ICLR

Dynamic Sparse Training with Structured Sparsity

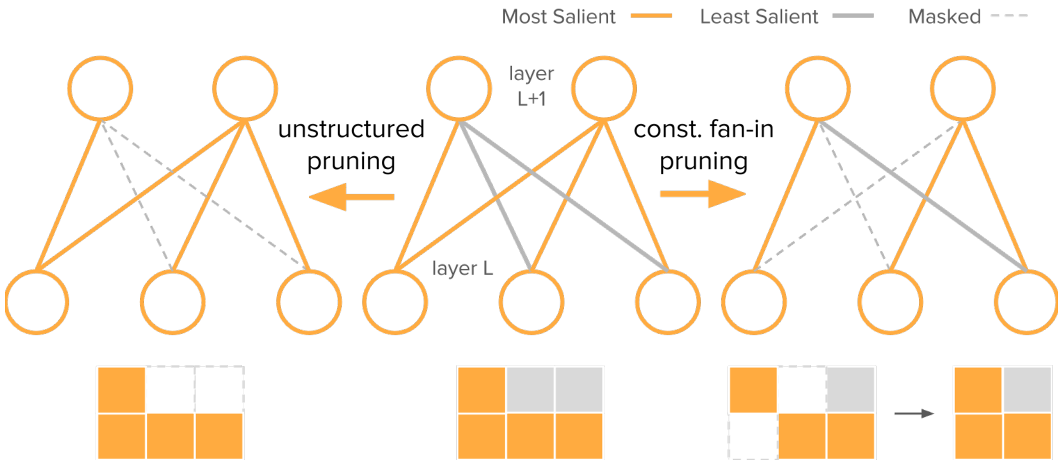
Mike Lasby[1], Anna Golubeva[2,3], Utku Evci[4], Mihai Nica[5,6], Yani Ioannou[1]

[1]University of Calgary, [2]MIT, [3]IAIFI, [4]Google DeepMind, [5]University of Guelph, [6]Vector Institute for AI

Motivation

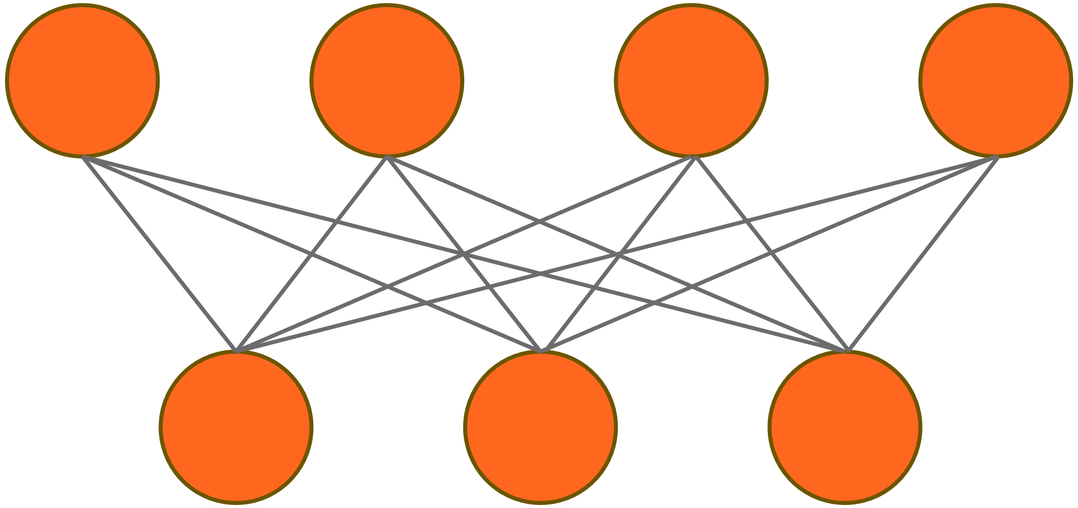
- Unstructured Dynamic Sparse Training (DST) matches the generalization performance of dense models with 85-95% fewer weights
- Accelerating unstructured Sparse Neural Networks (SNNs) is challenging
- Structured SNNs are easy to accelerate, but do not generalize as well as unstructured.
- **Can we use DST to learn a SNN with high generalization performance that is also amenable to acceleration?**

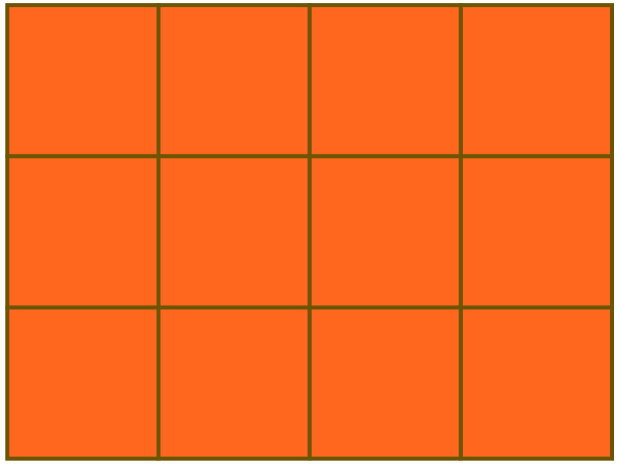
Method: Structured RigL (SRigL)



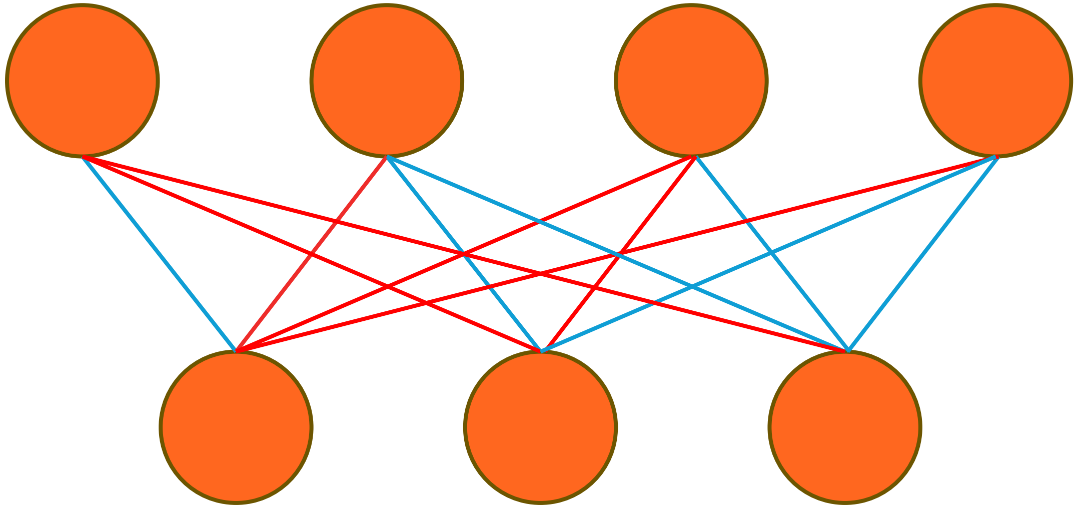
- Sparse-to-sparse DST method which extends RigL to learn a structured SNN
- Learns specific type of **N:M sparsity where M is dense fan-in**
- Constant fan-in constraint applied to each neuron within a given layer to enable efficient and compressible indexing of non-zero weights

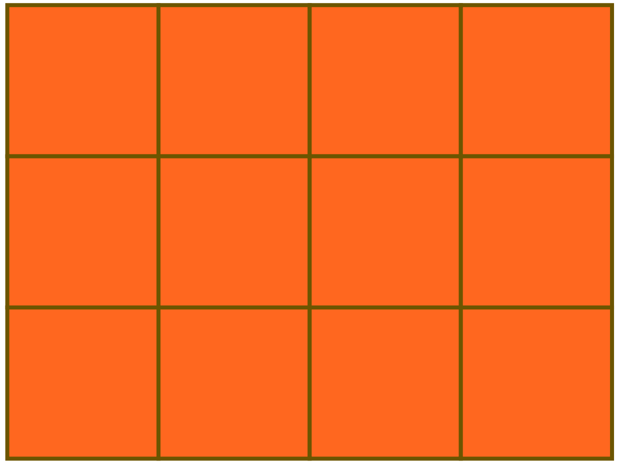
Unstructured vs. Constant Fan-In Sparsity





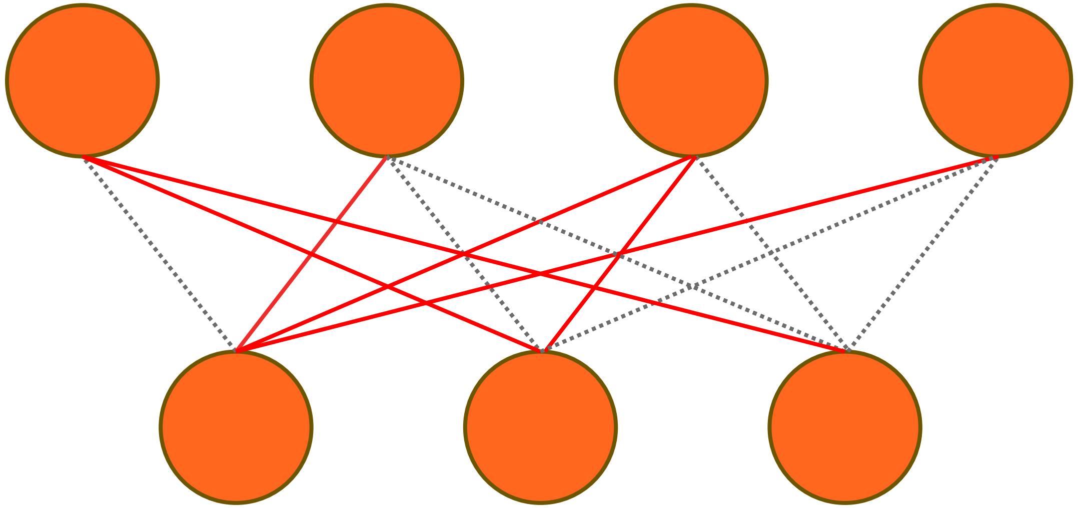
Unstructured vs. Constant Fan-In Sparsity

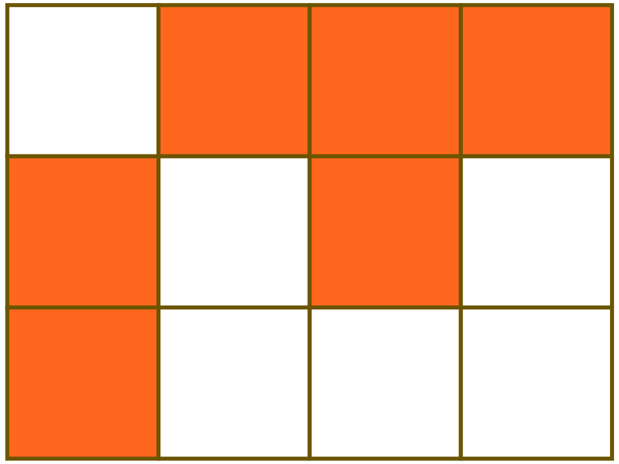




- More Salient
- Less Salient
- Masked

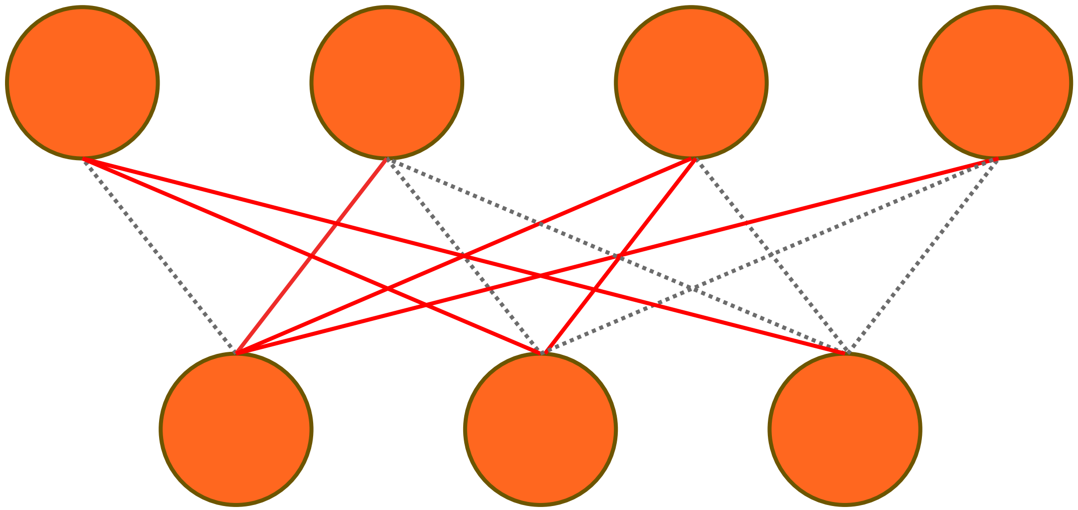
Unstructured vs. Constant Fan-In Sparsity

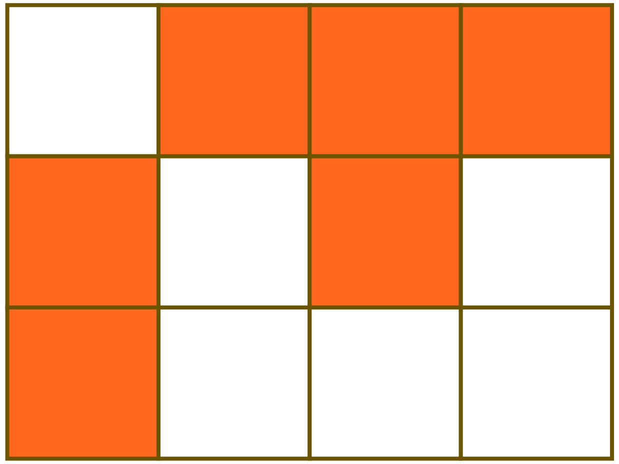




- More Salient
- Less Salient
- Masked

Unstructured vs. Constant Fan-In Sparsity

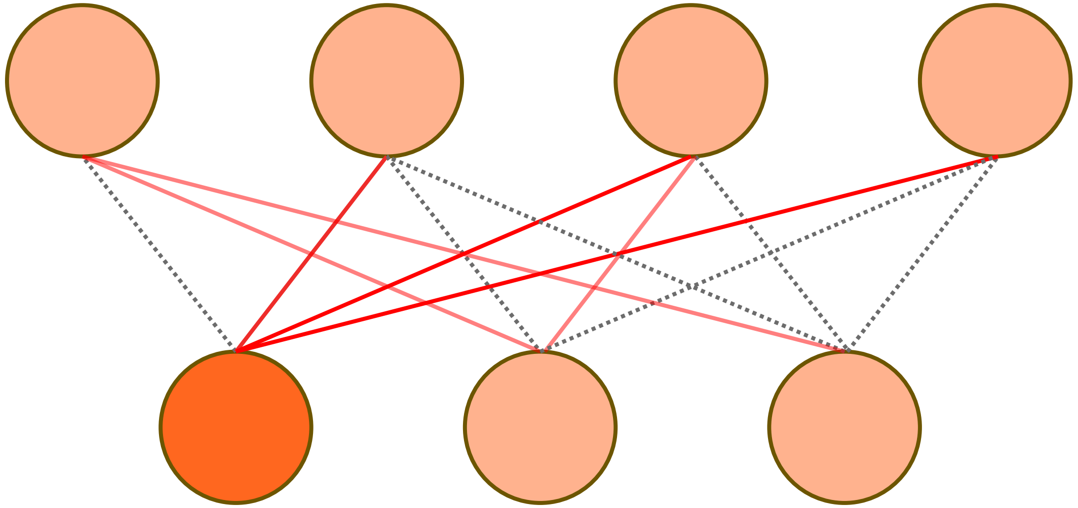


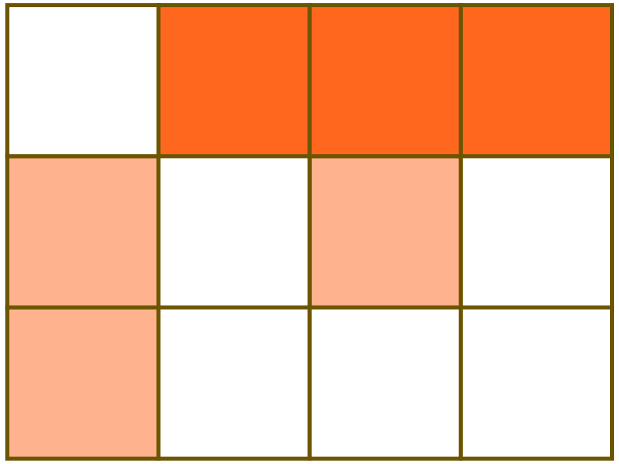


- More Salient
- Less Salient
- Masked

Noncontiguous weights make acceleration a challenge!

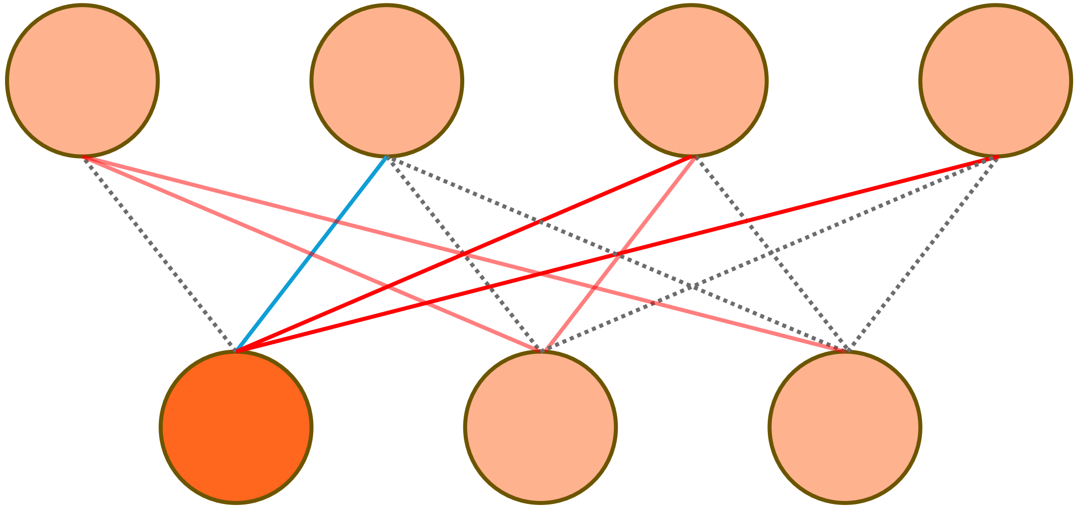
Unstructured vs. **Constant Fan-In Sparsity**

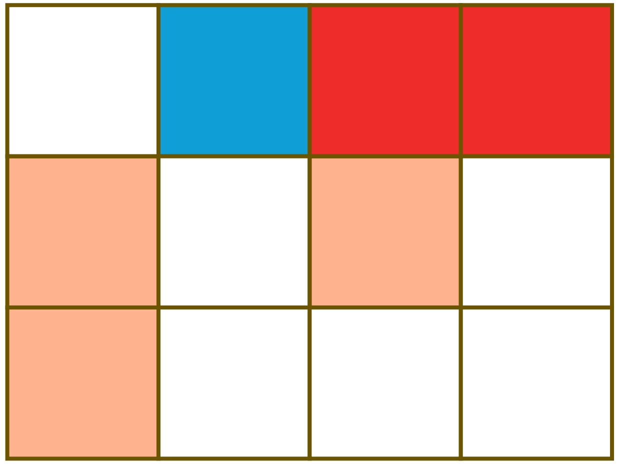




- More Salient
- Less Salient
- Masked

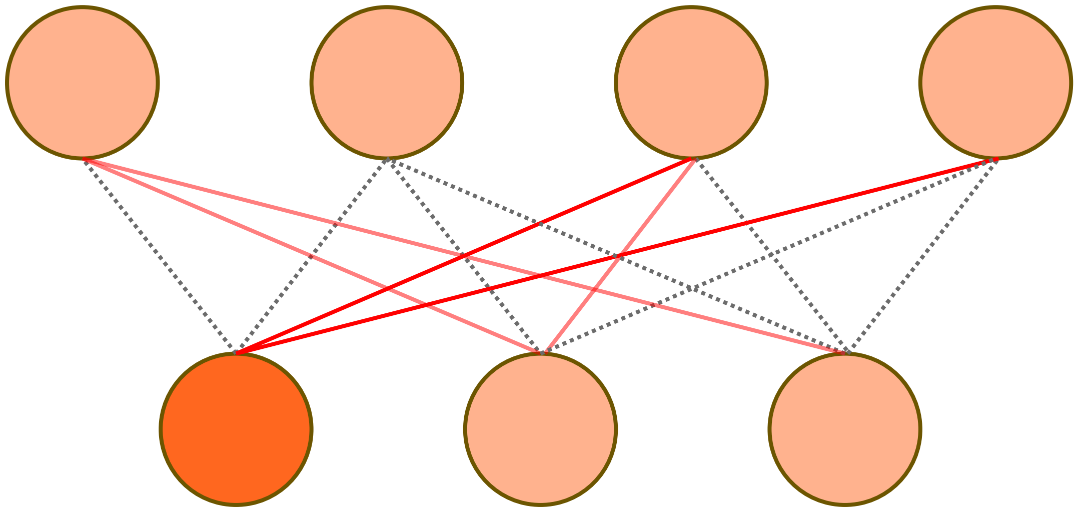
Unstructured vs. **Constant Fan-In Sparsity**

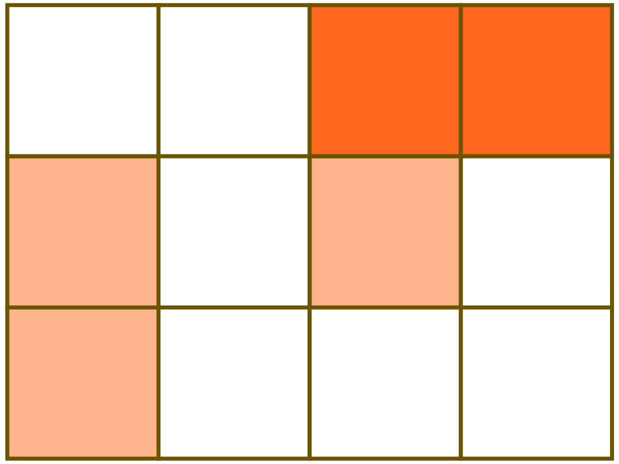




- More Salient
- Less Salient
- Masked

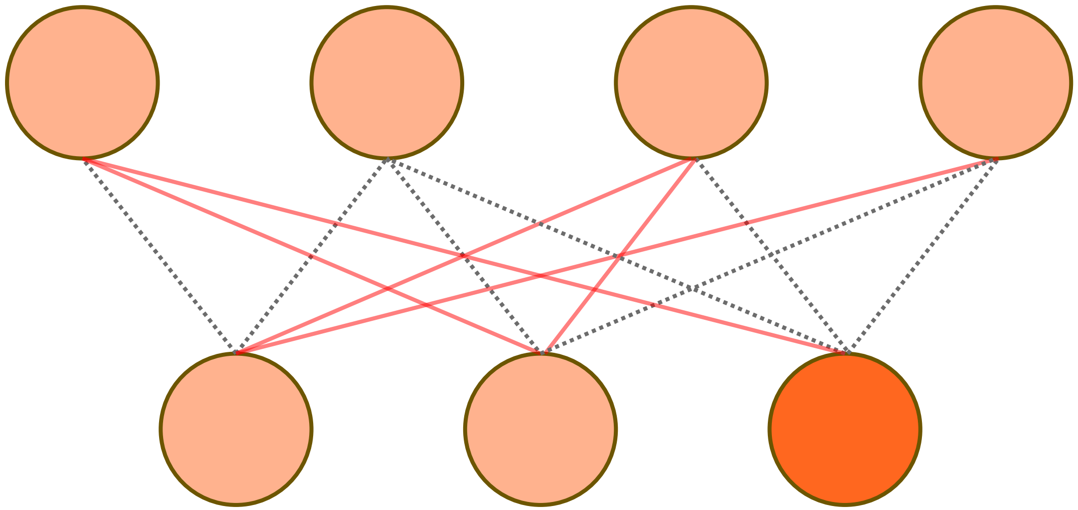
Unstructured vs. **Constant Fan-In Sparsity**

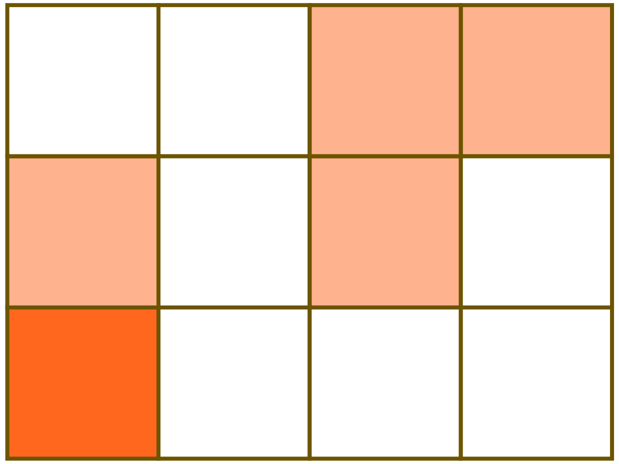




- More Salient
- Less Salient
- Masked

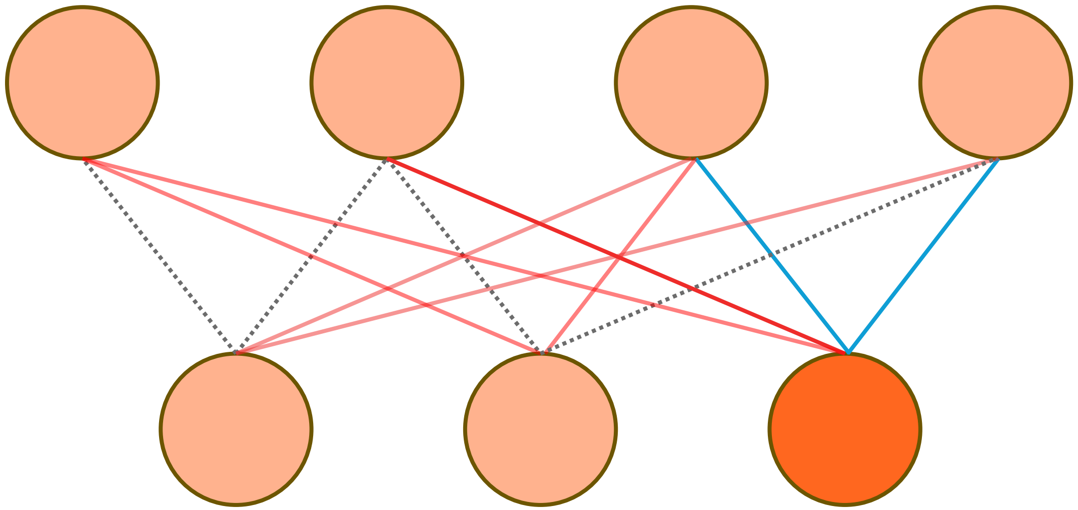
Unstructured vs. **Constant Fan-In Sparsity**

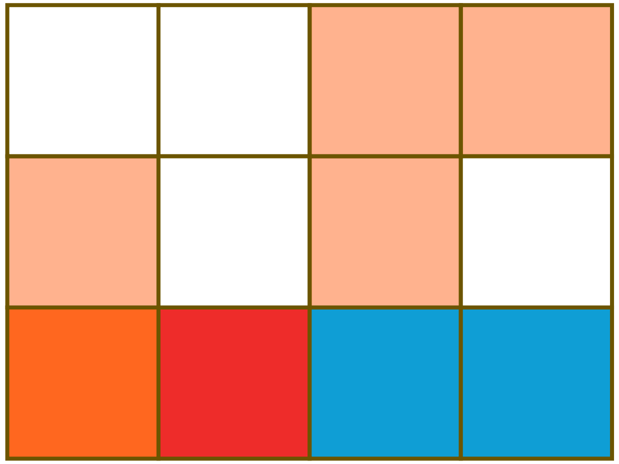




- More Salient
- Less Salient
- Masked

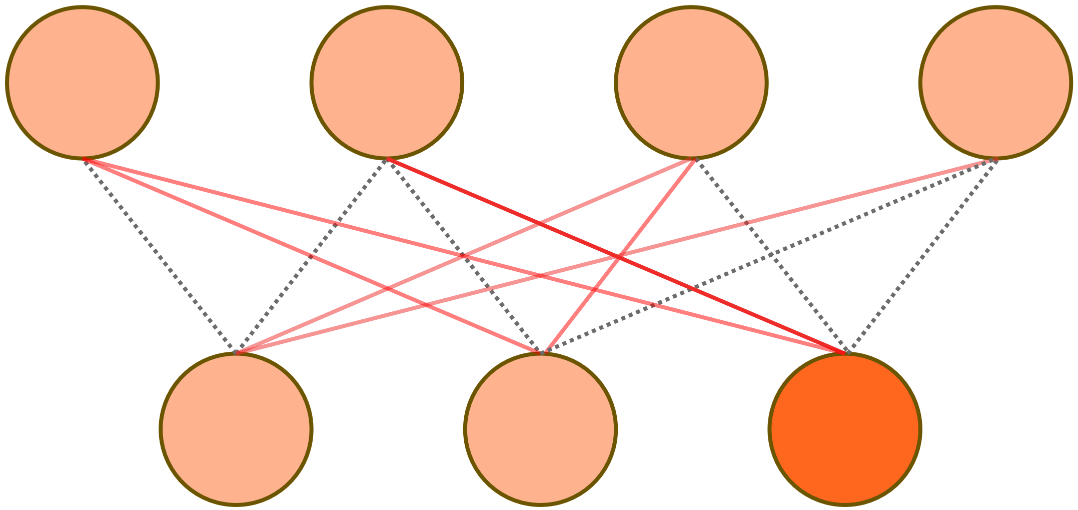
Unstructured vs. **Constant Fan-In Sparsity**

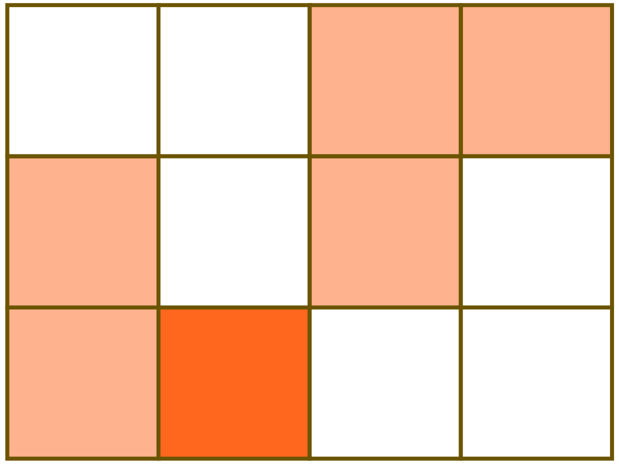




- More Salient
- Less Salient
- Masked

Unstructured vs. **Constant Fan-In Sparsity**

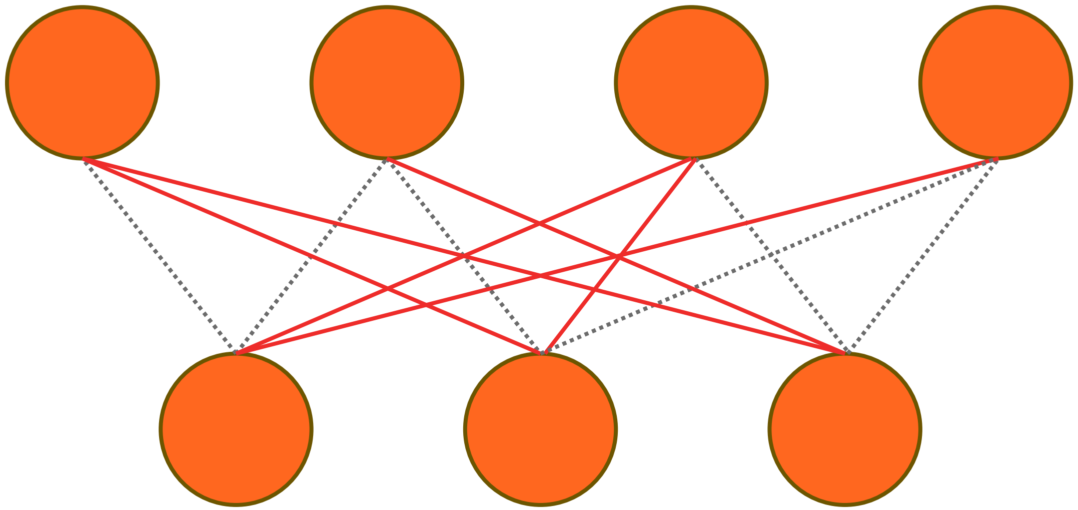


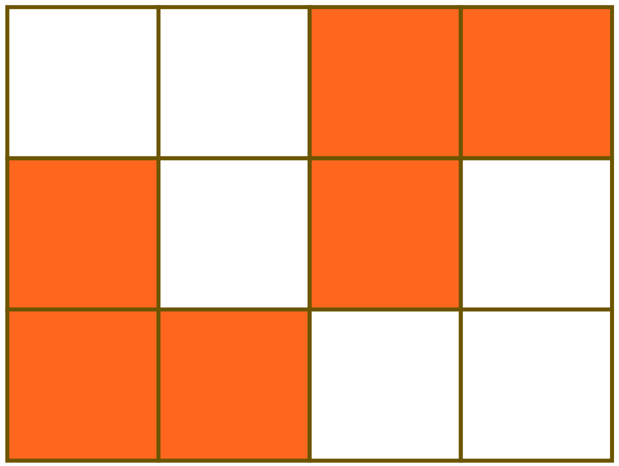


- More Salient
- Less Salient
- Masked

CMI.. UNIVERSITY OF CALGARY

Unstructured vs. **Constant Fan-In Sparsity**

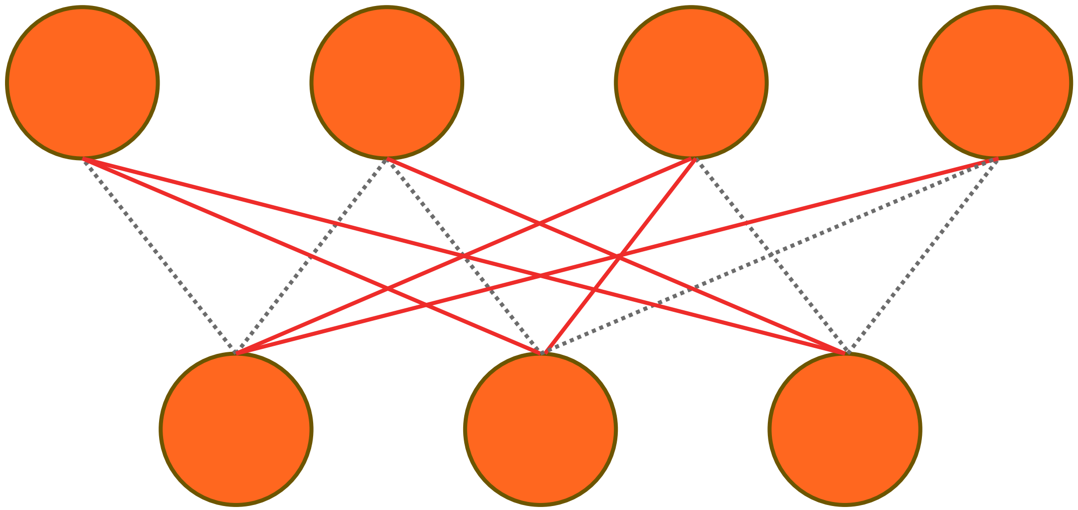


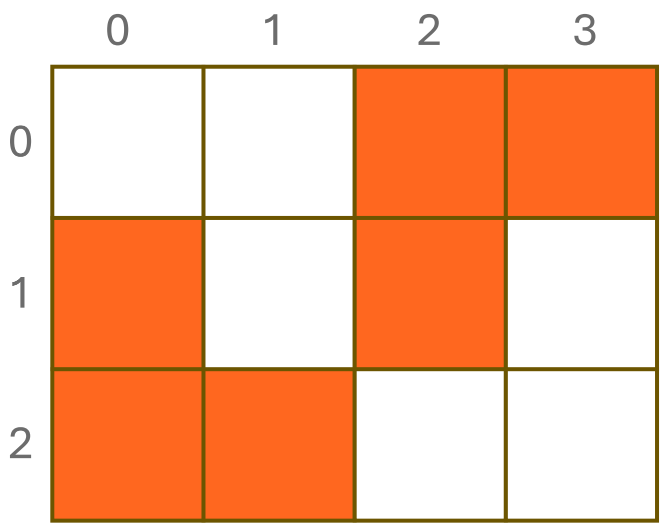


- More Salient
- Less Salient
- Masked

CMI UNIVERSITY OF CALGARY

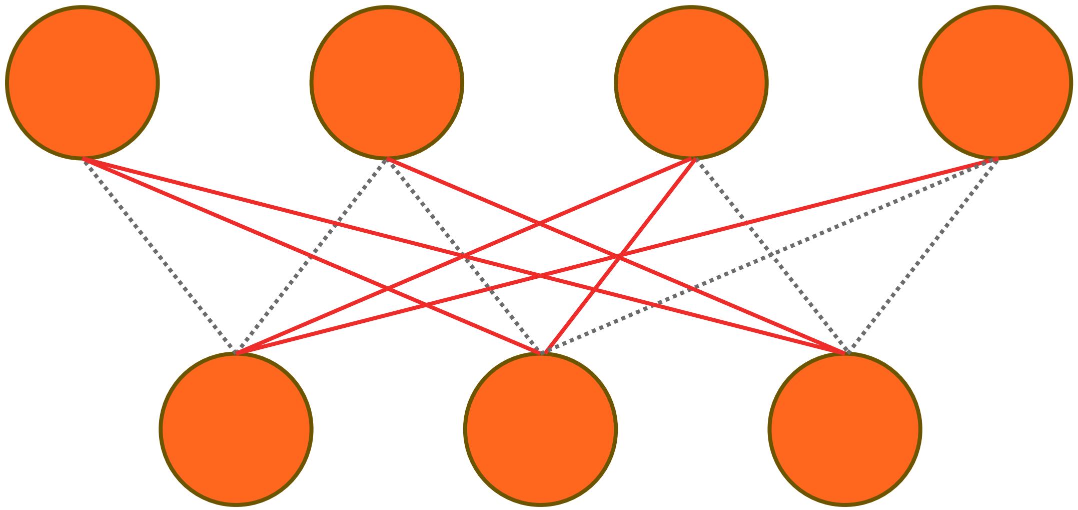
Unstructured vs. **Constant Fan-In Sparsity**





- More Salient
- Less Salient
- Masked

Unstructured vs. **Constant Fan-In Sparsity**



$W_{0,2}$	$W_{0,3}$
$W_{1,0}$	$W_{1,2}$
$W_{2,0}$	$W_{2,1}$

- More Salient
- Less Salient
- Masked

UNIVERSITY OF CALGARY

Constant Fan-In Matrix Multiplication

$$\begin{bmatrix} 0 & 0 & W_{0,2} & W_{0,3} \\ W_{1,0} & 0 & W_{1,2} & 0 \\ W_{2,0} & W_{2,1} & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

Sparse Matrix w/
Constant Fan-in=2

Constant Fan-In Matrix Multiplication

$$
\begin{bmatrix} 0 & 0 & W_{0,2} & W_{0,3} \\ W_{1,0} & 0 & W_{1,2} & 0 \\ W_{2,0} & W_{2,1} & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 0*X_0+0*X_1+W_{0,2}*X_2+W_{0,3}*X_3 \\ W_{1,0}*X_0+0*X_1+W_{1,2}*X_2+0*X_3 \\ W_{2,0}*X_0+W_{2,1}*X_1+0*X_2+0*X_3 \end{bmatrix}
$$

Sparse Matrix w/
Constant Fan-in=2

Constant Fan-In Matrix Multiplication

$$\begin{bmatrix} 0 & 0 & W_{0,2} & W_{0,3} \\ W_{1,0} & 0 & W_{1,2} & 0 \\ W_{2,0} & W_{2,1} & 0 & 0 \end{bmatrix} \bullet \begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} W_{0,2}*X_2+W_{0,3}*X_3 \\ W_{1,0}*X_0+W_{1,2}*X_2 \\ W_{2,0}*X_0+W_{2,1}*X_1 \end{bmatrix}$$

Sparse Matrix w/
Constant Fan-in=2

Constant Fan-In Matrix Multiplication

$$\begin{bmatrix} 0 & 0 & W_{0,2} & W_{0,3} \\ W_{1,0} & 0 & W_{1,2} & 0 \\ W_{2,0} & W_{2,1} & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} W_{0,2}*X_2+W_{0,3}*X_3 \\ W_{1,0}*X_0+W_{1,2}*X_2 \\ W_{2,0}*X_0+W_{2,1}*X_1 \end{bmatrix} = \begin{bmatrix} W_{0,2} & W_{0,3} \\ W_{1,0} & W_{1,2} \\ W_{2,0} & W_{2,1} \end{bmatrix}$$

Sparse Matrix w/ Constant Fan-in=2

Condensed Matrix

Constant Fan-In Matrix Multiplication

$$\begin{bmatrix} 0 & 0 & W_{0,2} & W_{0,3} \\ W_{1,0} & 0 & W_{1,2} & 0 \\ W_{2,0} & W_{2,1} & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} W_{0,2}*X_2+W_{0,3}*X_3 \\ W_{1,0}*X_0+W_{1,2}*X_2 \\ W_{2,0}*X_0+W_{2,1}*X_1 \end{bmatrix} = \begin{bmatrix} W_{0,2} & W_{0,3} \\ W_{1,0} & W_{1,2} \\ W_{2,0} & W_{2,1} \end{bmatrix} \odot \begin{bmatrix} X_2 \\ X_0 \\ X_0 \end{bmatrix} \begin{bmatrix} X_3 \\ X_2 \\ X_1 \end{bmatrix}$$

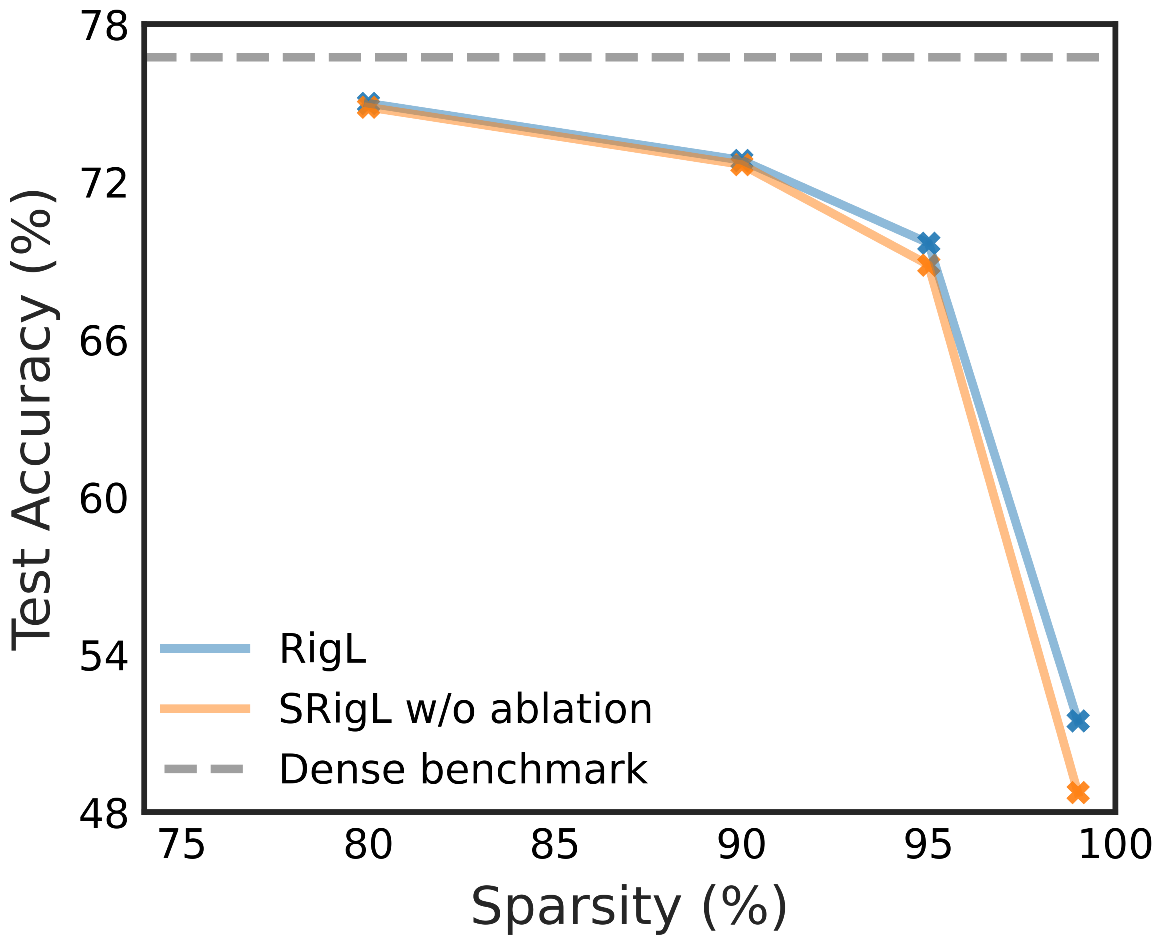
Sparse Matrix w/
Constant Fan-in=2

Condensed Matrix

Input Recombination
Vectors

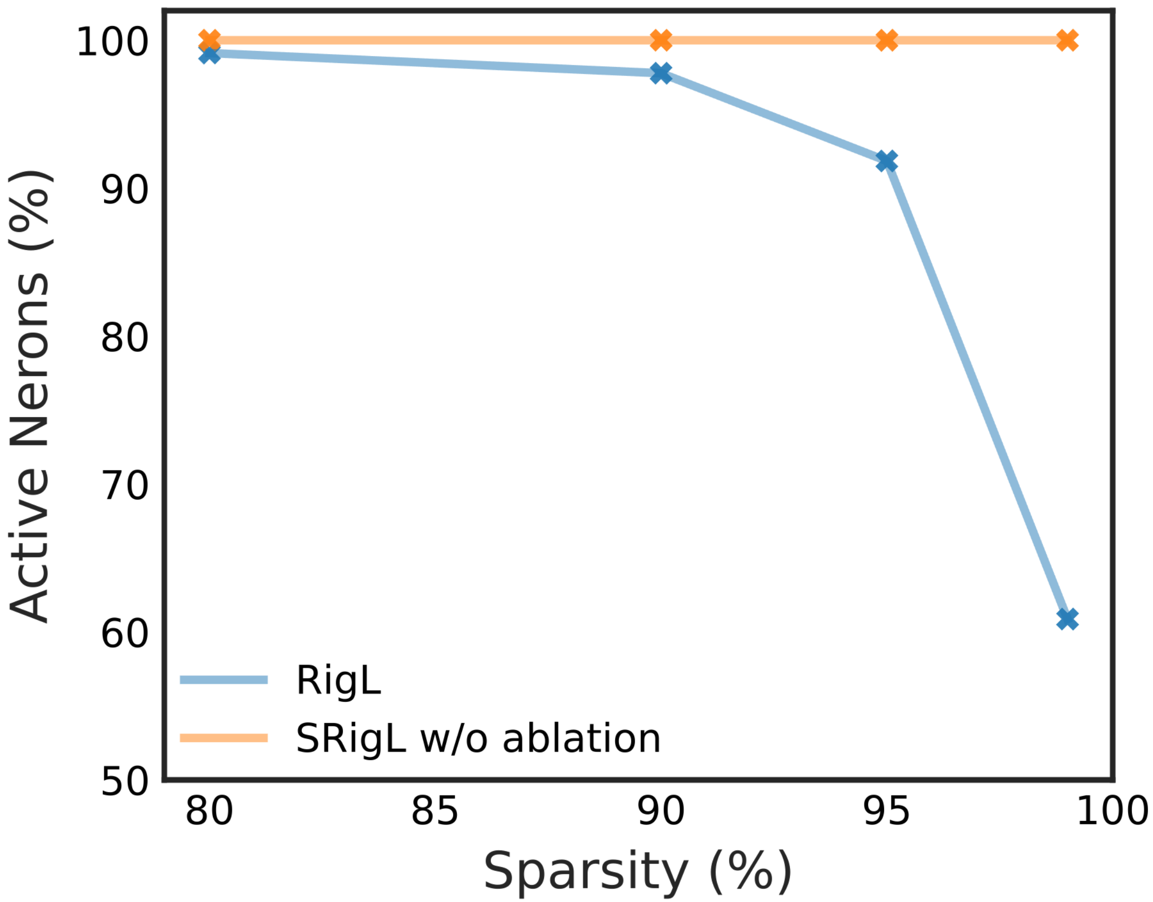
Initial Results (ImageNet/ResNet50)

- We saw similar generalization with constant fan-in as RigL up to **90% sparsity**
- At high sparsities (>= 90%) we found constant fan-in did not match RigL results…

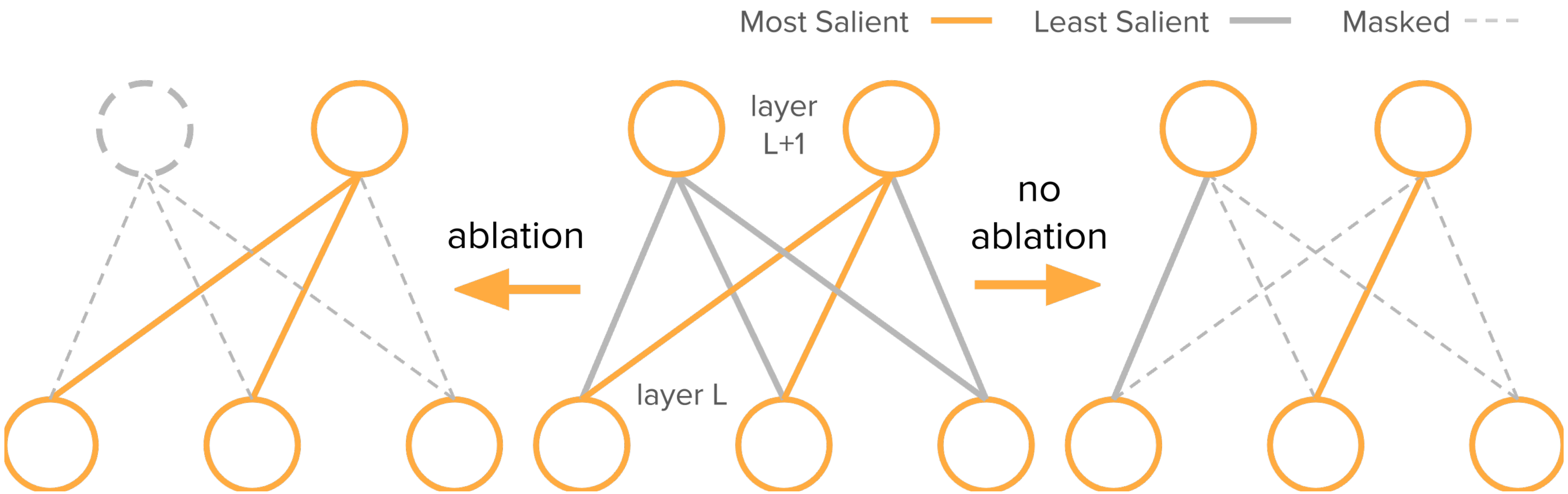


Neuron Ablation

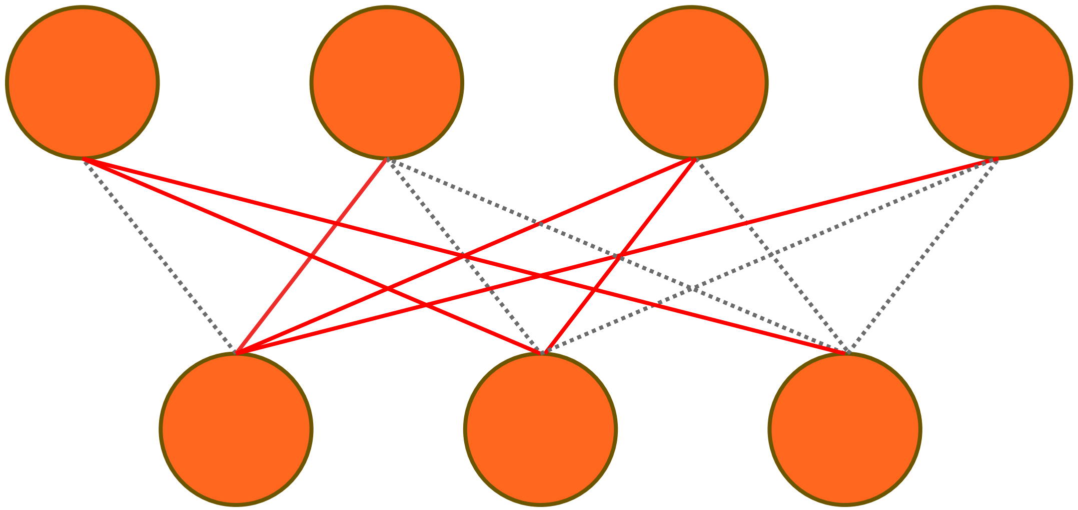
- At high sparsities (>= 90%) we found that RigL **ablates many neurons**
- Effectively RigL at high sparsity **learns to reduce the width of layers!**
- However, a naïve constant fan-in constraint **prohibits removal of neurons**, decreasing performance

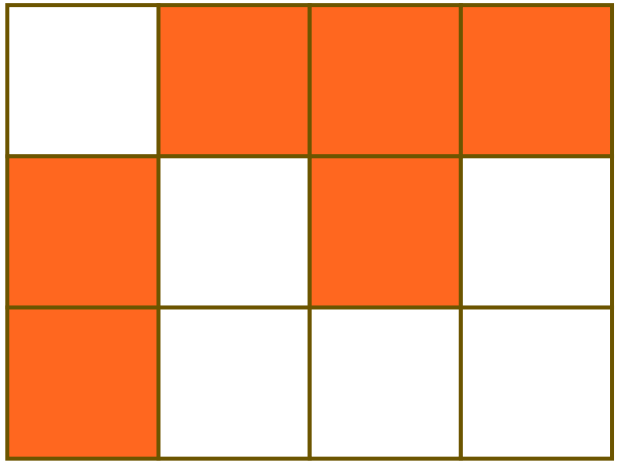


Neuron Ablation



Neuron Ablation

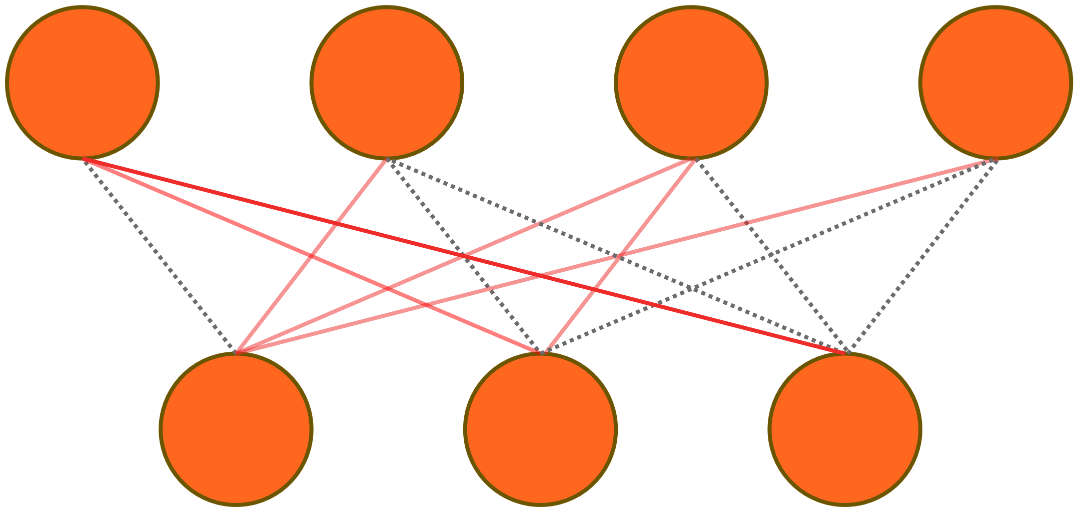


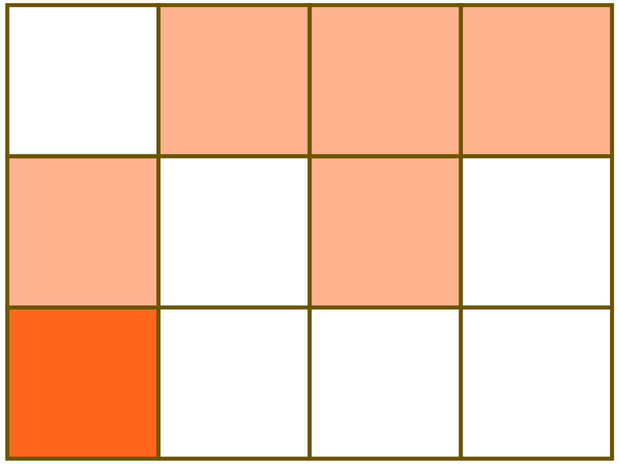


- More Salient
- Less Salient
- Masked

UNIVERSITY OF CALGARY

Neuron Ablation

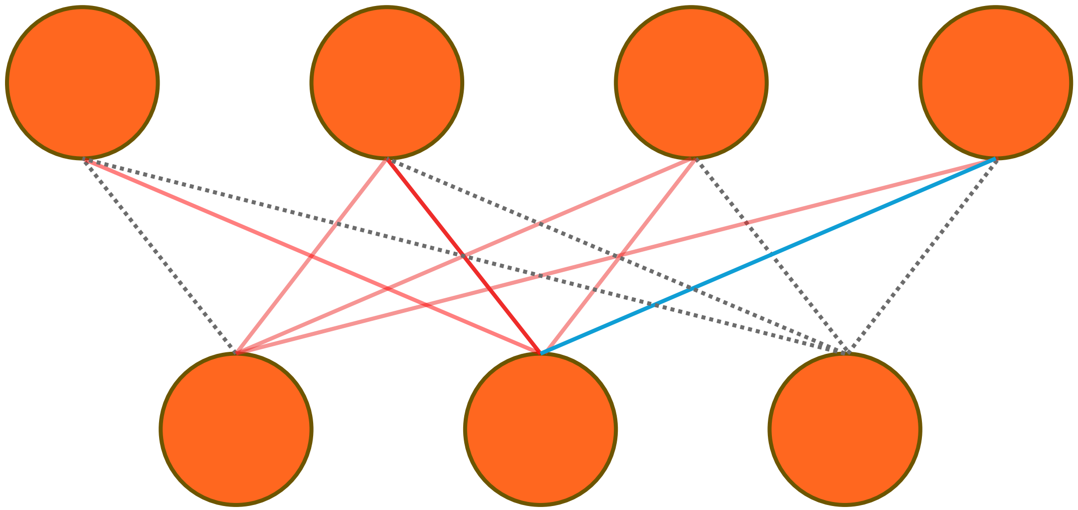


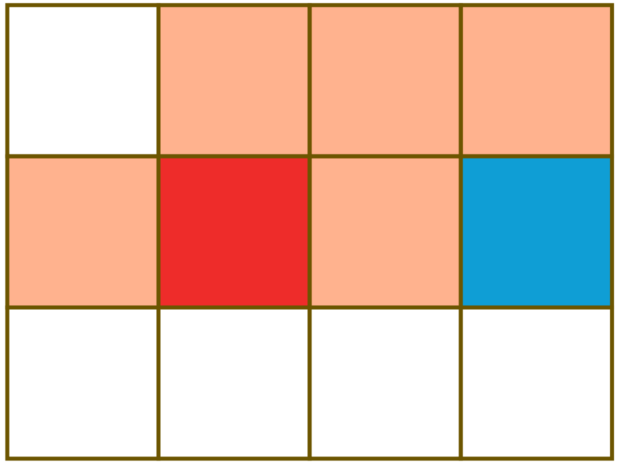


- More Salient
- Less Salient
- Masked

UNIVERSITY OF CALGARY

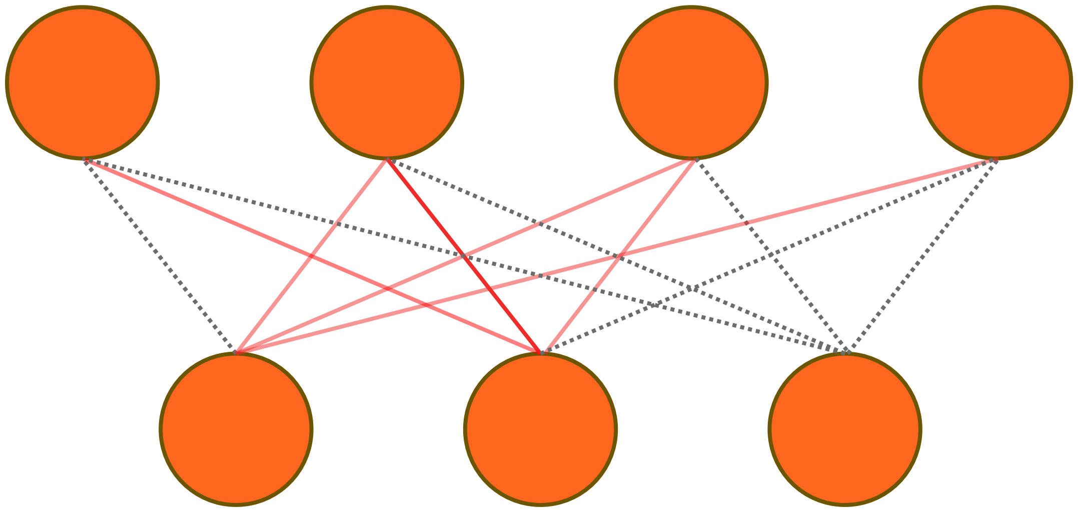
Neuron Ablation

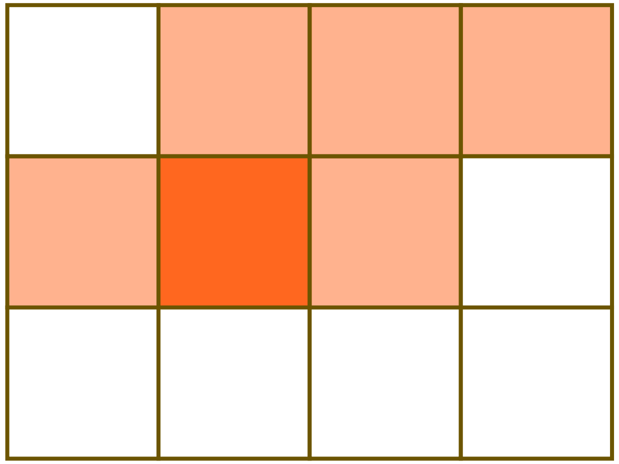




- More Salient
- Less Salient
- Masked

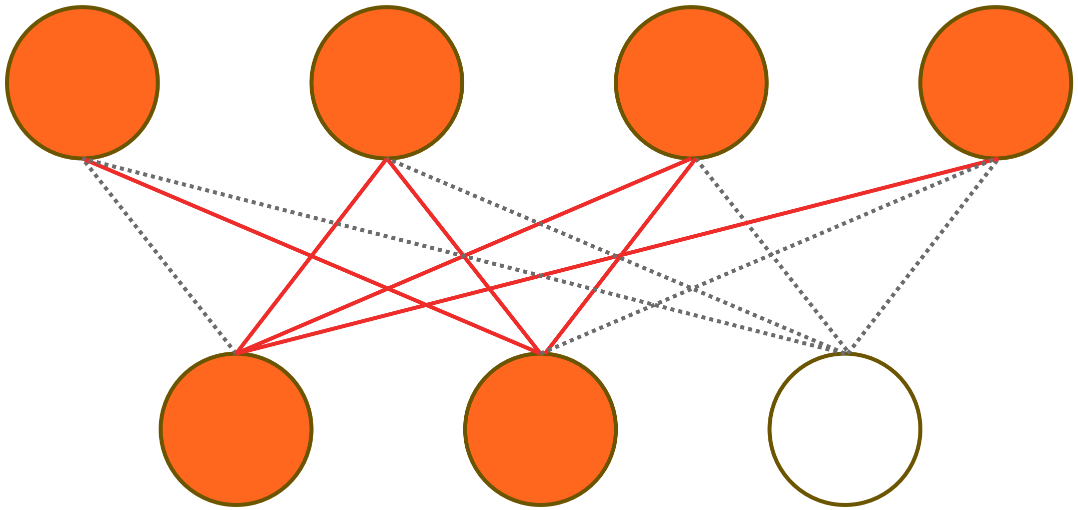
Neuron Ablation

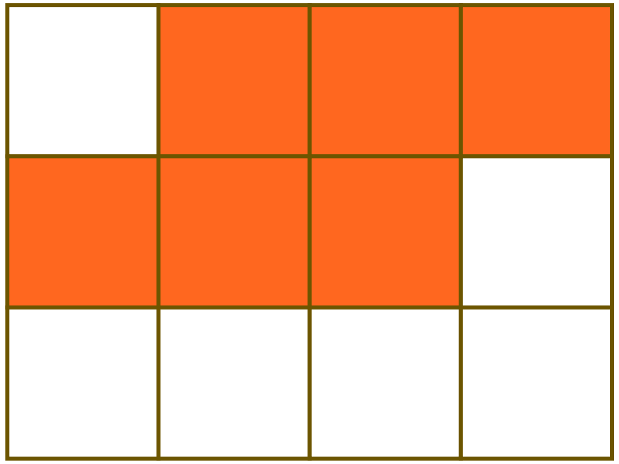




- More Salient
- Less Salient
- Masked

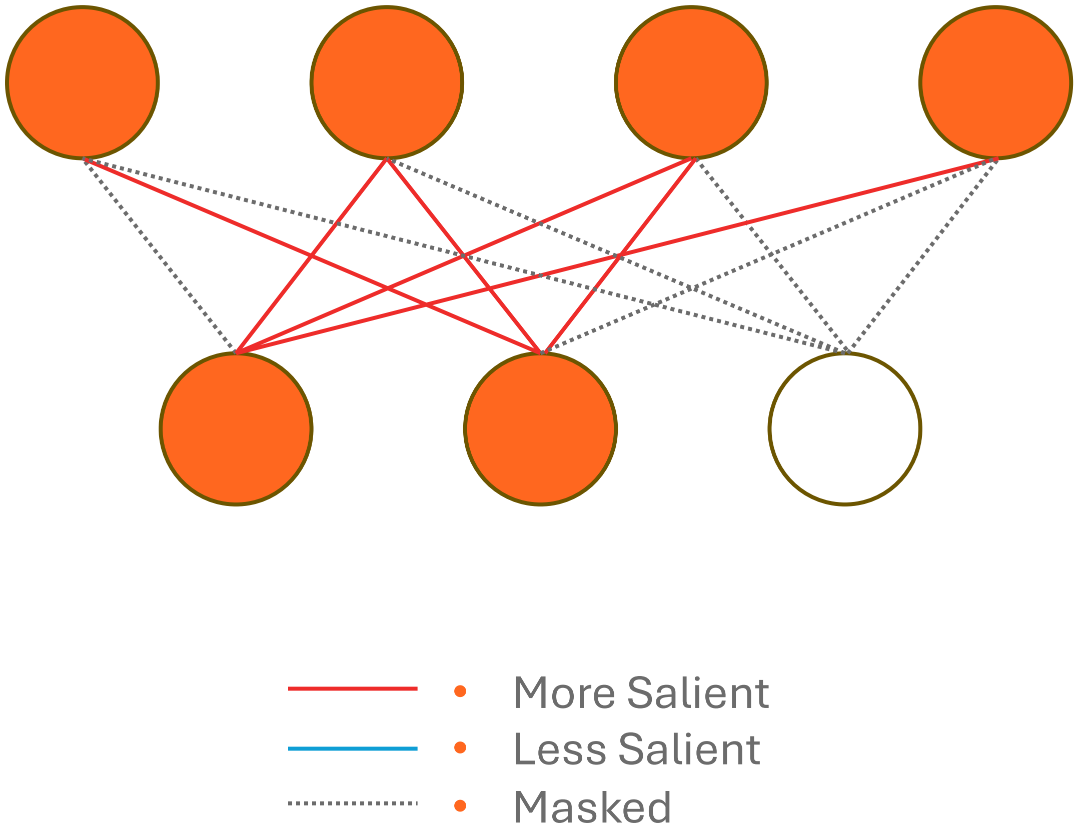
Neuron Ablation





- More Salient
- Less Salient
- Masked

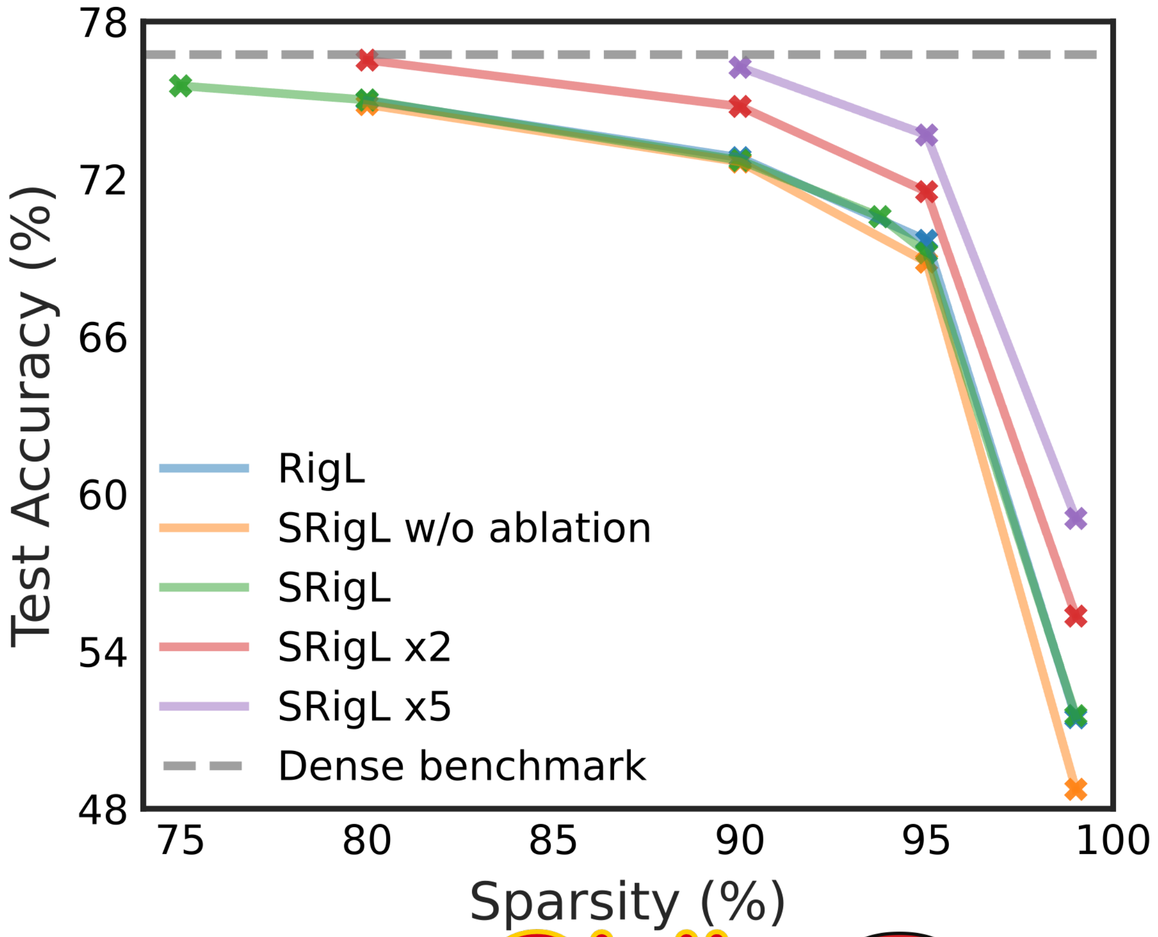
Neuron Ablation



$W_{0,1}$	$W_{0,2}$	$W_{0,3}$
$W_{1,0}$	$W_{1,1}$	$W_{1,2}$

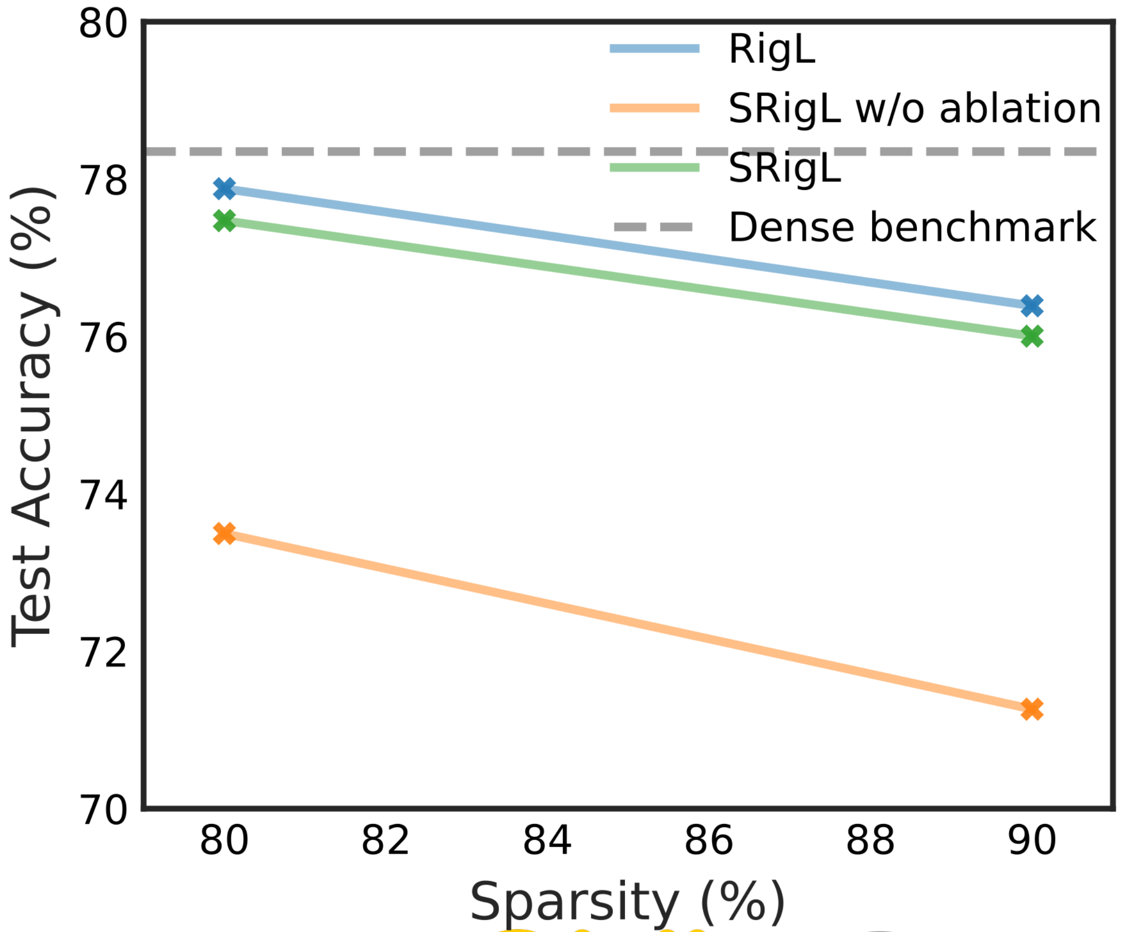
ImageNet/ResNet-50

- SRigL **matches the performance** of RigL at modest sparsities
- At high sparsities, **ablation is required** to maintain generalization
- Extended training of SRigL w/ablation **matches dense benchmark**, even at 90% sparsity (like RigL)!

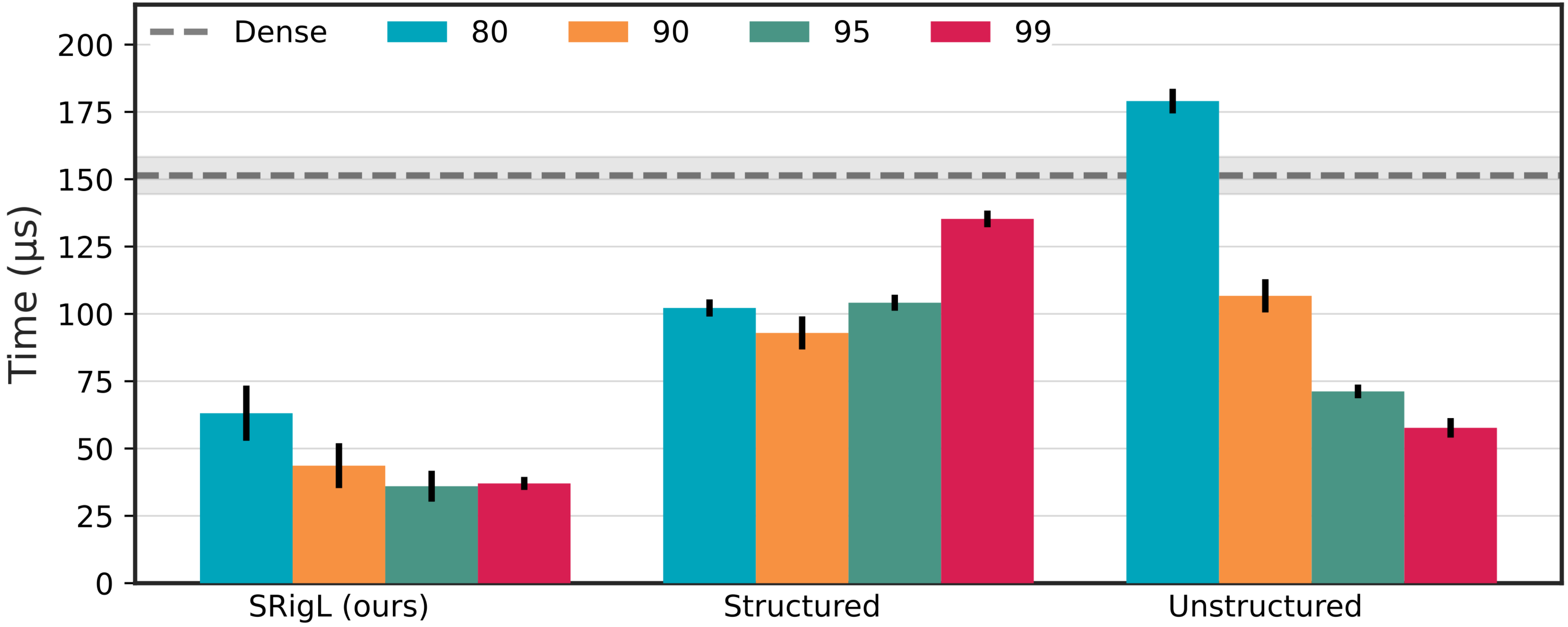


ImageNet/ViT-B-16

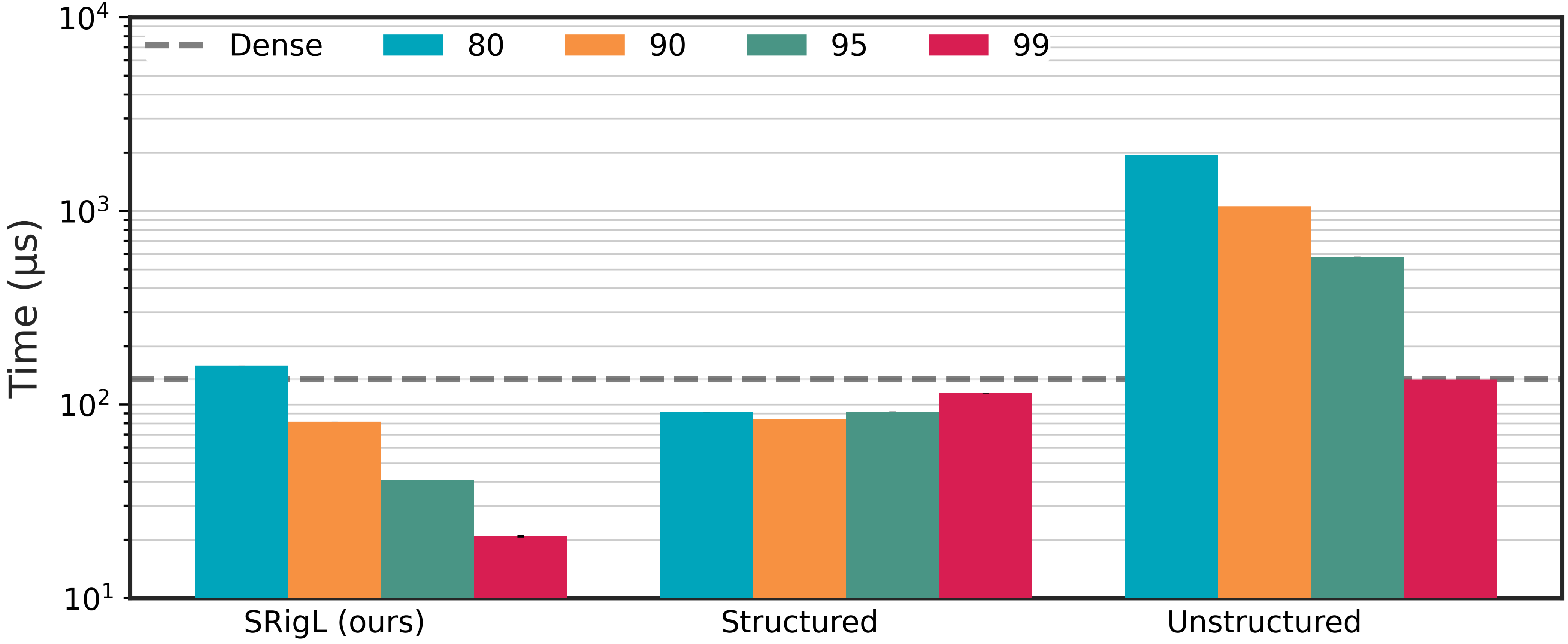
- SRigL also works well with transformer models
- **Neuron ablation is even more effective** with ViT compared to convolutional models



Acceleration - CPU, batch size = 1



Acceleration - GPU, batch size = 2048



ICLR

Dynamic Sparse Training with Structured Sparsity

Thank you!

Mike Lasby · Anna Golubeva · Utku Evci · Mihai Nica · Yani Ioannou

Google · UNIVERSITY of GUELPH · VECTOR INSTITUTE · CMI · UNIVERSITY OF CALGARY