Assessing the role of feedback connections in artificial and biological neural networks

MBCT Seminar

2020.05.18

Aran Nayebi
Neurosciences PhD Program
Stanford University

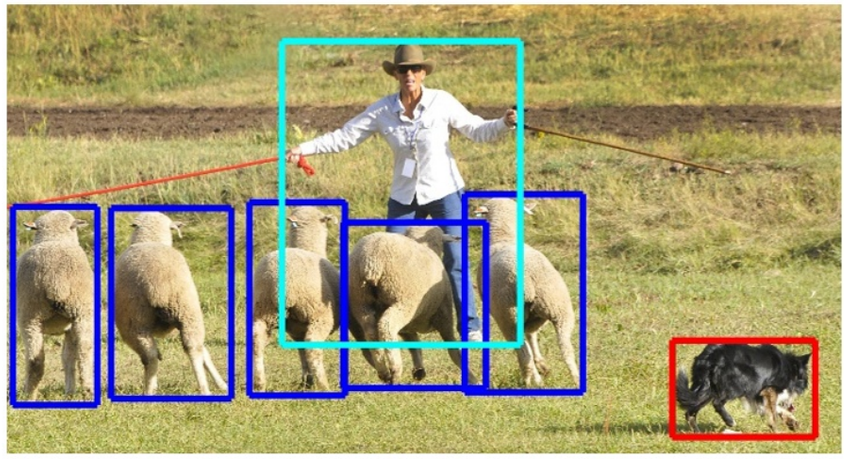
Object recognition is easy for us

Classification, Segmentation, Localization, …

(a) **Image classification**

Classification, Segmentation, Localization, …

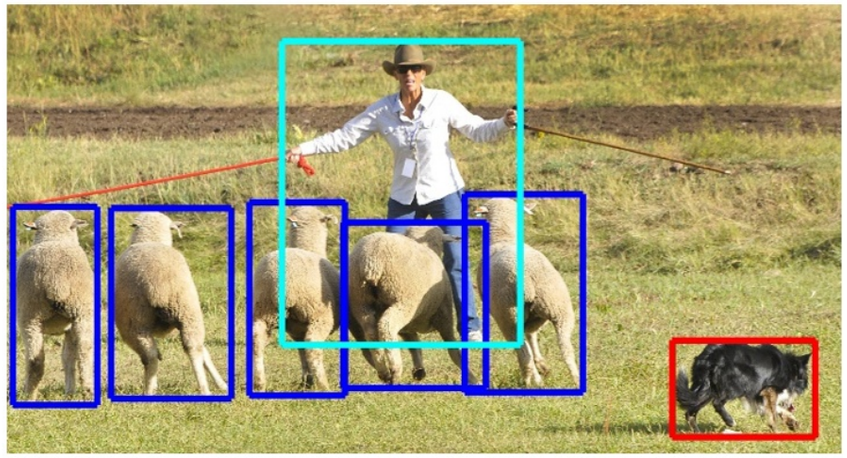
(a) **Image classification**



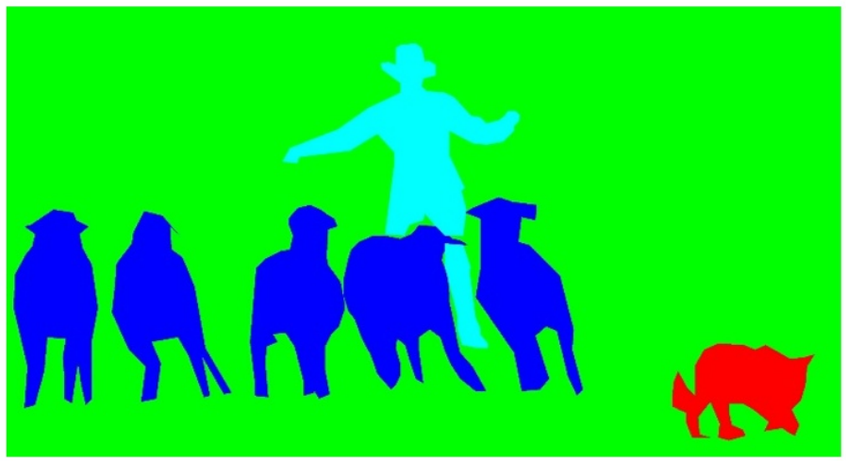
(b) **Object localization**

Classification, Segmentation, Localization, …

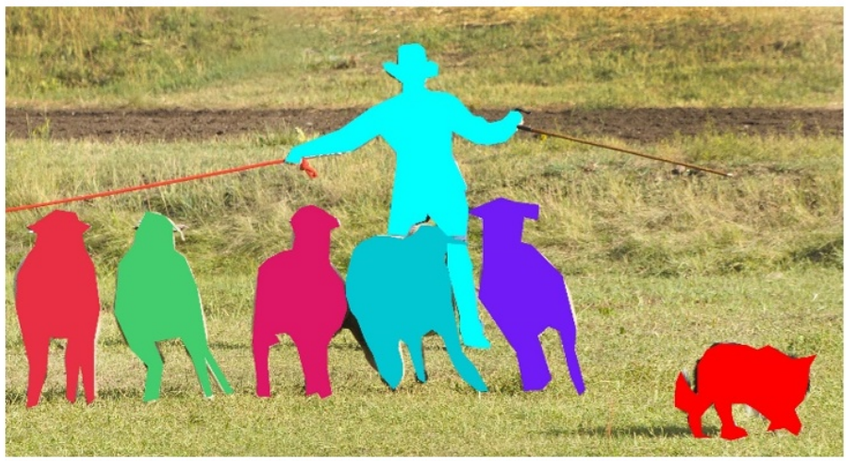
(a) **Image classification**



(b) **Object localization**



(c) **Semantic segmentation**

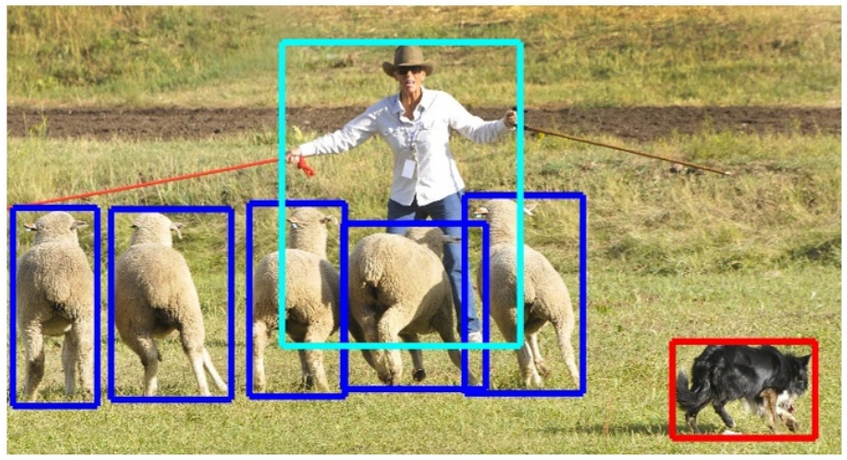


(d) **Instance segmentation**

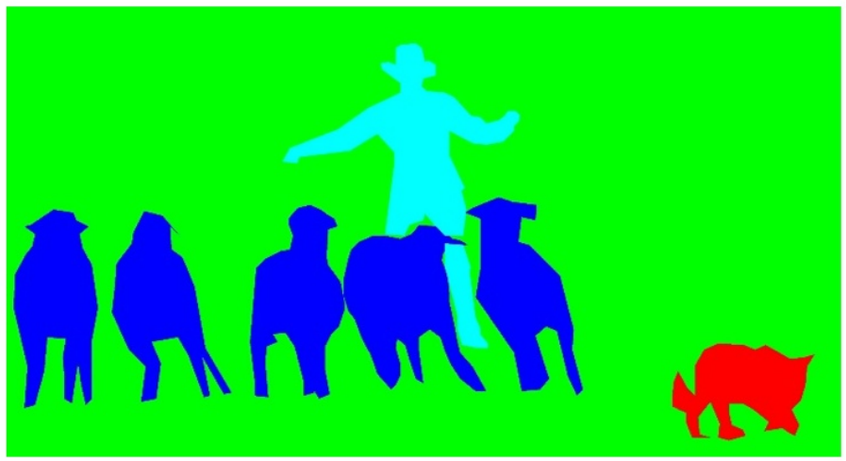
Lin et al. 2014

Classification, Segmentation, Localization, …

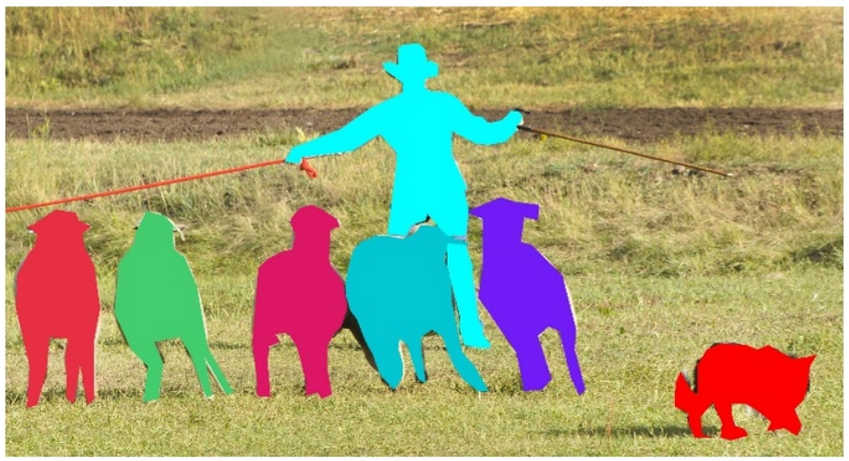
(a) **Image classification**



(b) **Object localization**



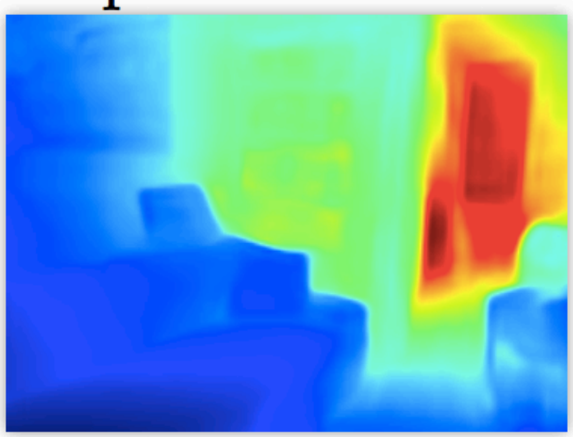
(c) **Semantic segmentation**



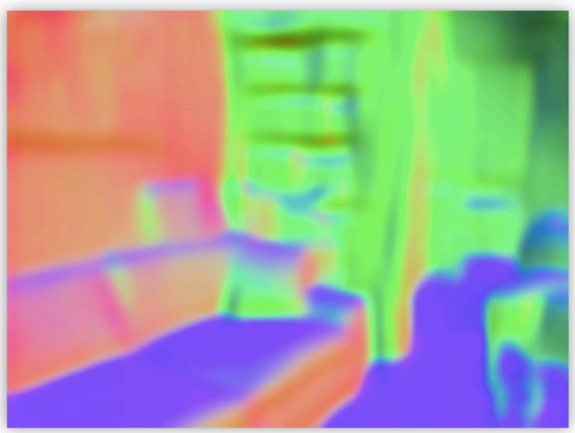
(d) **Instance segmentation**

Lin et al. 2014

Depth



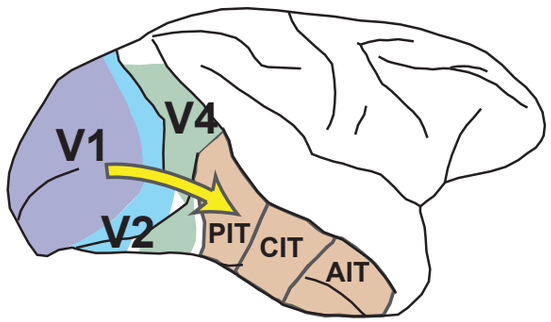
Normals



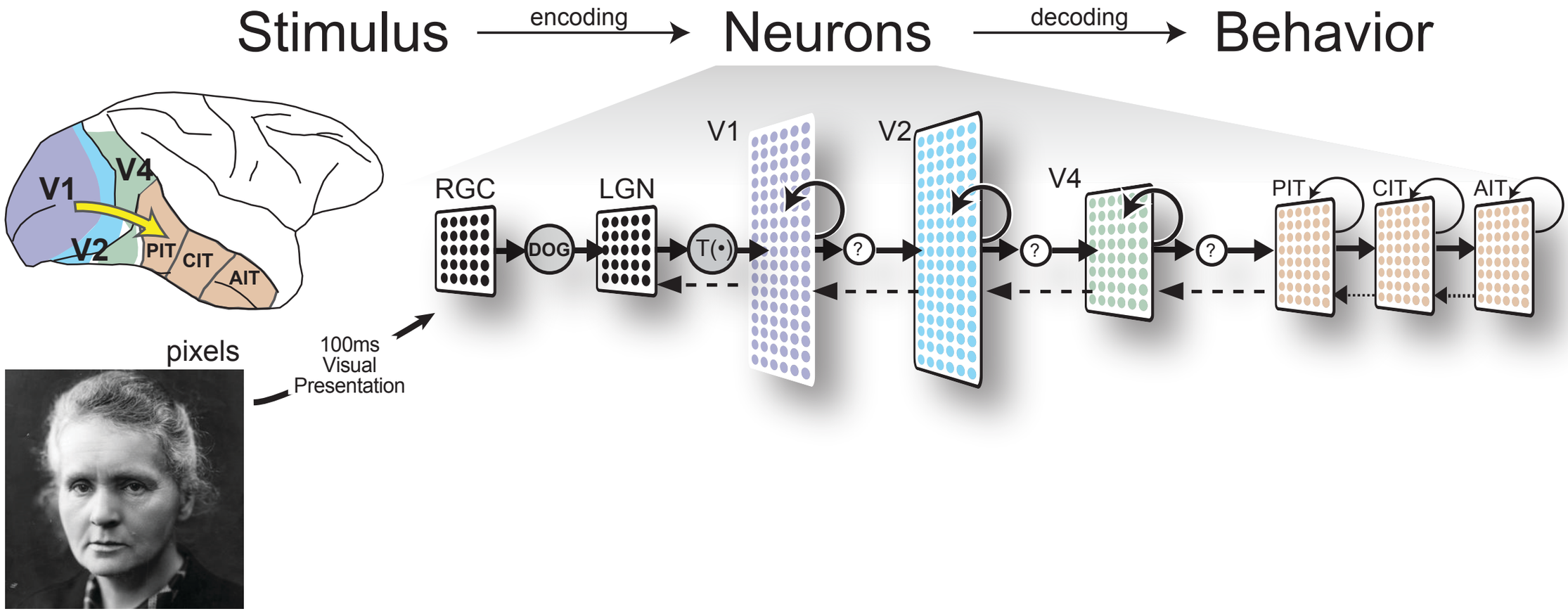
Eigen and Fergus 2015

…thanks to the Ventral Stream

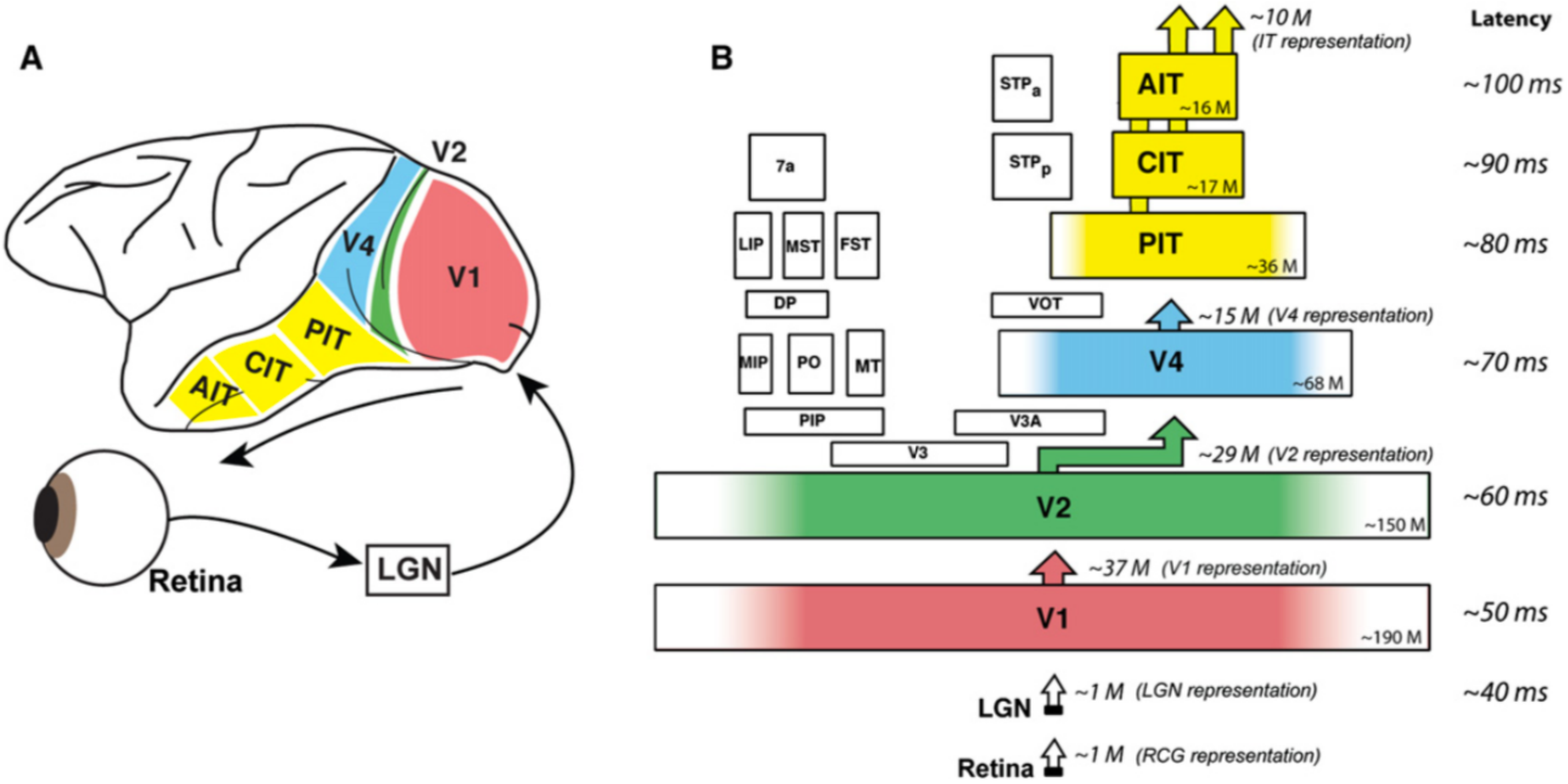
Stimulus —encoding→ Neurons —decoding→ Behavior



…which is somewhat feedforward



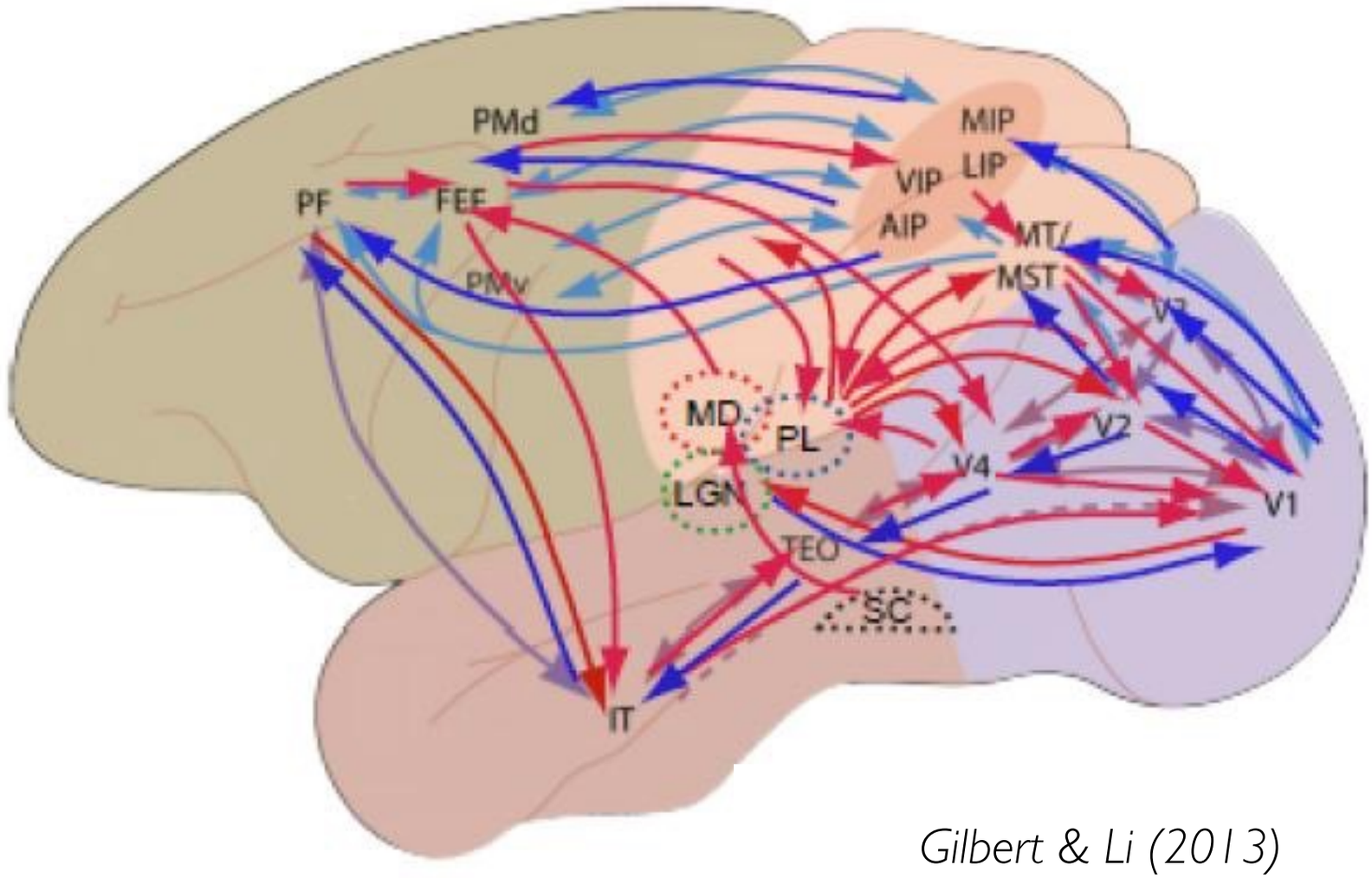
…which is somewhat feedforward



DiCarlo, Zoccolan, & Rust 2012

…but also not just feedforward

Feedbacks are everywhere anatomically:



Gilbert & Li (2013)

…but what are they for?

Many hypotheses for recurrent connections

Many hypotheses for recurrent connections

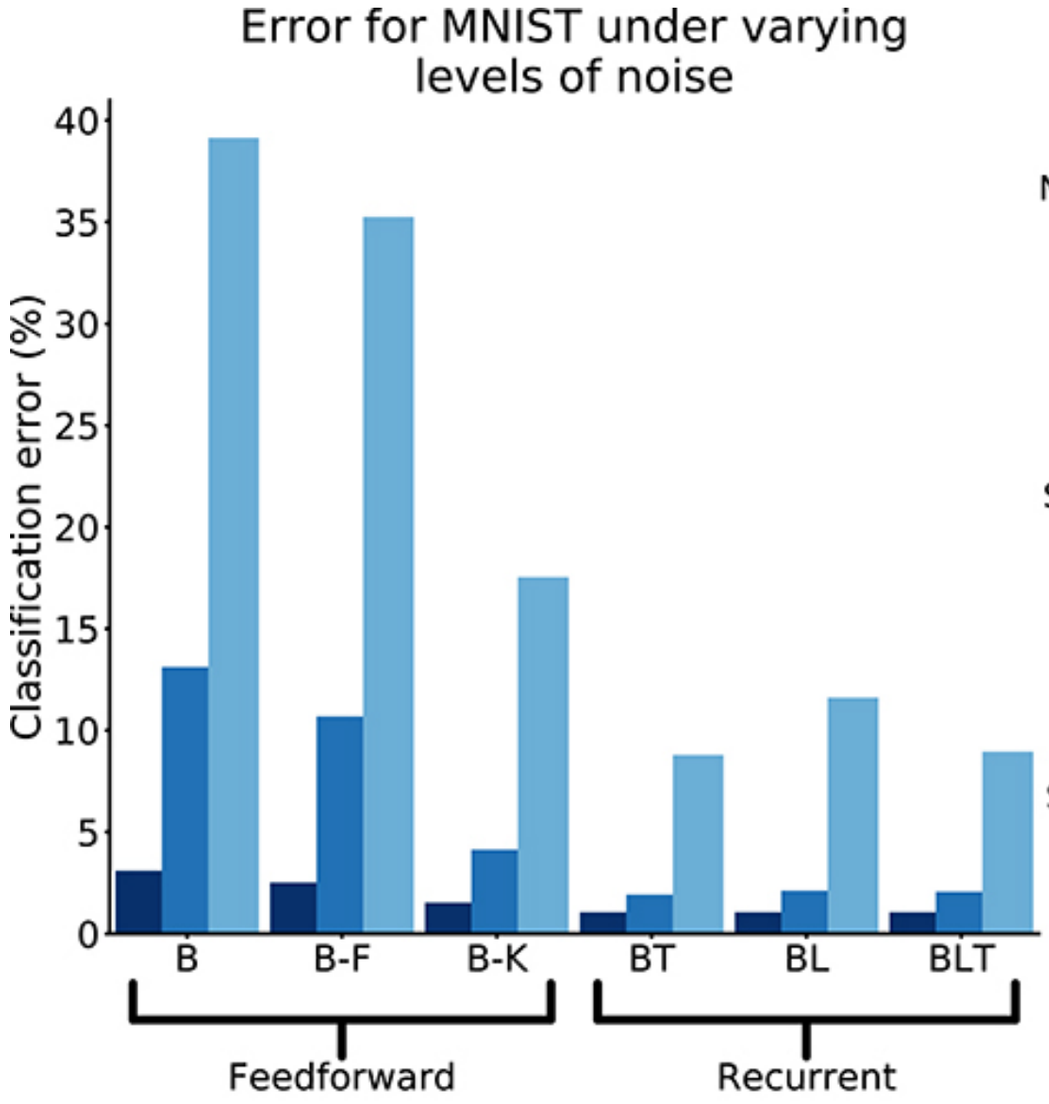
- Feedback connections primarily used during inference?

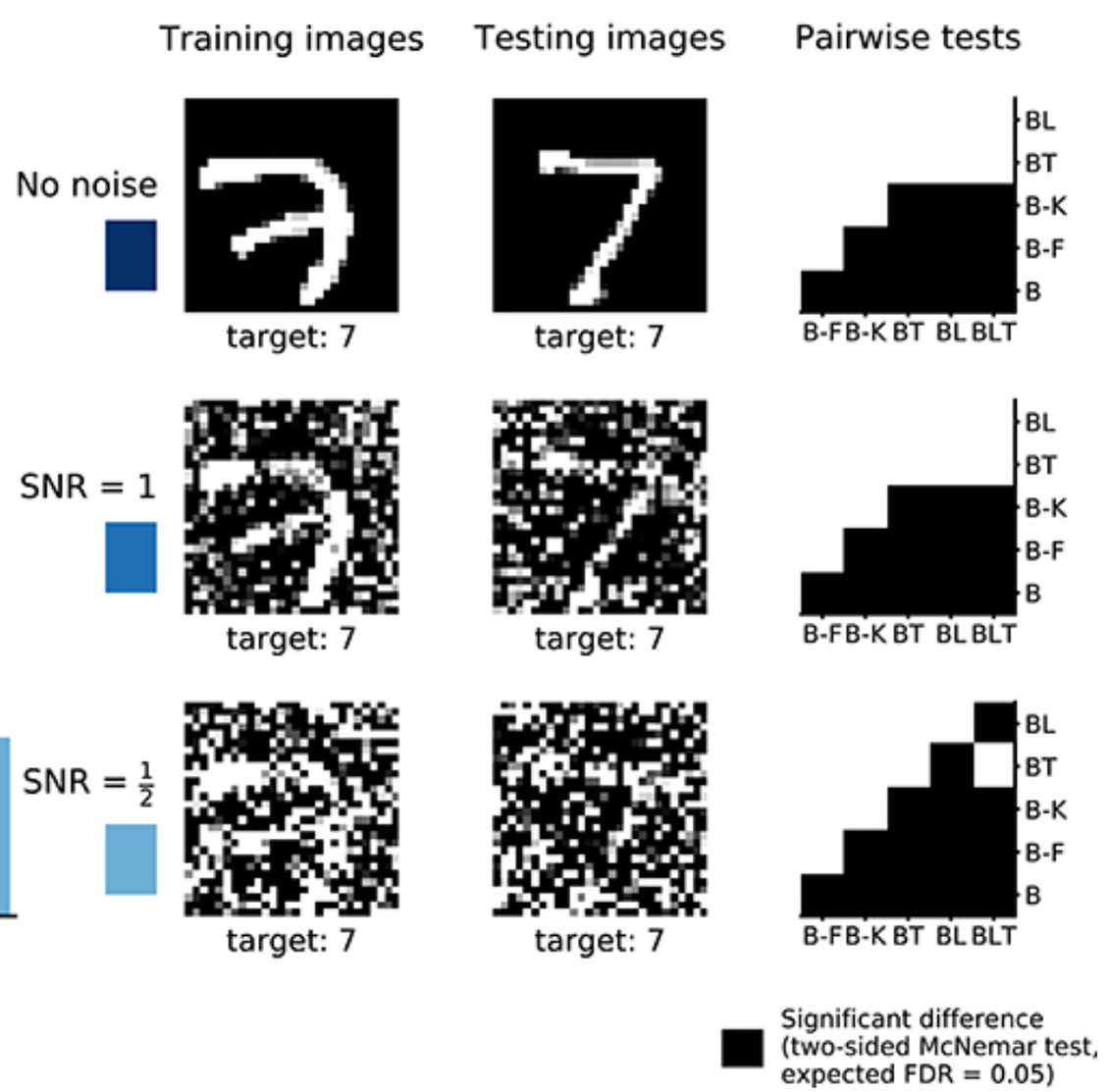
Hypotheses for recurrence - Occlusions

Recurrent convolutional neural networks suppress occluders and enhance targets in occluded object recognition

Courtney J. Spoerer (courtney.spoerer@mrc-cbu.cam.ac.uk)
Medical Research Council Cognition and Brain Sciences Unit,
15 Chaucer Road, Cambridge, CB2 7EF, UK

Nikolaus Kriegeskorte (nikokriegeskorte@gmail.com)
Medical Research Council Cognition and Brain Sciences Unit,
15 Chaucer Road, Cambridge, CB2 7EF, UK





Hypotheses for recurrence - Top Down Feature Attention

CBMM Memo No. 047 April 12, 2016

Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex

by

Qianli Liao and Tomaso Poggio
Center for Brains, Minds and Machines, McGovern Institute, MIT

Feedback Networks

Amir R. Zamir[1,3*] Te-Lin Wu[1*] Lin Sun[1,2] William B. Shen[1] Bertram E. Shi[2]
Jitendra Malik[3] Silvio Savarese[1]
[1] Stanford University [2] HKUST [3] University of California, Berkeley
http://feedbacknet.stanford.edu/

Accepted as a workshop contribution at ICLR 2015

ATTENTION FOR FINE-GRAINED CATEGORIZATION

Pierre Sermanet, Andrea Frome, Esteban Real
Google, Inc.
{sermanet,afrome,ereal,}@google.com

Hypotheses for recurrence - Top Down Feature Attention

CBMM Memo No. 047 April 12, 2016

Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex

by

Qianli Liao and Tomaso Poggio
Center for Brains, Minds and Machines, McGovern Institute, MIT

Performance gains only on quite small datasets

Feedback Networks

Amir R. Zamir[1,3*] Te-Lin Wu[1*] Lin Sun[1,2] William B. Shen[1] Bertram E. Shi[2]
Jitendra Malik[3] Silvio Savarese[1]
[1] Stanford University [2] HKUST [3] University of California, Berkeley
http://feedbacknet.stanford.edu/

Accepted as a workshop contribution at ICLR 2015

ATTENTION FOR FINE-GRAINED CATEGORIZATION

Pierre Sermanet, Andrea Frome, Esteban Real
Google, Inc.
{sermanet,afrome,ereal,}@google.com

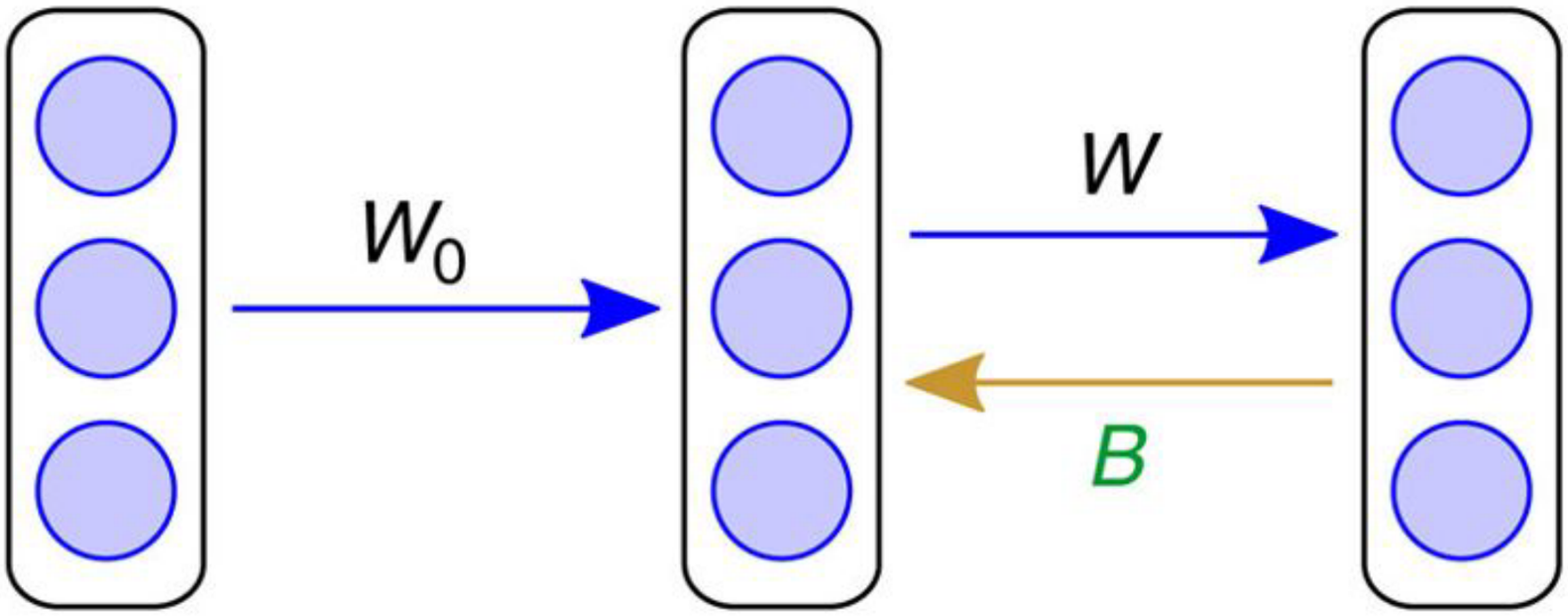
Many hypotheses for recurrent connections

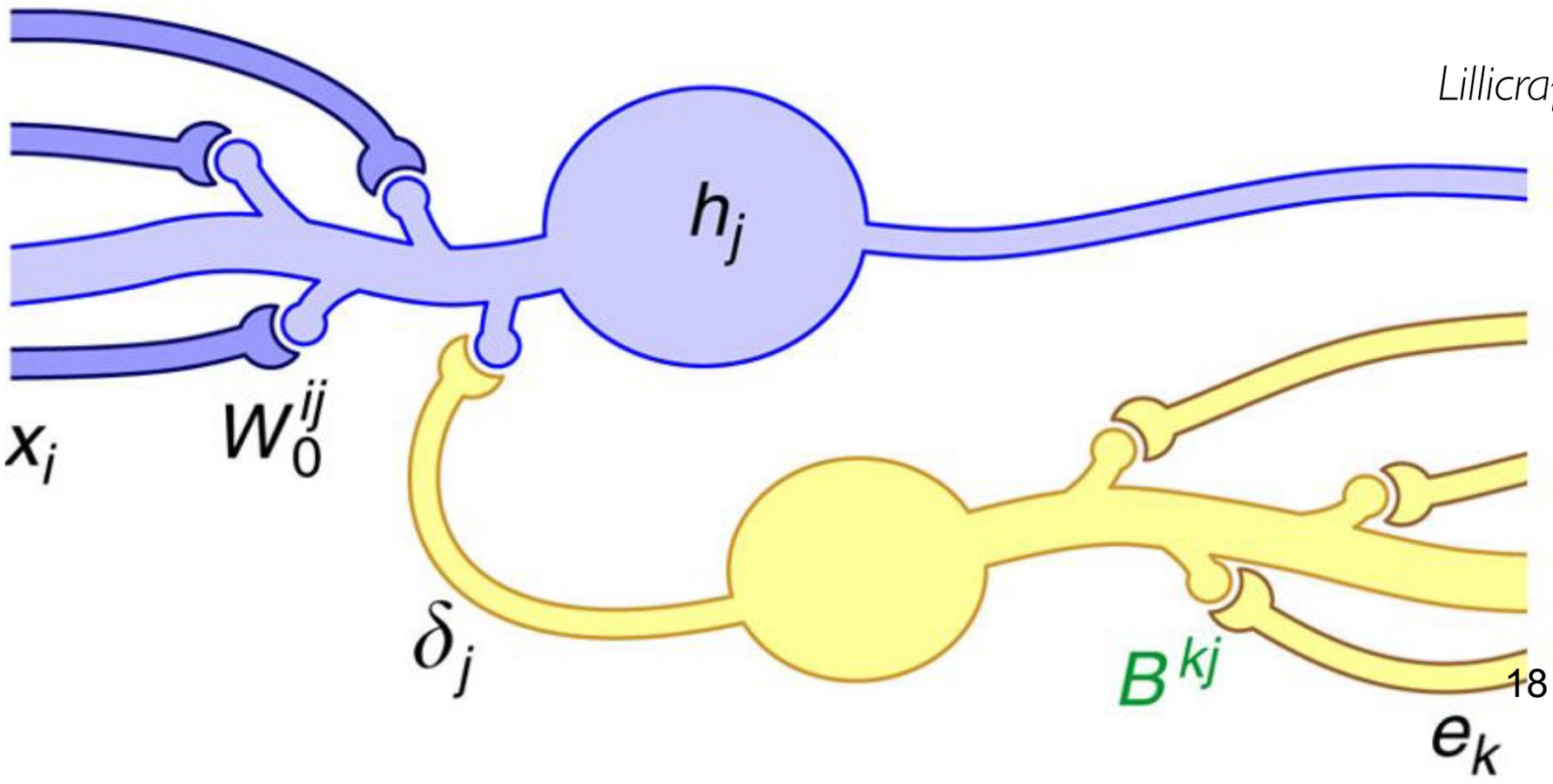
- Feedback connections primarily used during inference?

Many hypotheses for recurrent connections

- **Feedback connections primarily used during inference?**
- Feedback connections primarily used for propagating error signals?

Rationale - Mechanisms for Credit Assignment





Lillicrap et al. (2016)

Many hypotheses for recurrent connections

- Feedback connections primarily used during inference?

- Feedback connections primarily used for propagating error signals?

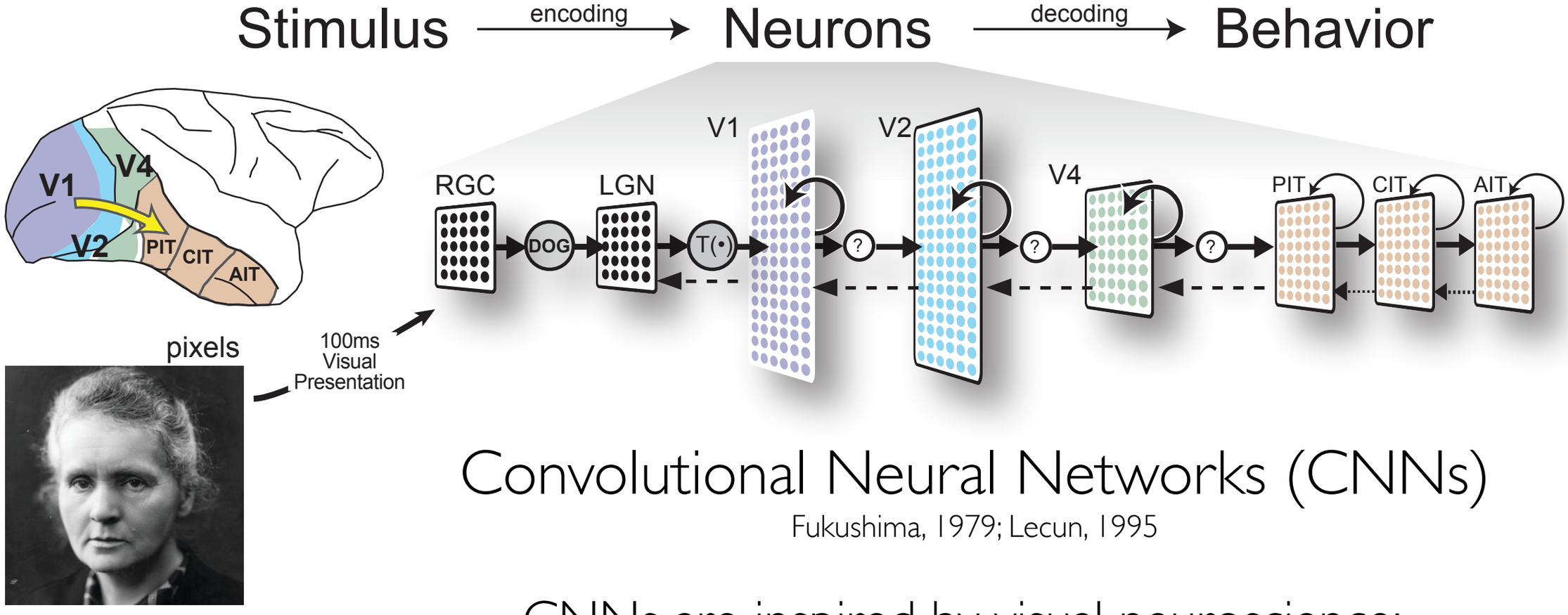
How can we adjudicate between these possibilities?

Is recurrence useful during inference?

- Feedback connections primarily used during inference?

- Feedback connections primarily used for propagating error signals?

CNNs as Models of Object Recognition



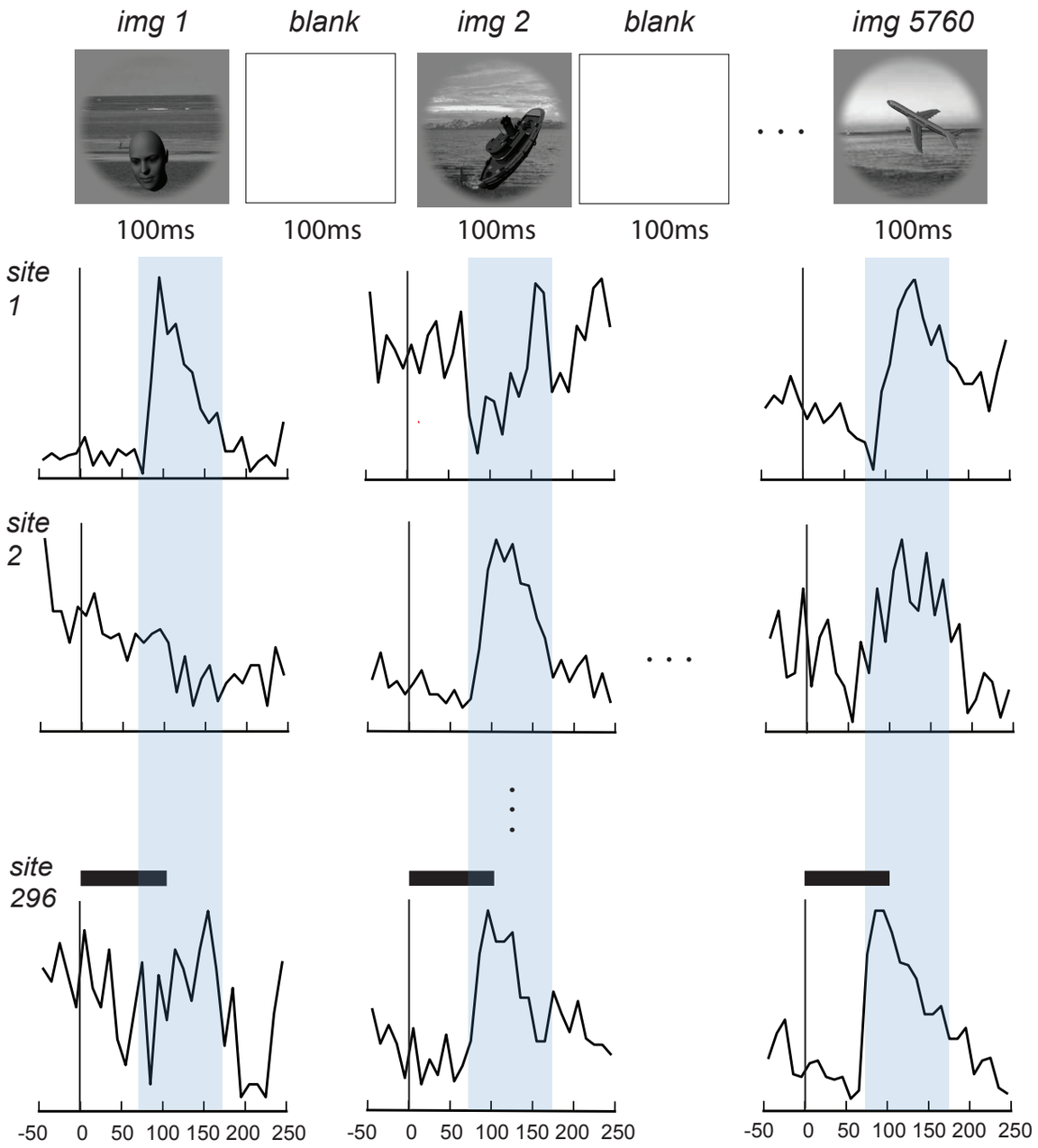
Convolutional Neural Networks (CNNs)

Fukushima, 1979; Lecun, 1995

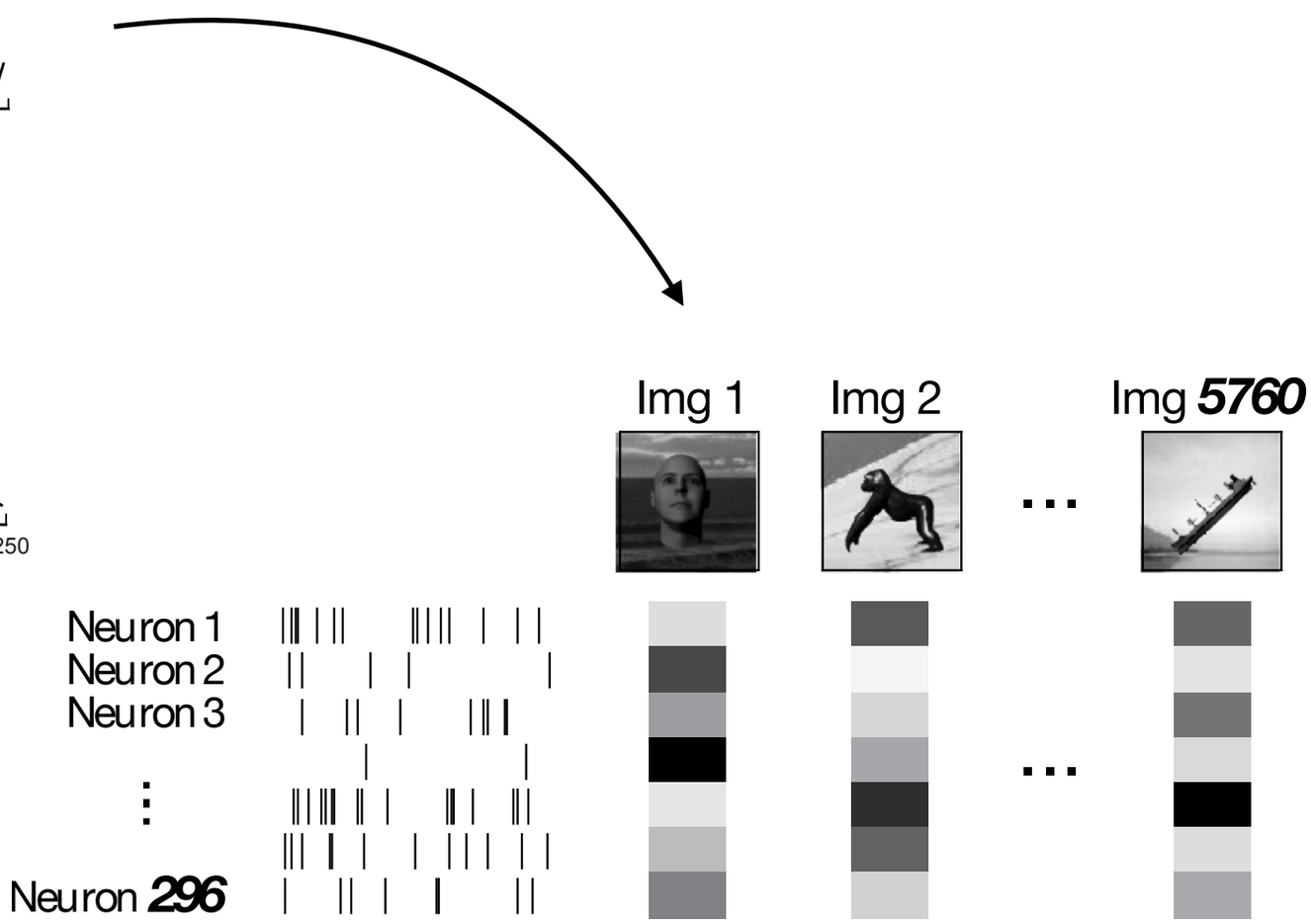
CNNs are inspired by visual neuroscience:

1) **hierarchy**
2) **retinotopy** (spatially tiled)

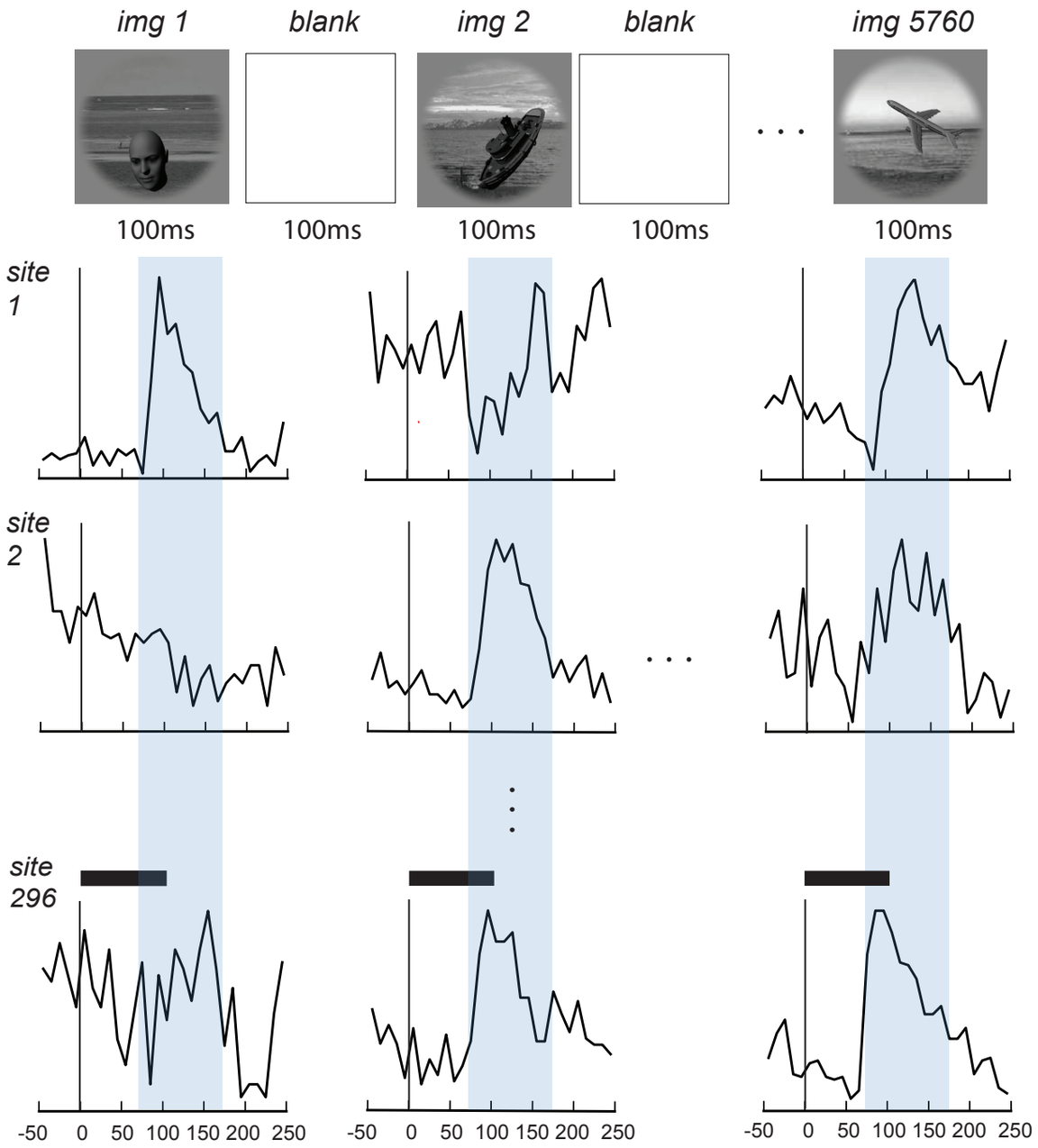
So far, only explaining temporal average of responses



e.g. Binned spike counts 70ms-170ms post stimulus presentation

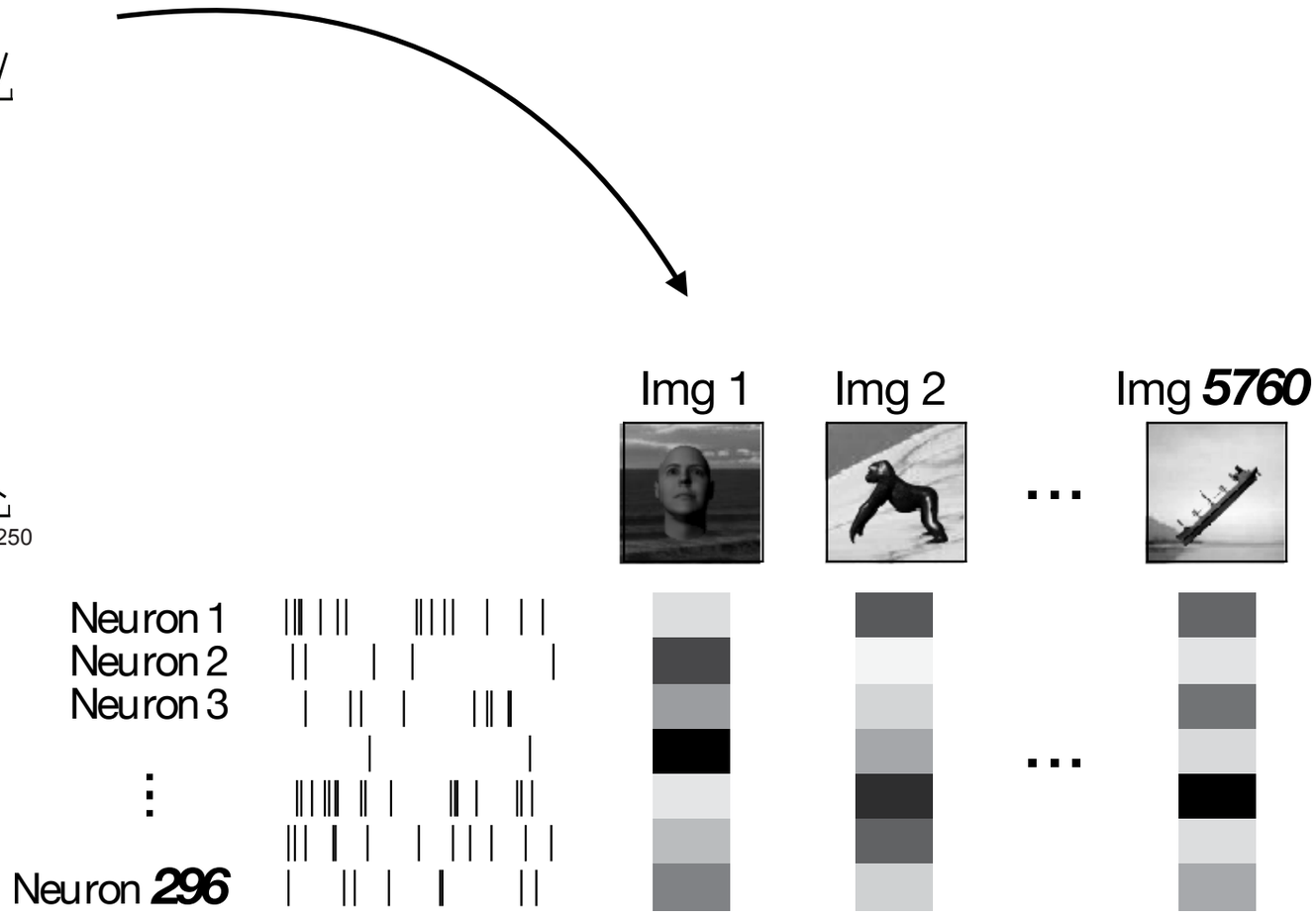


So far, only explaining temporal average of responses



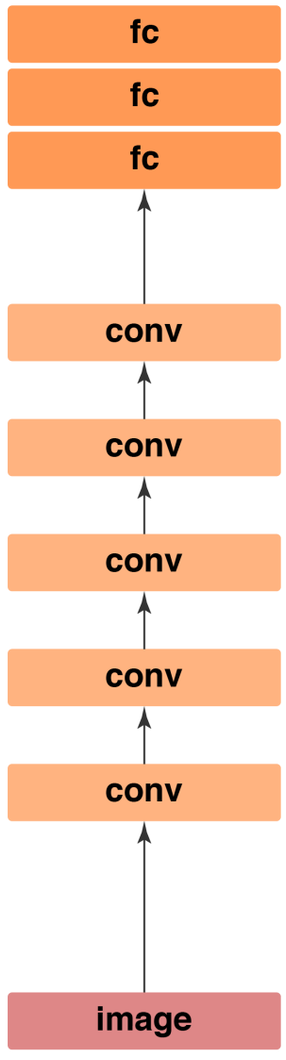
e.g. Binned spike counts 70ms-170ms post stimulus presentation

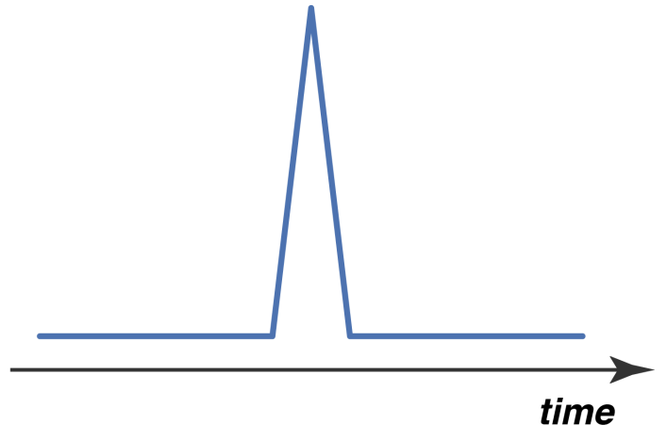
but actually the data is highly reliable at much finer grain —10ms bins



Trajectory Possibilities

Simple feedforward networks simple dynamics:

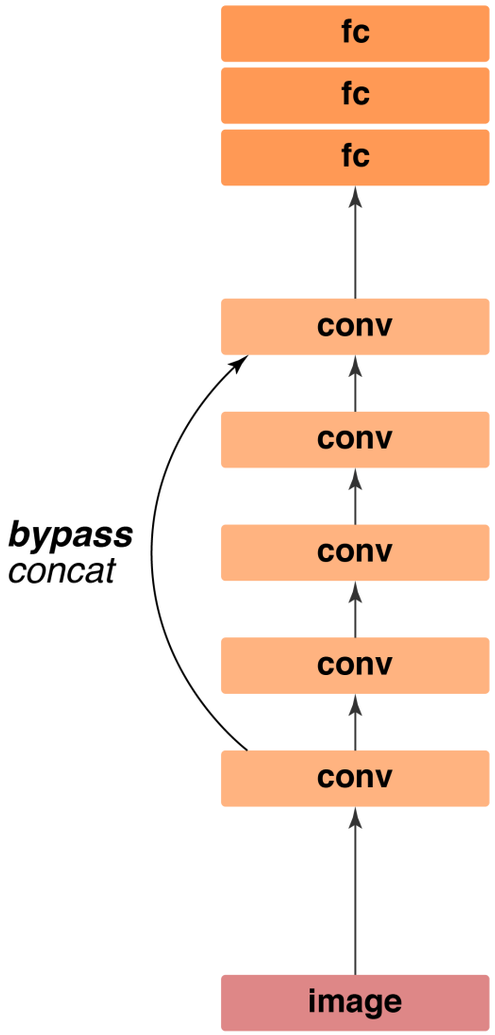


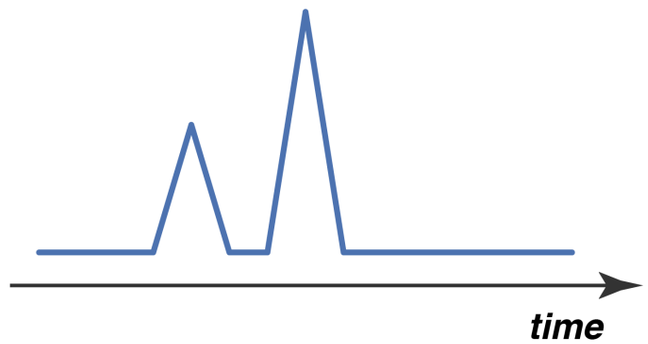


courtesy Jonas Kubilius

Trajectory Possibilities

Dynamics more interesting with bypasses:

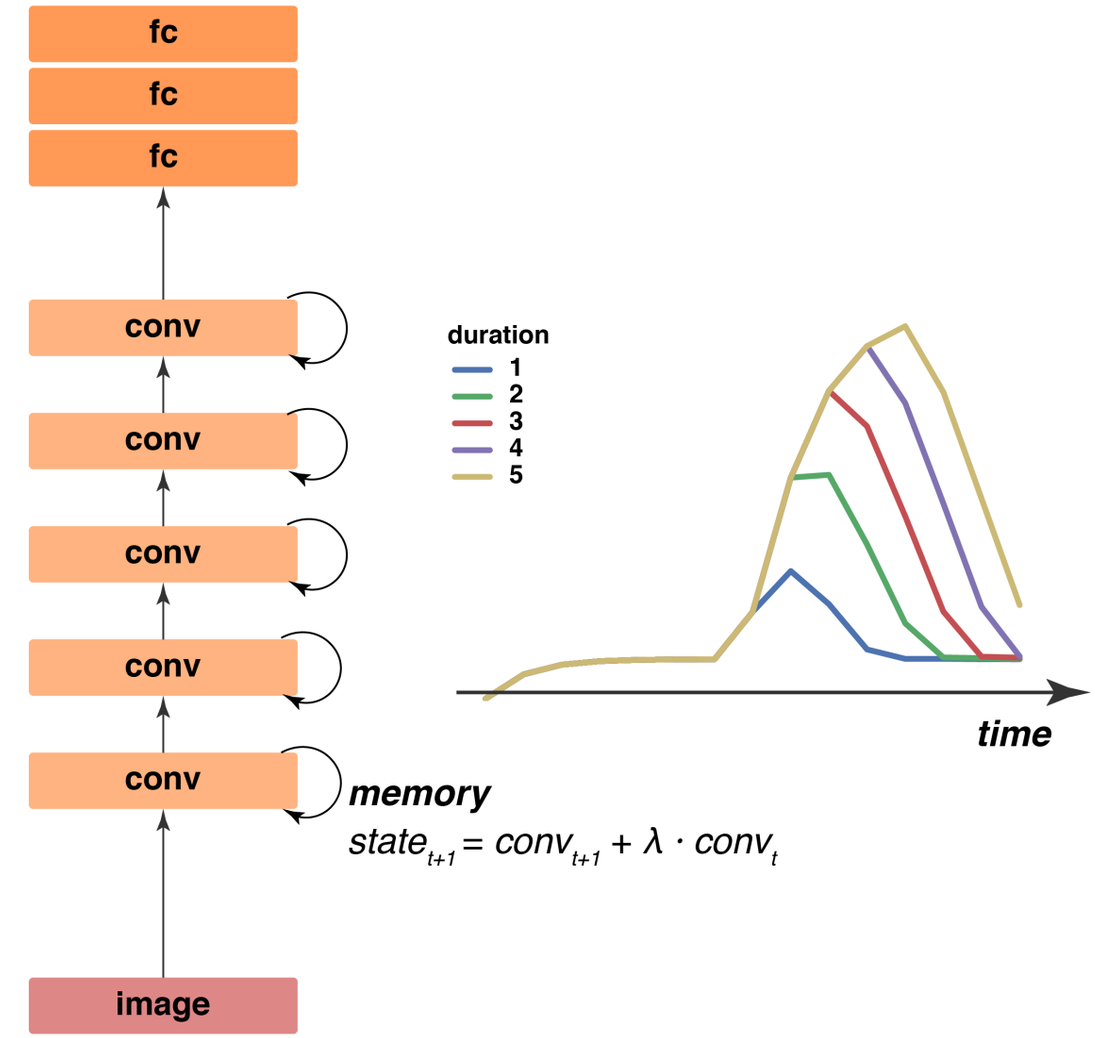




courtesy Jonas Kubilius

Trajectory Possibilities

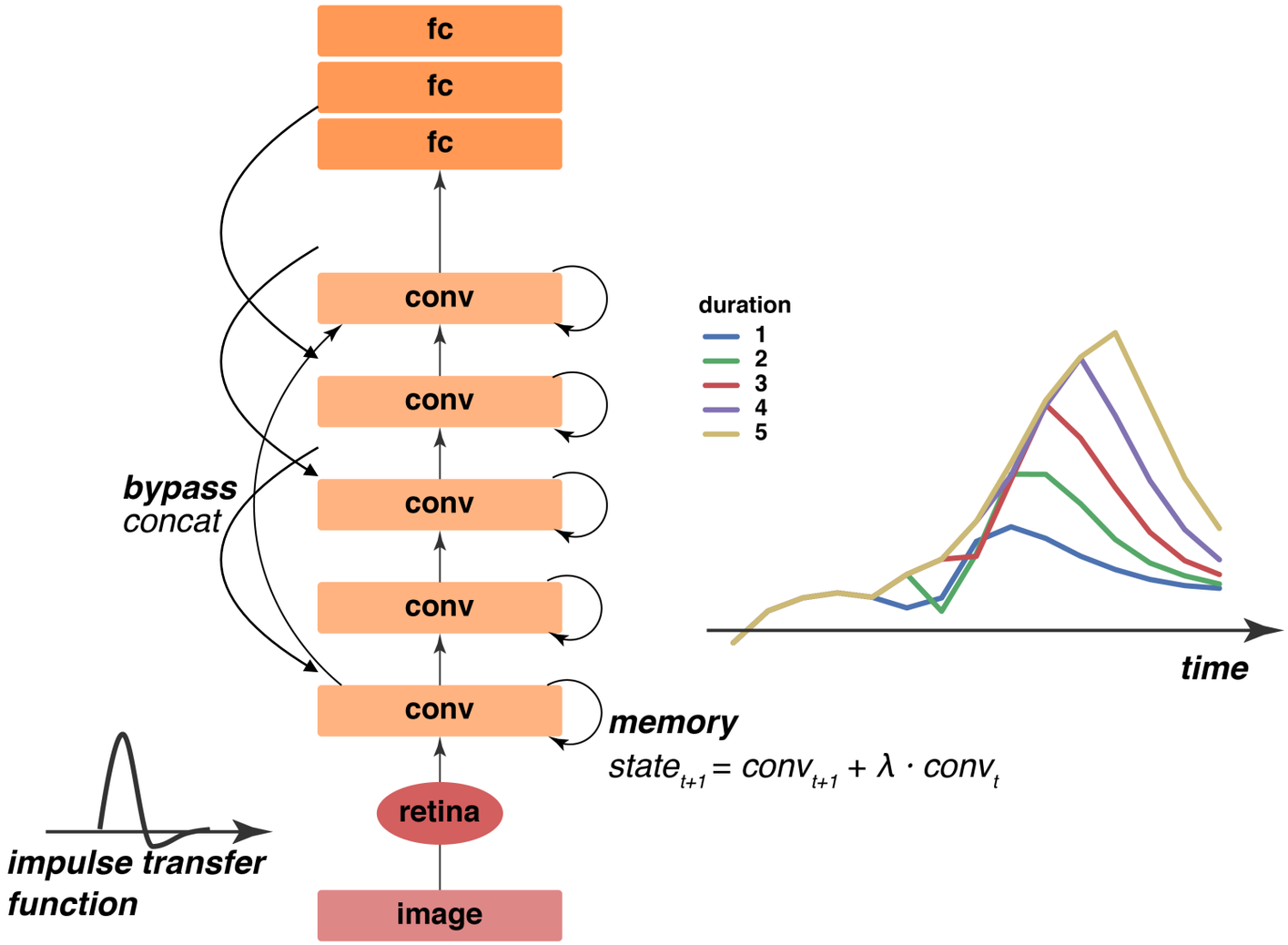
Dynamics more interesting with bypasses, local recurrence:



courtesy Jonas Kubilius

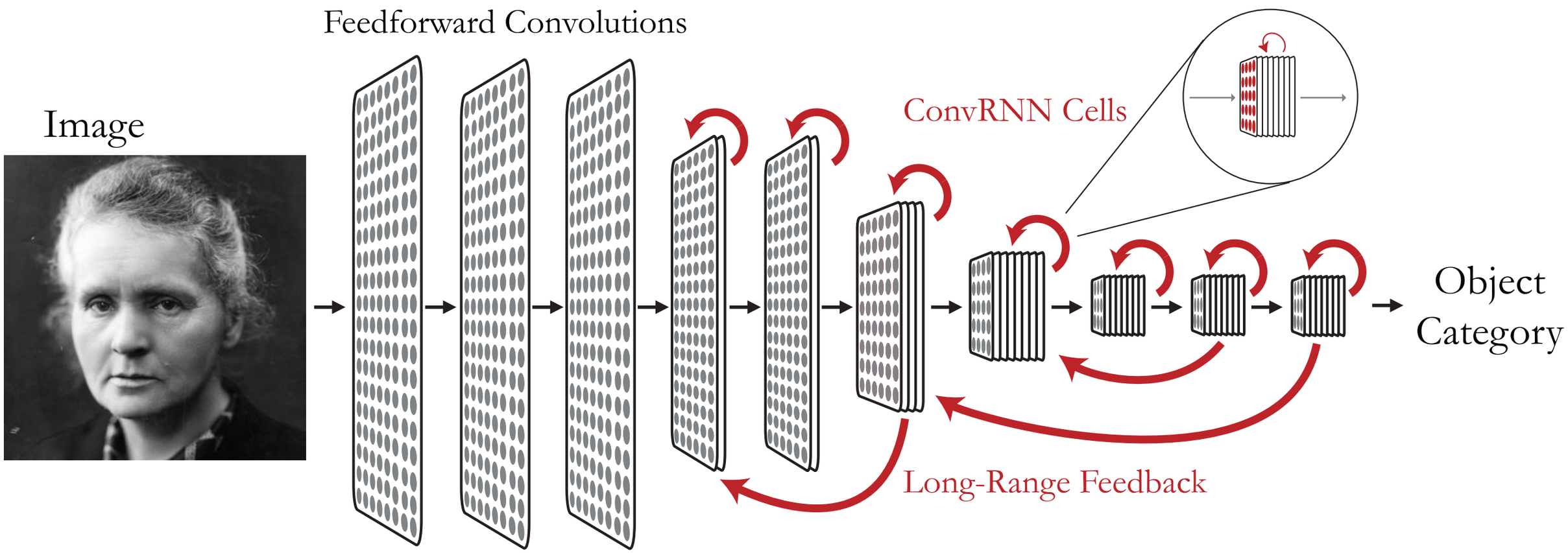
Trajectory Possibilities

Dynamics more interesting with bypasses, local recurrence, long-range feedback:

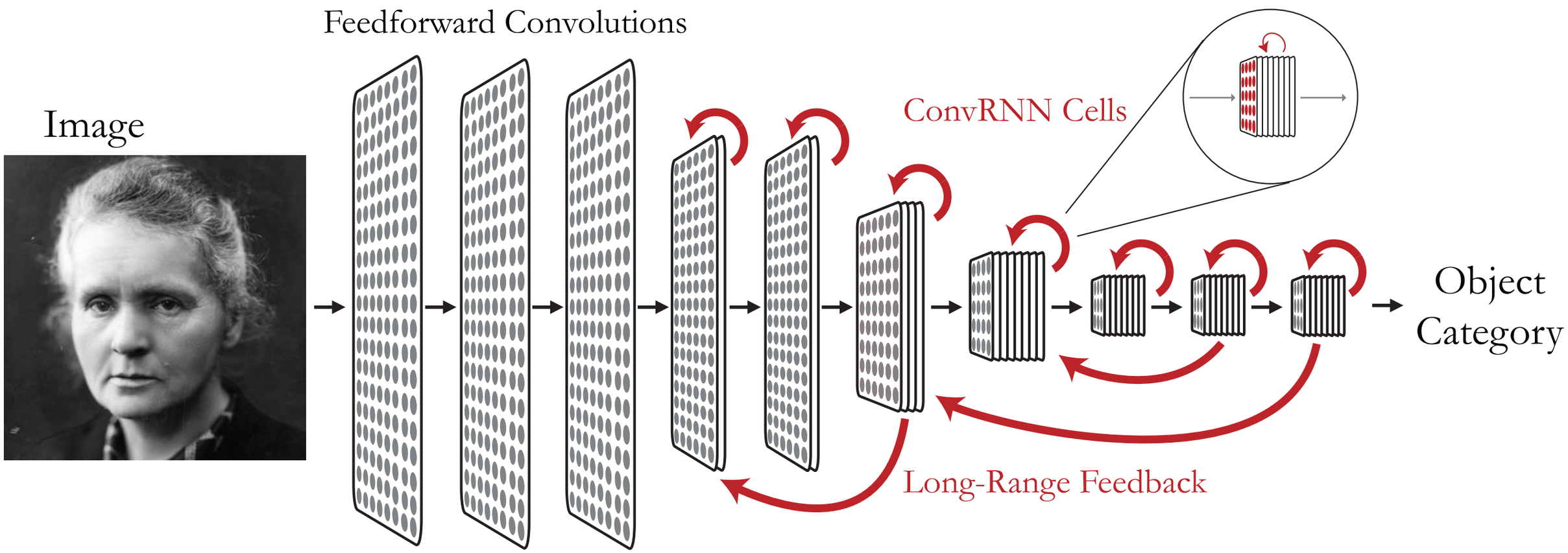


courtesy Jonas Kubilius

Convolutional Recurrent Neural Networks (ConvRNNs)



Improving ImageNet Performance with ConvRNNs



Each time-step (10 ms) is treated equally — including feedforward steps

Approach

- Expand architecture class (local and global recurrence)
- Parametrize local and global feedback motifs and optimize for performance on ImageNet
- Evaluate neural predictivity in V4 and IT temporal responses

Nayebi, A.*, Bear, D.*, Kubilius, J.*, et al. *Task-Driven Convolutional Recurrent Models of the Visual System*, NeurIPS 2018

Daniel Bear
(Stanford)

Jonas Kubilius
(MIT)

Approach

- Expand architecture class (local and global recurrence)

Nayebi, A.*, Bear, D.*, Kubilius, J.*, et al. *Task-Driven Convolutional Recurrent Models of the Visual System*, NeurIPS 2018

- Parametrize local and global feedback motifs and optimize for performance on ImageNet

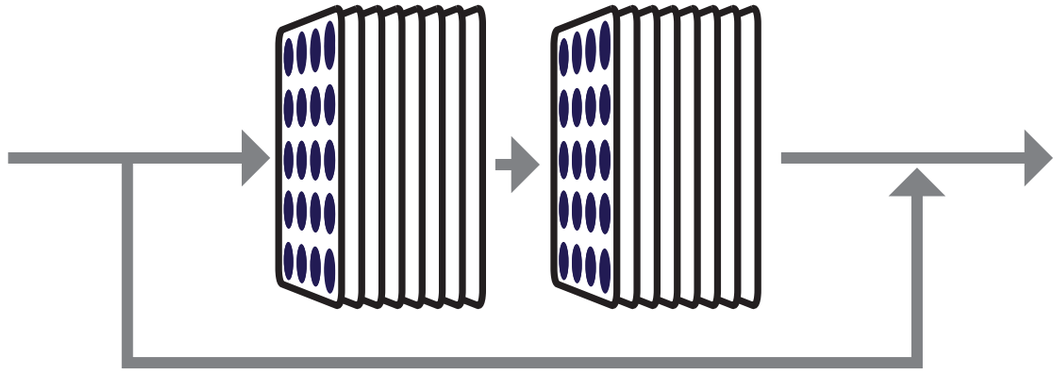
- Evaluate neural predictivity in V4 and IT temporal responses

Daniel Bear
(Stanford)

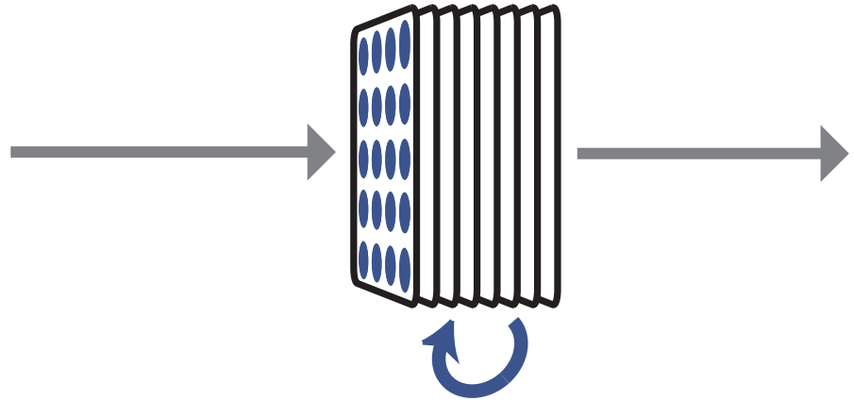
Jonas Kubilius
(MIT)

Many Choices of Local Recurrence

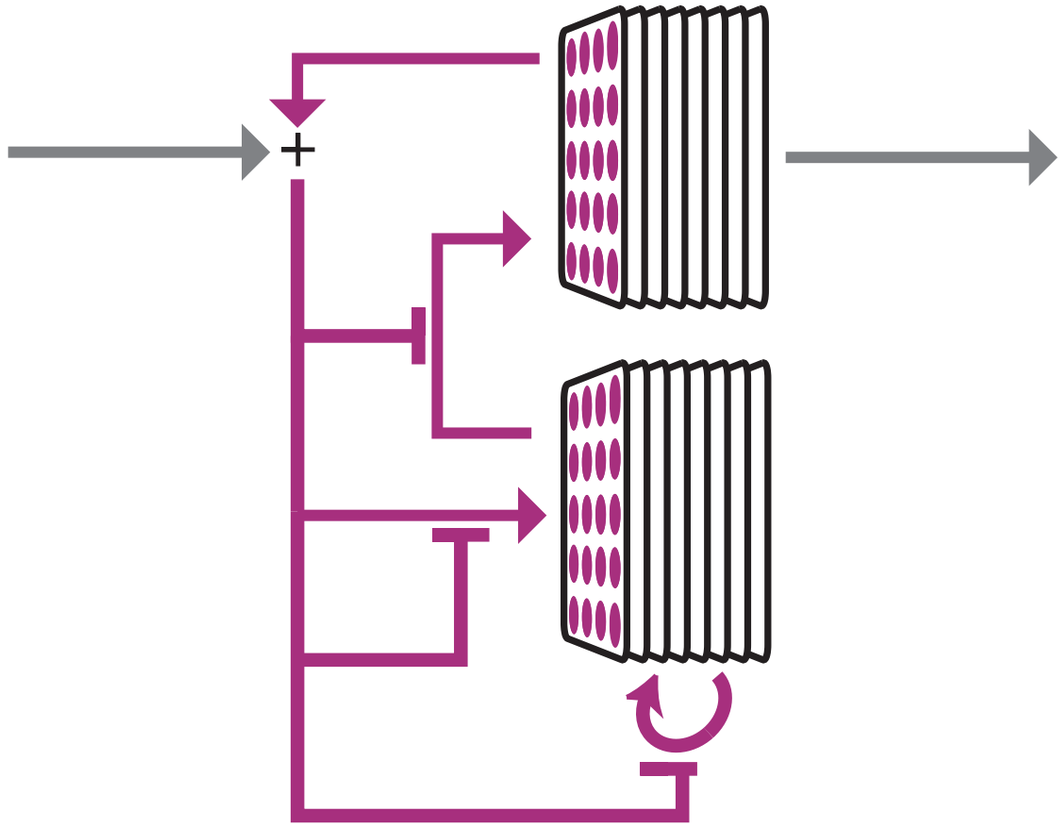
ResNet Block



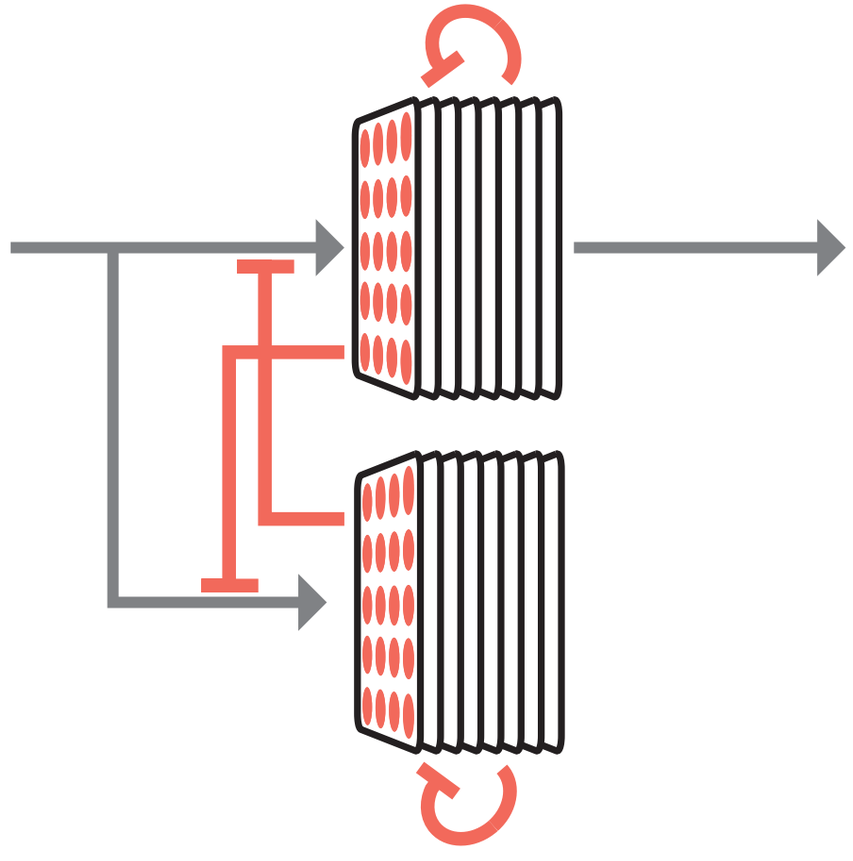
Vanilla RNN Cell



LSTM Cell



Reciprocal Gated Cell



Principles of Local Recurrence

Two complementary principles:

(1) gating = multiplication by input-dependent tensor w/ values in [0, 1]

(2) bypassing = when recurrent cell is in 0 state, input is unchanged ("performance preserving")

Principles of Local Recurrence

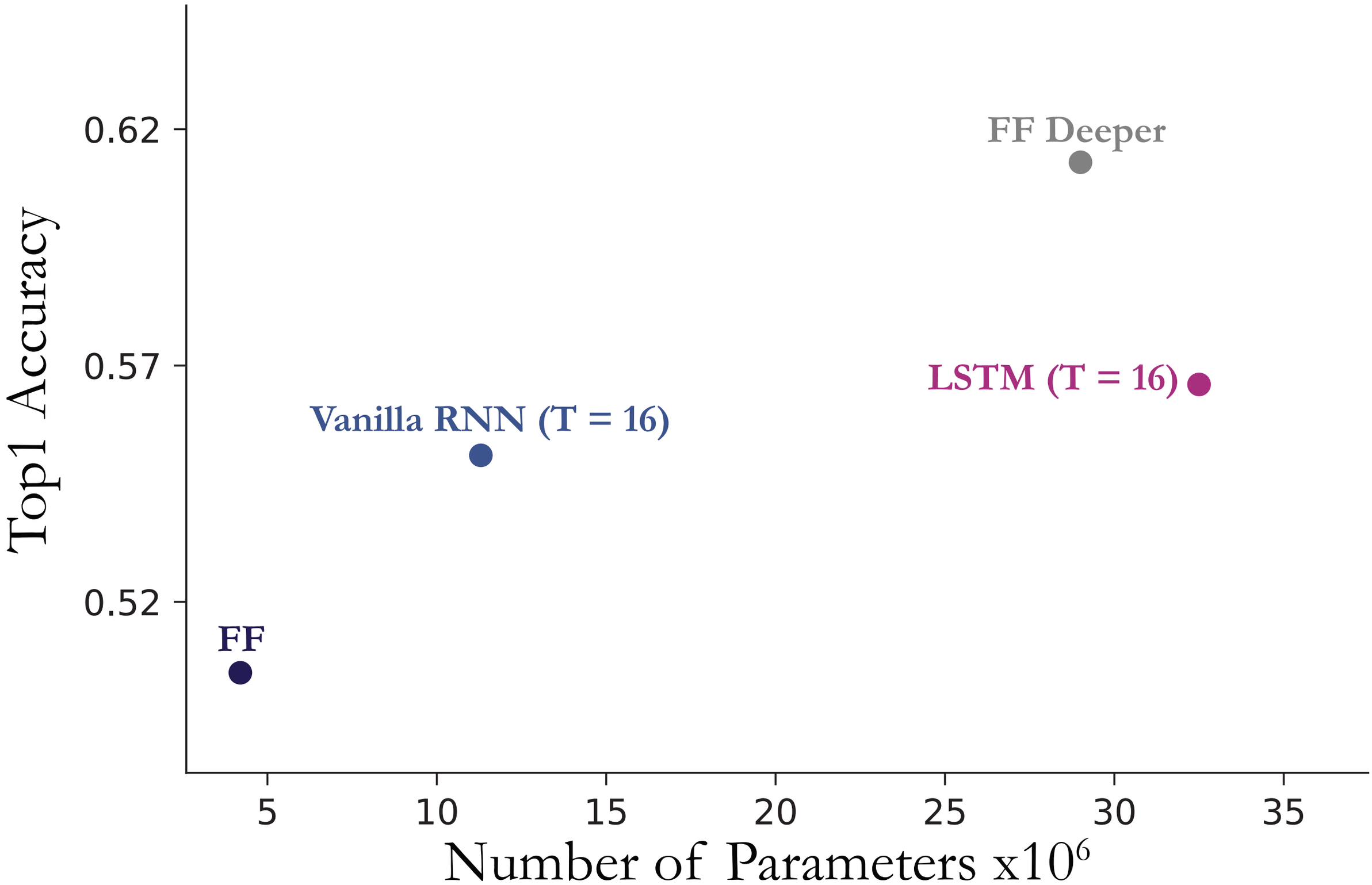
Two complementary principles:

(1) gating = multiplication by input-dependent tensor w/ values in [0, 1]

(2) bypassing = when recurrent cell is in 0 state, input is unchanged ("ResNet-like")

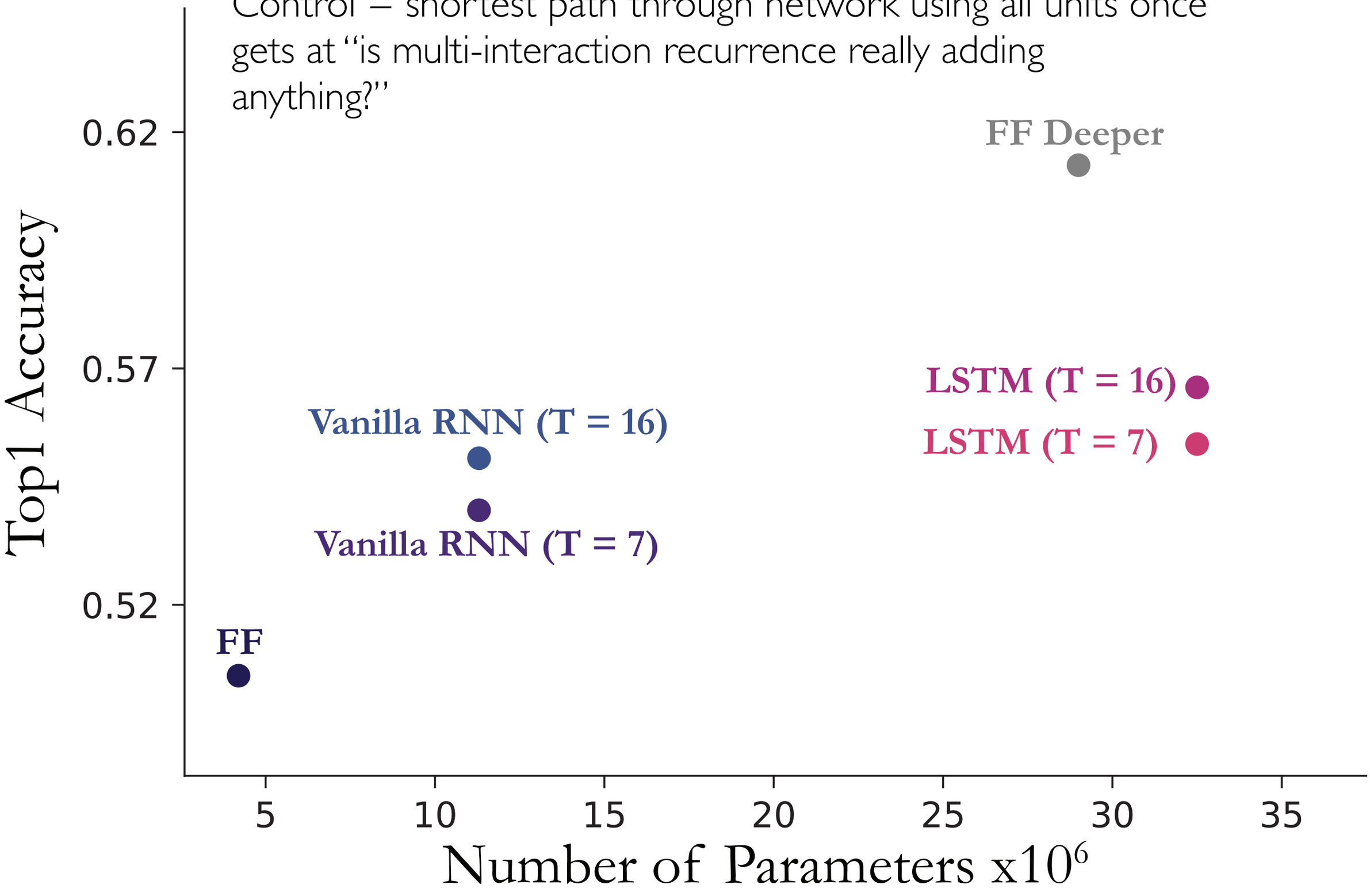
LSTM has **(1)** but not **(2)**;VanillaRNN has **(2)** but not **(1)**

Not All Local Recurrence is Equal

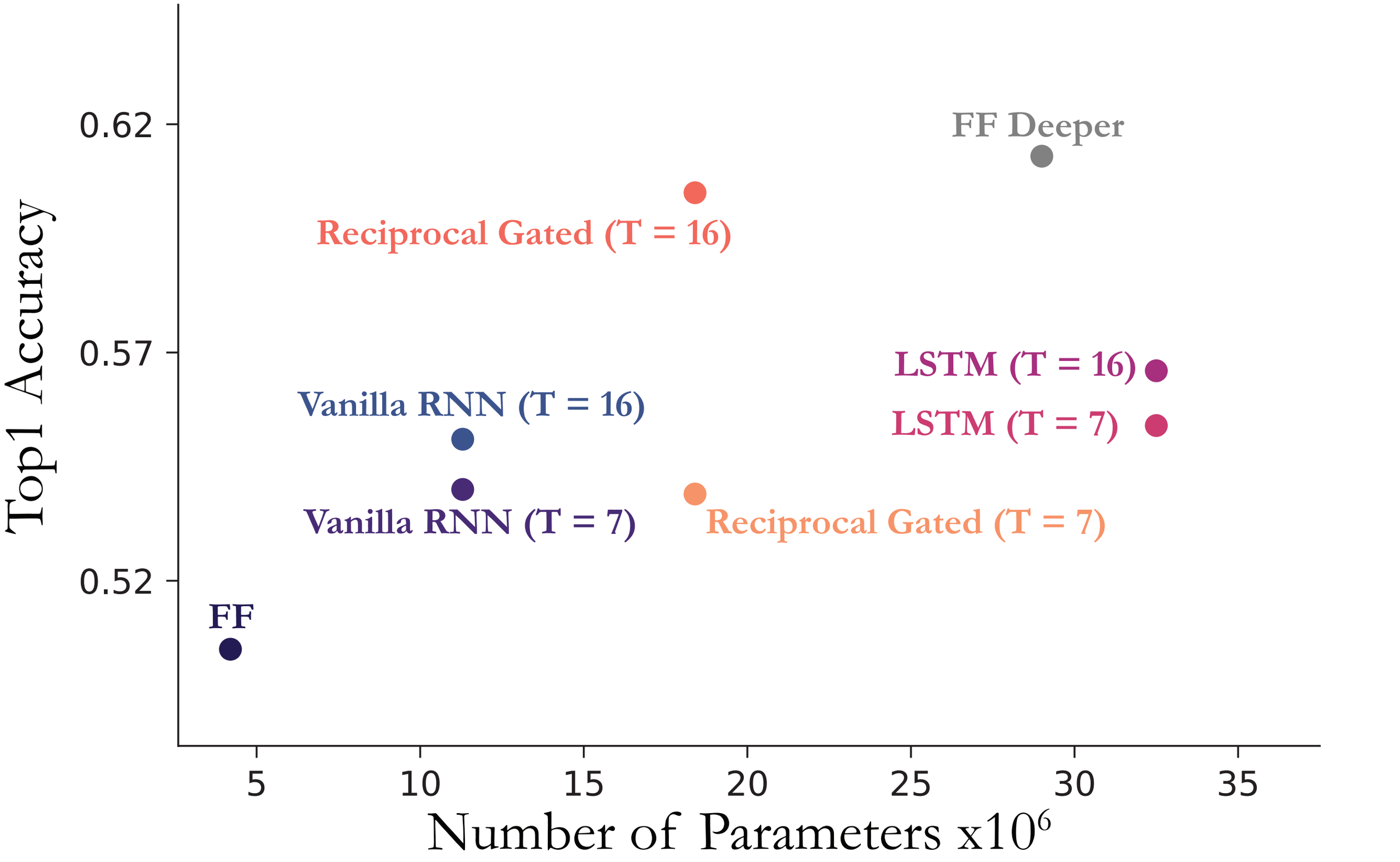


Not All Local Recurrence is Equal

Control = shortest path through network using all units once gets at "is multi-interaction recurrence really adding anything?"



Not All Local Recurrence is Equal



Large-Scale Search over Deep Recurrent Architectures

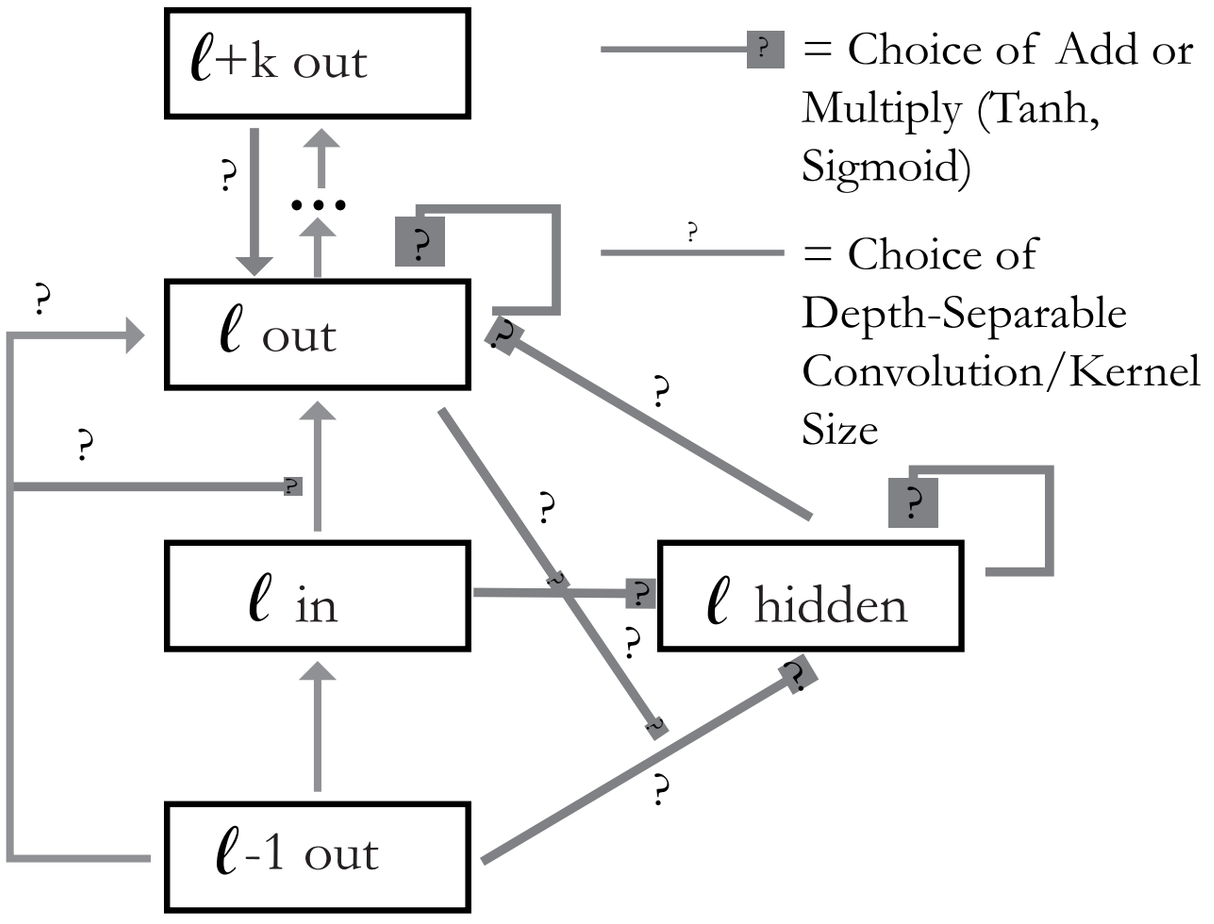
- Expand architecture class (local and global recurrence)
- Parametrize local and global feedback motifs and optimize for performance on ImageNet
- Evaluate neural predictivity in V4 and IT temporal responses

Nayebi, A.*, Bear, D.*, Kubilius, J.*, et al. *Task-Driven Convolutional Recurrent Models of the Visual System*, NeurIPS 2018

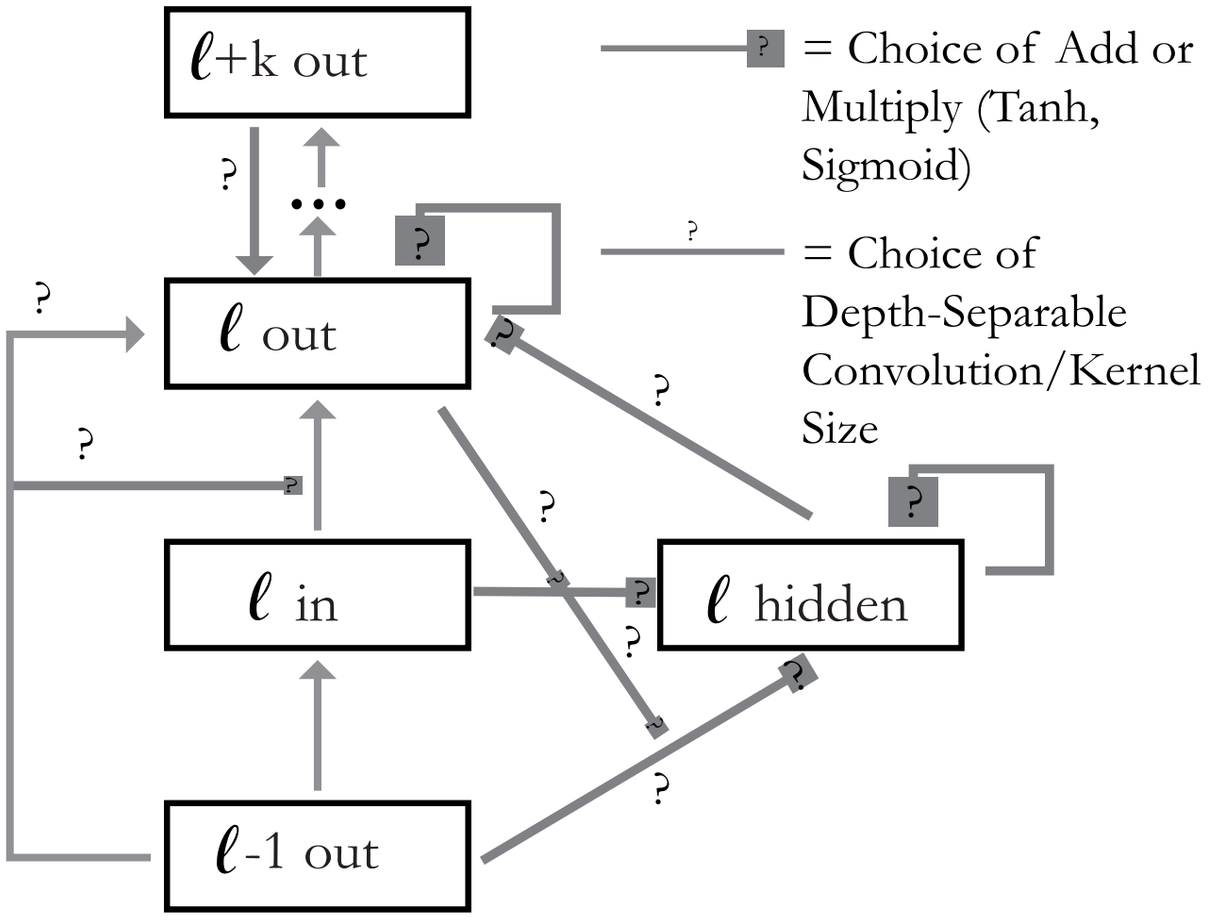
Daniel Bear
(Stanford)

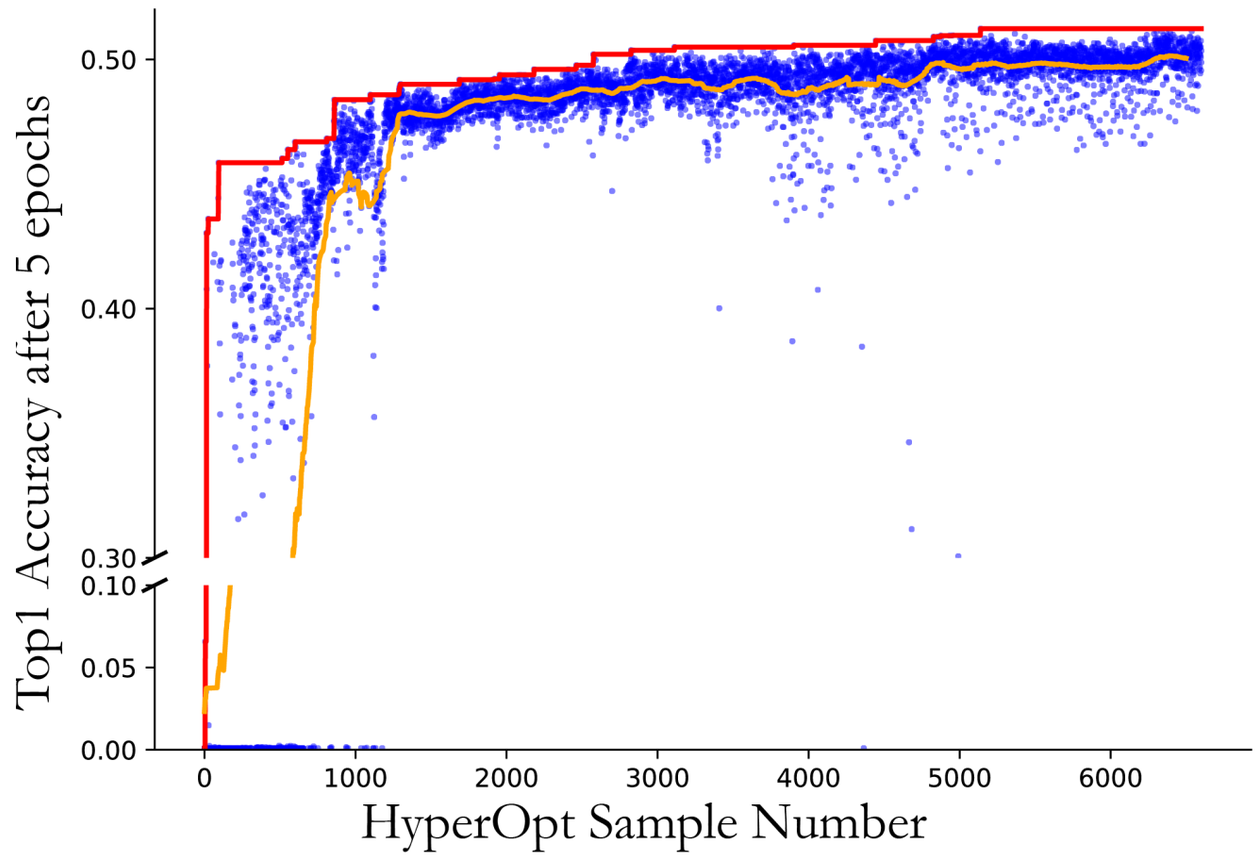
Jonas Kubilius
(MIT)

Search Over Local and Global Recurrence

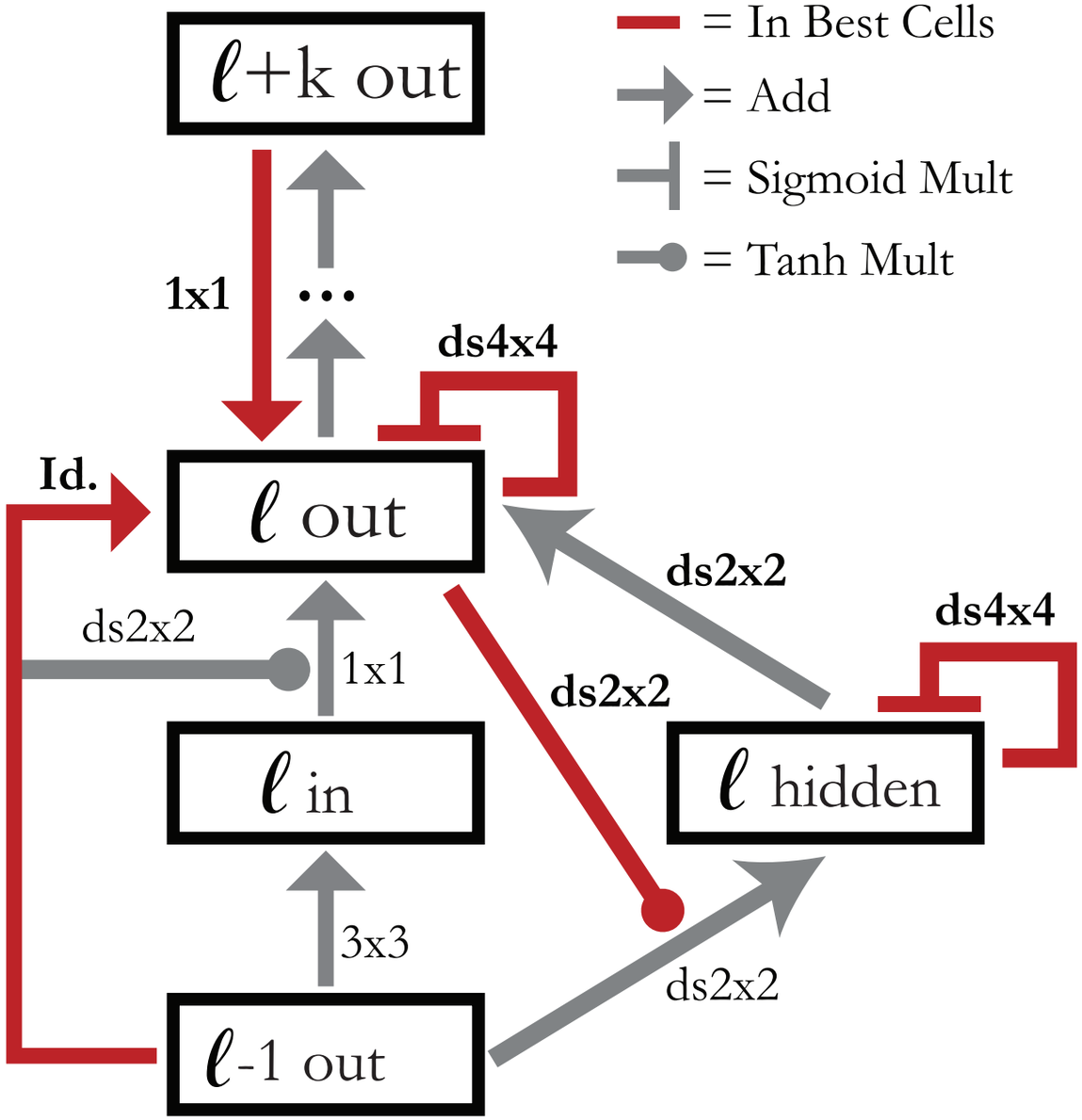


Search Over Local and Global Recurrence

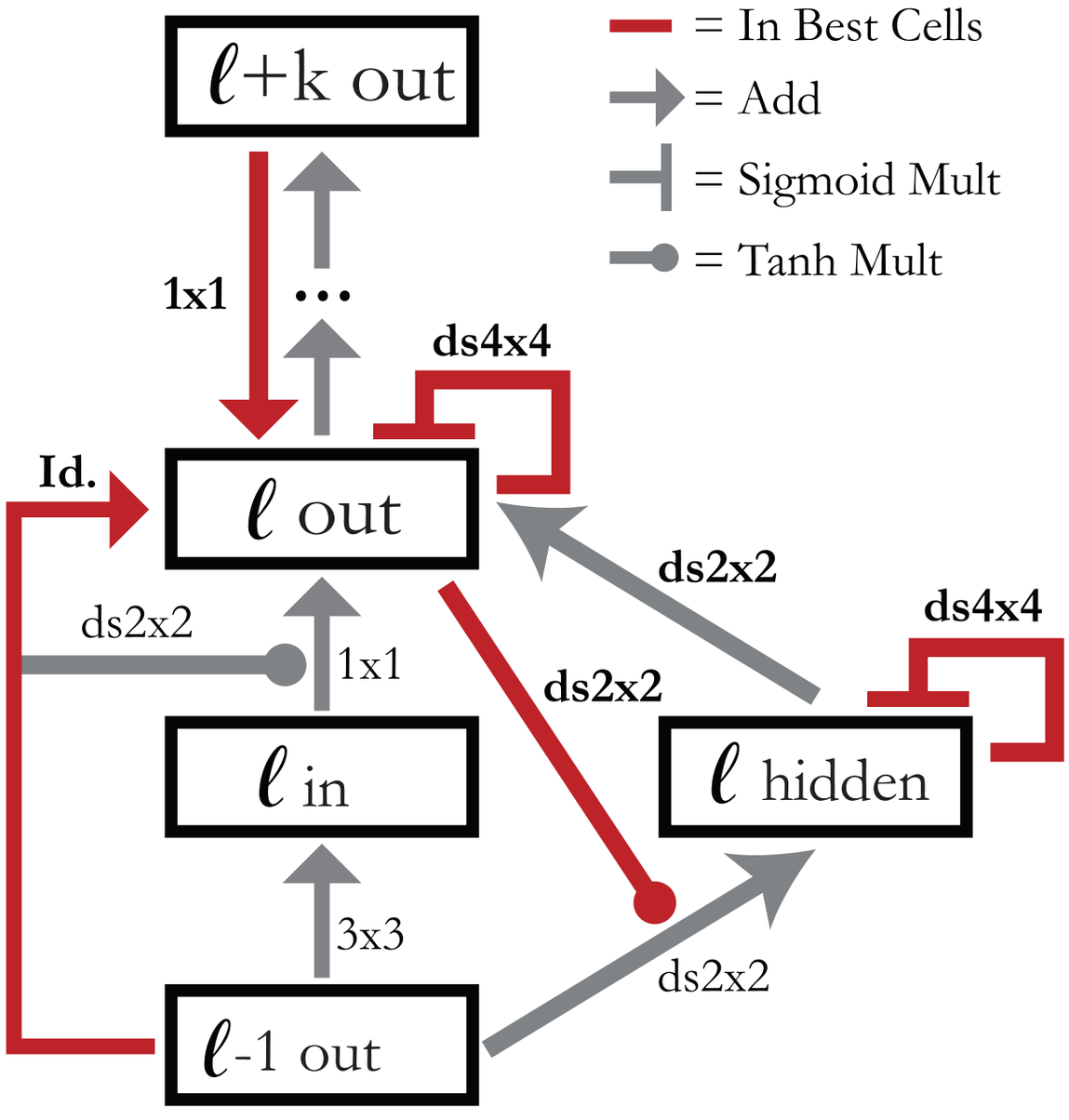


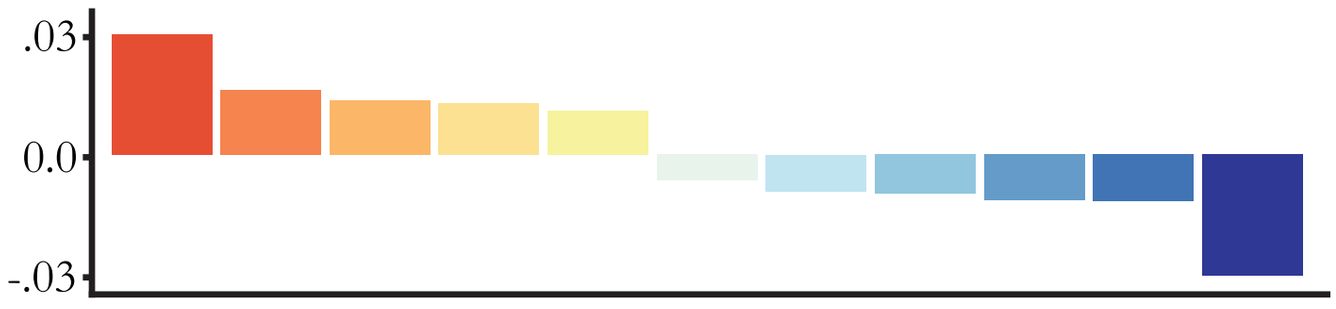


Emergent Local and Global Connectivity Patterns



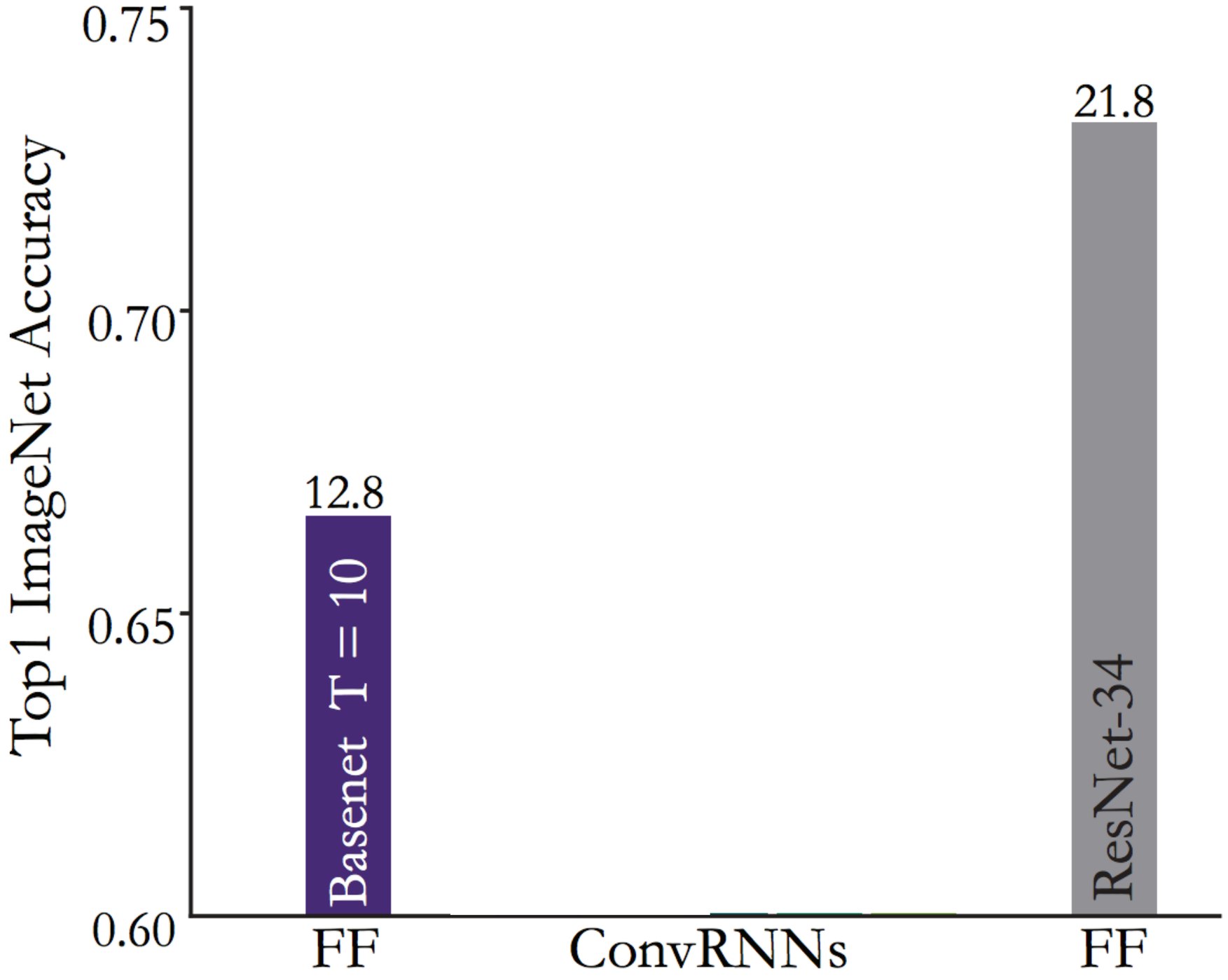
Emergent Local and Global Connectivity Patterns



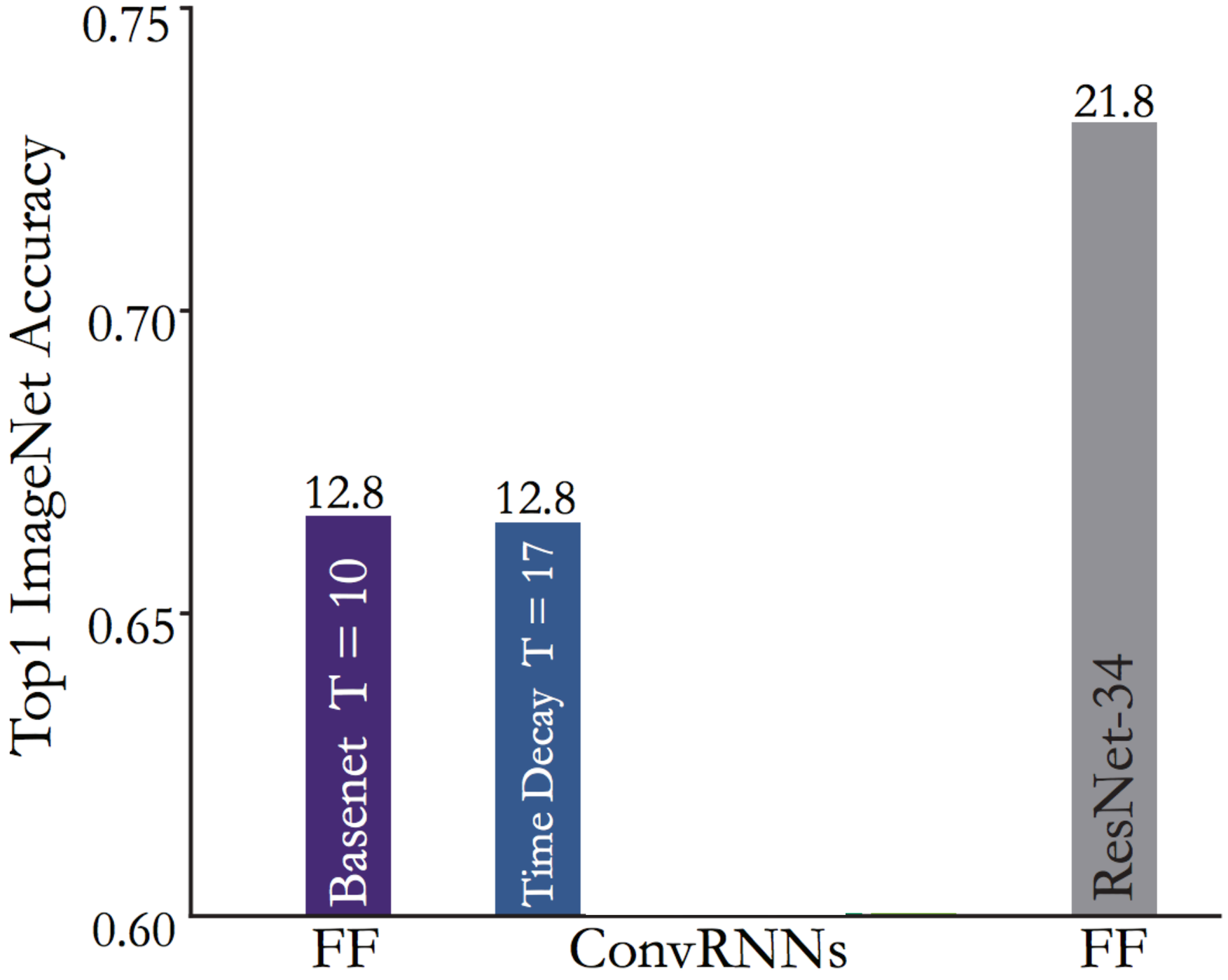


Top1 Gain from Feedback (Best to Worst)

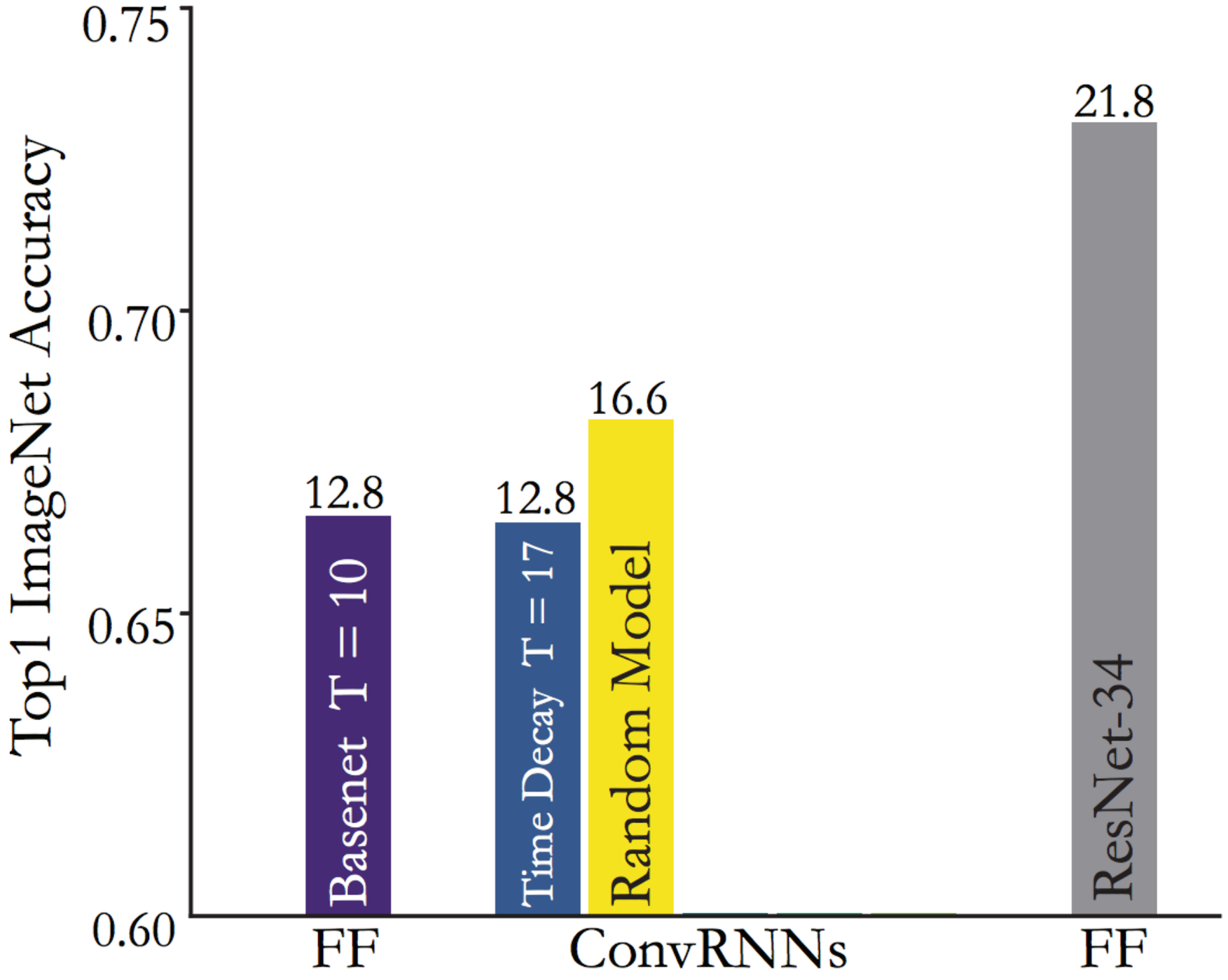
Improving ImageNet Performance with ConvRNNs



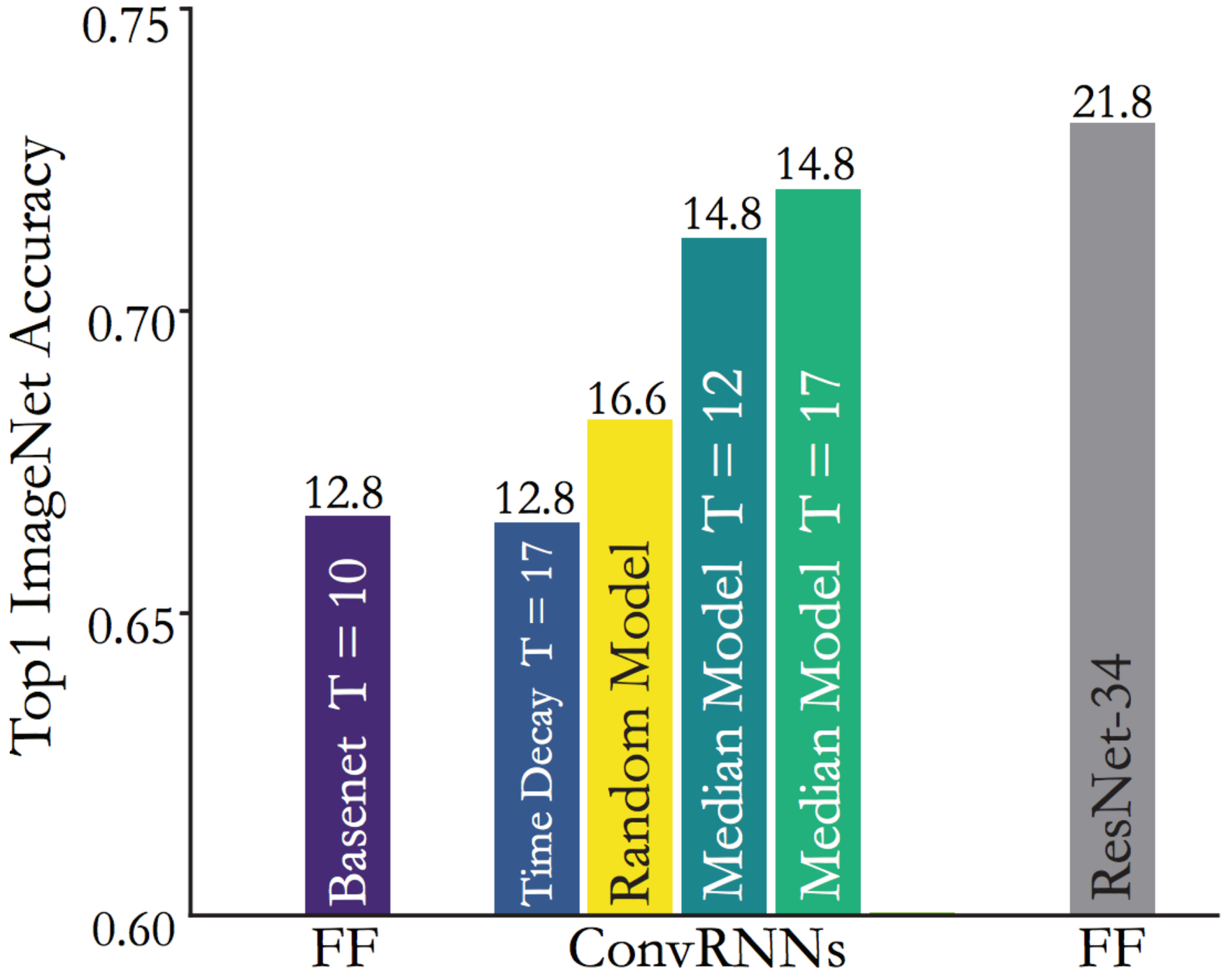
Improving ImageNet Performance with ConvRNNs



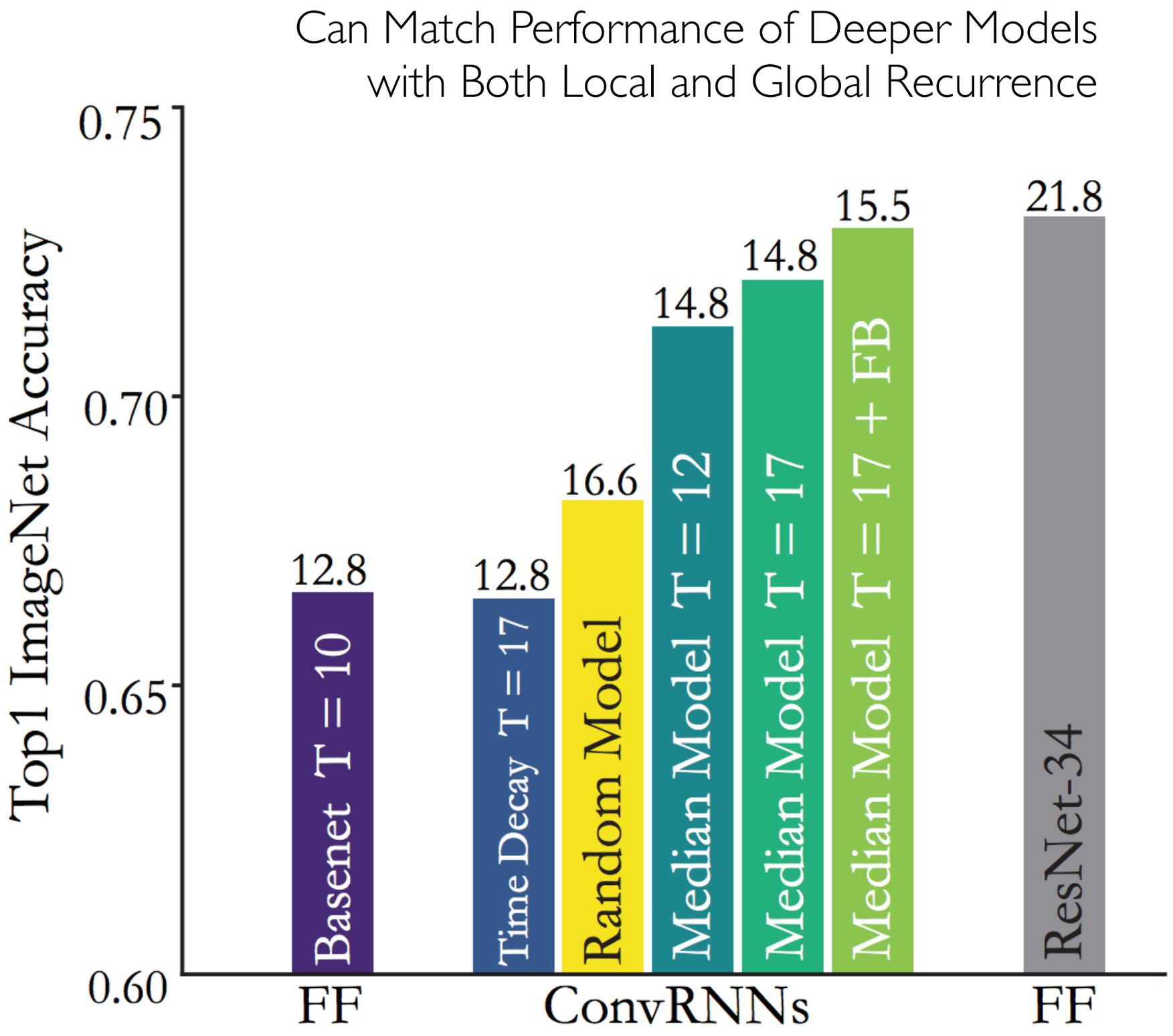
Improving ImageNet Performance with ConvRNNs



Improving ImageNet Performance with ConvRNNs



Improving ImageNet Performance with ConvRNNs



Role of Recurrence in Core Object Recognition

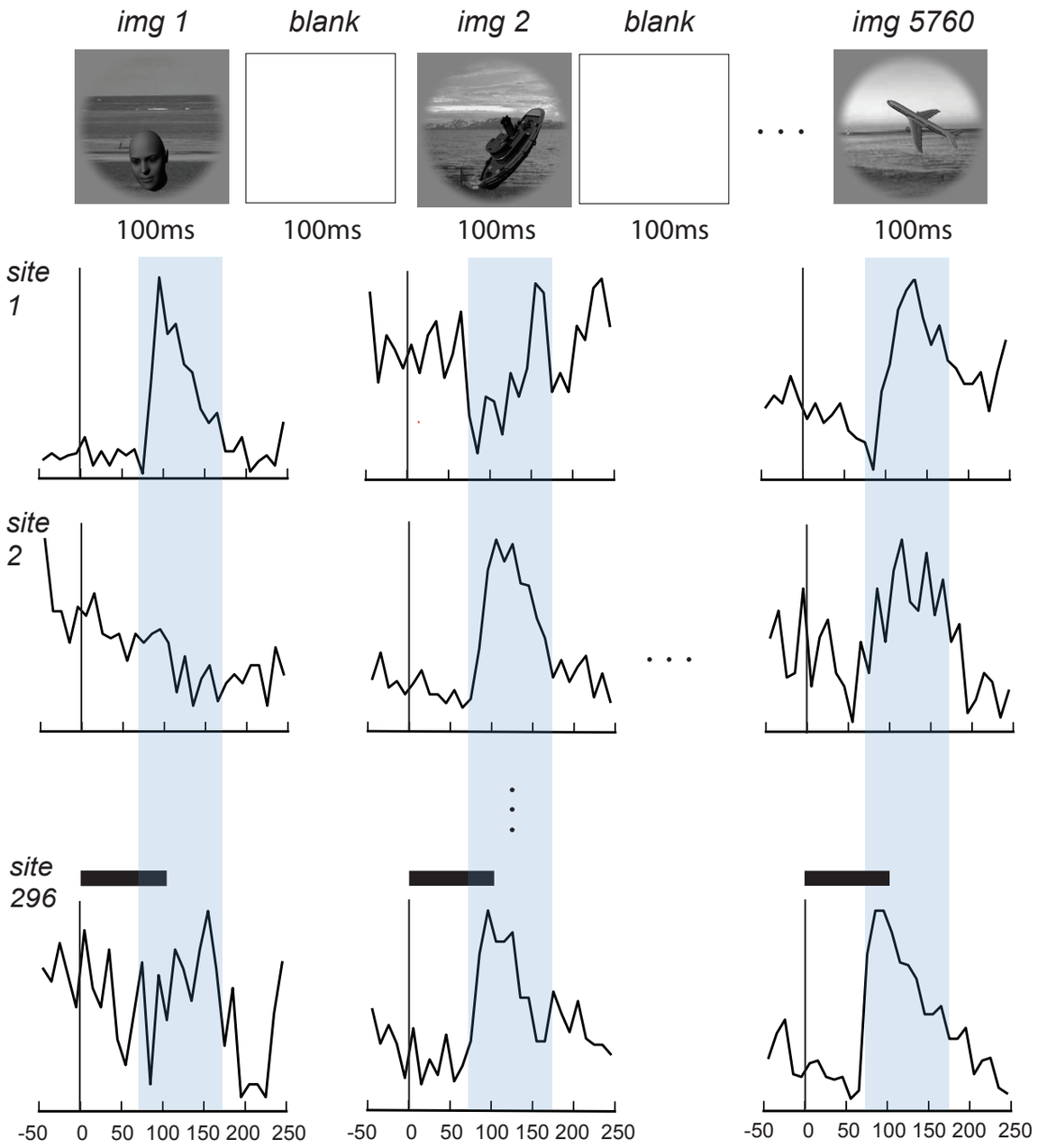
Nayebi, A.*, Bear, D.*, Kubilius, J.*, et al. *Task-Driven Convolutional Recurrent Models of the Visual System*, NeurIPS 2018

- Expand architecture class (local and global recurrence)
- Parametrize local and global feedback motifs and optimize for performance on ImageNet
- Evaluate neural predictivity in V4 and IT temporal responses

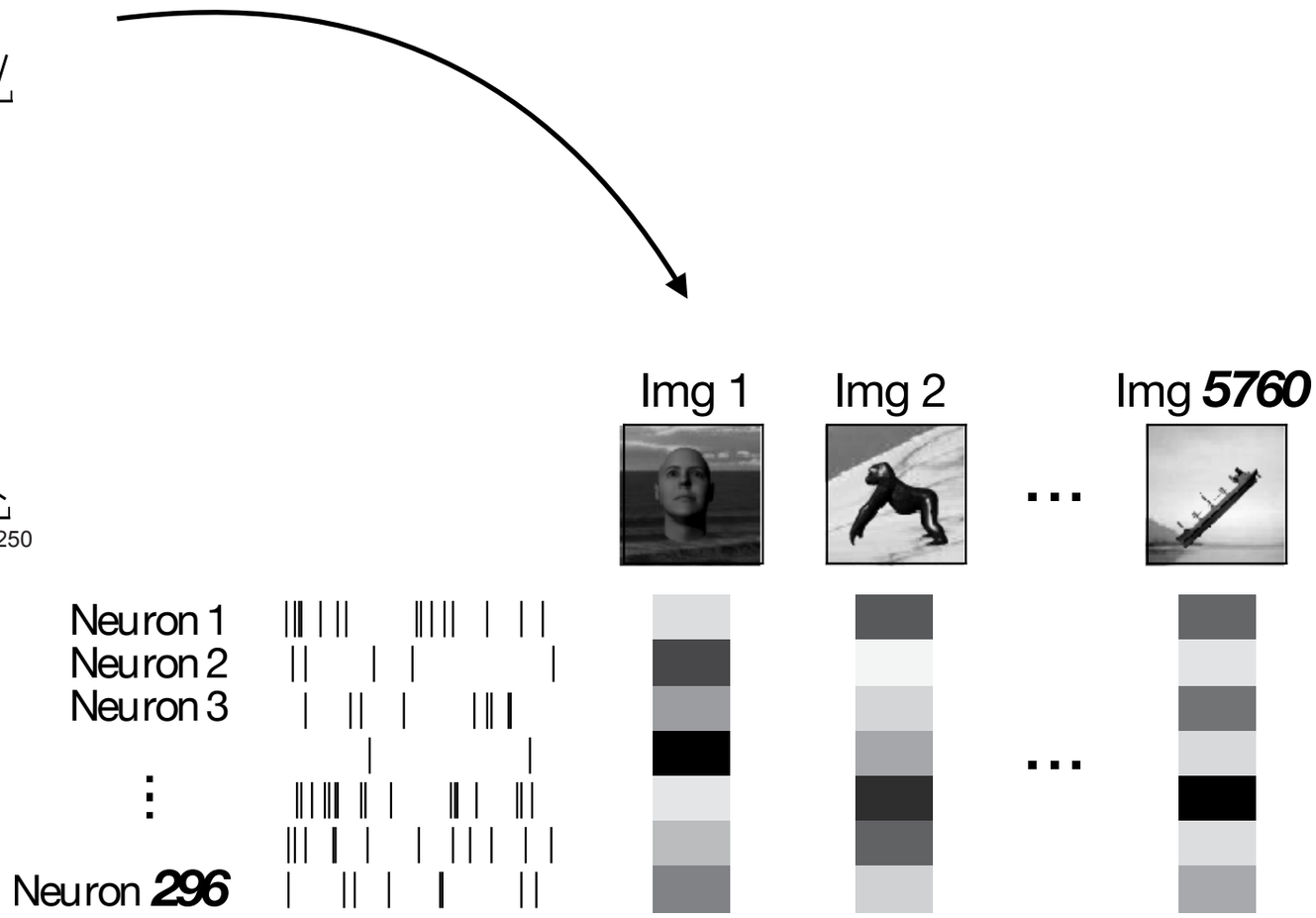
Daniel Bear
(Stanford)

Jonas Kubilius
(MIT)

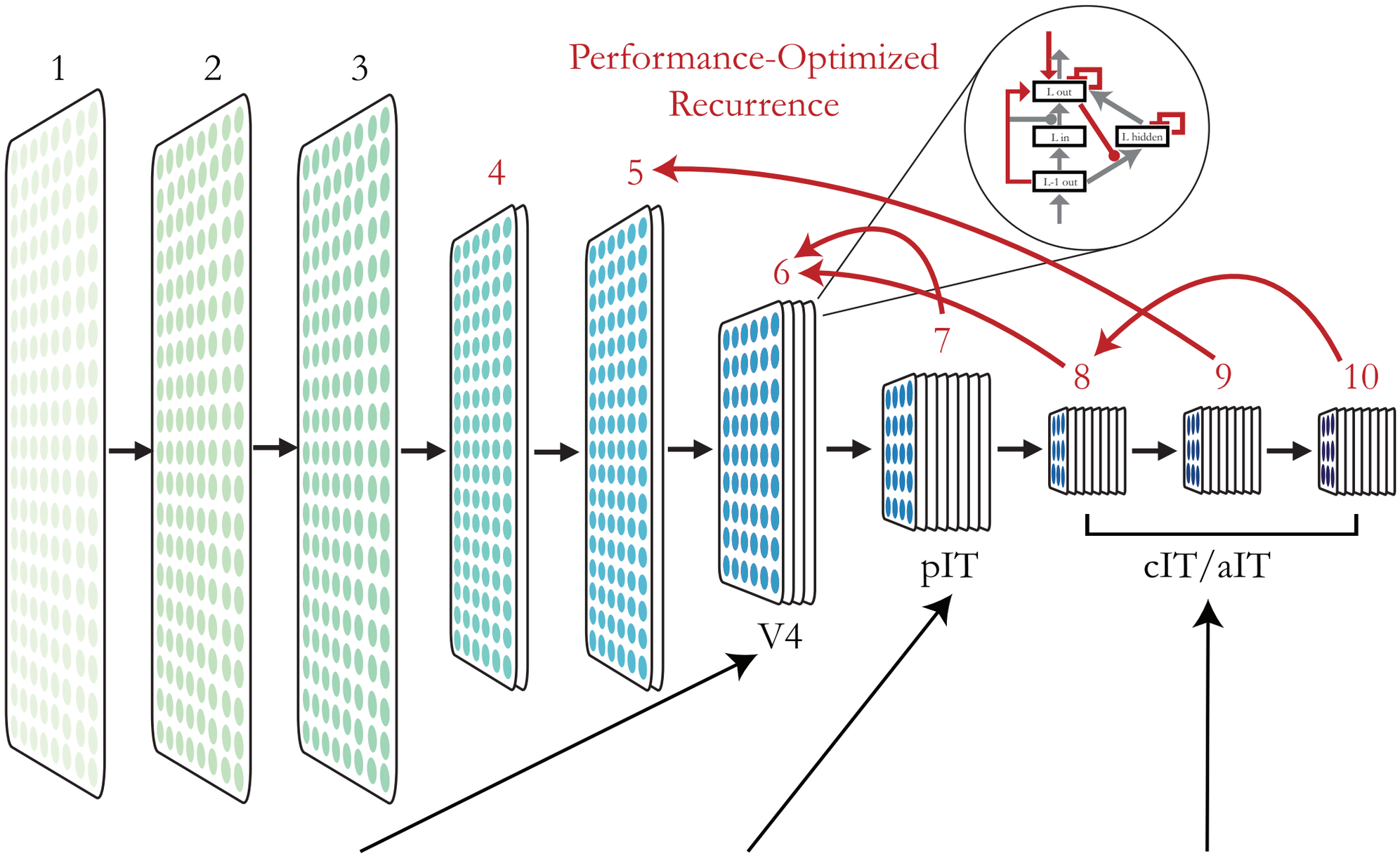
Neural Predictivity with ConvRNNs



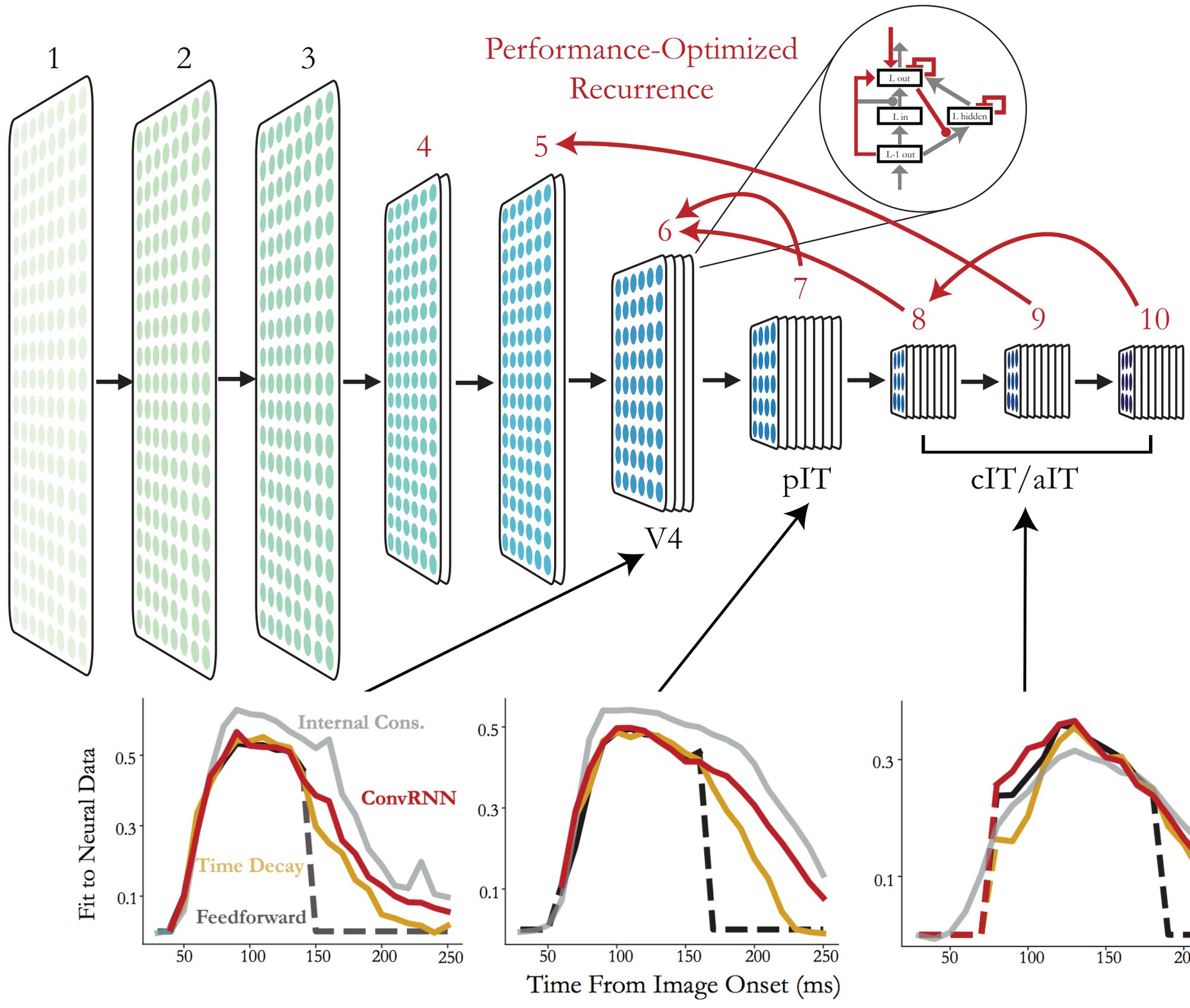
Binned spike counts 70ms-170ms post stimulus presentation



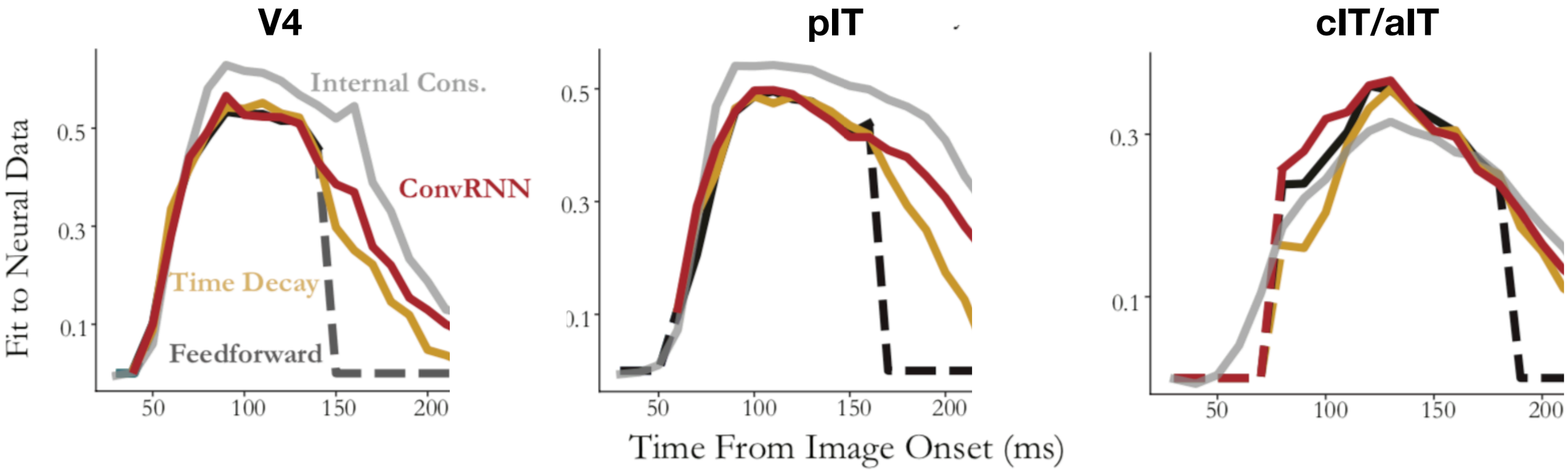
Neural Fit with ConvRNNs



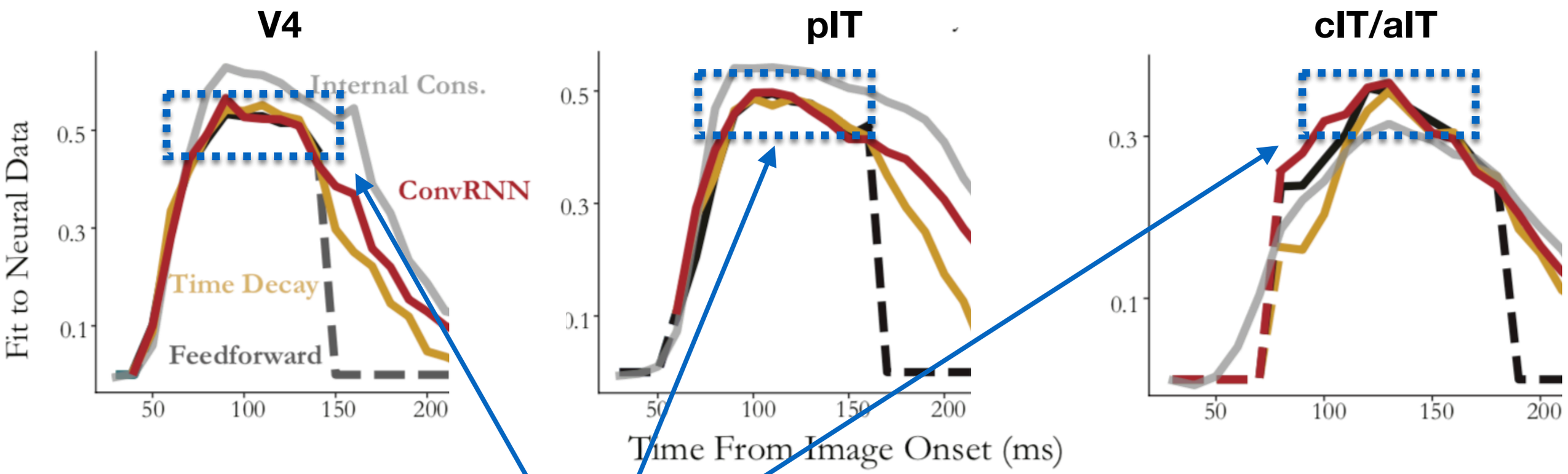
ConvRNNs explain entire temporal trajectory



ConvRNNs explain entire temporal trajectory



But no better than feedforward models during stimulus



ConvRNNs (red) are comparable to feedforward models (black) when stimulus is on

Is recurrence useful during inference?

- Feedback connections primarily used during inference?

- Feedback connections primarily used for propagating error signals?

Is recurrence useful during inference? Maybe not.

- **Feedback connections primarily used during inference?**

Currently not much evidence in the affirmative.
More consistent with the hypothesis that feedback connections allow a shallower network to "approximate" a deeper network that could not otherwise physically fit.

- **Feedback connections primarily used for propagating error signals?**

Is recurrence useful during learning?

- **Feedback connections primarily used during inference?**

Currently not much evidence in the affirmative.
More consistent with the hypothesis that feedback connections allow a shallower network to "approximate" a deeper network that could not otherwise physically fit.

- Feedback connections primarily used for propagating error signals?

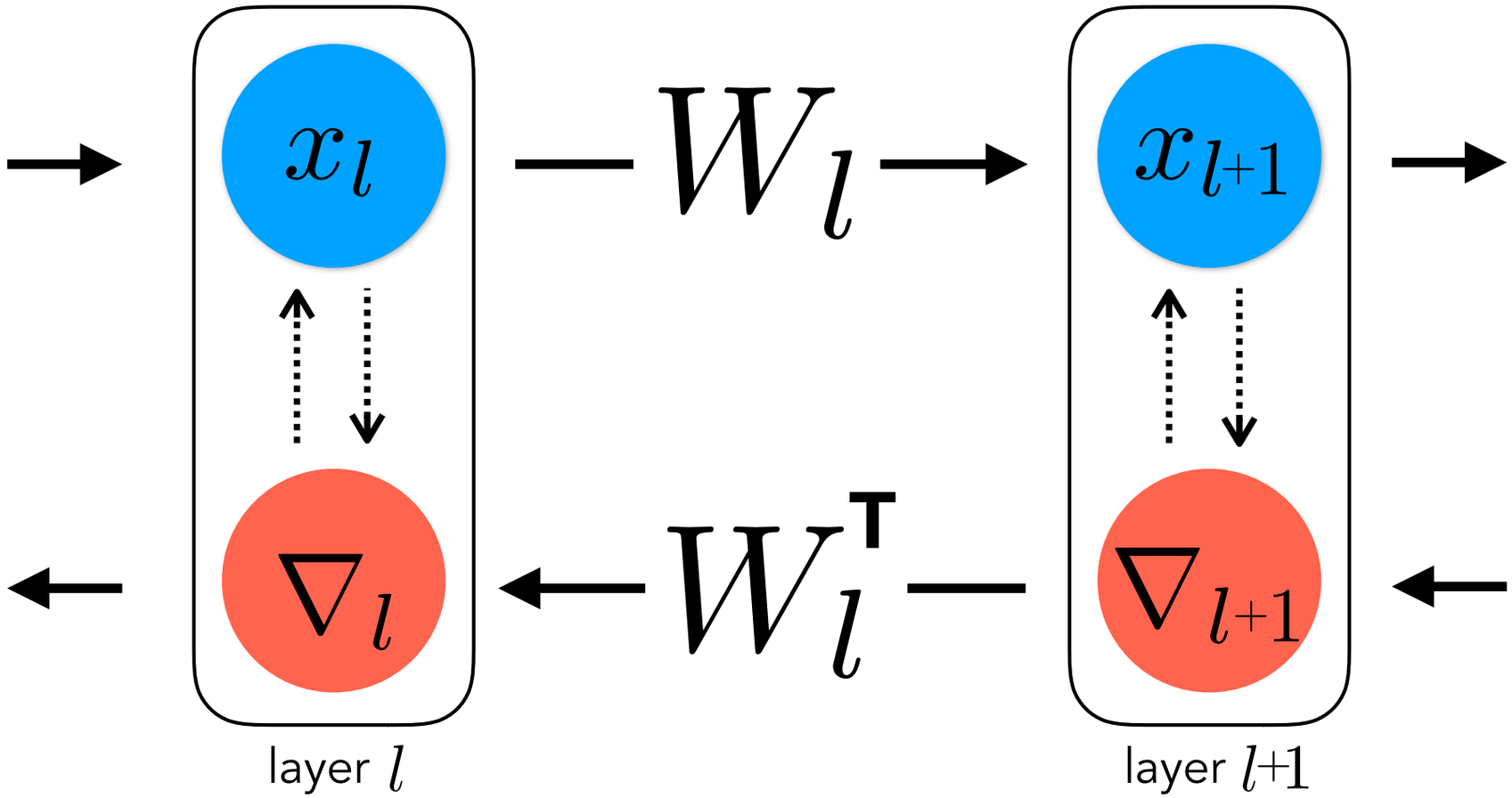
The problems with backpropagation

- Requires the derivatives of the activation functions.
- Requires separate forward and backward passes.
- The weight updates require access to transposes of the feedforward weights.

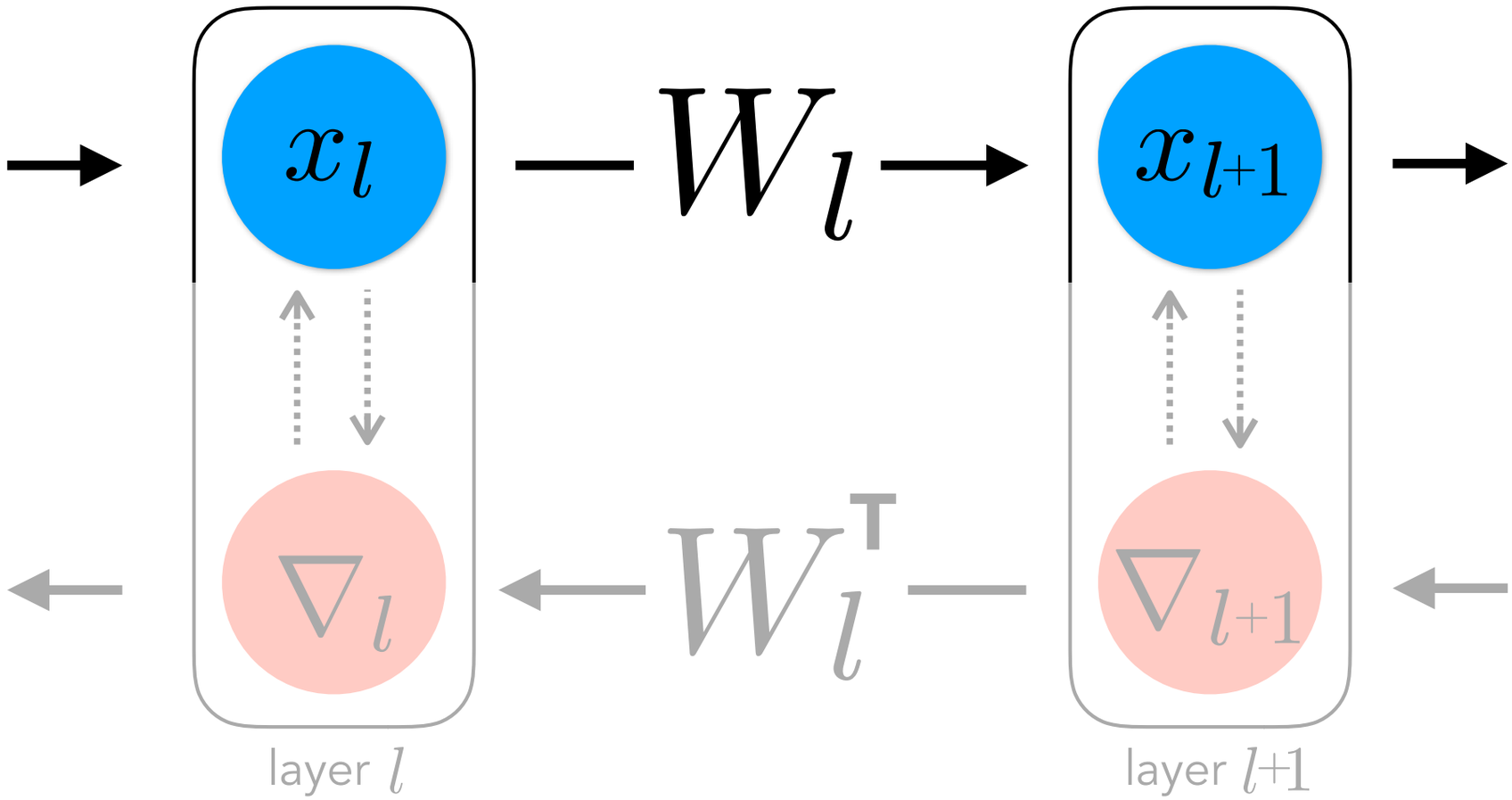
Weight Symmetry Problem

- Requires the derivatives of the activation functions.
- Requires separate forward and backward passes.
- **The weight updates require access to transposes of the feedforward weights.**

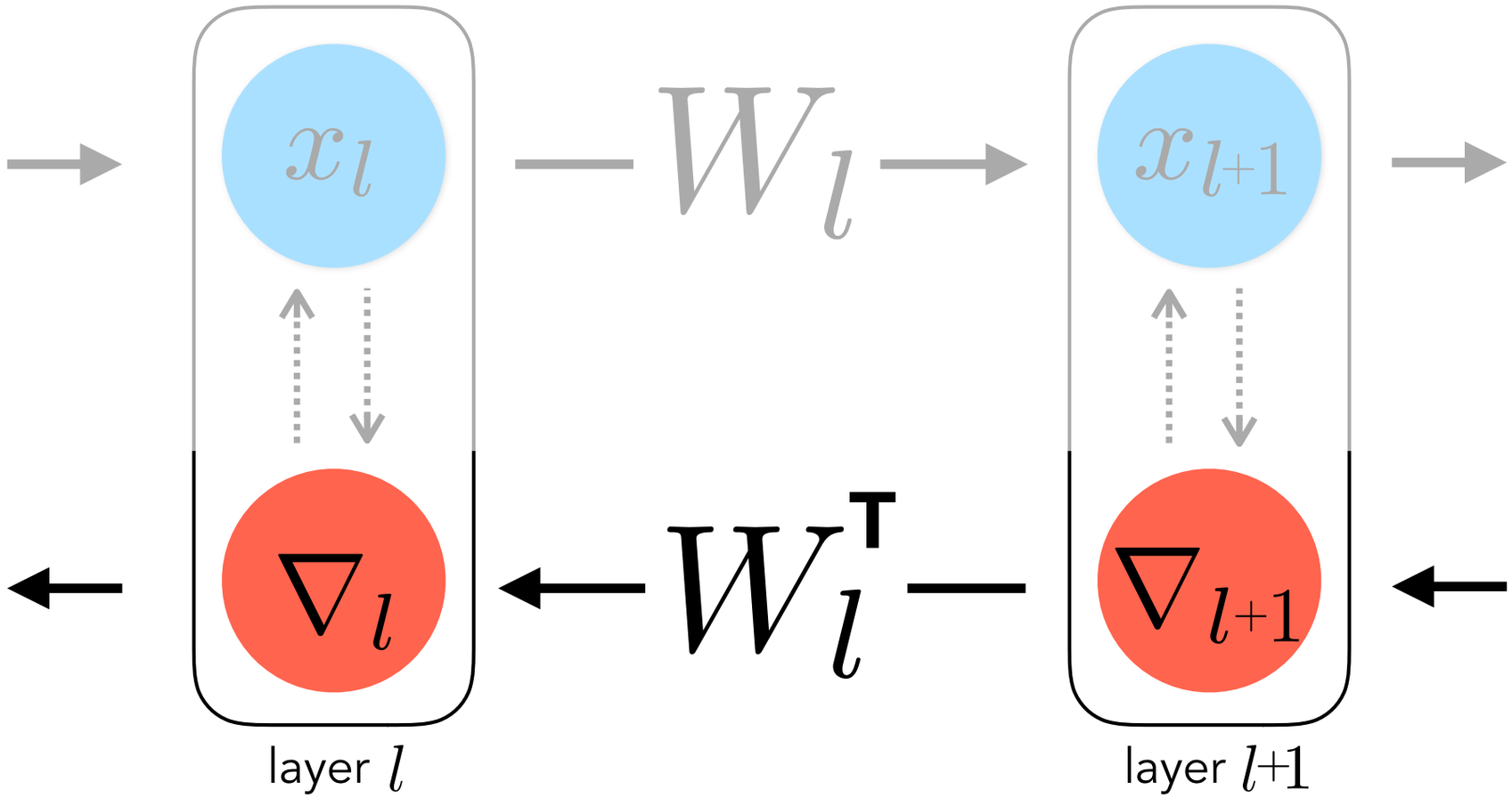
Breaking the backpropagation weight symmetry constraint



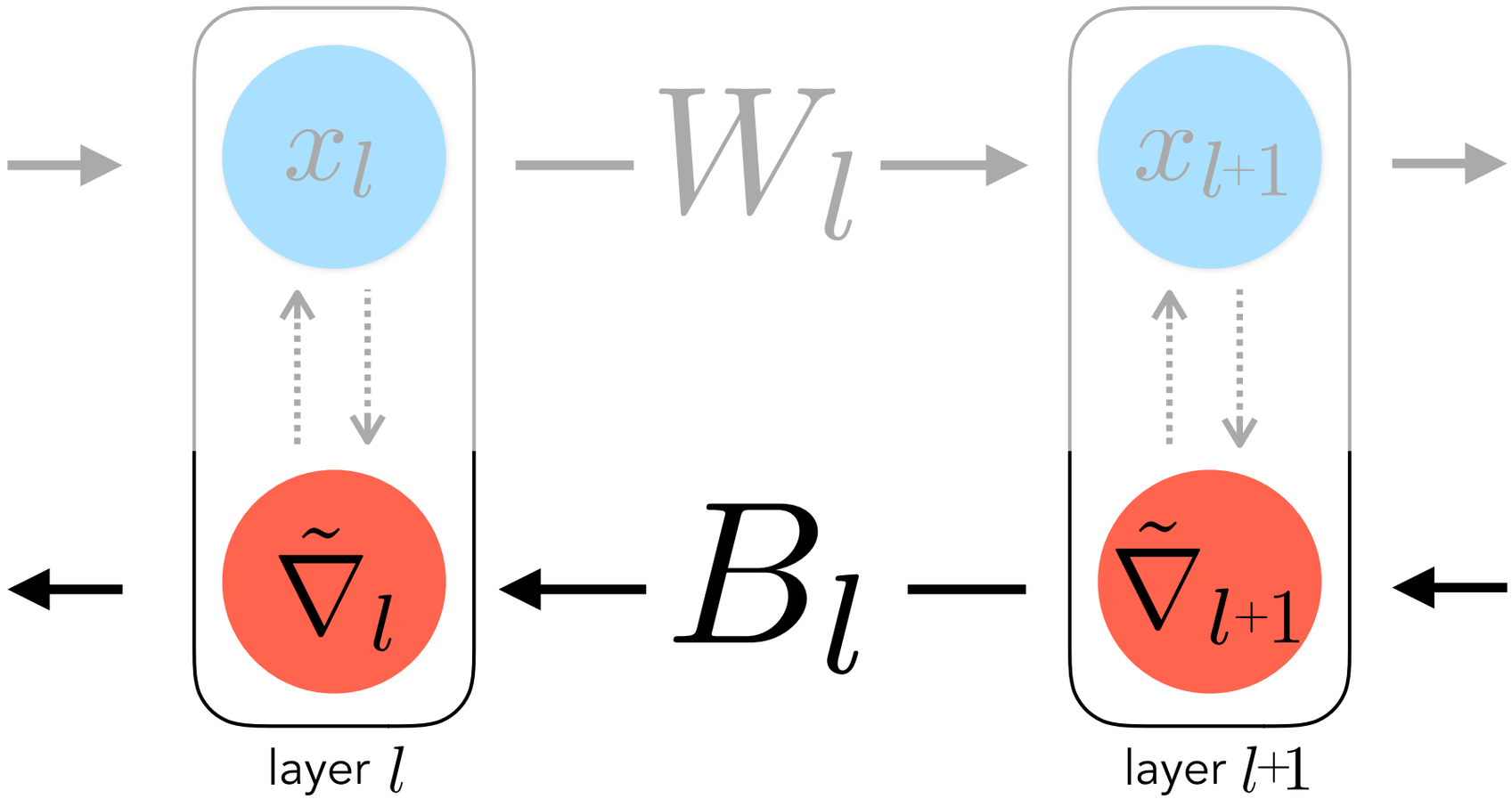
Breaking the backpropagation weight symmetry constraint



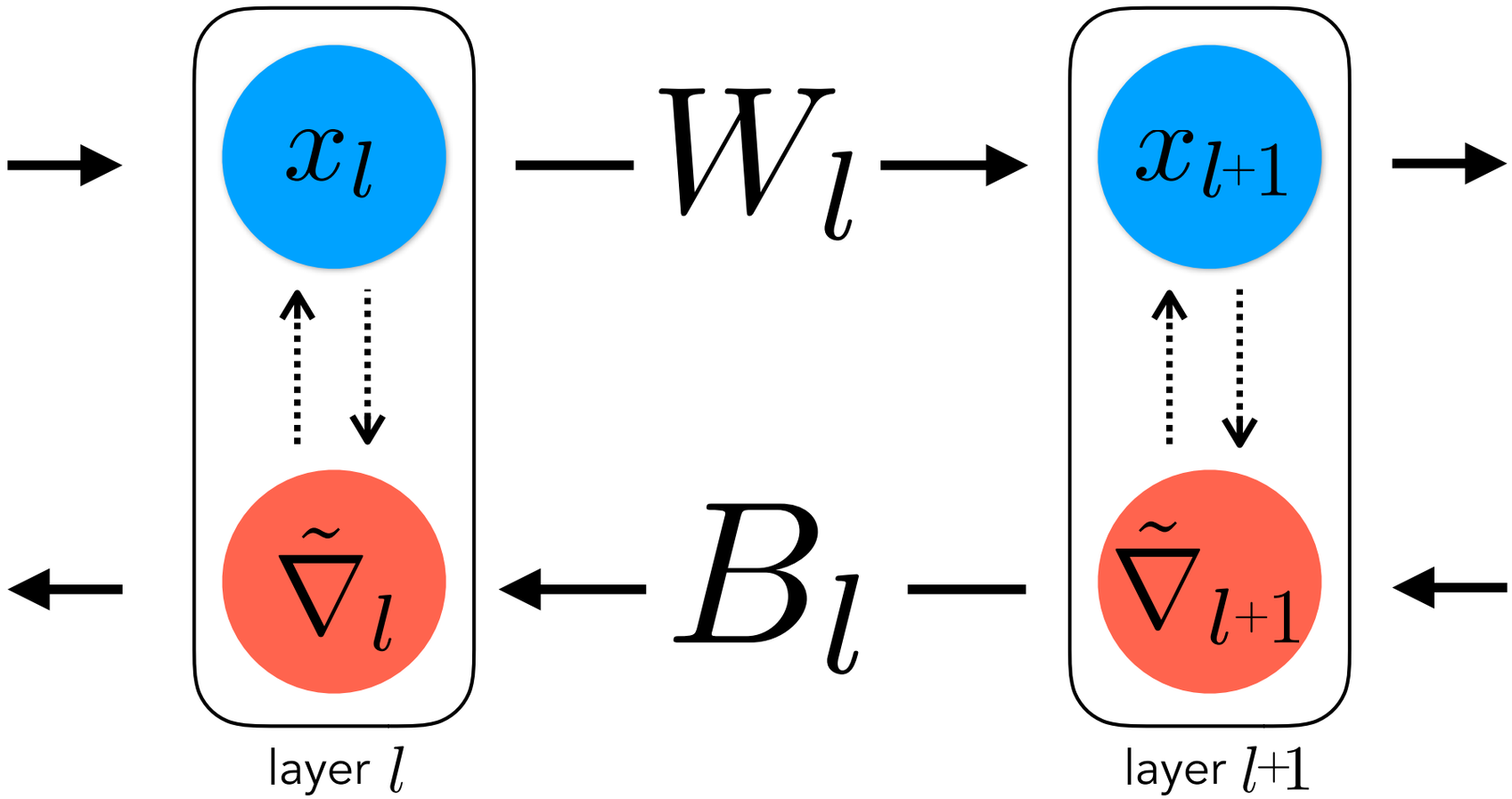
Breaking the backpropagation weight symmetry constraint



Breaking the backpropagation weight symmetry constraint



Breaking the backpropagation weight symmetry constraint



- What should the dynamics on the backward weights be?

Feedback Alignment

- Some previous proposals:
  - **Feedback Alignment [1]:** no dynamics. B is fixed, random

$$\Delta B_l = 0$$

[1] Lillicrap, Timothy P., et al. "Random synaptic feedback weights support error backpropagation for deep learning." *Nature communications* 7.1 (2016): 1-10.
[2] Akrout, Mohamed, et al. "Deep learning without weight transport." *Advances in Neural Information Processing Systems*. 2019.
[3] Kolen, John F., and Jordan B. Pollack. "Backpropagation without weight transport." *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*. Vol. 3. IEEE, 1994.

Comparing Feedback Alignment to Backprop

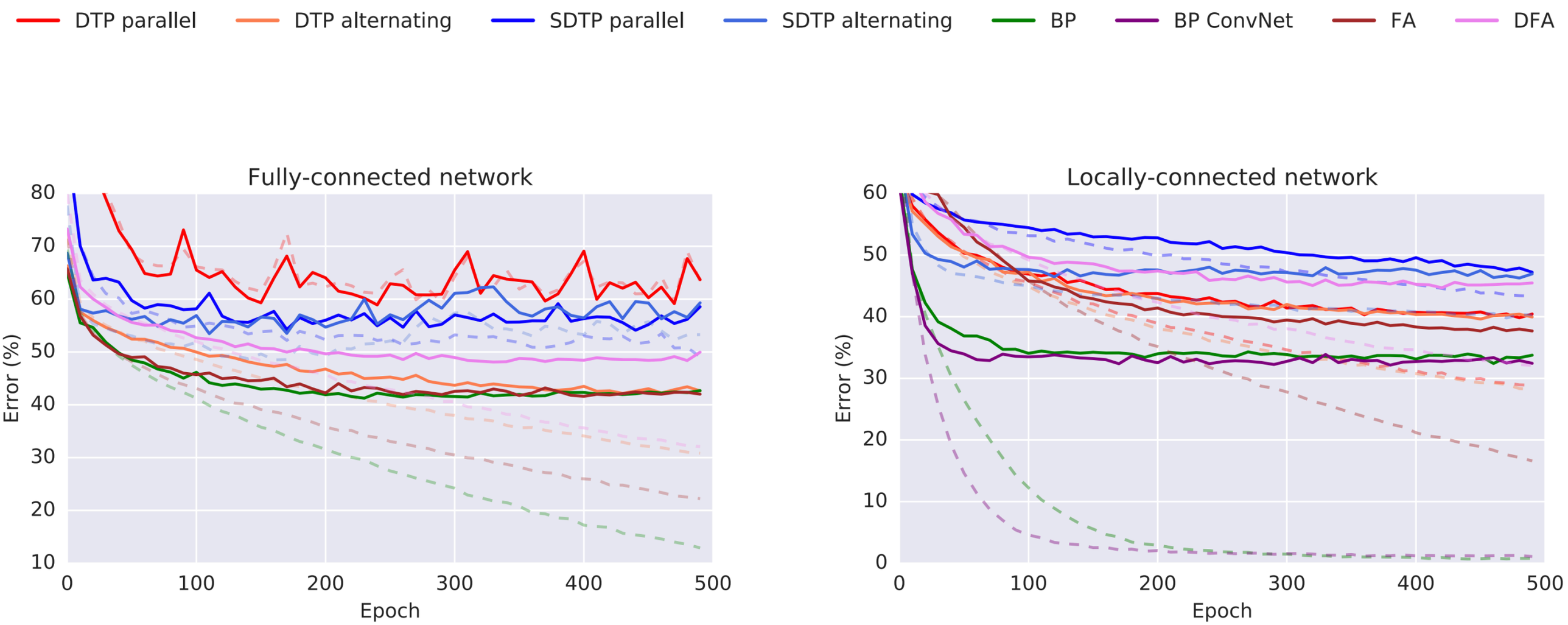


Figure 2: Train (dashed) and test (solid) classification errors on CIFAR.

Bartunov et al.
NeurIPS *(2018)*

Scales as Backprop does on simple tasks

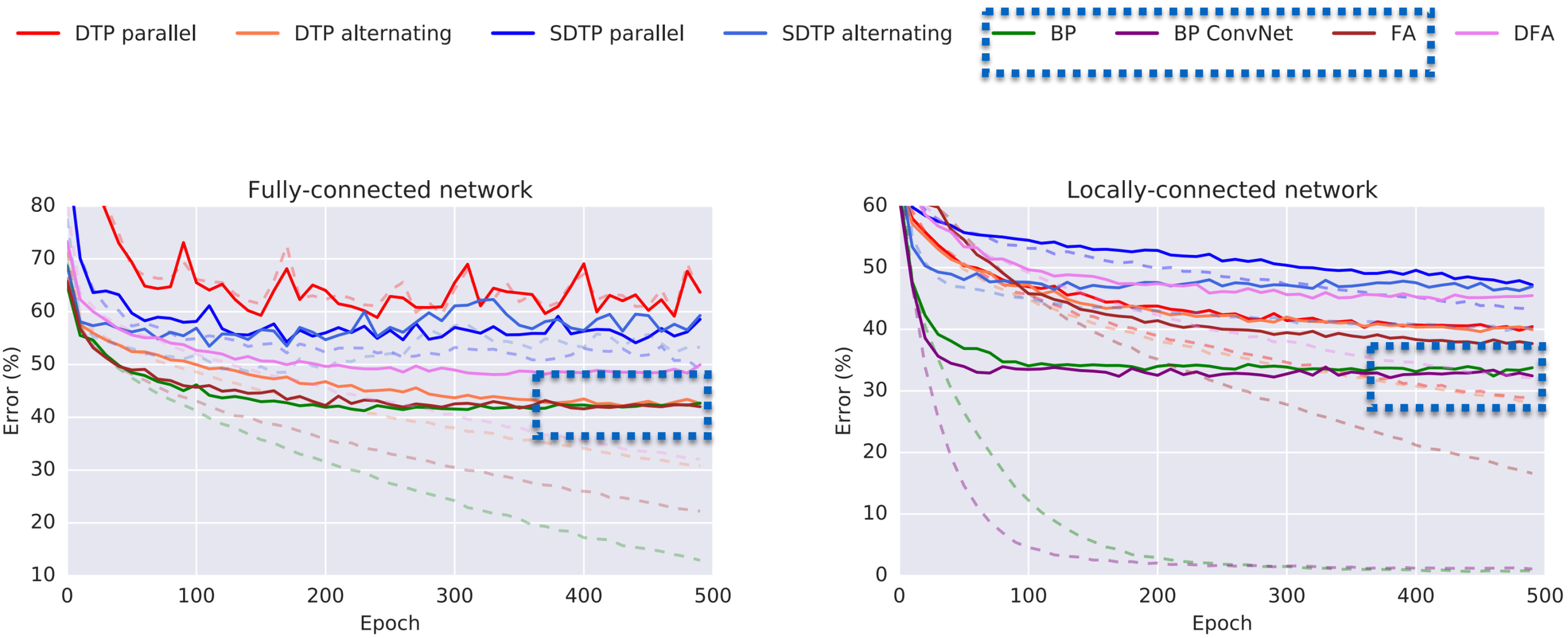
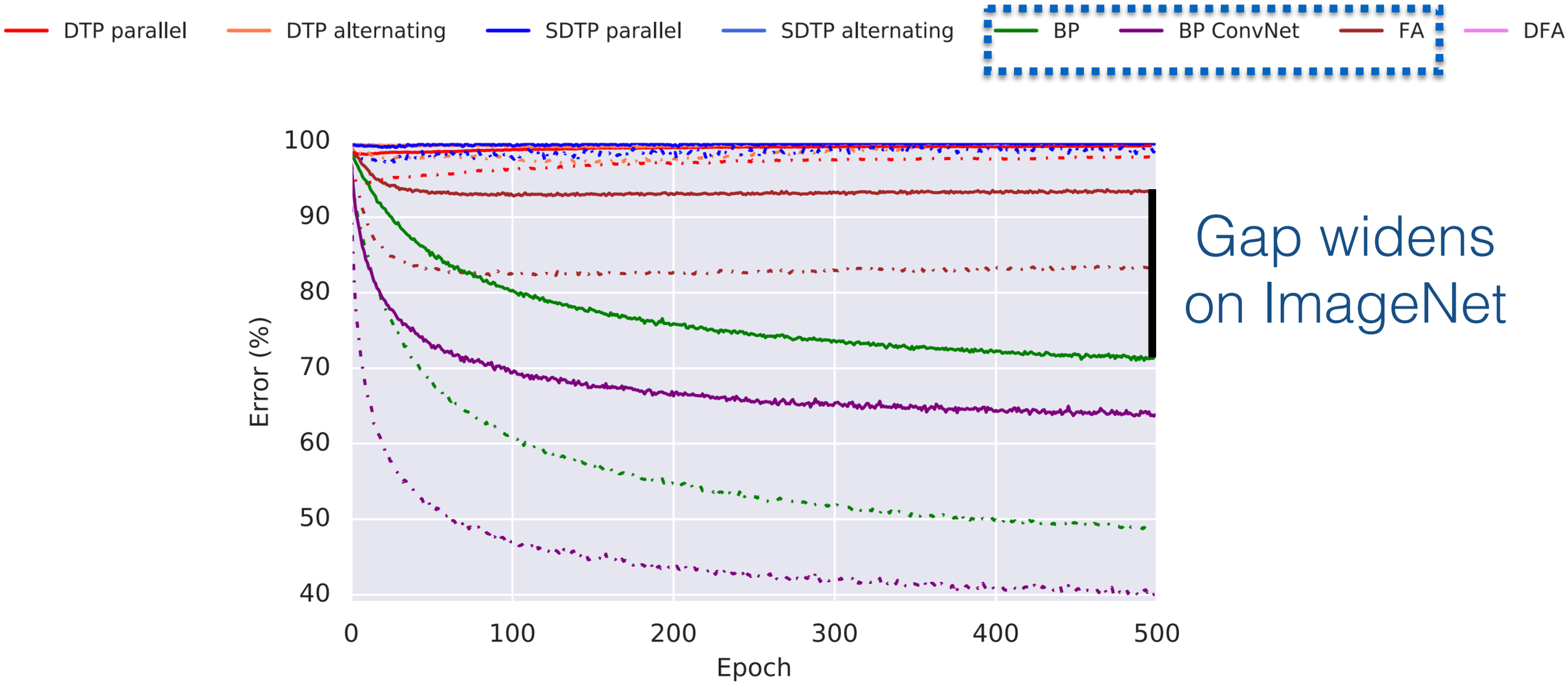


Figure 2: Train (dashed) and test (solid) classification errors on CIFAR.

Similar performance between FA and BP on small tasks.

Bartunov et al.
***NeurIPS** (2018)*

Does *not* scale as Backprop does on harder tasks



Gap widens on ImageNet

Figure 3: Top-1 (solid) and Top-5 (dotted) test errors on ImageNet. Color legend is the same as for figure 2.

Bartunov et al.
***NeurIPS** (2018)*

What can we do to fix this issue?

- Some previous proposals:
  - **Feedback Alignment [1]:** no dynamics. B is fixed, random

$$\Delta B_l = 0$$

[1] Lillicrap, Timothy P., et al. "Random synaptic feedback weights support error backpropagation for deep learning." *Nature communications* 7.1 (2016): 1-10.
[2] Akrout, Mohamed, et al. "Deep learning without weight transport." *Advances in Neural Information Processing Systems*. 2019.
[3] Kolen, John F., and Jordan B. Pollack. "Backpropagation without weight transport." *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*. Vol. 3. IEEE, 1994.

Imposing dynamics on the backwards weights

- Some previous proposals:
  - **Feedback Alignment [1]:** no dynamics. B is fixed, random

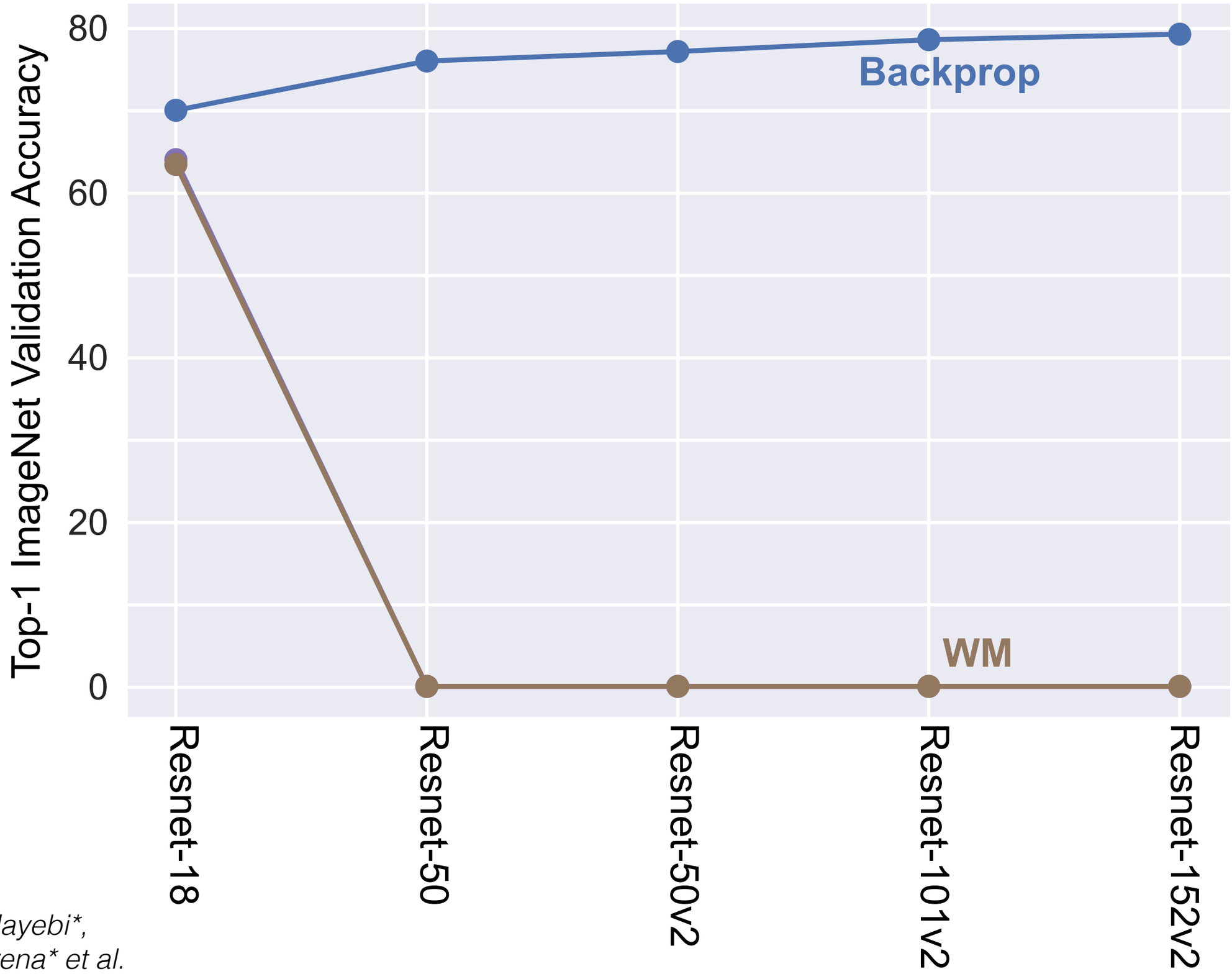
$$\Delta B_l = 0$$

  - **Weight Mirror [2]:** feedforward neurons noisily discharge onto the backward path. Use a Hebbian update with this noise and add weight decay.

$$\Delta B_l = \eta x_l x_{l+1}^\intercal - \lambda_{\mathrm{WM}} B_l$$

[1] Lillicrap, Timothy P., et al. "Random synaptic feedback weights support error backpropagation for deep learning." *Nature communications* 7.1 (2016): 1-10.
[2] Akrout, Mohamed, et al. "Deep learning without weight transport." *Advances in Neural Information Processing Systems*. 2019.

Weight Mirror does not transfer across architecture



Kunin, Nayebi*, Sagastuy-Brena* et al. (2020)*

Approach

- Parametrize the space of dynamics allowed on the backward weights to allow for backpropagation alternatives to be more stable

- Search over learning metaparameters in this enlarged space

- Compare ImageNet model performance and transfer across architectures to backpropagation

Kunin, D.*, Nayebi, A.*, Sagastuy-Brena*, J. et al. *Two Routes to Scalable Credit Assignment without Weight Symmetry*, 2020

Daniel Kunin

Javier Sagastuy-Brena

Approach

- Parametrize the space of dynamics allowed on the backward weights to allow for backpropagation alternatives to be more stable

- Search over learning metaparameters in this enlarged space

- Compare ImageNet model performance and transfer across architectures to backpropagation

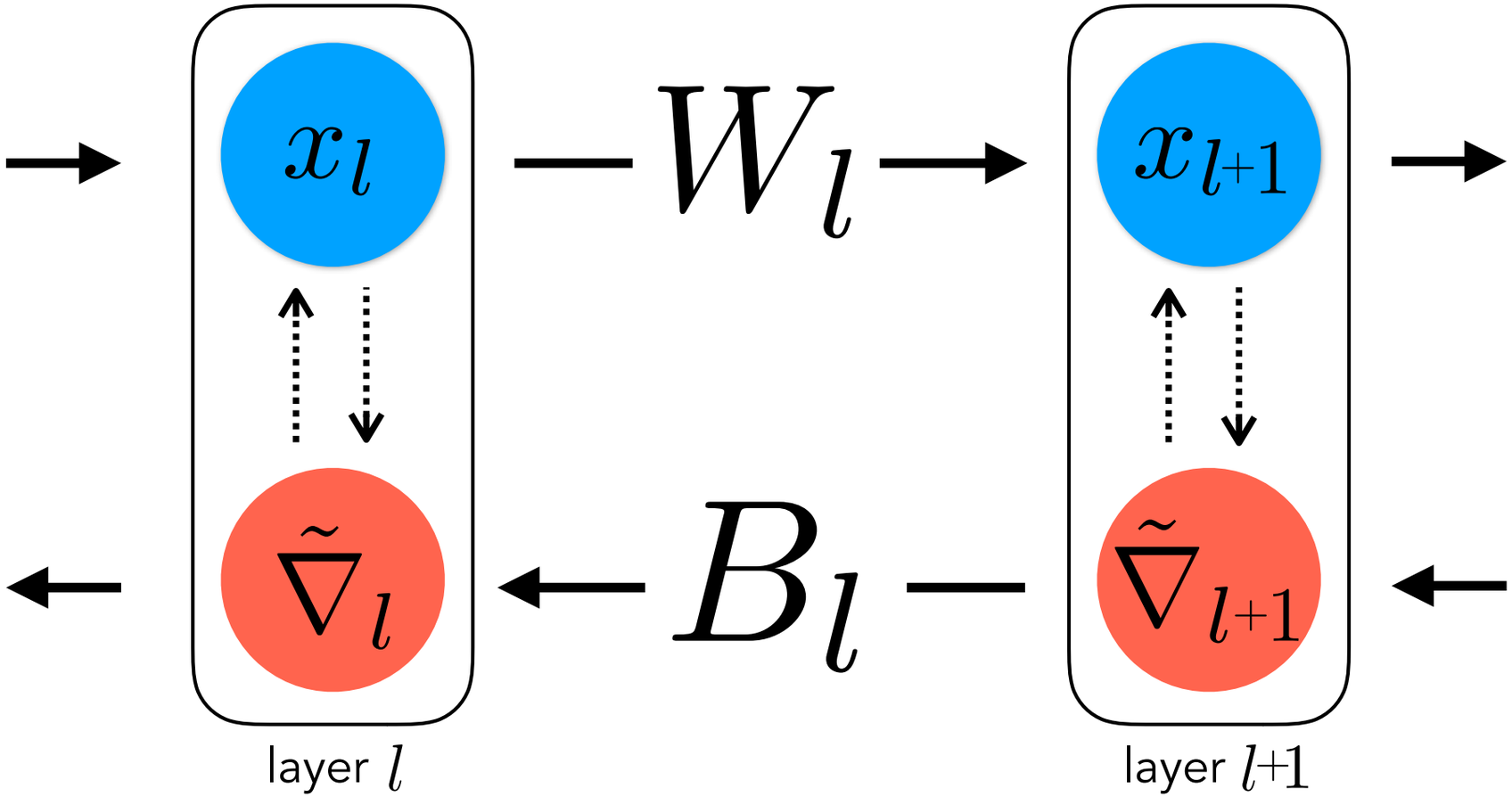
Kunin, D.*, Nayebi, A.*, Sagastuy-Brena*, J. et al. *Two Routes to Scalable Credit Assignment without Weight Symmetry*, 2020

Daniel Kunin

Javier Sagastuy-Brena

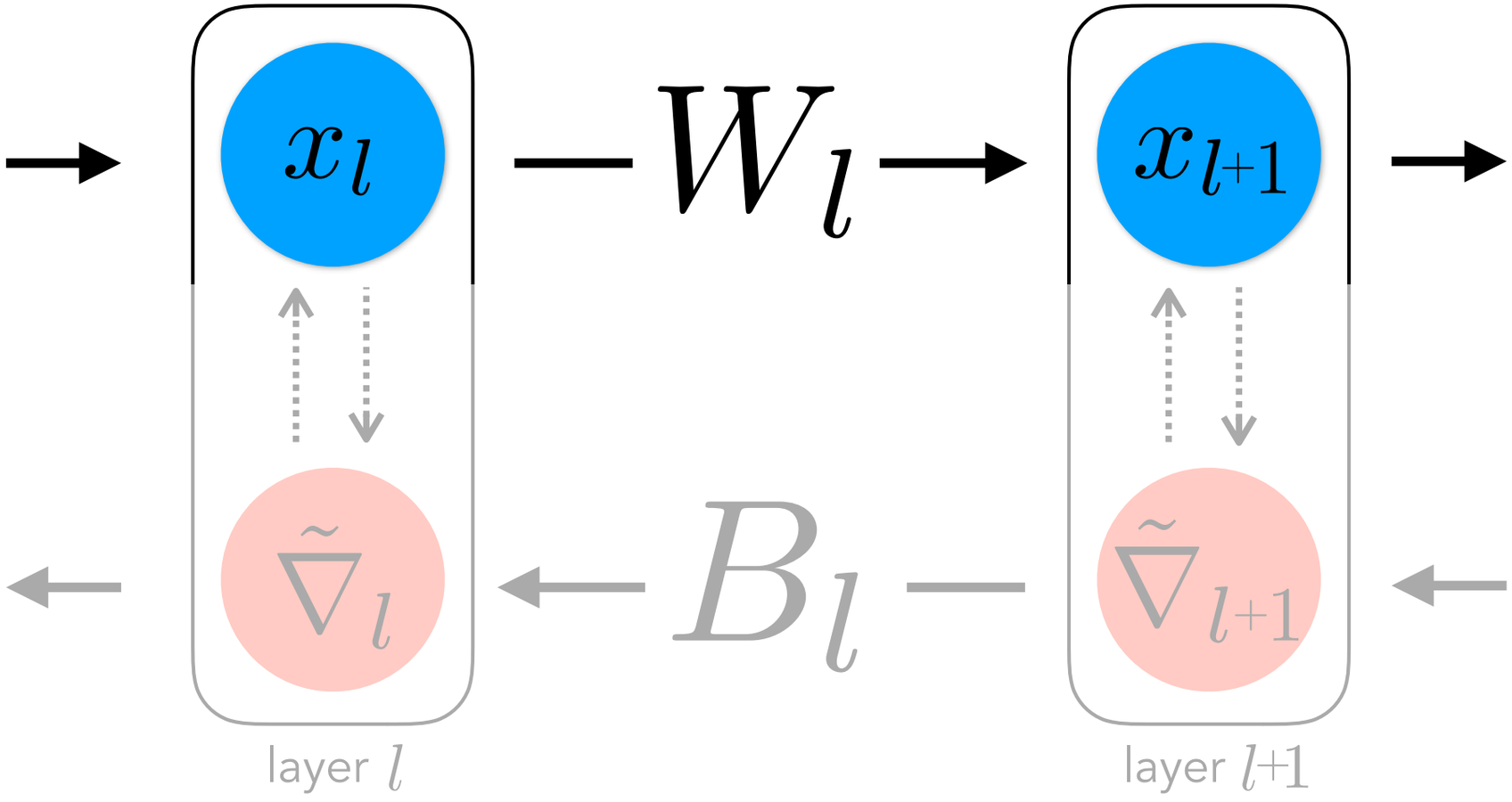
Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$



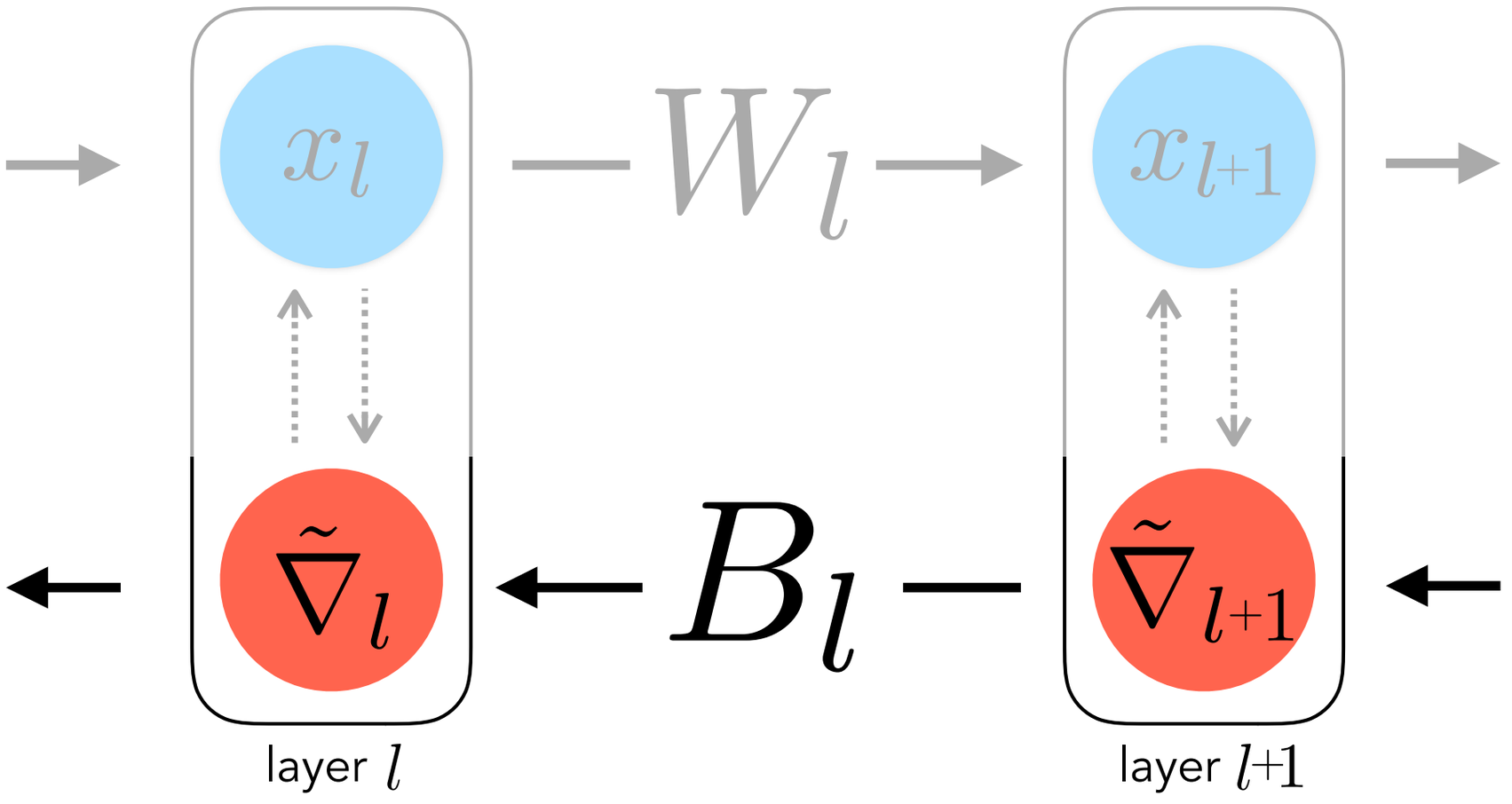
Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$



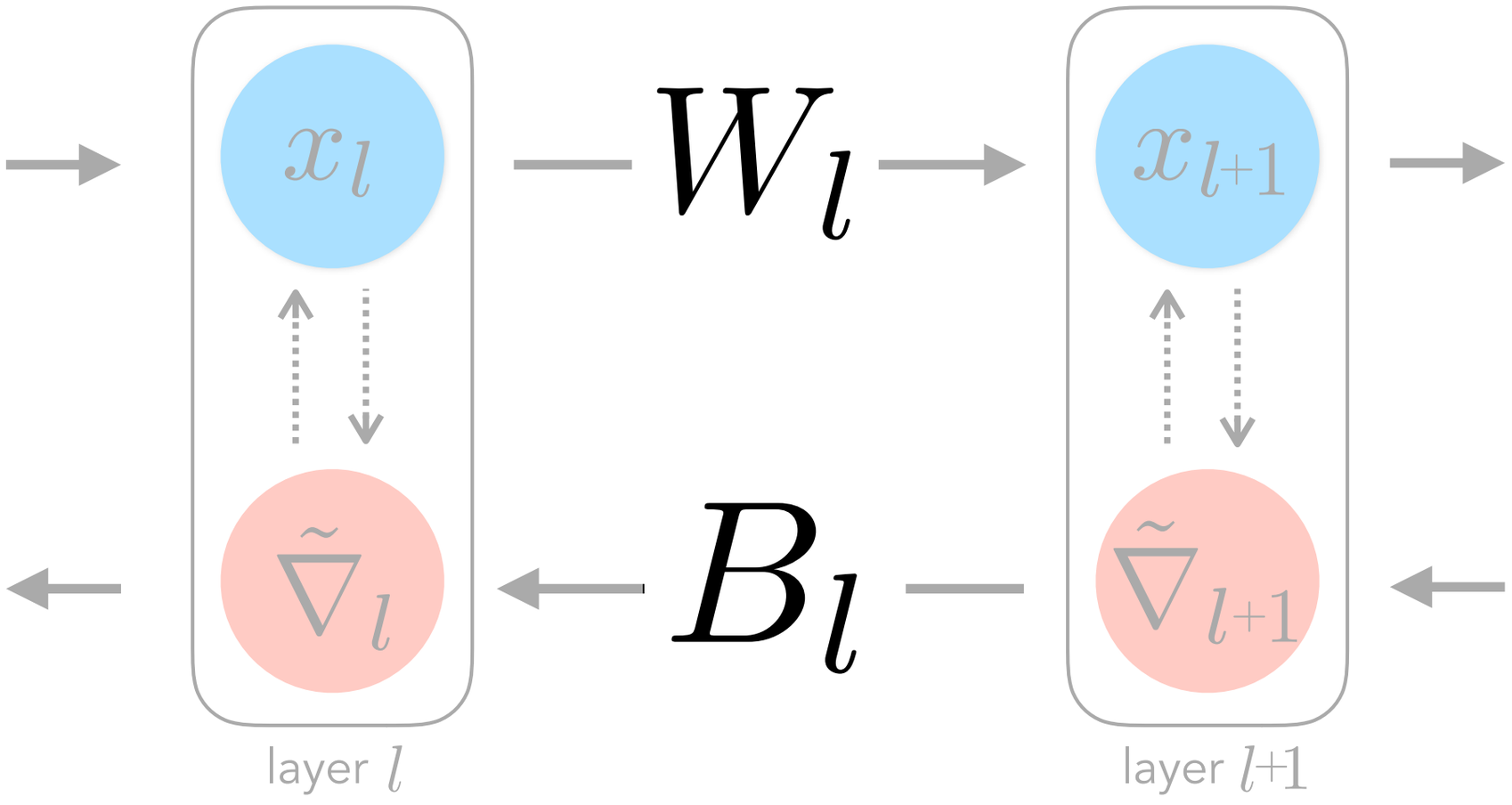
Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$



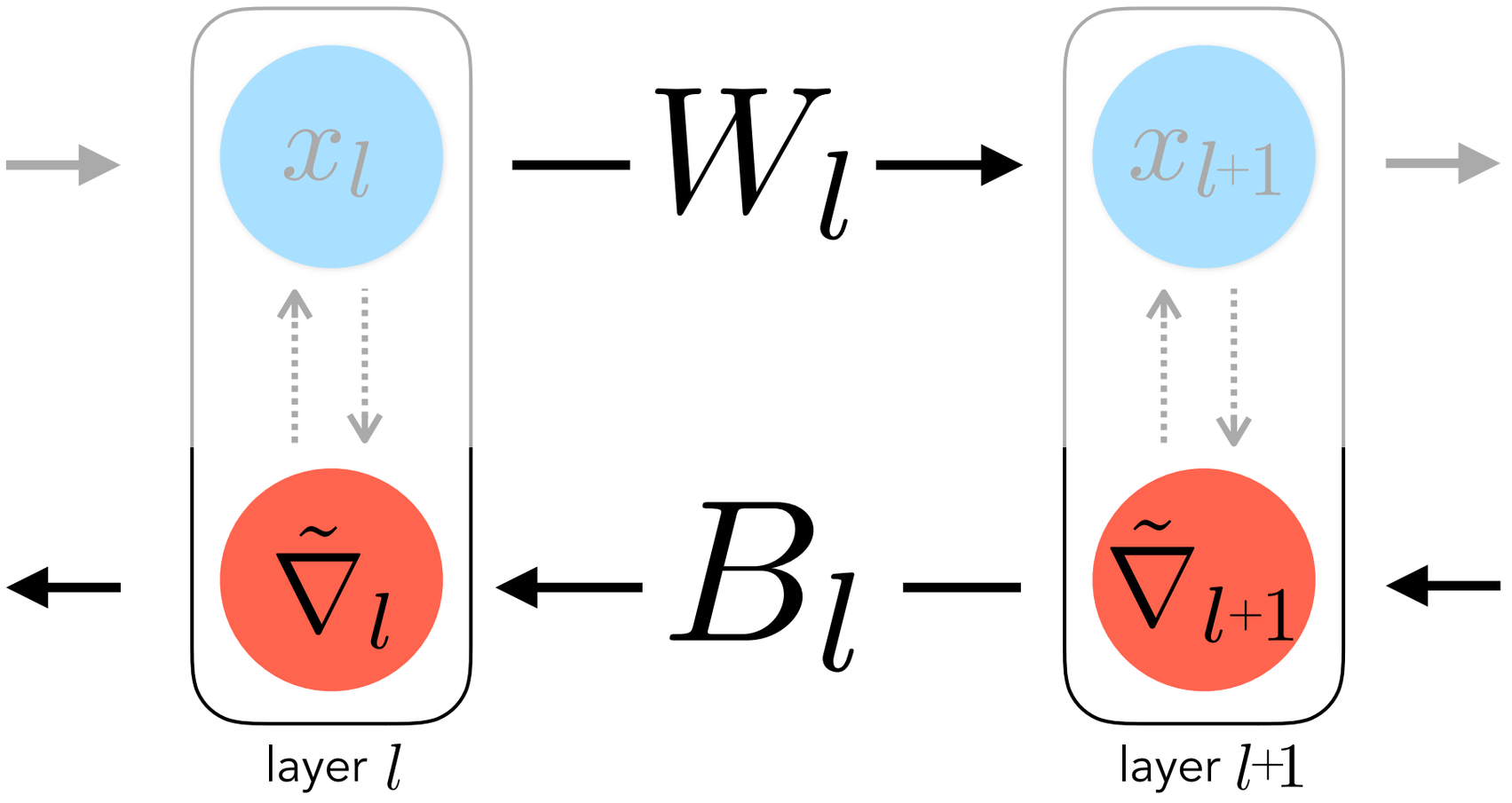
Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$



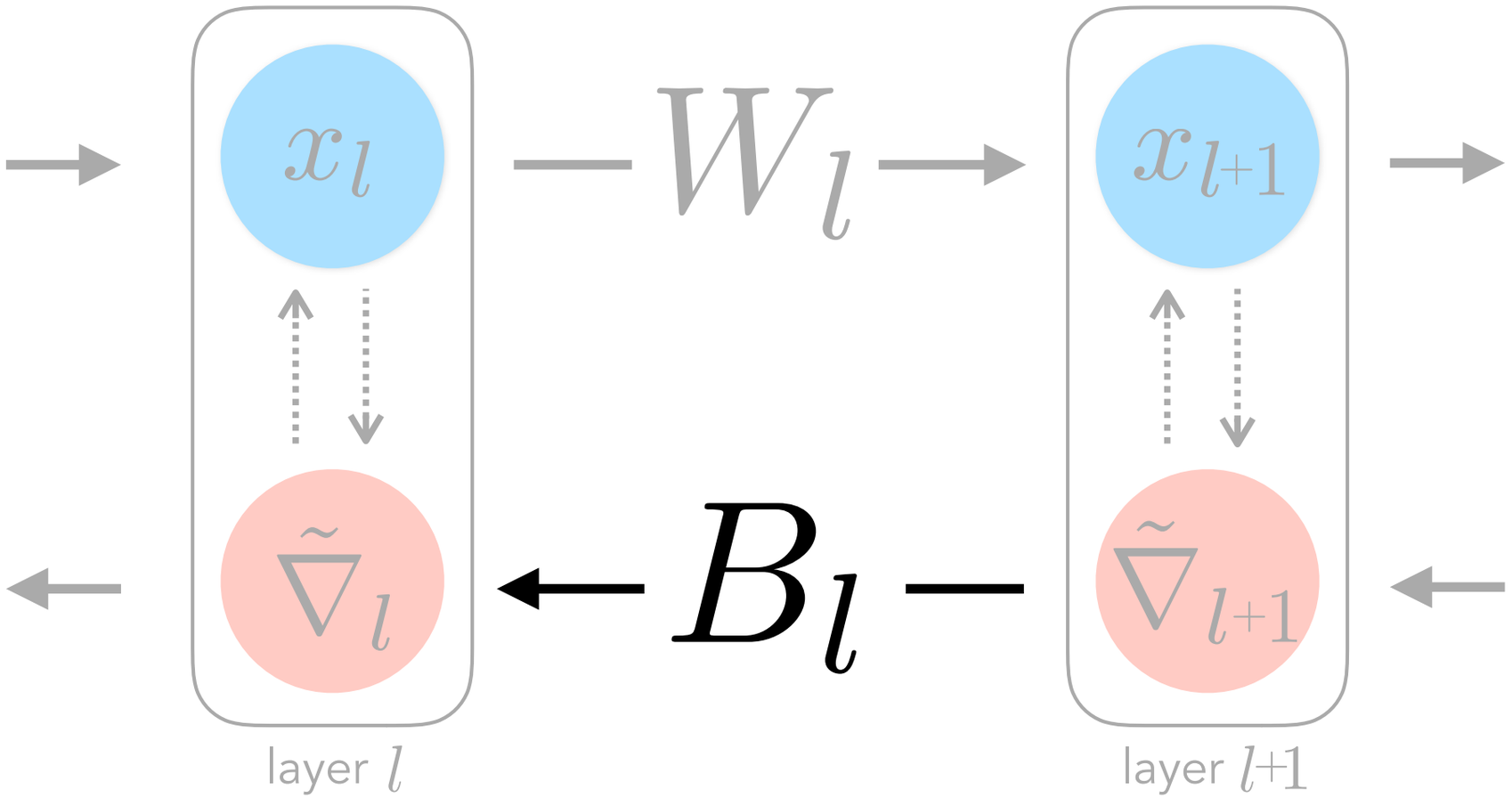
Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$



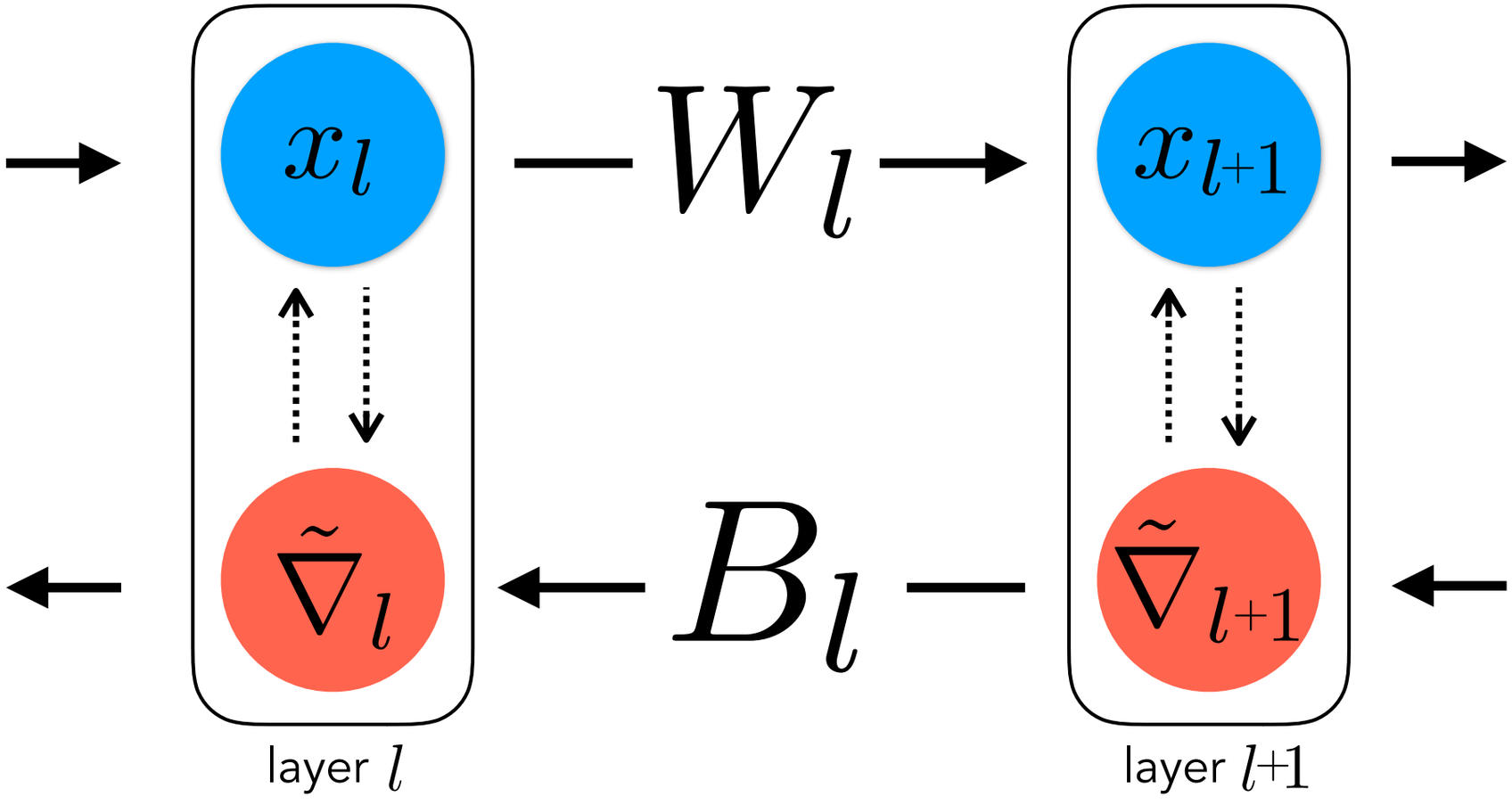
Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$



Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$



Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$

Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$

Local	$\mathcal{P}_l$	$\nabla \mathcal{P}_l$
decay	$\frac{1}{2}\|\|B_l\|\|^2$	B_l
amp	$-\mathrm{tr}(x_l^\intercal B_l x_{l+1})$	$-x_l x_{l+1}^\intercal$
null	$\frac{1}{2}\|\|B_l x_{l+1}\|\|^2$	$B_l x_{l+1} x_{l+1}^\intercal$
Non-local	$\mathcal{P}_l$	$\nabla \mathcal{P}_l$
sparse	$\frac{1}{2}\|\|x_l^\intercal B_l\|\|^2$	$x_l x_l^\intercal B_l$
self	$-\mathrm{tr}(B_l W_l)$	$-W_l^\intercal$

Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$

$$\mathcal{R}_{\text{FA}} \equiv 0$$

Local	$\mathcal{P}_l$	$\nabla \mathcal{P}_l$
decay	$\frac{1}{2}\|\|B_l\|\|^2$	B_l
amp	$-\text{tr}(x_l^\intercal B_l x_{l+1})$	$-x_l x_{l+1}^\intercal$
null	$\frac{1}{2}\|\|B_l x_{l+1}\|\|^2$	$B_l x_{l+1} x_{l+1}^\intercal$
Non-local	$\mathcal{P}_l$	$\nabla \mathcal{P}_l$
sparse	$\frac{1}{2}\|\|x_l^\intercal B_l\|\|^2$	$x_l x_l^\intercal B_l$
self	$-\text{tr}(B_l W_l)$	$-W_l^\intercal$

Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$

Local	$\mathcal{P}_l$	$\nabla\mathcal{P}_l$
decay	$\frac{1}{2}\|\|B_l\|\|^2$	B_l
amp	$-\mathrm{tr}(x_l^\intercal B_l x_{l+1})$	$-x_l x_{l+1}^\intercal$
null	$\frac{1}{2}\|\|B_l x_{l+1}\|\|^2$	$B_l x_{l+1} x_{l+1}^\intercal$
Non-local	$\mathcal{P}_l$	$\nabla\mathcal{P}_l$
sparse	$\frac{1}{2}\|\|x_l^\intercal B_l\|\|^2$	$x_l x_l^\intercal B_l$
self	$-\mathrm{tr}(B_l W_l)$	$-W_l^\intercal$

$$\mathcal{R}_{\mathrm{FA}} \equiv 0$$

$$\mathcal{R}_{\mathrm{WM}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{amp}} + \beta \mathcal{P}_l^{\mathrm{decay}}$$

Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$

Local	$\mathcal{P}_l$	$\nabla\mathcal{P}_l$
decay	$\frac{1}{2}\|\|B_l\|\|^2$	B_l
amp	$-\mathrm{tr}(x_l^\intercal B_l x_{l+1})$	$-x_l x_{l+1}^\intercal$
null	$\frac{1}{2}\|\|B_l x_{l+1}\|\|^2$	$B_l x_{l+1} x_{l+1}^\intercal$
Non-local	$\mathcal{P}_l$	$\nabla\mathcal{P}_l$
sparse	$\frac{1}{2}\|\|x_l^\intercal B_l\|\|^2$	$x_l x_l^\intercal B_l$
self	$-\mathrm{tr}(B_l W_l)$	$-W_l^\intercal$

Previous proposals

$$\mathcal{R}_{\mathrm{FA}} \equiv 0$$

$$\mathcal{R}_{\mathrm{WM}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{amp}} + \beta \mathcal{P}_l^{\mathrm{decay}}$$

Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$

Local	$\mathcal{P}_l$	$\nabla\mathcal{P}_l$
decay	$\frac{1}{2}\|\|B_l\|\|^2$	B_l
amp	$-\mathrm{tr}(x_l^\intercal B_l x_{l+1})$	$-x_l x_{l+1}^\intercal$
null	$\frac{1}{2}\|\|B_l x_{l+1}\|\|^2$	$B_l x_{l+1} x_{l+1}^\intercal$
Non-local	$\mathcal{P}_l$	$\nabla\mathcal{P}_l$
sparse	$\frac{1}{2}\|\|x_l^\intercal B_l\|\|^2$	$x_l x_l^\intercal B_l$
self	$-\mathrm{tr}(B_l W_l)$	$-W_l^\intercal$

Previous proposals

$$\mathcal{R}_{\mathrm{FA}} \equiv 0$$

$$\mathcal{R}_{\mathrm{WM}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{amp}} + \beta \mathcal{P}_l^{\mathrm{decay}}$$

$$\mathcal{R}_{\mathrm{IA}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{amp}} + \beta \mathcal{P}_l^{\mathrm{decay}} + \gamma \mathcal{P}_l^{\mathrm{null}}$$

Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$

Local	$\mathcal{P}_l$	$\nabla\mathcal{P}_l$
decay	$\frac{1}{2}\|\|B_l\|\|^2$	B_l
amp	$-\mathrm{tr}(x_l^\intercal B_l x_{l+1})$	$-x_l x_{l+1}^\intercal$
null	$\frac{1}{2}\|\|B_l x_{l+1}\|\|^2$	$B_l x_{l+1} x_{l+1}^\intercal$
Non-local	$\mathcal{P}_l$	$\nabla\mathcal{P}_l$
sparse	$\frac{1}{2}\|\|x_l^\intercal B_l\|\|^2$	$x_l x_l^\intercal B_l$
self	$-\mathrm{tr}(B_l W_l)$	$-W_l^\intercal$

Previous proposals

$$\mathcal{R}_{\mathrm{FA}} \equiv 0$$

$$\mathcal{R}_{\mathrm{WM}} = \sum_{l \in \mathrm{layers}} \alpha\mathcal{P}_l^{\mathrm{amp}} + \beta\mathcal{P}_l^{\mathrm{decay}}$$

$$\mathcal{R}_{\mathrm{IA}} = \sum_{l \in \mathrm{layers}} \alpha\mathcal{P}_l^{\mathrm{amp}} + \beta\mathcal{P}_l^{\mathrm{decay}} + \gamma\mathcal{P}_l^{\mathrm{null}}$$

$$\mathcal{R}_{\mathrm{SA}} = \sum_{l \in \mathrm{layers}} \alpha\mathcal{P}_l^{\mathrm{self}} + \beta\mathcal{P}_l^{\mathrm{decay}}$$

Approach

- Parametrize the space of dynamics allowed on the backward weights to allow for backpropagation alternatives to be more stable

- Search over learning metaparameters in this enlarged space

- Compare ImageNet model performance and transfer across architectures to backpropagation

Kunin, D.*, Nayebi, A.*, Sagastuy-Brena*, J. et al. *Two Routes to Scalable Credit Assignment without Weight Symmetry*, 2020

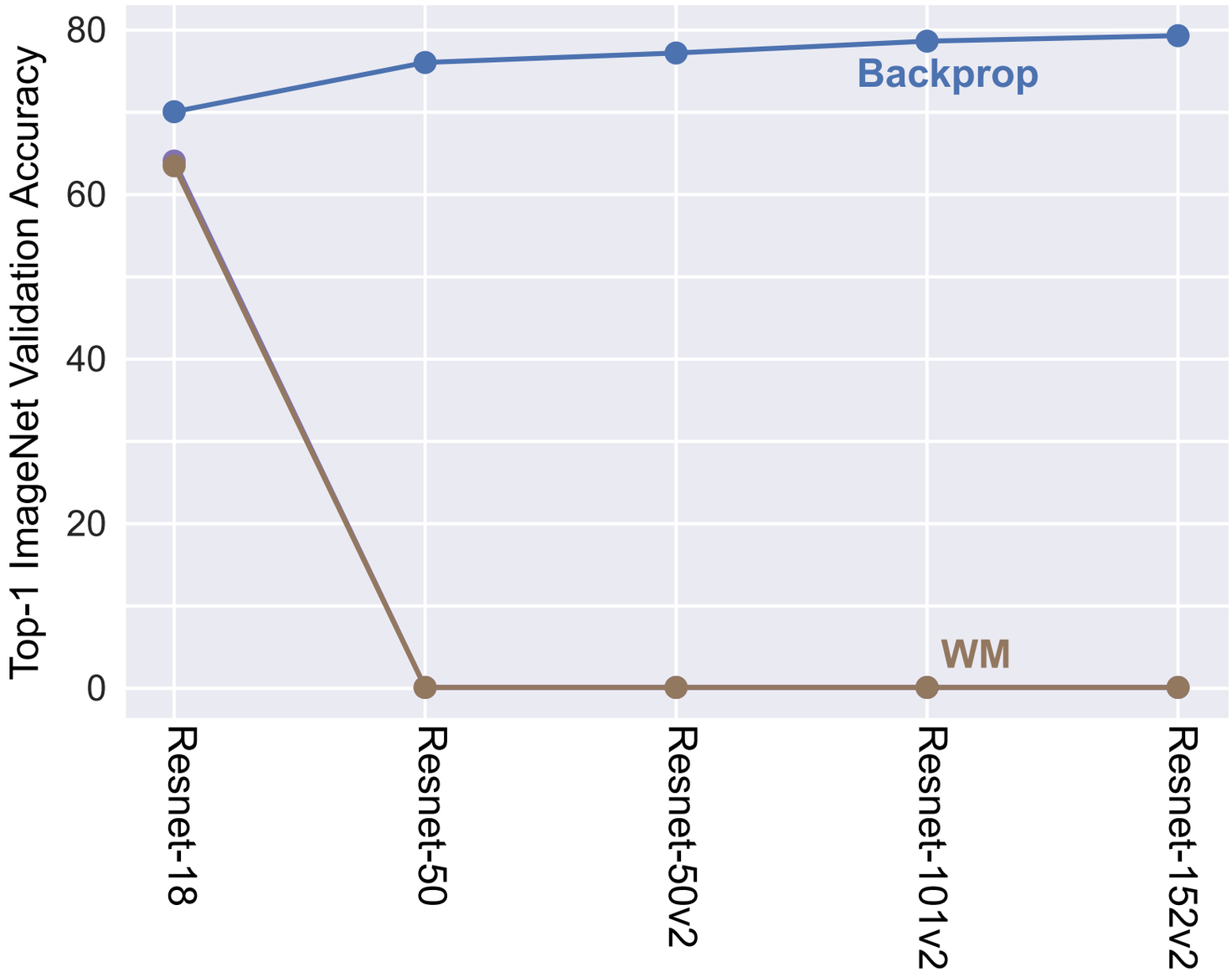
Daniel Kunin

Javier Sagastuy-Brena

Weight Mirror does not transfer across architecture

Weight Mirror: literature

$$\mathcal{R}_{\mathrm{WM}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{amp}} + \beta \mathcal{P}_l^{\mathrm{decay}}$$

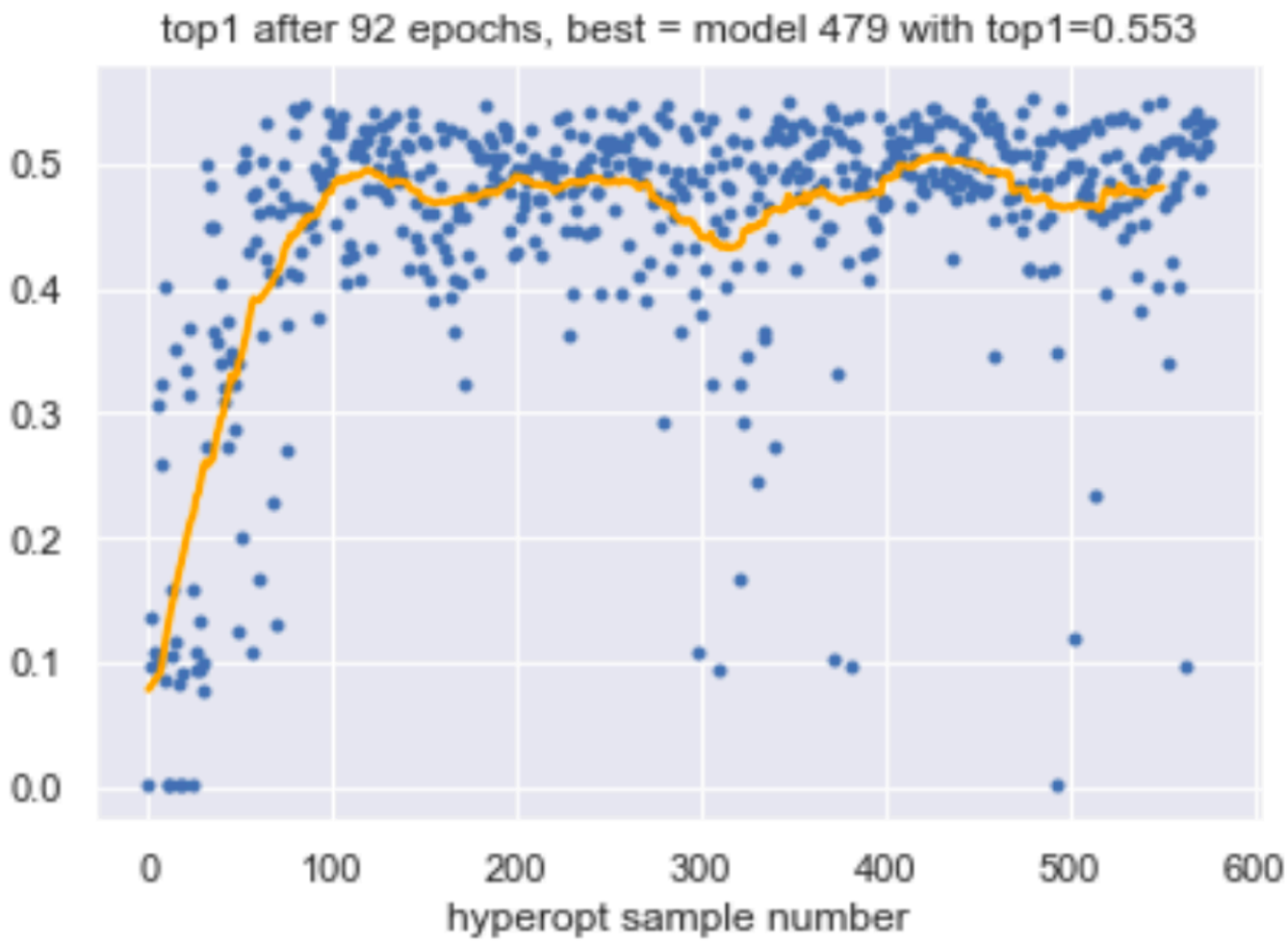


Local Learning Rules: Weight Mirror

Weight Mirror: optimized metaparameters

$$\mathcal{R}_{\mathrm{WM}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{amp}} + \beta \mathcal{P}_l^{\mathrm{decay}}$$

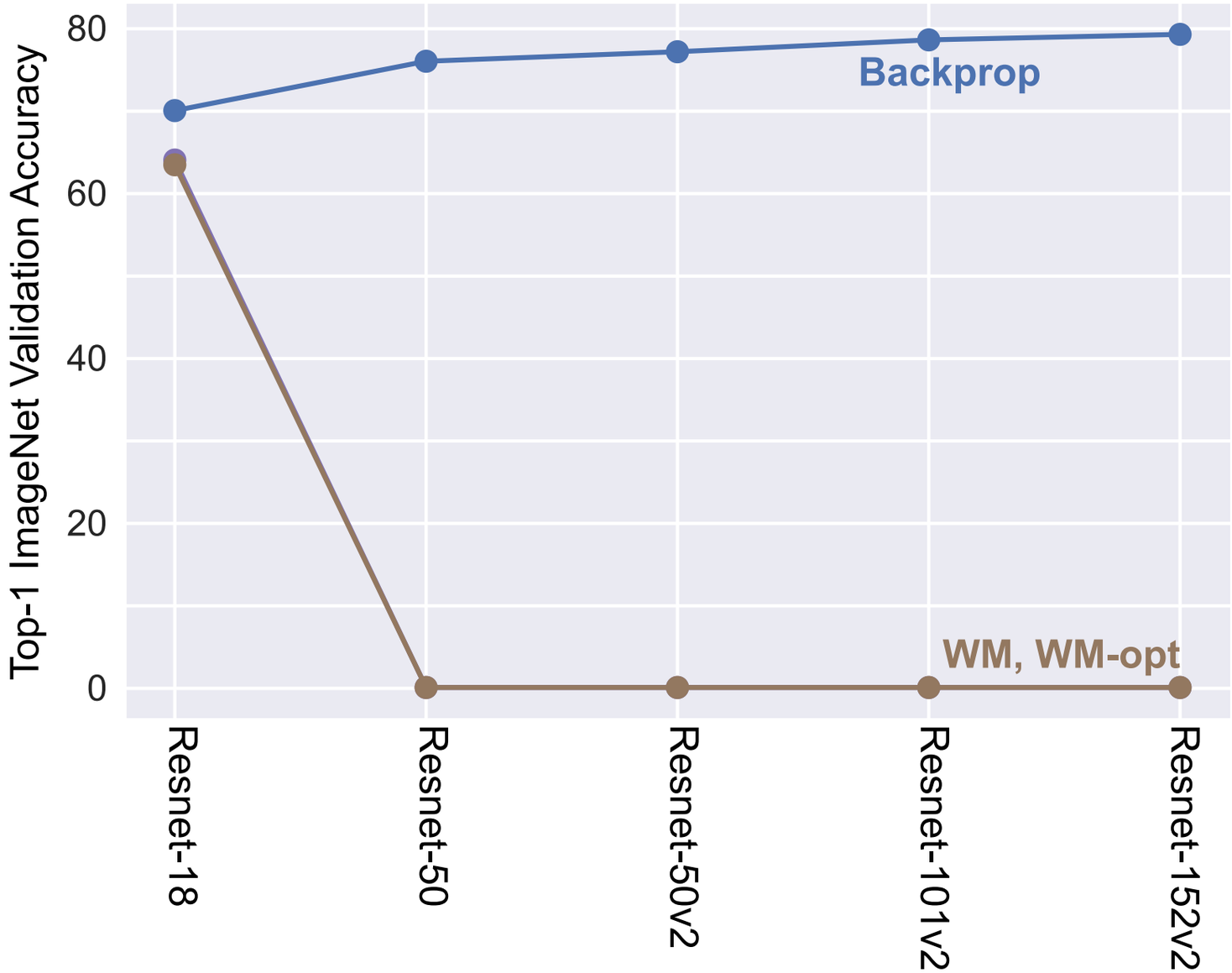
- TPE Search over alpha, beta and the variance of the noise used in mirror mode on ResNet-18.



Weight Mirror *still* does not transfer across architecture

Weight Mirror: optimized metaparameters

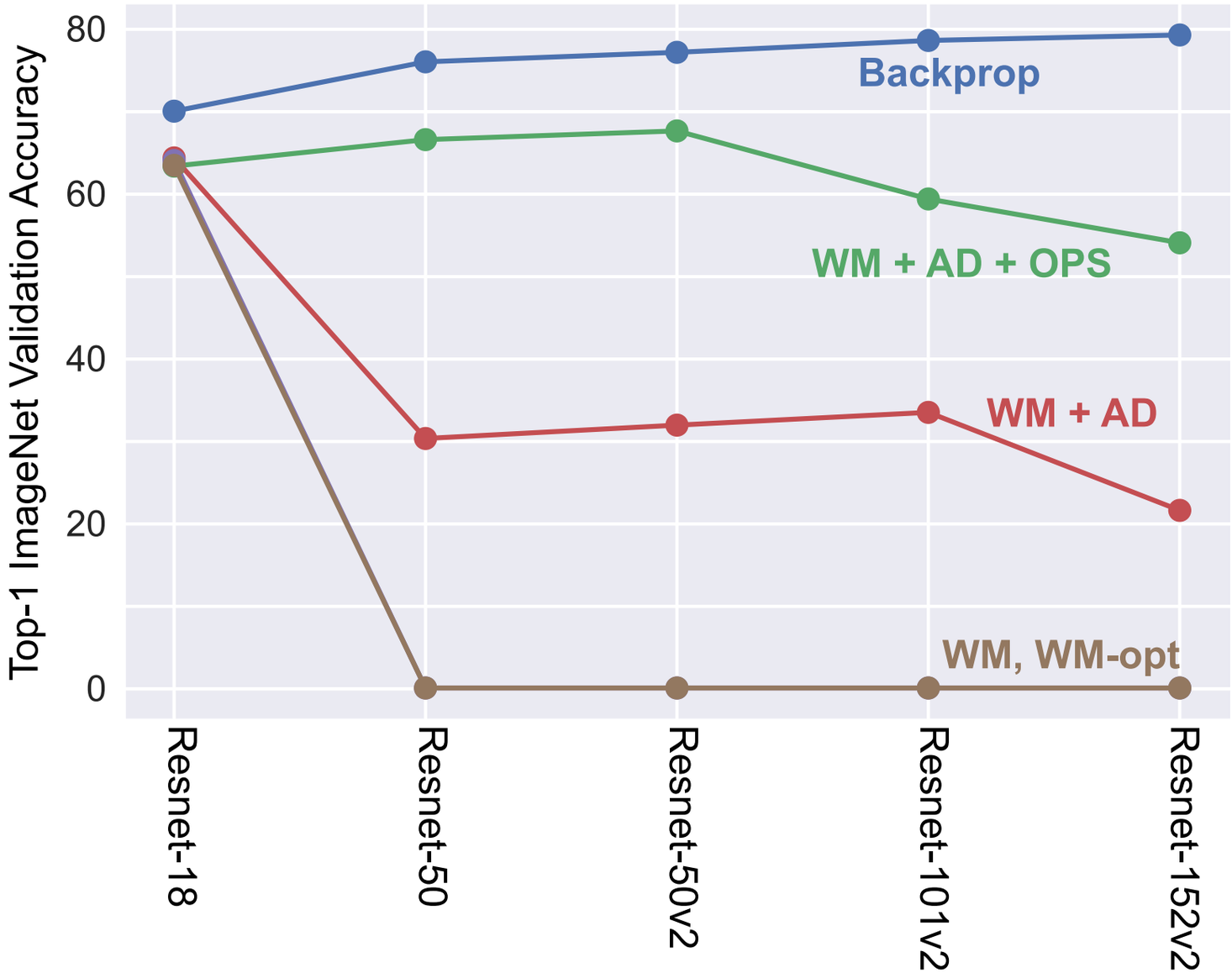
$$\mathcal{R}_{\mathrm{WM}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{amp}} + \beta \mathcal{P}_l^{\mathrm{decay}}$$



Local Learning Rules: Improved Metaparameter Robustness

Weight Mirror: adding an adaptive optimizer and normalizing operations

$$\mathcal{R}_{\mathrm{WM}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{amp}} + \beta \mathcal{P}_l^{\mathrm{decay}}$$



Instability of Weight Mirror

- The update given by WM (without decay) is Hebbian
- Purely Hebbian learning rules are unstable
- WM adds weight decay to prevent diverging norms
- An alternative strategy to stabilizing Hebbian dynamics given by Oja (1982) for learning dynamics of linear neurons

Normalizing operation:

$$B_l^{(t+1)} = \frac{B_l^{(t)} + \eta x_l x_{l+1}^\intercal}{||B_l^{(t)} + \eta x_l x_{l+1}^\intercal||}$$

Taylor series expansion:

$$B_l^{(t+1)} = B_l^{(t)} + \eta \left(x_l x_{l+1}^\intercal - B_l^{(t)} x_l^\intercal B_l^{(t)} x_{l+1} \right) + O(\eta^2)$$

New update:

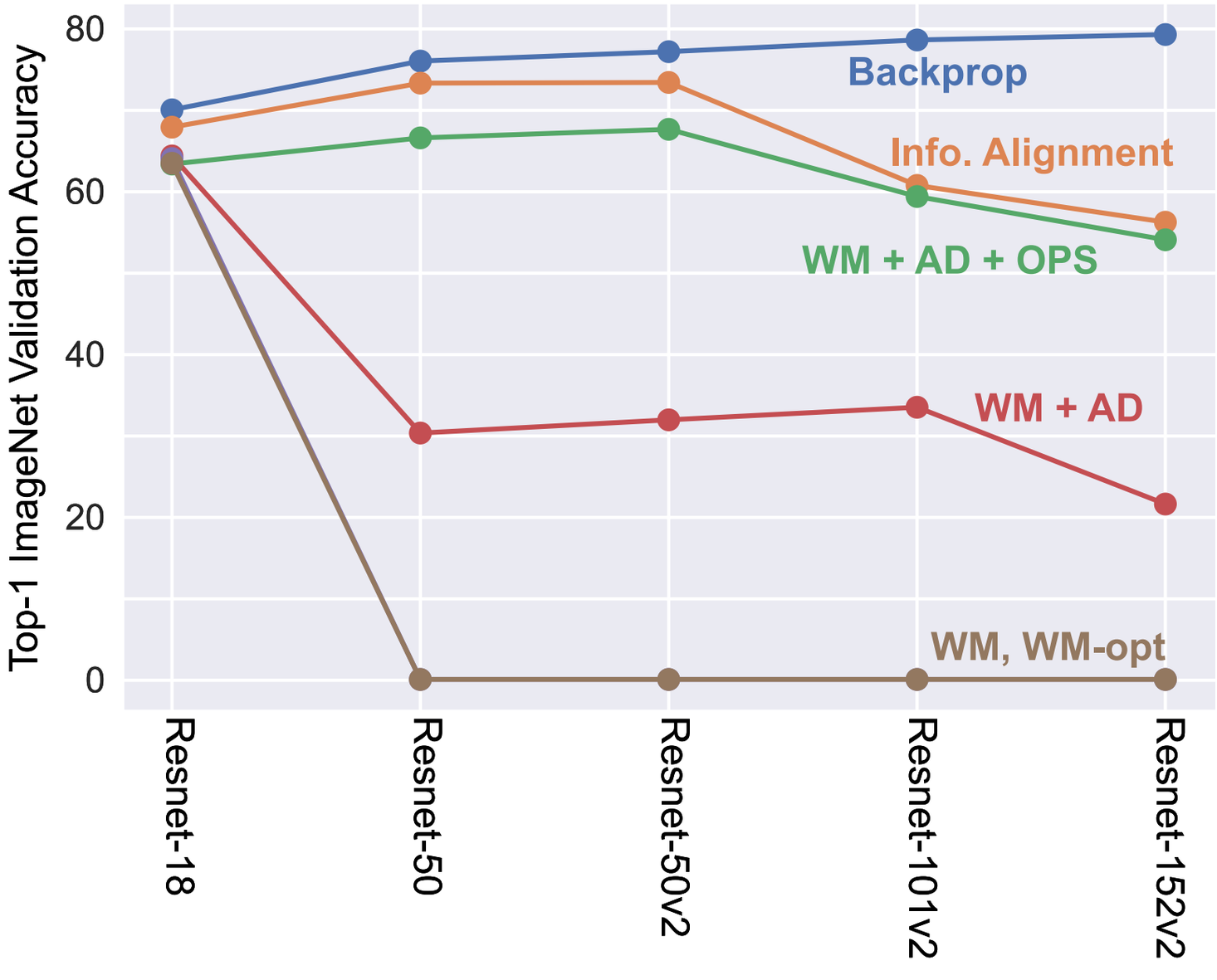
$$\Delta B_l = \eta \left(x_l x_{l+1}^\intercal - \underbrace{B_l x_l^\intercal B_l x_{l+1}}_{\approx \nabla \mathcal{P}_{\text{null}}} \right)$$

Local Learning Rules: Performance

Weight Mirror

$$\mathcal{R}_{\text{WM}} = \sum_{l \in \text{layers}} \alpha \mathcal{P}_l^{\text{amp}} + \beta \mathcal{P}_l^{\text{decay}}$$

Information Alignment



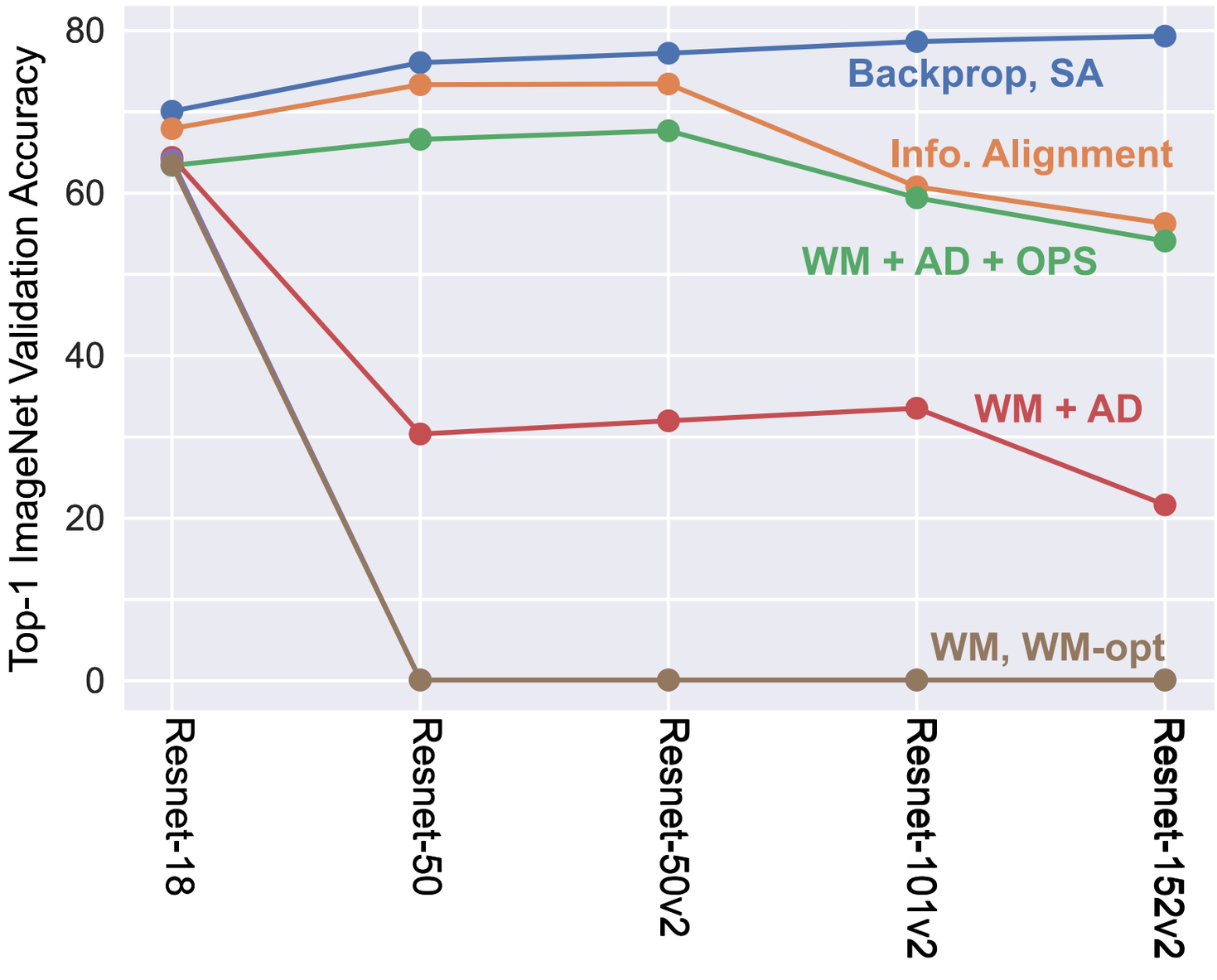
Non-Local Learning Rules

Weight Mirror

$$\mathcal{R}_{\mathrm{WM}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{amp}} + \beta \mathcal{P}_l^{\mathrm{decay}}$$

Symmetric Alignment

$$\propto \sum_{l \in \mathrm{layers}} \frac{1}{2} ||W_l - B_l^{\intercal}||^2$$



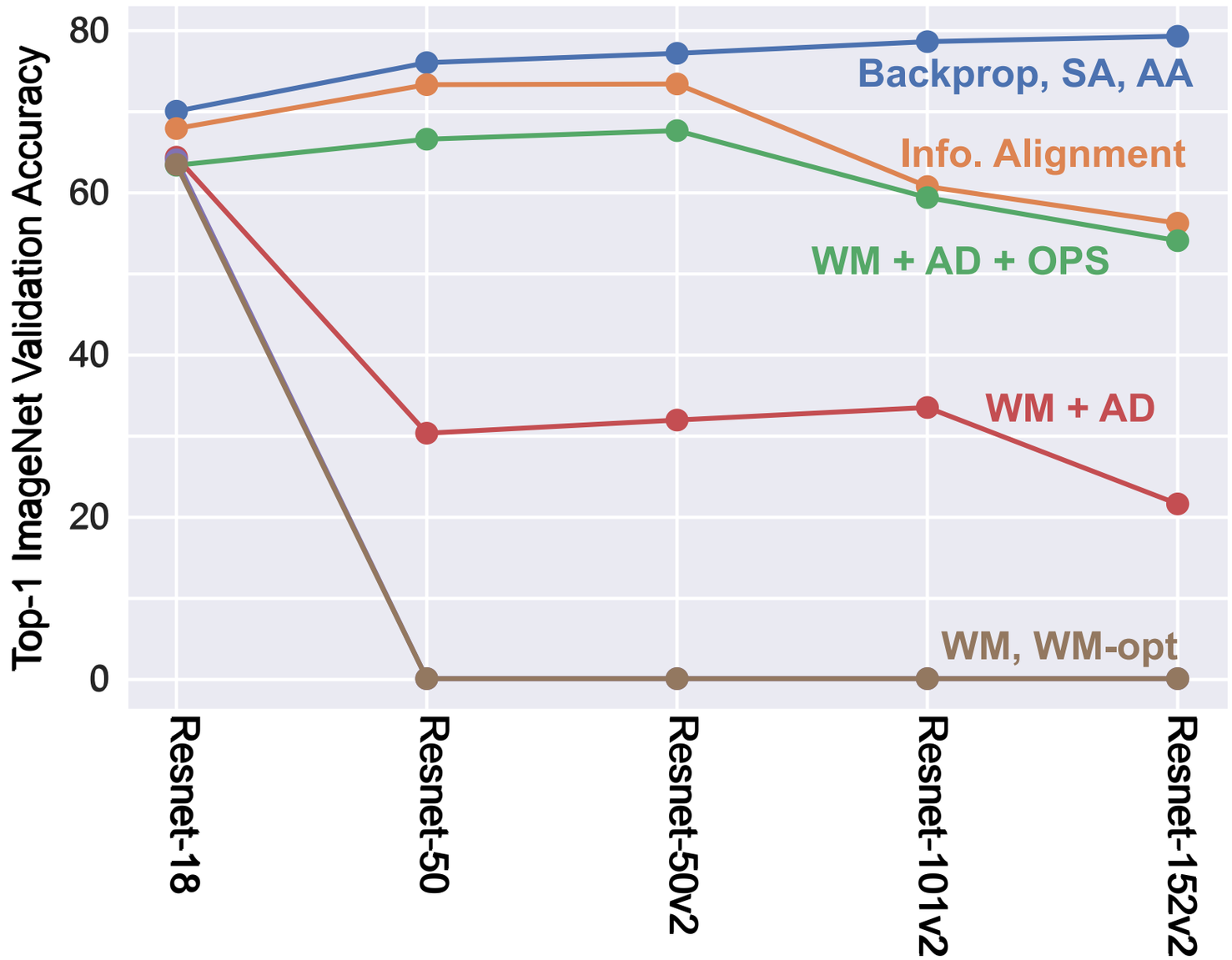
Non-Local Learning Rules

Weight Mirror

$$\mathcal{R}_{\text{WM}} = \sum_{l \in \text{layers}} \alpha \mathcal{P}_l^{\text{amp}} + \beta \mathcal{P}_l^{\text{decay}}$$

Activation Alignment

$$\propto \sum_{l \in \text{layers}} \frac{1}{2} ||W_l x_l - B_l^\intercal x_l||^2$$



Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$

Local	$\mathcal{P}_l$	$\nabla\mathcal{P}_l$
decay	$\frac{1}{2}\|\|B_l\|\|^2$	B_l
amp	$-\mathrm{tr}(x_l^\intercal B_l x_{l+1})$	$-x_l x_{l+1}^\intercal$
null	$\frac{1}{2}\|\|B_l x_{l+1}\|\|^2$	$B_l x_{l+1} x_{l+1}^\intercal$
Non-local	$\mathcal{P}_l$	$\nabla\mathcal{P}_l$
sparse	$\frac{1}{2}\|\|x_l^\intercal B_l\|\|^2$	$x_l x_l^\intercal B_l$
self	$-\mathrm{tr}(B_l W_l)$	$-W_l^\intercal$

Previous proposals

$$\mathcal{R}_{\mathrm{FA}} \equiv 0$$

$$\mathcal{R}_{\mathrm{WM}} = \sum_{l \in \mathrm{layers}} \alpha\mathcal{P}_l^{\mathrm{amp}} + \beta\mathcal{P}_l^{\mathrm{decay}}$$

$$\mathcal{R}_{\mathrm{IA}} = \sum_{l \in \mathrm{layers}} \alpha\mathcal{P}_l^{\mathrm{amp}} + \beta\mathcal{P}_l^{\mathrm{decay}} + \gamma\mathcal{P}_l^{\mathrm{null}}$$

$$\mathcal{R}_{\mathrm{SA}} = \sum_{l \in \mathrm{layers}} \alpha\mathcal{P}_l^{\mathrm{self}} + \beta\mathcal{P}_l^{\mathrm{decay}}$$

$$\mathcal{R}_{\mathrm{AA}} = \sum_{l \in \mathrm{layers}} \alpha\mathcal{P}_l^{\mathrm{amp}} + \beta\mathcal{P}_l^{\mathrm{sparse}}$$

Regularization Inspired Learning Rule Framework

$$\mathcal{L}(W, B) = \mathcal{J}(W) + \mathcal{R}(B)$$

Local	$\mathcal{P}_l$	$\nabla\mathcal{P}_l$
decay	$\frac{1}{2}\|\|B_l\|\|^2$	B_l
amp	$-\mathrm{tr}(x_l^\intercal B_l x_{l+1})$	$-x_l x_{l+1}^\intercal$
null	$\frac{1}{2}\|\|B_l x_{l+1}\|\|^2$	$B_l x_{l+1} x_{l+1}^\intercal$
Non-local	$\mathcal{P}_l$	$\nabla\mathcal{P}_l$
sparse	$\frac{1}{2}\|\|x_l^\intercal B_l\|\|^2$	$x_l x_l^\intercal B_l$
self	$-\mathrm{tr}(B_l W_l)$	$-W_l^\intercal$

Previous proposals

$$\mathcal{R}_{\mathrm{FA}} \equiv 0$$

$$\mathcal{R}_{\mathrm{WM}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{amp}} + \beta \mathcal{P}_l^{\mathrm{decay}}$$

$$\mathcal{R}_{\mathrm{IA}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{amp}} + \beta \mathcal{P}_l^{\mathrm{decay}} + \gamma \mathcal{P}_l^{\mathrm{null}}$$

$$\mathcal{R}_{\mathrm{SA}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{self}} + \beta \mathcal{P}_l^{\mathrm{decay}}$$

$$\mathcal{R}_{\mathrm{AA}} = \sum_{l \in \mathrm{layers}} \alpha \mathcal{P}_l^{\mathrm{amp}} + \beta \mathcal{P}_l^{\mathrm{sparse}}$$

Novel proposals

Is recurrence useful during learning?

- **Feedback connections primarily used during inference?**

Currently not much evidence in the affirmative.
More consistent with the hypothesis that feedback connections allow a shallower network to "approximate" a deeper network that could not otherwise physically fit.

- Feedback connections primarily used for propagating error signals?

Is recurrence useful during learning? More likely.

- **Feedback connections primarily used during inference?**

Currently not much evidence in the affirmative.
More consistent with the hypothesis that feedback connections allow a shallower network to "approximate" a deeper network that could not otherwise physically fit.

- **Feedback connections primarily used for propagating error signals?**

With the proper dynamics on the feedback weights, one can achieve backpropagation-level performance without requiring weight symmetry.

Acknowledgements

Daniel Yamins**

Surya Ganguli**

Thanks!

Daniel Bear

Jonas Kubilius

Daniel Kunin

Javier Sagastuy-Brena

Contact:
anayebi@stanford.edu

Funding:

Neurosciences PhD Program

Stanford Mind, Brain, Computation and Technology Training Program,
Wu Tsai Neurosciences Institute