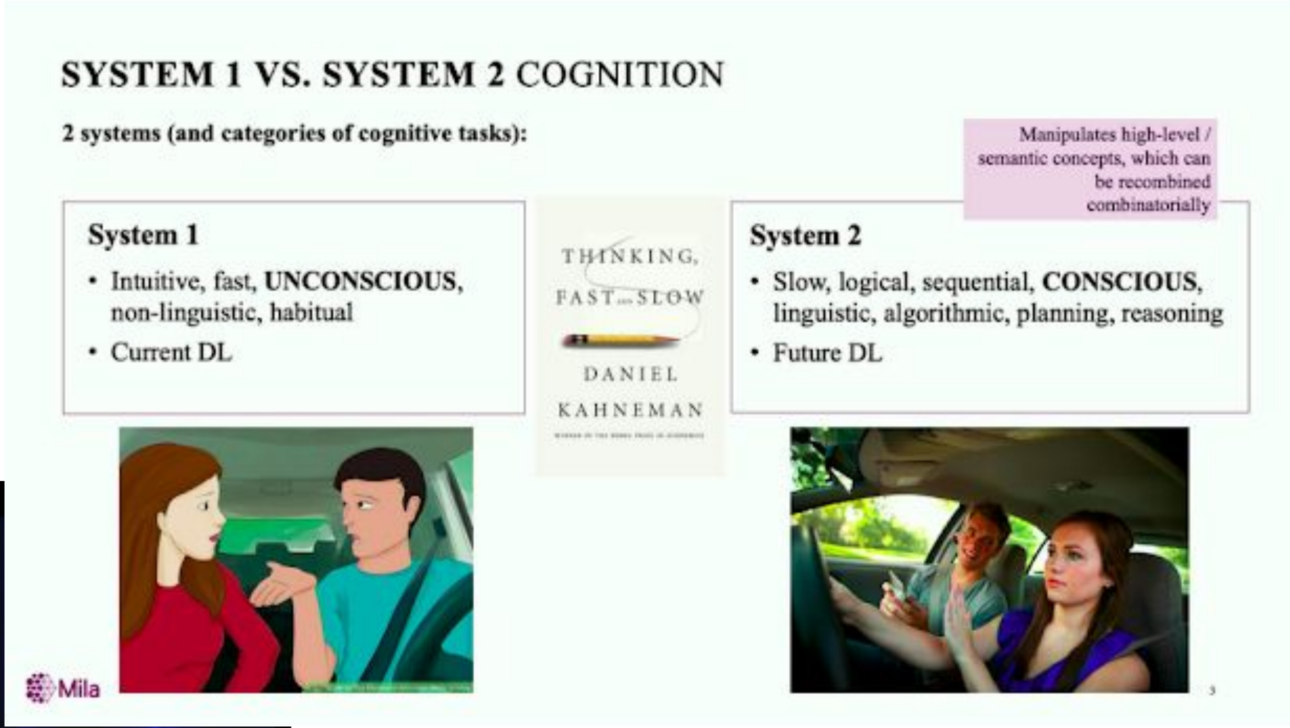
Recurrent Independent Mechanisms

https://arxiv.org/pdf/1909.10893.pdf

Flexible Learning Reading Group, 8th Jan

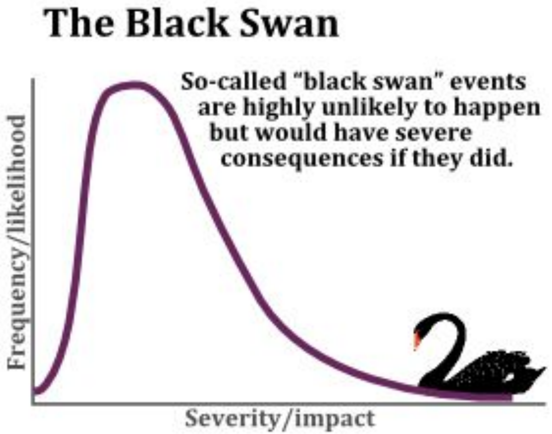
NeurIPS 2019



https://slideslive.com/38921750/from-system-1-deep-learning-to-system-2-deep-learning

Challenge: Handle Changes in Distribution

- Assumption of IID data is necessary for generalization



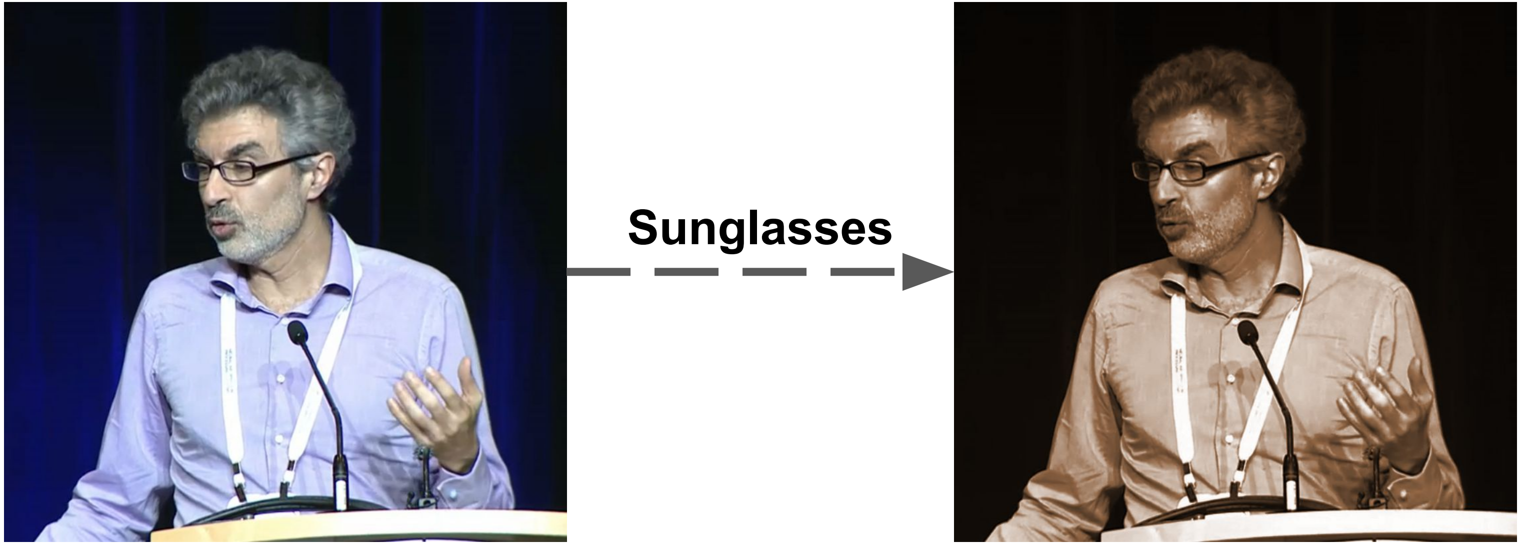
http://sites.utexas.edu/climatesecurity/2019/12/16/climate-change-and-black-swan-a-case-for-alarmism/

- But: most data is not IID,
  Agents face non-stationarities

- Compositionality helps systematic IID and OOD generalization
  - Dynamically recombine existing concepts

Changes are Consequences of Interventions

Builds on informationally ***independent mechanisms*** (Schölkopf et al. 2012)



Good representation of mechanisms ⇒ Only few bits/observations needed

SOTA struggles with fully independent mechanisms

Fully-connected layers are used over entire hidden state

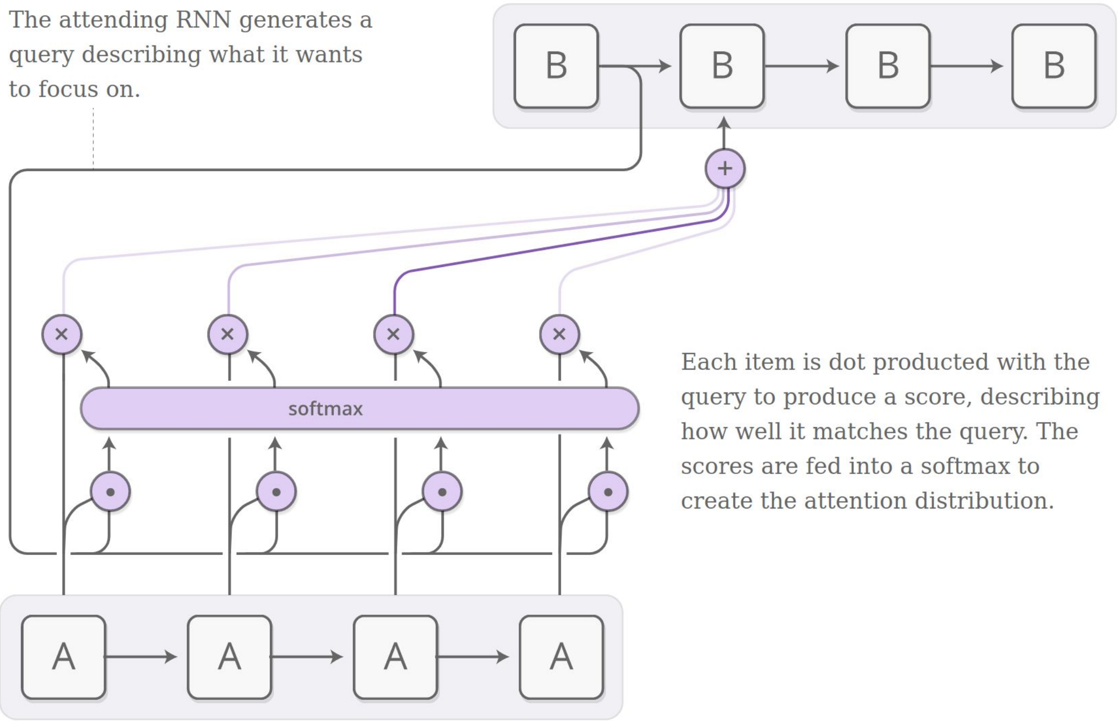
➔ Information separation between states only possible if most entries are zero

➔ No perfect modularity (just well enough for training data)

➔ Poor generalization to changing environment

Learning independent mechanisms that can flexibly be reused, composed and re-purposed is desirable!

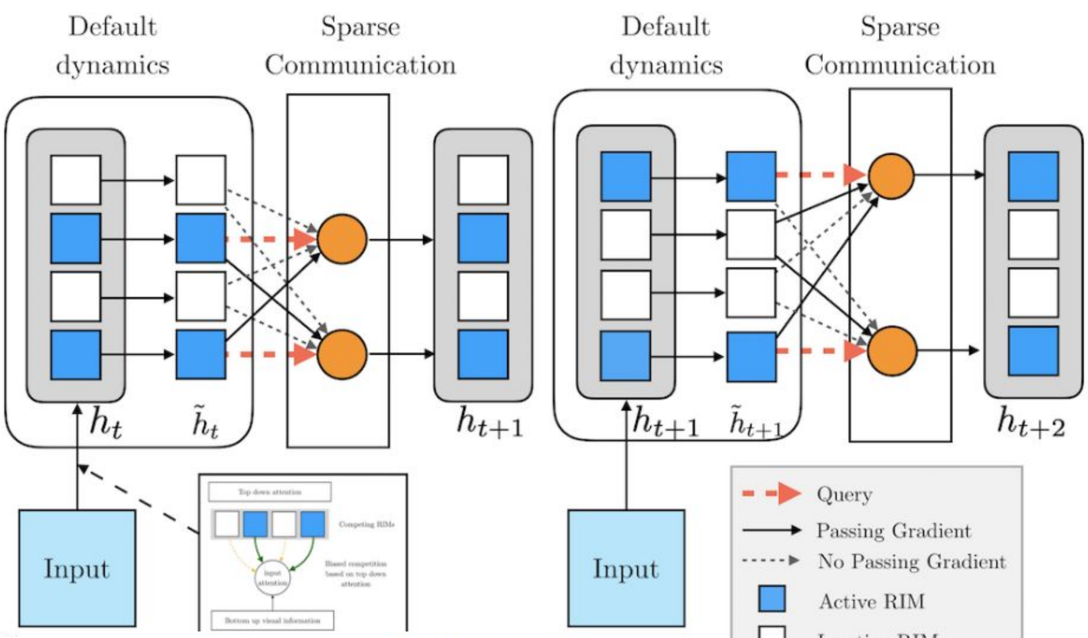
Computation through Attention (...the transformer network, machine translation type…)

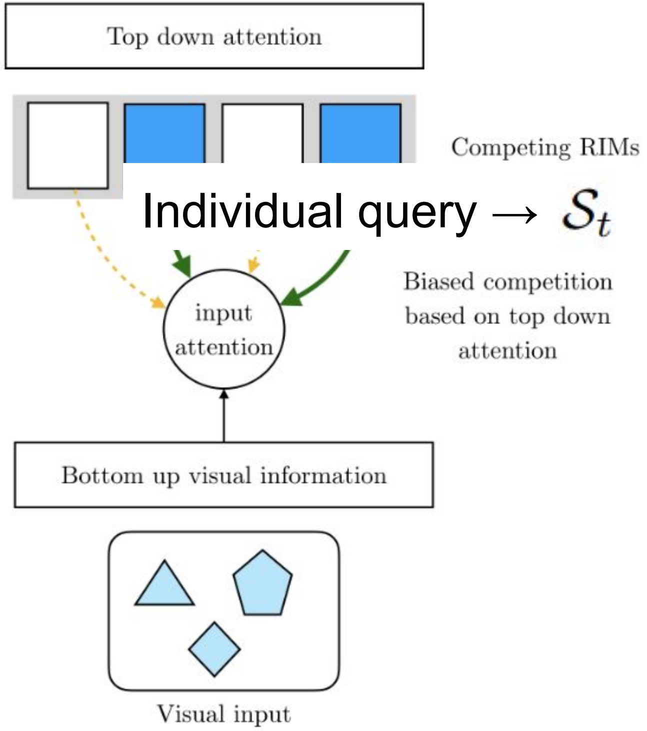


https://distill.pub/2016/augmented-rnns/

- Consider all options in parallel but focus only few relevant elements
  - learn where to attend
  - sequentially use the right words
- Attention as dynamic connection
  - allows developing (fairly) independent subsystems
- Information bottleneck!

Recurrent Independent Mechanisms





Individual query → $\mathcal{S}_t$

$$\tilde{h}_{t+1,k} = h_{t,k} \qquad \forall k \notin \mathcal{S}_t$$

$$\tilde{h}_{t+1,k} = D_k(h_{t,k}) = LSTM(h_{t,k}, A_k^{(in)}; \theta_k^{(D)}) \qquad \forall k \in \mathcal{S}_t$$

Key-Value Attention and Top-Down Modulation

- Soft-attention: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V,$
    - Queries come from the RIMs, while keys and values come from the current input

Key-Value Attention and Top-Down Modulation

- Soft-attention: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V,$
    - Queries come from the RIMs, while keys and values come from the current input
- Parameters of attention mechanism itself are separate for each RIM
    - Input at time t is set of elements x_t. $X = \emptyset \oplus x_t$. $K = XW^k, V = XW^v, Q = RW_k^q$

Key-Value Attention and Top-Down Modulation

- Soft-attention: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V,$
  - Queries come from the RIMs, while keys and values come from the current input
- Parameters of attention mechanism itself are separate for each RIM
  - Input at time t is set of elements x_t. $X = \emptyset \oplus x_t$. $K = XW^k, V = XW^v, Q = RW_k^q$
- Attention: $A_k^{(in)} = \text{softmax}\left(\frac{RW_k^q(XW^k)^T}{\sqrt{d_e}}\right)XW^v, \text{ where } \theta_k^{(in)} = (W_k^q, W^e, W^v).$

Key-Value Attention and Top-Down Modulation

- Soft-attention: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V,$
  - Queries come from the RIMs, while keys and values come from the current input
- Parameters of attention mechanism itself are separate for each RIM
  - Input at time t is set of elements x_t. $X = \emptyset \oplus x_t$. $K = XW^k, V = XW^v, Q = RW_k^q$
- Attention: $A_k^{(in)} = \text{softmax}\left(\frac{RW_k^q (XW^k)^T}{\sqrt{d_e}}\right)XW^v$, where $\theta_k^{(in)} = (W_k^q, W^e, W^v)$.
- At each step, select the top k_A RIMs to be activated $\rightarrow \mathcal{S}_t$

Communication between RIMs

- Attention mechanism allows sharing of information among the RIMs
  - activated RIMs can read from all other RIMs

$$Q_{t,k} = \tilde{W}_k^q \tilde{h}_{t,k}, \quad \forall k \in \mathcal{S}_t$$

$$K_{t,k} = \tilde{W}_k^e \tilde{h}_{t,k}, \quad \forall k$$

$$V_{t,k} = \tilde{W}_k^v \tilde{h}_{t,k}, \quad \forall k$$

$$h_{t+1,k} = \mathrm{softmax}\left(\frac{Q_{t,k}(K_{t,:})^T}{\sqrt{d_e}}\right) V_{t,:} + \tilde{h}_{t,k} \quad \forall k \in \mathcal{S}_t, \text{ where } \theta_k^{(c)} = (\tilde{W}_k^q, \tilde{W}_k^e, \tilde{W}_k^v).$$

Experiments

- Drop-in replacement for an LSTM or GRU cell, following the exact same interface with the exact same inputs and outputs.
    - No change to the loss function which results from using RIMs

- Diverse tasks, focus on changing environments

Temporal Task: Copying

- ➔ Receiving short sequence of characters
- ➔ Blank inputs for large number of steps
- ➔ Task: Reproduce the original sequence

Copying	k_T	k_A	h_{size}	**Train(50)** CE	**Test(200)** CE
RIMs	6	5	600	0.01	3.5
	6	4	600	**0.00**	**0.00**
	6	3	600	**0.00**	**0.00**
	6	2	600	**0.00**	**0.00**
	5	3	500	**0.00**	**0.00**
LSTM	-	-	300	0.00	2.28
	-	-	600	0.00	3.56
NTM	-	-	-	0.00	2.54
RMC	-	-	-	0.00	0.13
Transformers	-	-	-	0.00	0.54

⇒ RIMs can specialize over distinct patterns in the data and improve generalization to settings where these patterns change

Temporal Task: Sequential MNIST

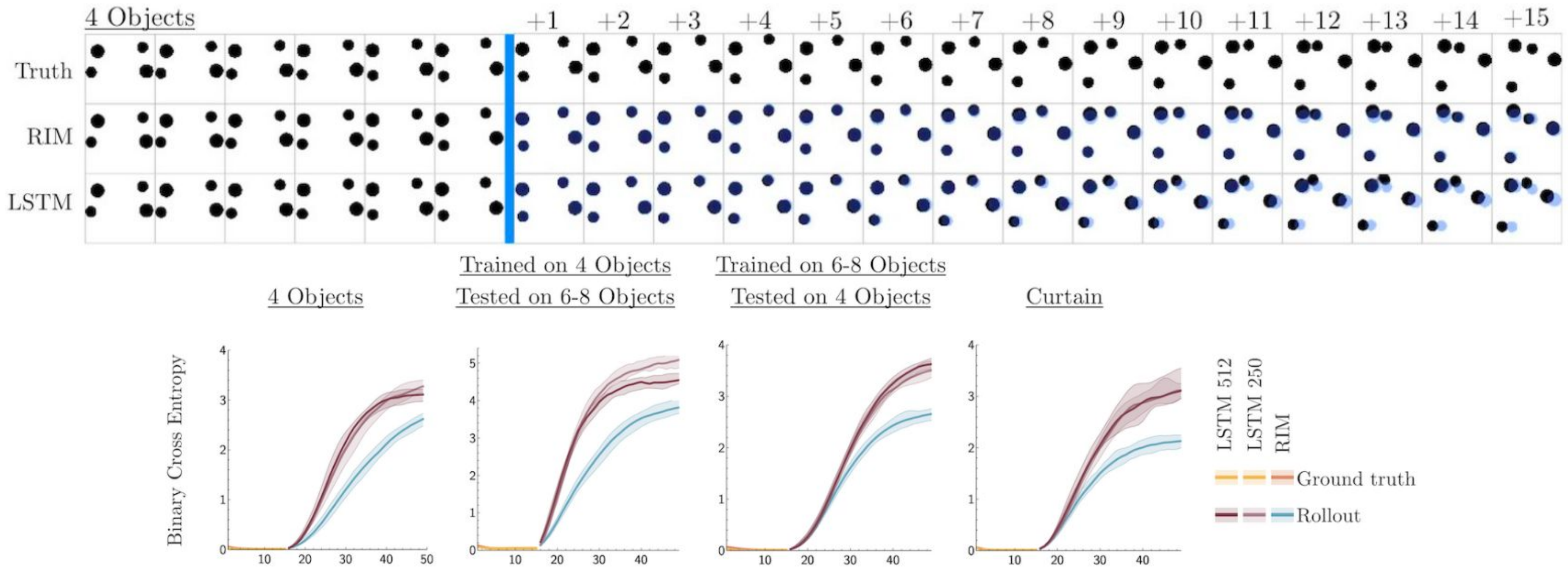
- ➔ Receiving sequence of pixels
- ➔ Train on 14x14 resolution
- ➔ Task: Classify MNIST digits

Sequential MNIST	k_T	k_A	h_{size}	**16 x 16** Accuracy	**19 x 19** Accuracy	**24 x 24** Accuracy
RIMs	6	6	600	85.5	56.2	30.9
	6	5	600	88.3	43.1	22.1
	6	4	600	**90.0**	**73.4**	**38.1**
LSTM	-	-	300	86.8	42.3	25.2
	-	-	600	84.5	52.2	21.9
EntNet	-	-	-	89.2	52.4	23.5
RMC	-	-	-	89.58	54.23	27.75
DNC	-	-	-	87.2	44.1	19.8
Transformers	-		-	**91.2**	51.6	22.9

⇒ RIM model shows higher robustness to changing the sequence length and is able to pass gradients through large empty regions

Object Task: Bouncing Ball Environment

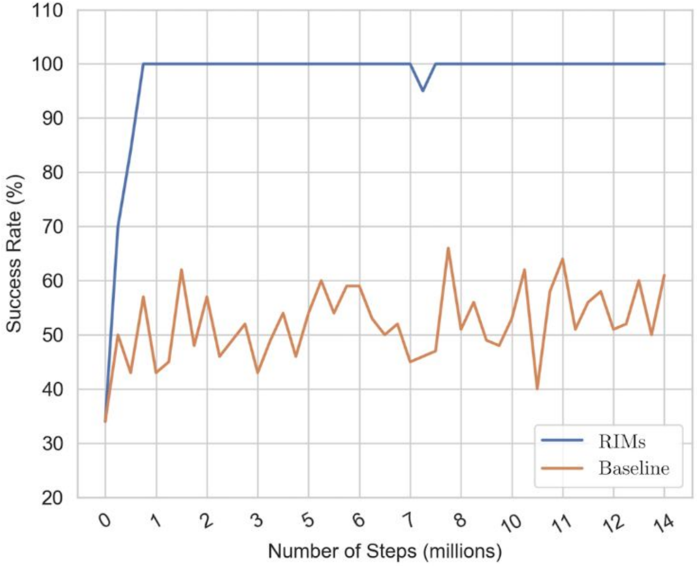
- ➔ Multiple balls with independent movement, except for collision events
- ➔ Task: Predict future motion



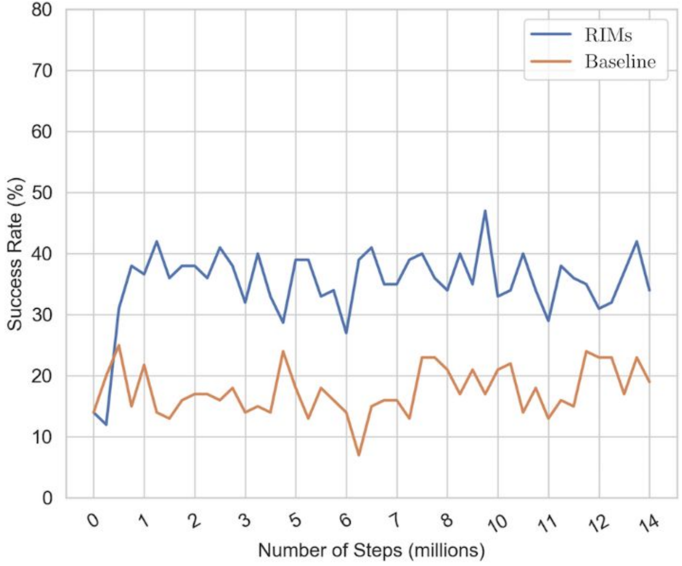
Object Task: Environment with Novel Distractors

- ➔ Task: Agent must retrieve a specific object in the presence of distractors
- ➔ Challenges: Partial observability of environment and sparsity of reward

Only distractors from training

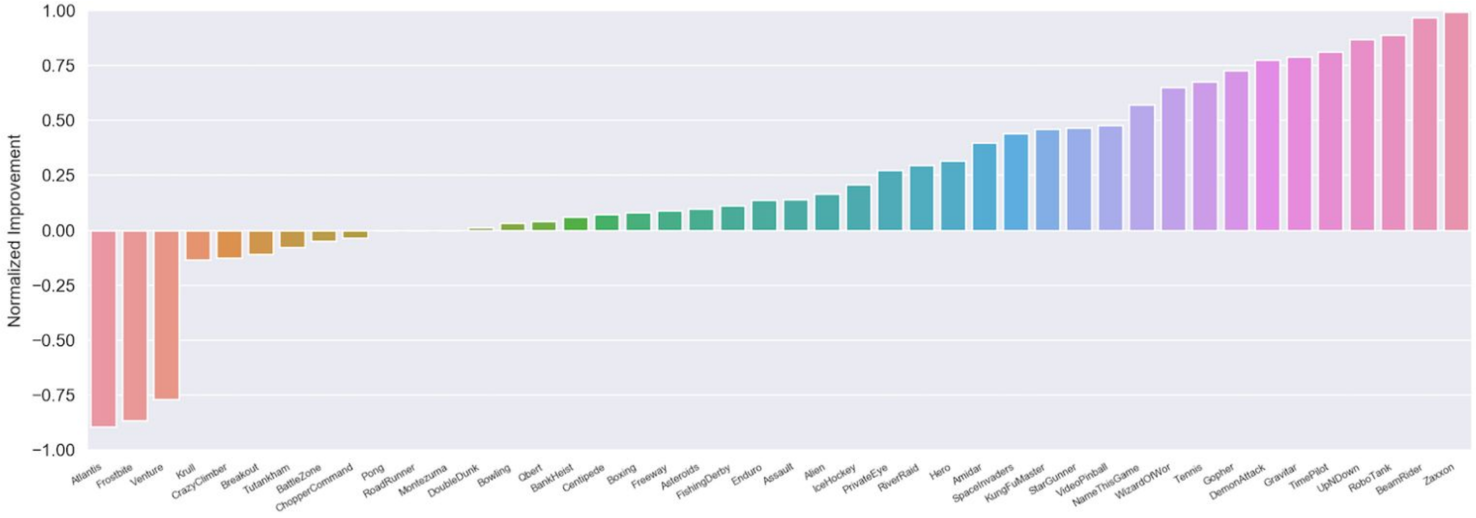


Novel distractors added for testing



Generalization in Complex Environments Task

- ➔ RL agent trained using PPO with a recurrent network producing the policy
  - ➔ Task: Learn to play Atari games (with same PPO settings)



⇒ RIM improves most in environments with a dynamic combination of risks and opportunities, since it allows to rapidly adapt information processing

Conclusion

- New architecture that reflects independence of mechanisms in the real world
  - Inductive bias from SOTA RNN models: FC ↔ all processes interact
- RIMs with sparse interaction (attention) lead to improved OOD generalization
  - Experiments show specialization and improved robustness to changing task distributions
- Support for consciousness prior...?

operating on sets of pointable objects with dynamically recombined modules

modularize computation and operate on sets of named and typed objects in a DL way

apply idea of independent mechanisms to RNNs