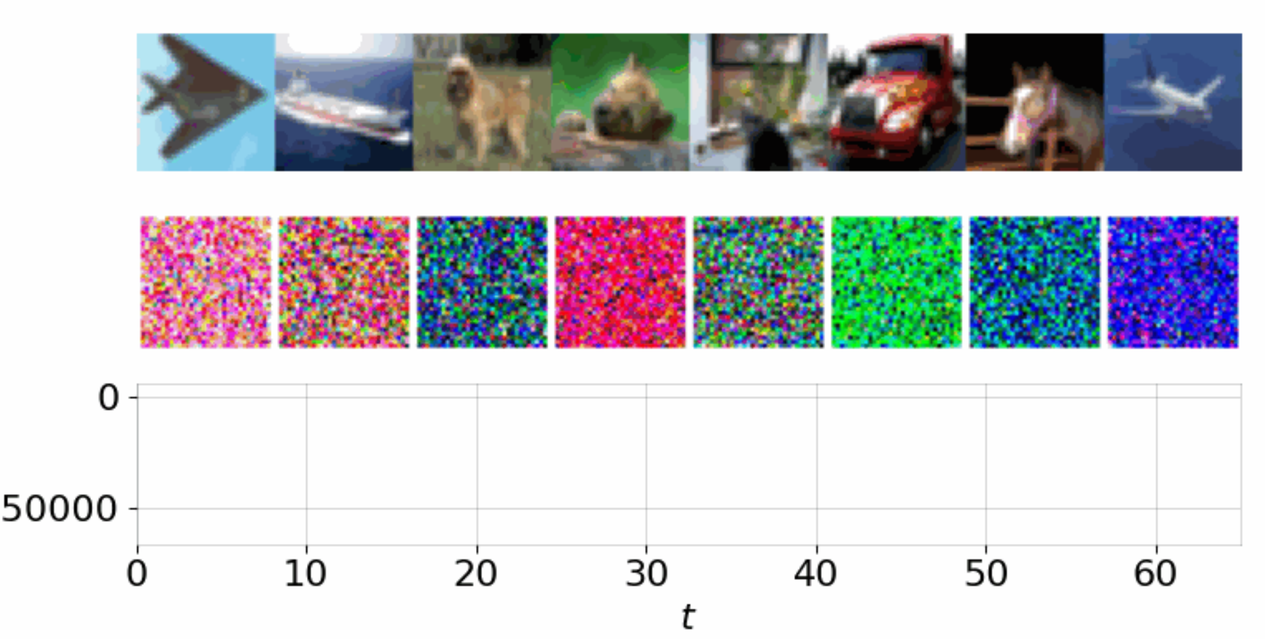


- The use of natural frequency improves the robustness



Rationale for robustness?

1st layers' oscillators highly fluctuate



Gaussian noise + ResNet does not get the same robustness property as AKOrN

Model	↑ Accuracy Clean	Adv	CC	↓ ECE CC
ResNet ($\sigma = 0.2$)	85.2	22.3	75.1	2.3
ResNet ($\sigma = 0.225$)	83.9	25.5	73.6	2.6
AKOrN ($N = 2$)	84.6	**64.9**	**78.3**	**1.8**

Table 8: Robustness comparison with ResNets trained to resist Gaussian noises. σ indicates the standard deviation of the noise added during training.

AKOrN's output is well-calibrated

- Heavily-regularized models tend to be overly conservative, while non-regularized models tend to be overconfident.

- AKOrN achieves almost perfect correlation between confidence and accuracy
  - without any explicit regularization

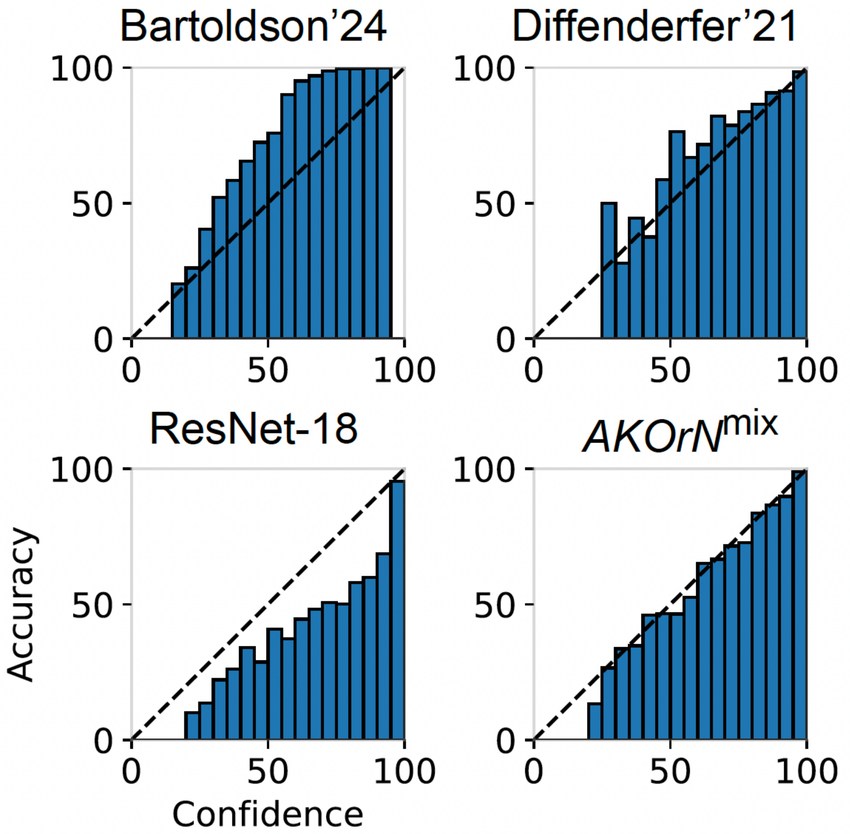


Figure 9: Confidence vs Accuracy plots on CIFAR10 with common corruptions.

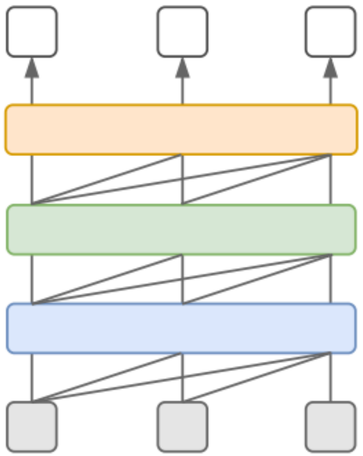
Fast-Slow Recurrence for Train-Short, Test-Long Generalization

Shota Takashiro[1], Masanori Koyama[1], Takeru Miyato[2], Yusuke Iwasawa[1], Yutaka Matsuo[1], Kohei Hayashi[1]

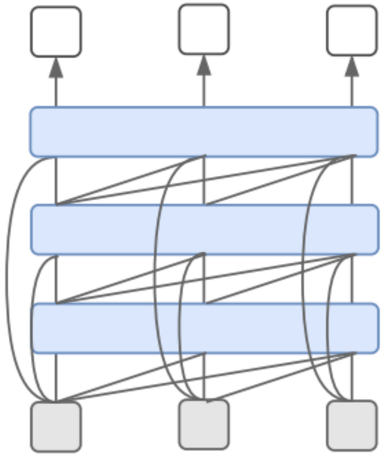
1: University of Tokyo, 2: University of Tübingen,

時間方向の再帰と汎化について

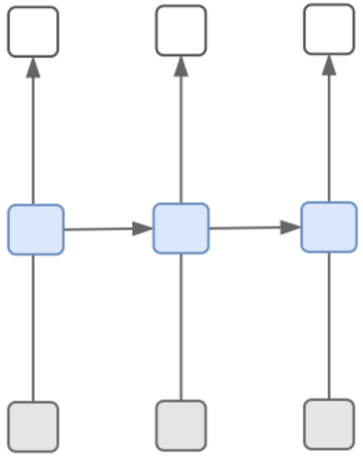
- 再帰の回数を時間方向に増やすこと（longer thinking in time direction) がどう汎化に影響するか?
- 潜在状態は各観測ごとにリセットされるのではなく、複数回の更新によって反復的に洗練される -> ストリーム全体にわたってcoherentな状態



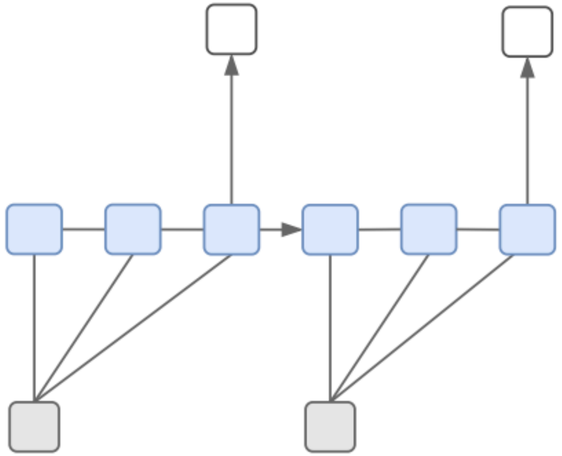
Transformer



Iterative Transformer



RNN/SSM



Fast-slow recurrent models (FSRM, ours)

実験

- Dyck strings, Ego centric maze, and several grid reinforcement learning tasks

input ( ) [{ } ( ) < ( ) >] [. . .

target) *] }])] >) >] *] . . .

(a) ID (random pattern)

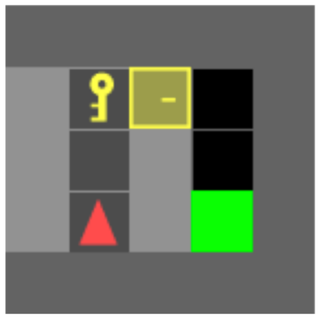
input [( ) ( ) ( ) ( ) ( ) ( ) . . .

target])])])])])] . . .

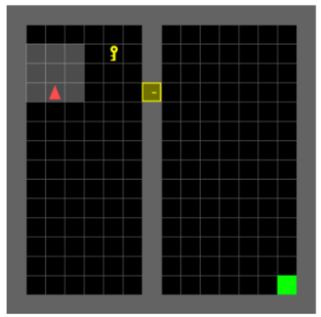
(b) OOD (1-regular run)

(a) ID (19 × 19)

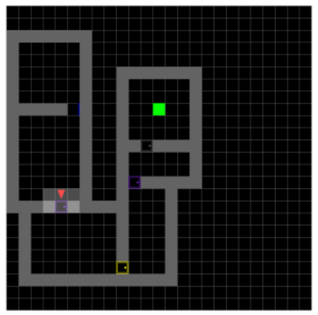
(b) OOD (39 × 39)



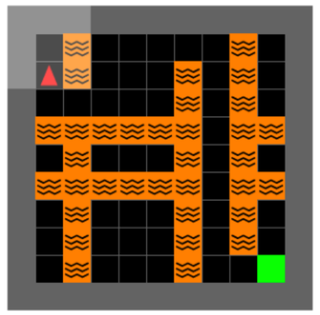
(a) DoorKey (ID)



(b) DoorKey (OOD)

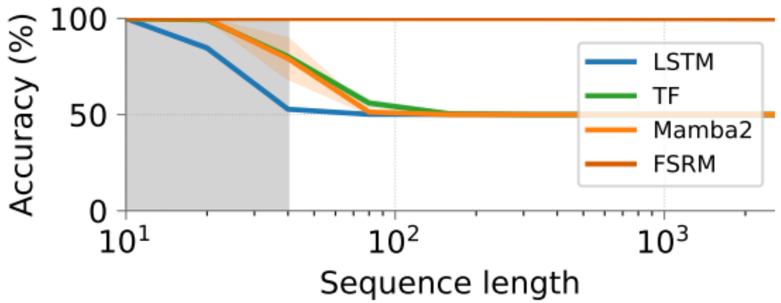


(c) MultiRoom (OOD)

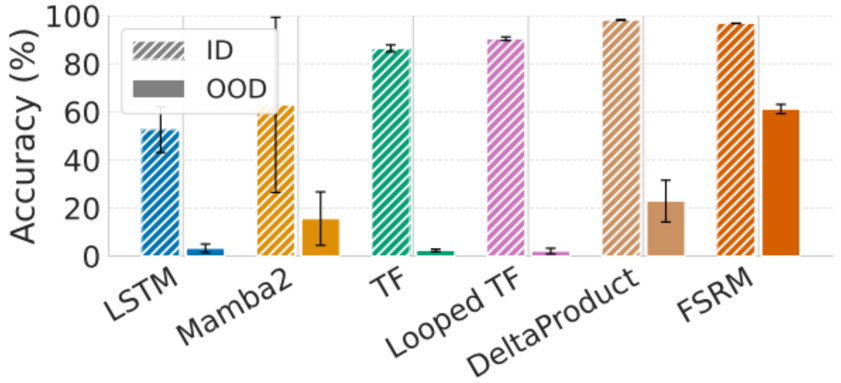


(d) LavaCrossing (OOD)

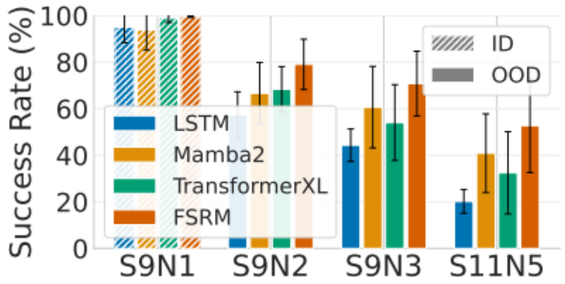
トランスフォーマー等との比較



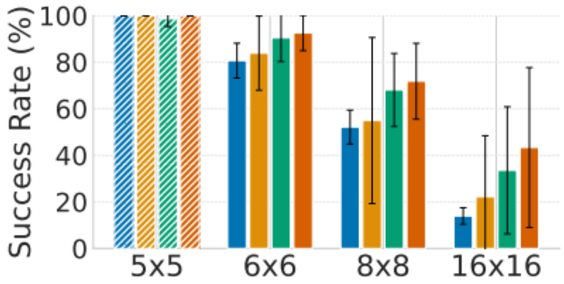
(a) Dyck-$(30, 5)$ OOD accuracy.



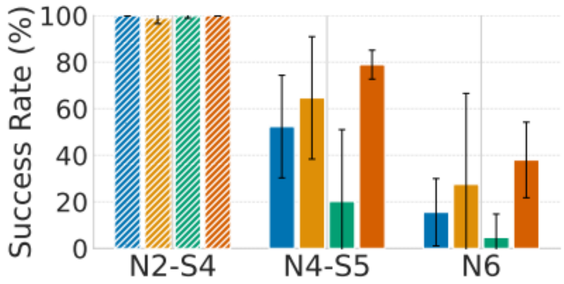
(b) Local-Maze ID and OOD accuracy



(c) LavaCrossing



(d) DoorKey



(e) MultiRoom

ループを増やすvs層を増やす

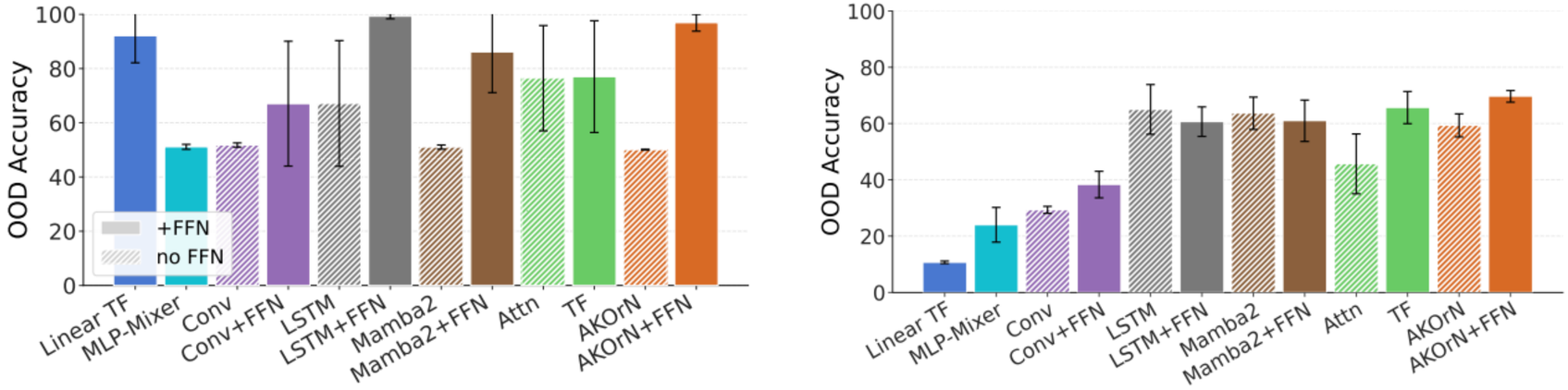
- ループを増やすOODの性能は上がるが、層を増やすことでは逆に性能は下がる。

ループを増やす

層を増やす

Model	Params/FLOPs	OOD Acc.
Shared loop: train-time T sweep		
Base $(1 \otimes 1)$	$1\times/1\times$	0.306 ± 0.135
Loop $(1 \otimes 2)$	$1\times/2\times$	0.431 ± 0.186
Loop $(1 \otimes 4)$	$1\times/4\times$	0.595 ± 0.036
Loop $(1 \otimes 5)$	$1\times/5\times$	0.612 ± 0.019
Loop $(1 \otimes 8)$	$1\times/8\times$	$\mathbf{0.673 \pm 0.006}$
Loop $(1 \otimes 16)$	$1\times/16\times$	0.673 ± 0.032
Weight-sharing control		
Stack $(5 \otimes 1)$	$2\times/5\times$	0.509 ± 0.098
Stack $(10 \otimes 1)$	$3\times/10\times$	0.445 ± 0.164

再帰コアの比較



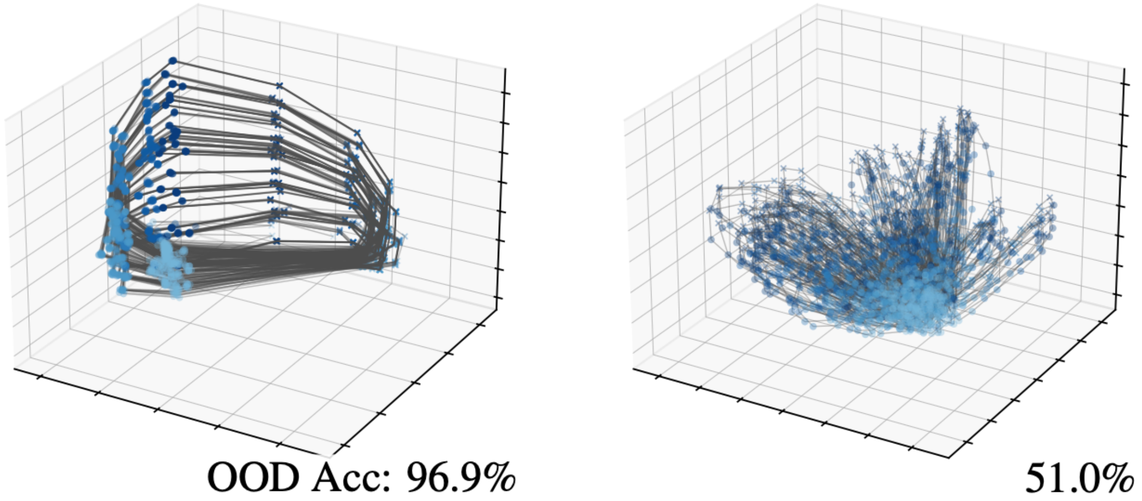
(a) Dyck-$(30, 5)$ OOD accuracy.

(b) Local-Maze OOD accuracy (39×39).

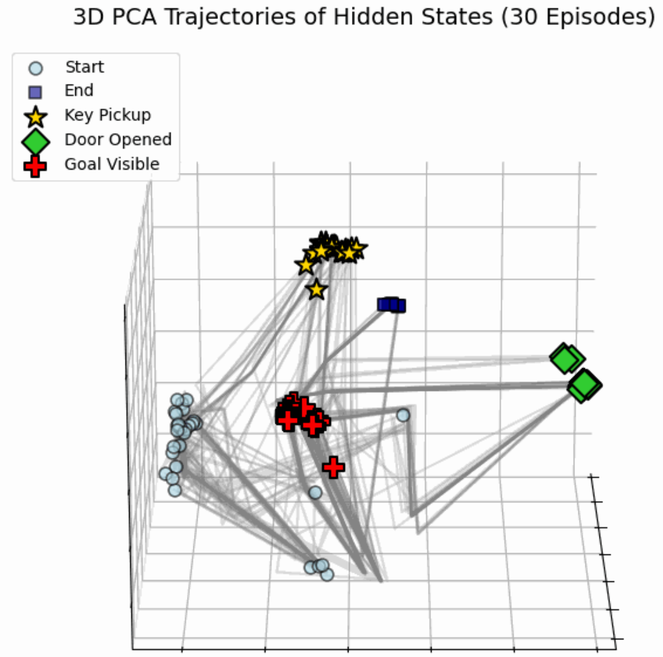
Figure 5: Core ablations on (a) Dyck-$(30, 5)$ at lengths 1,280–2,560 and (b) Local-Maze under OOD evaluation. Solid bars include the native or augmented FFN block; hatched bars ablate it. ID results are in C.1.4 and C.2.1.

内部状態のアトラクター的挙動

- より良く汎化するモデルほど、内部状態にイベント依存のclusterができる



(a) FSRM-AKOrN+FFN　　(b) FSRM-Mamba2



FSRM-AKOrN

内部状態のアトラクター的挙動

- AKOrN のEnergy関数をみてみると一種のfrustrationのような挙動が見られる

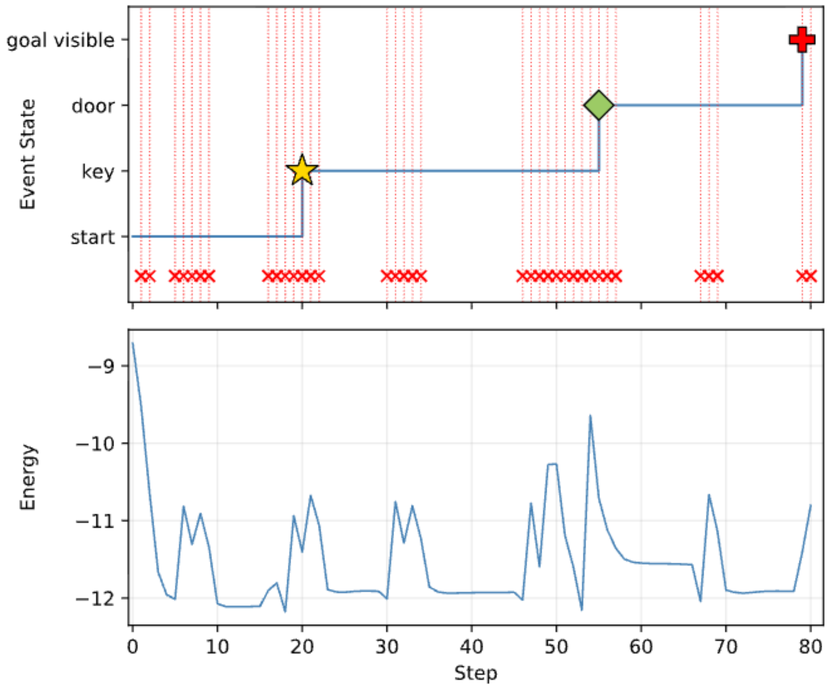


Figure 21: Energy traces along an episode in the DoorKey-16x16 (OOD) environment. **Top:** event history along the trajectory; red crosses mark changes in the agent's view. **Bottom:** energy trajectories.

Spontaneous symmetry breaking and Goldstone modes for deep information propagation

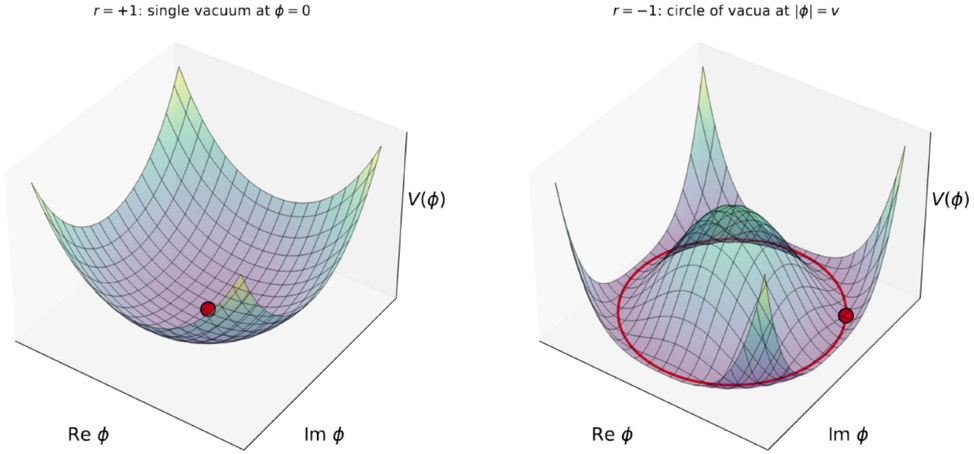
https://arxiv.org/abs/2605.14685

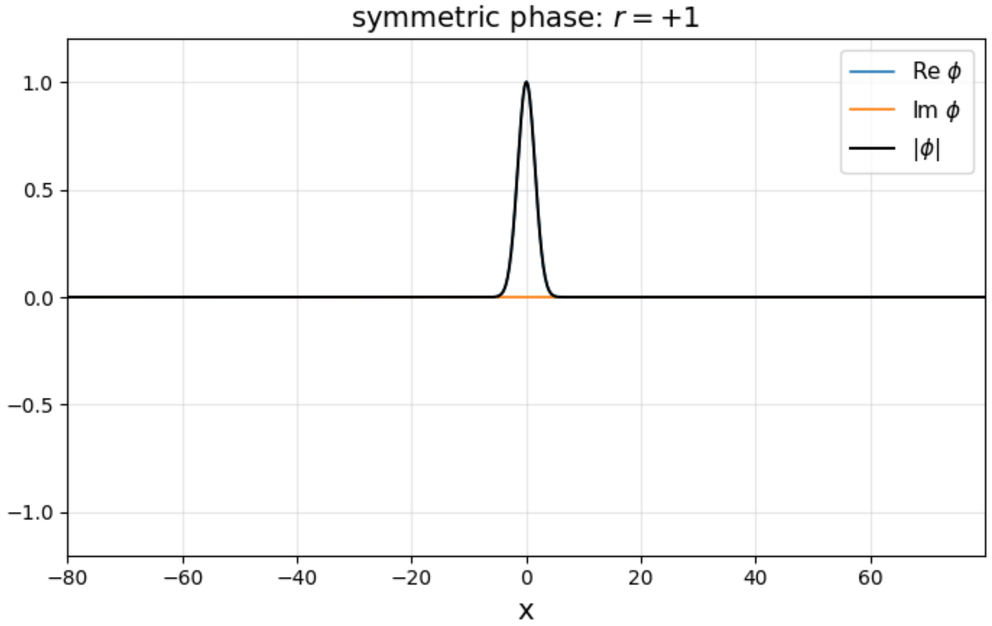
Nabil Iqbal[1,2], Andy Keller[3], Yue Song[4], Takeru Miyato[5], Max Welling[2]

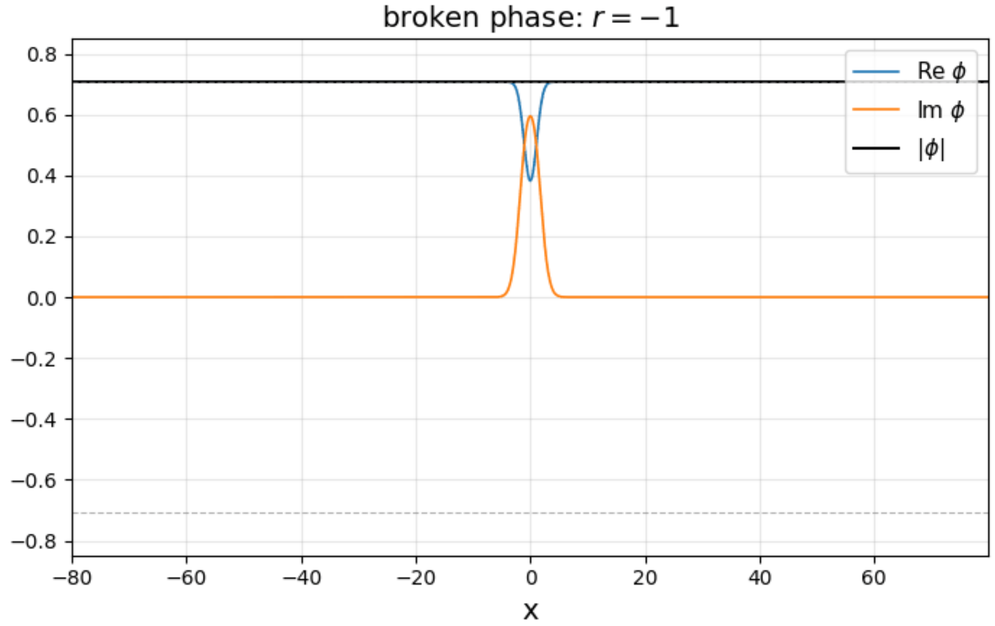
1:Durham University, 2:University of Amsterdam 3:Harvard University

4:Tsinghua University 5:University of Tübingen

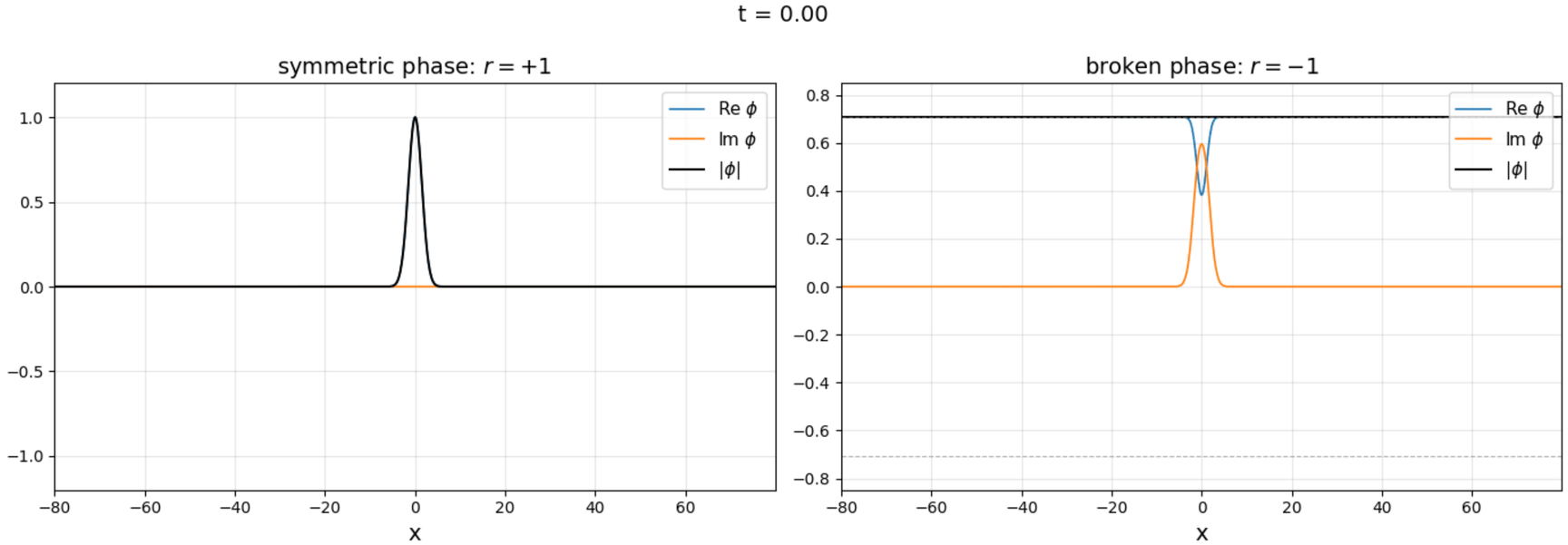
Goldstone modeと情報伝播







Goldstone modeと情報伝播



いろんな場面でgoldstone mode的な情報伝播は起こる

- 音波、磁場、光
- ではニューラルネットでは？

同変モデルによるSSB

- 同変モデルでは、**Goldstone mode 的な情報伝播**が可能になる
- **同変モデルは、通常モデルとは異なる相転移を示す**
  - 汎用モデルは、edge of chaos においてのみ学習できる
  - 一方同変モデルは、同変方向に沿った保護されたチャネルにより、臨界点を超えても学習できる
- 実験では、この Goldstone mode 的な伝播が長期記憶に有効であり、長系列モデリングにおける RNN の性能を改善することを示した

同変モデル (Equivariant models)

- 各層がU(1), もしくは O(k) 同変性をもつニューラルネットワーク

$$\rho_g f^l(\mathbf{x}^l) = f^l(\rho_g \mathbf{x}^l)$$

  - x: 入力ベクトル
  - f: 層の関数 (線形層+非線形活性化関数)

$$z_i^{l+1} = \sum_j W_{ij}^l \phi(z_j^l) \quad \phi(z) = \frac{\tanh(|z|)}{|z|} z$$

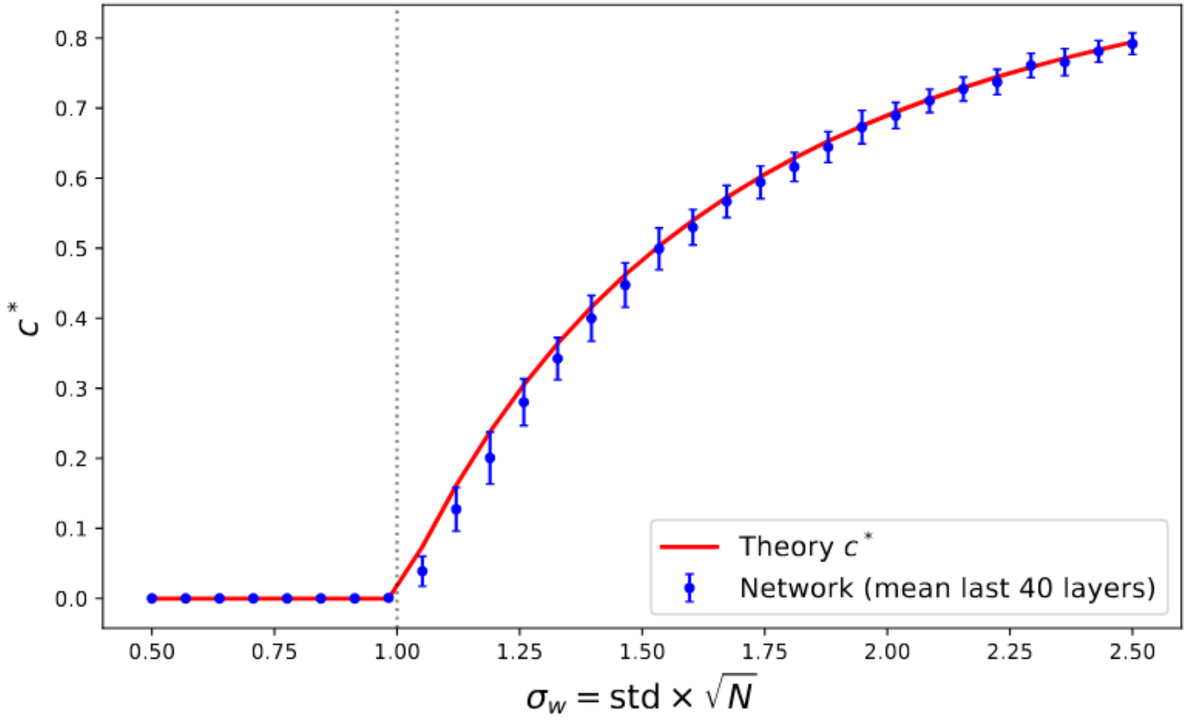
Phase of symmetry

- 秩序変数を以下のように定義

$$c^l \equiv \frac{1}{N}\mathbb{E}\left(\sum_j \phi(z_j^l)\phi^\dagger(z_j^l)\right)$$

各ニューロンノルムの平均
(期待値は重みのランダム性に対して)

Phase of symmetry



<-symmetry unbroken

symmetry broken->

重み要素の分散が**小さい**時
-> 対称性が壊れない(入力に対する回転に対して出力が不変)

重み要素の分散が**大きい**時
-> 対称性が壊れたフェーズに切り替わり、入力における回転が出力にnon-trivialに作用する

異なるデータ間の共分散について

$$C^l_{ab} \equiv \frac{1}{N}\mathbb{E}\left(\sum_j \phi(z^l_{j,a})\phi^\dagger(z^l_{j,b})\right) = \begin{pmatrix} c^l & \Delta^l e^{i\varphi^l} \\ \Delta^l e^{-i\varphi^l} & c^l \end{pmatrix}$$

位相

$$\boxed{\varphi^{l+1} = \varphi^l}$$

スケール

$$\boxed{\Delta^l \sim \exp\left(-\frac{l}{\xi_\Delta}\right)}$$

異なるデータ間の**位相は保存**され、データの相関は(比較的)ゆっくり減衰していく

通常NN vs 同変NN

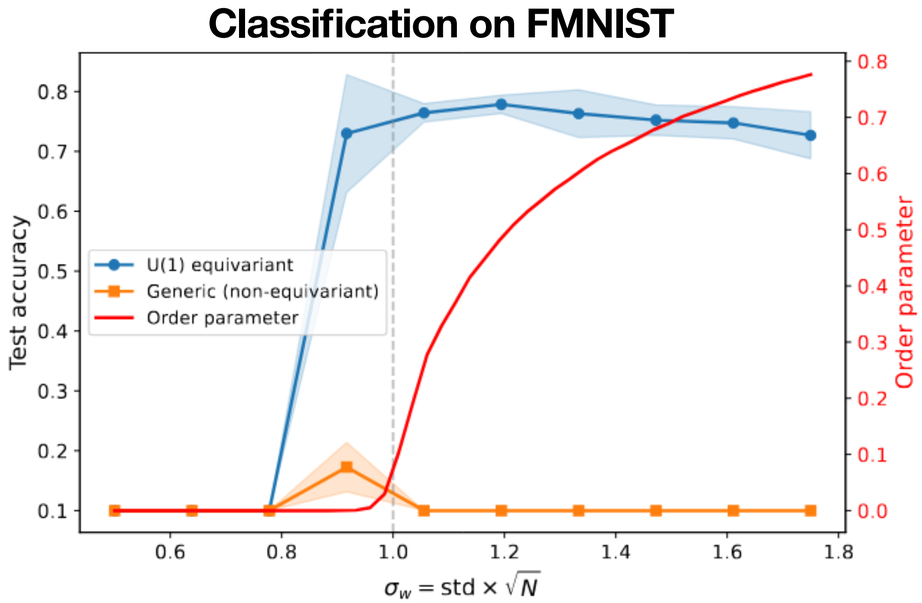
- 通常のモデル: カオスの縁でのみ学習可能
  - その他の相では層を増やすごとに異なるデータの相関が急速に小さくなる。
- 同変モデル : critical pointを超えても回転対称の方向は壊れない

Model (Theory)	**Before critical point**	**At critical point**	**After critical point**
Generic (DIP)	Ordered phase; inputs collapse to a fixed point, so **information is lost**.	Edge of chaos; long-range propagation is possible.	Chaotic phase; input correlations decorrelate, so **information is lost**.
Equivariant (SSB / Goldstone)	Unbroken phase; $c^{\ell} \to 0$, so **information is lost**.	Symmetry-breaking transition; $c^{\star}$ starts to become nonzero.	SSB phase; $c^{\star} > 0$, so **symmetry-direction information is protected**.

Table 1: Comparison between deep information propagation (DIP) and the equivariant Goldstone-mode mechanism.

critical point前後の通常/同変モデルの性能比較

- 通常のモデルはcritical point (カオスの縁)付近でのみ少しだけ学習
- U(1) Equivariant モデルはcritical point以降も学習可能

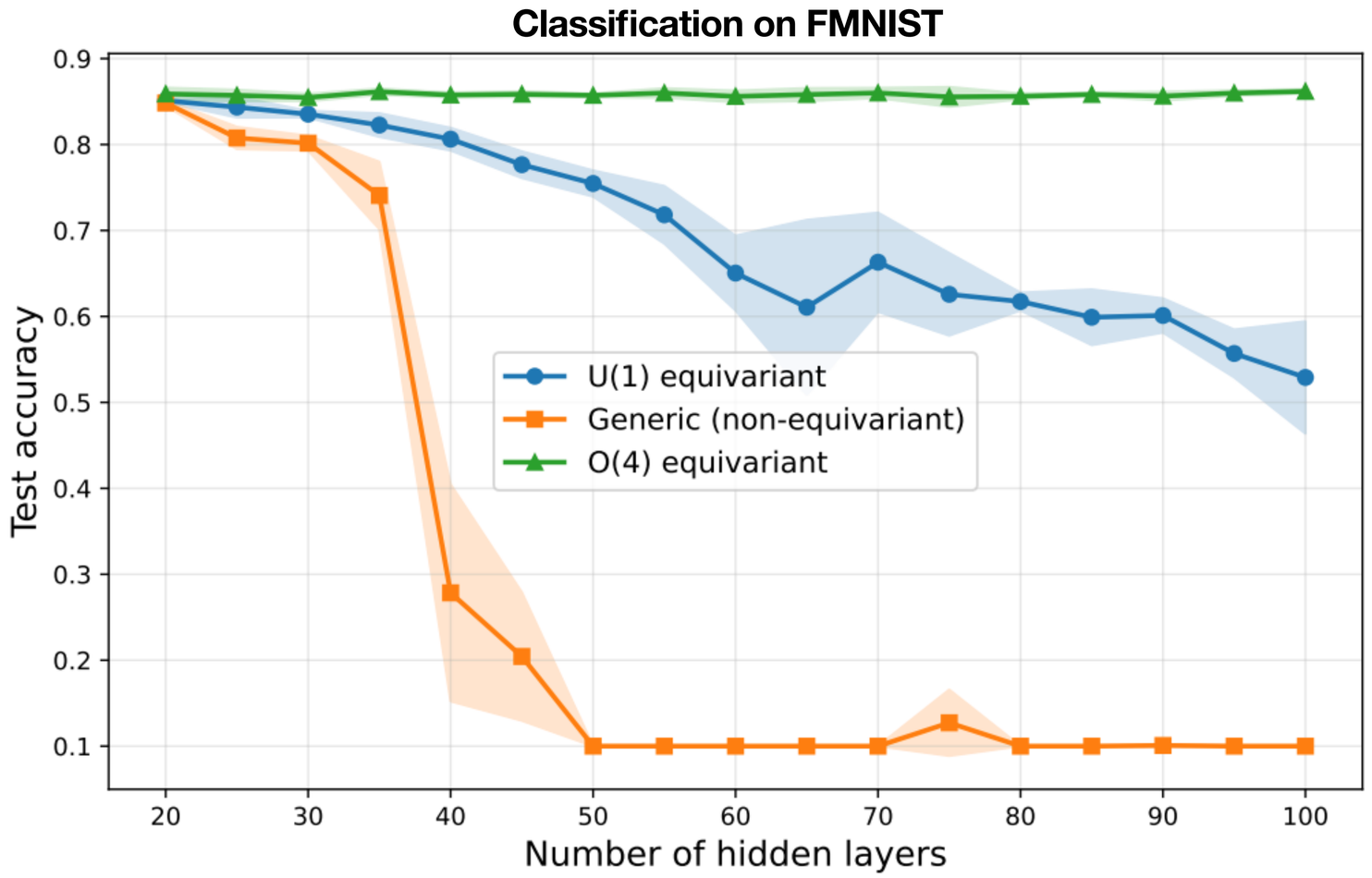


層を増やしたときの性能の比較

- 通常のニューラルネットワークでは、層を増やすにつれて性能が大きく低下

- 一方U(1)モデルは通常モデルに比べ深くしても性能低下が起きづらい

- O(4)モデルは層を増やしても性能が落ちない



どのように同変モデルが役に立つか？

- RNNを用いたcopy taskやsequential image classificationでの実験で同変モデルがより良い性能を達成

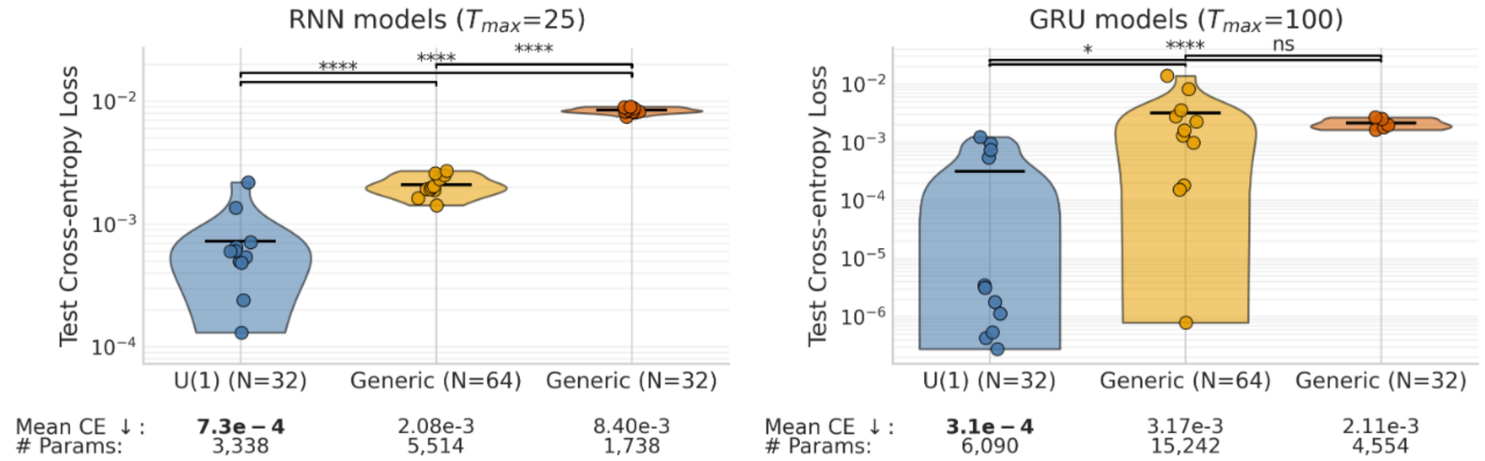


Figure 6: Comparison of RNNs and GRUs with $U(1)$ equivariant counterparts on copy tasks of max delay $T_{max} = 25$ and $T_{max} = 100$ respectively. We see that the $U(1)$ models significantly outperform the non-equivariant counterparts for both architectures, even when generic baselines have more real trainable params.

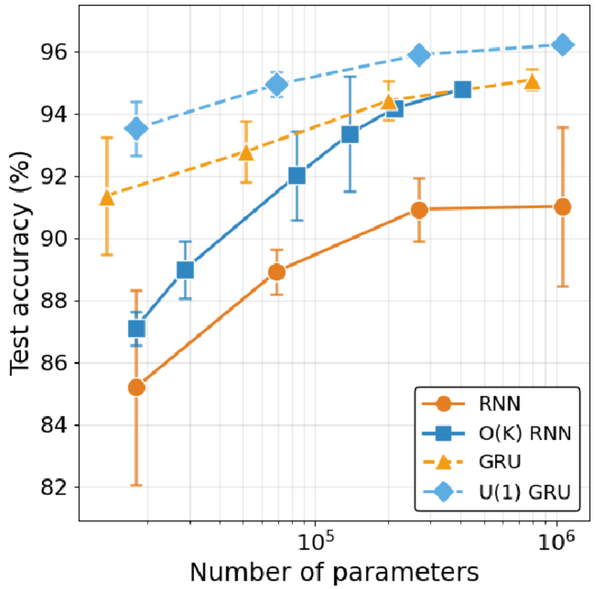
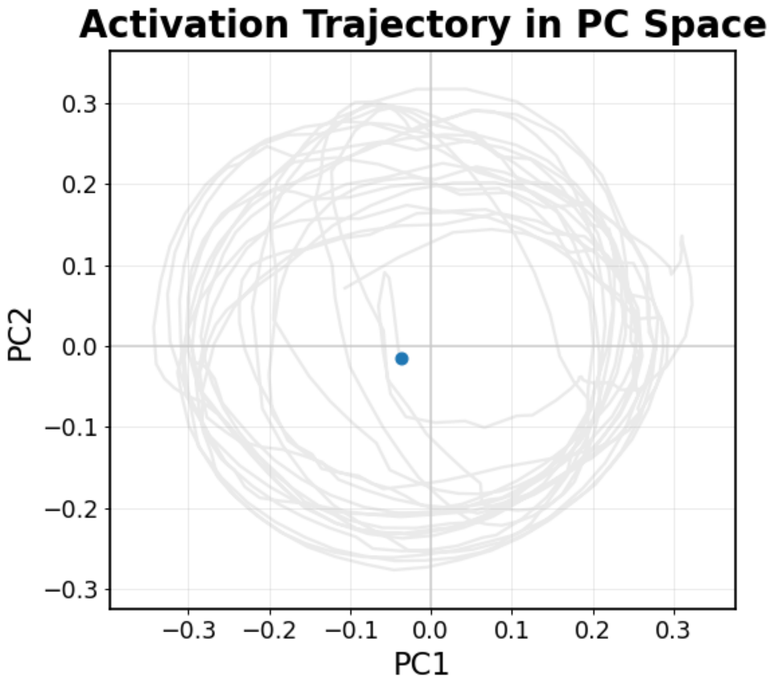
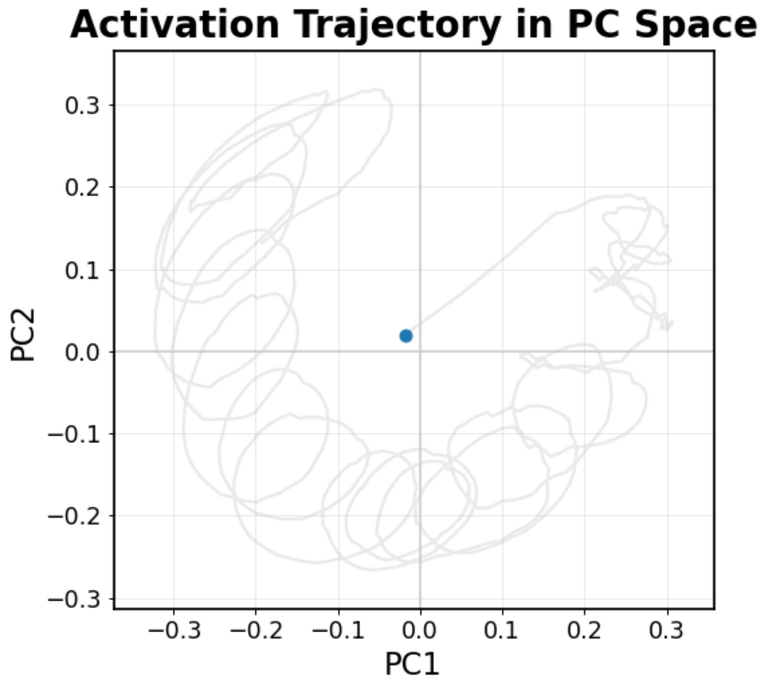


Figure 7: Test accuracy vs. # real trainable params on psMNIST. Equivariant RNNs and GRUs consistently outperform non-equivariant models at all parameter ranges.

ニューロンのダイナミクスの可視化

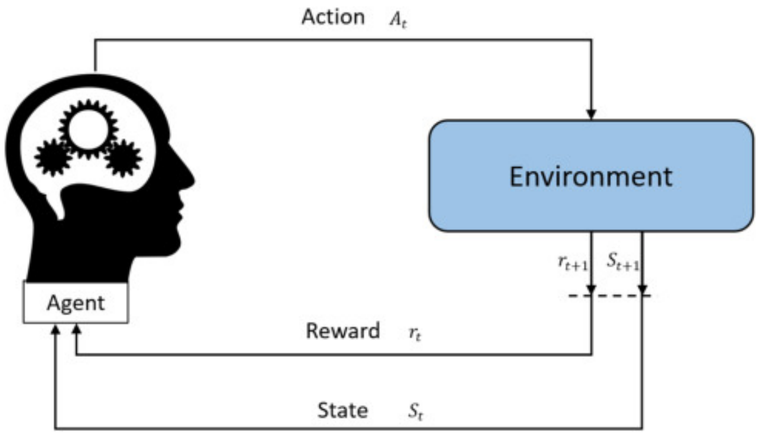
- PCA でRNNの潜在表現の可視化





再帰 = 現実世界に即した物理プロセス

- ルールの繰り返し適用でこの世界は生成されている
- ニューラルODE, SDE
- 脳のモデル
  - 入力を受け取りながら内部状態を継続的に更新
- 推論のアトラクター 的解釈
  - 状態が安定した構造へ向かう過程
- リカレントモデルのスケーリング則
- It’s still great time to join AI research from a physics background!



論文へのリンク

AKOrN: https://arxiv.org/abs/2410.13821
Fast-Slow Recurrence for Long-Horizon Sequential Modeling: https://arxiv.org/abs/2604.01577
Spontaneous symmetry breaking and Goldstone modes for deep information propagation https://arxiv.org/abs/2605.14685

その他参考文献

looped transformers:
- [2103.03206] Perceiver: General Perception with Iterative Attention
- [1807.03819] Universal Transformers
- [2301.13196] Looped Transformers as Programmable Computers
- [2502.05171] Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach
- [2512.19941] Block-Recurrent Dynamics in Vision Transformers
- [2605.12466] Solve the Loop: Attractor Models for Language and Reasoning
- [2605.20613] HRM-Text: Efficient Pretraining Beyond Scaling

“Time-recurrent” transformers:
- [2603.14360] M^{2}RNN: Non-Linear RNNs with Matrix-Valued States for Scalable Language Modeling
- [2604.21215] The Recurrent Transformer: Greater Effective Depth and Efficient Decoding

Investigating properties of sequential models:
- [2210.10749] Transformers Learn Shortcuts to Automata
- [2604.17121] The Topological Trouble With Transformers
- ON THE “INDUCTION BIAS” IN SEQUENCE MODELS