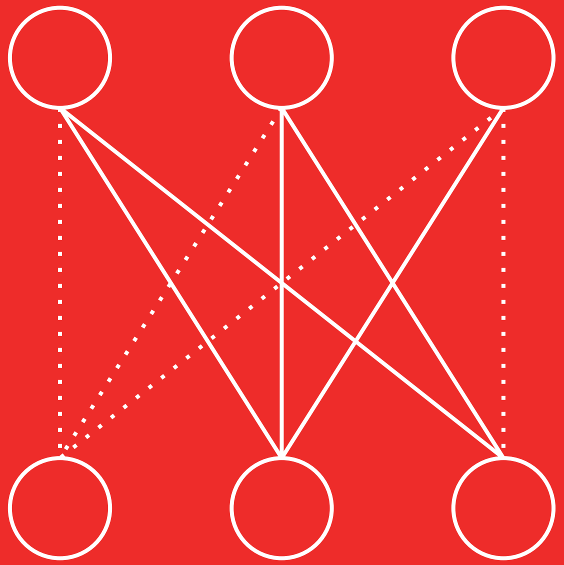
Training Sparse Neural Networks

Yani Ioannou
Schulich Research Chair / Assistant Professor
Dept. of Electrical & Software Engineering,
Schulich School of Engineering

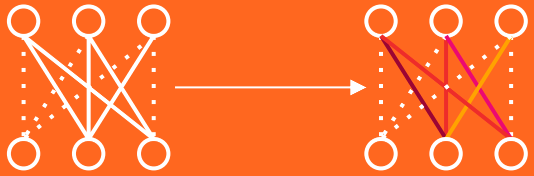
Computational Neuroscience Day

November 14th, 2024

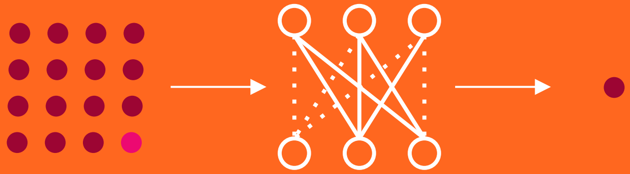



UNIVERSITY OF
CAMBRIDGE
Microsoft
VECTOR
INSTITUTE
NASA
GY69 UES
WAYVE
Google

Training Sparse Neural Networks

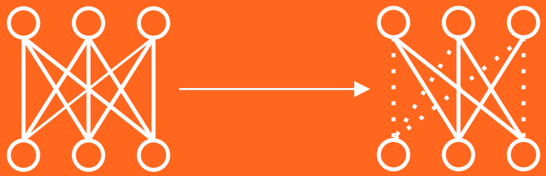


Trustworthy AI (Fairness / Security)



www.calgaryml.com

Learning Network Structure



Distribution Shift (Reinforcement Learning)

UNIVERSITY OF CALGARY

Training Sparse Neural Networks

Michael Lasby
PhD Student

Anna Golubeva
MIT / IAFI

Utku Evci
Google

Deep Reinforcement Learning

Muhammad Athar Ganaie
MSc Student

www.calgaryml.com

Learning Network Structure

Tejas Pote
MSc Student

Rohan Jain
MSc Student

Adnan Mohammad
PhD Student

Trustworthy AI

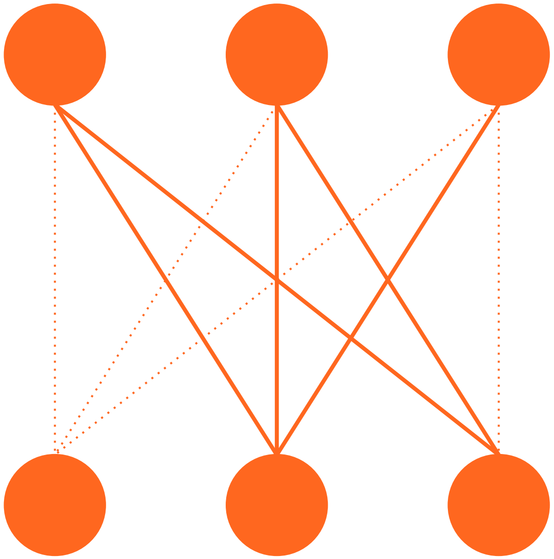
Yufan Feng
MSc Student

Aida Mohammadshahi
MSc Student

UNIVERSITY OF CALGARY

Why Sparse Neural Networks?

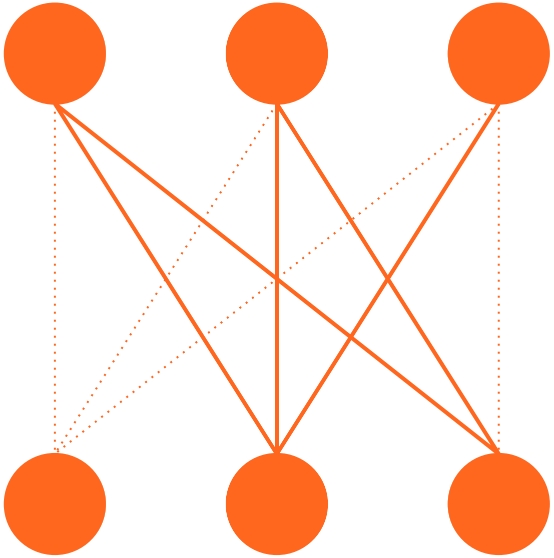
- For fixed number weights, **better generalization, FLOPs at inference**
- Potential to reduce the cost of training NNs
- **Learning the structure of NNs**



Learning Structure of NNs

Structure in Neural Networks:

- NNs are **fully-connected** by default
- I practice we rarely use fully-connected NNs for **learning representations…**
- Instead, we must use our domain knowledge to change **the structure** of the model
    - CNNs, Transformers, RNNs, GNNs, …
- Most of these architectures can be represented as subset of fully-connected NNs



Why Learn NN Structure?

Neural Network Architectures:

- NN architectures are more often built from **pre-designed blocks** that we know are good for some problems/domains
- We can learn how to fit together these blocks better with **Neural Architecture Search (NAS)**
- However, NAS is expensive and can only assemble pre-existing blocks
- **What about when we need a new block?**

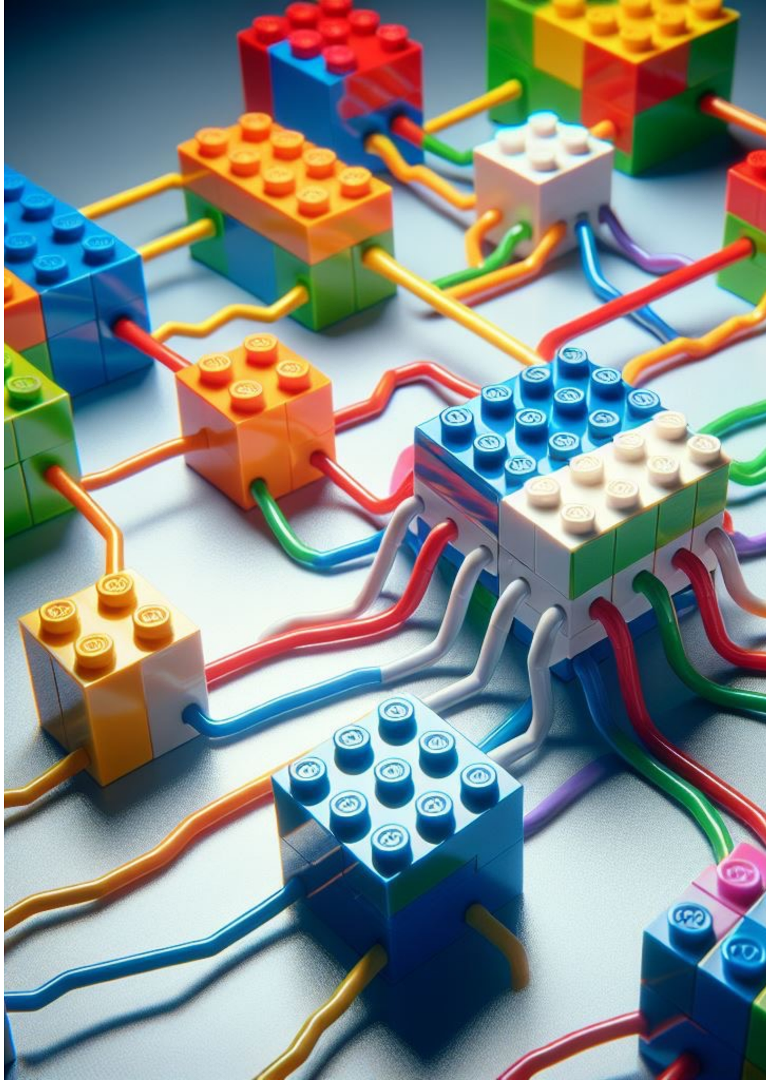


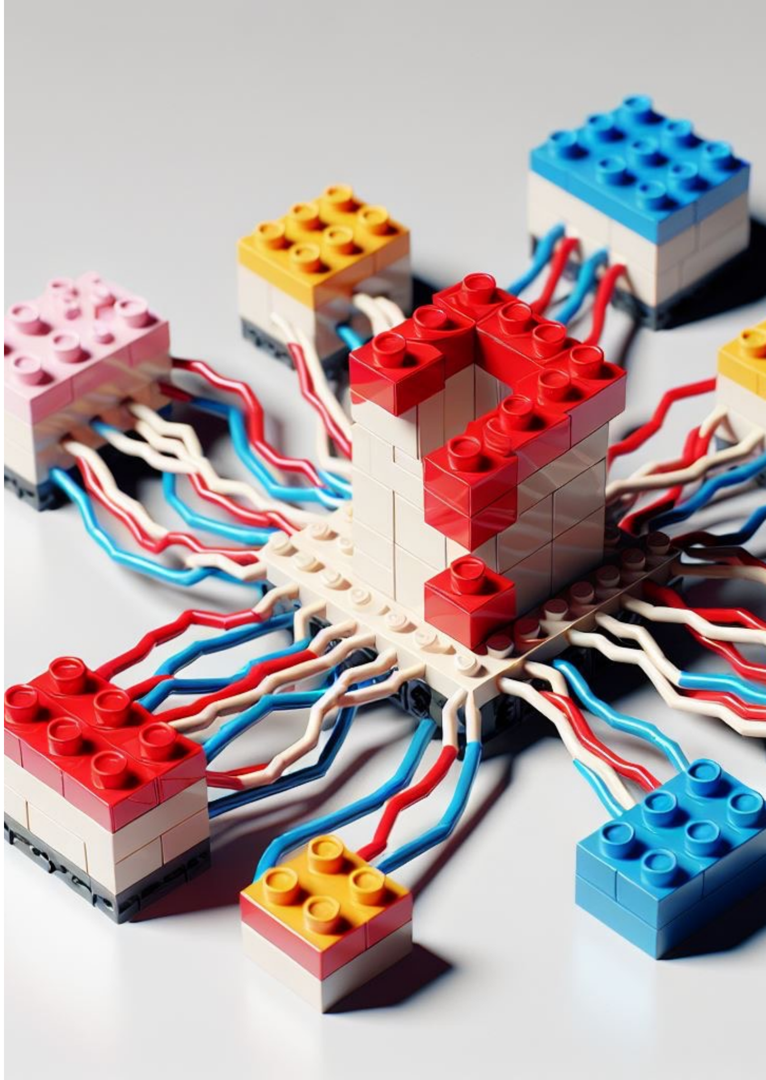
Image by Microsoft Co-pilot Designer

Why not Architecture Search?

Why not Neural Architecture Search:

- NAS doesn't help when we need **new blocks!**
  - i.e. non-NLP, CV, speech, graph, etc.
  - Novel data domain or application
- More important as AI is applied to a broader set of data domains and problems
  - e.g. AI for Science
- **Existing approach is to spend decades of research determining the blocks...**

Image by Microsoft Co-pilot Designer



Why Sparse Training?

Why Sparse Training

- Sparse NNs **learn structure** within dense NNs
  - Learn sparse masks, where weights are removed according to training data
  - Not a new idea! As old as NNs themselves.
- **But** sparse training is more difficult than dense training
  - Many sparse training approaches result in models that don't generalize well
  - Lottery Ticket Hypothesis, etc.

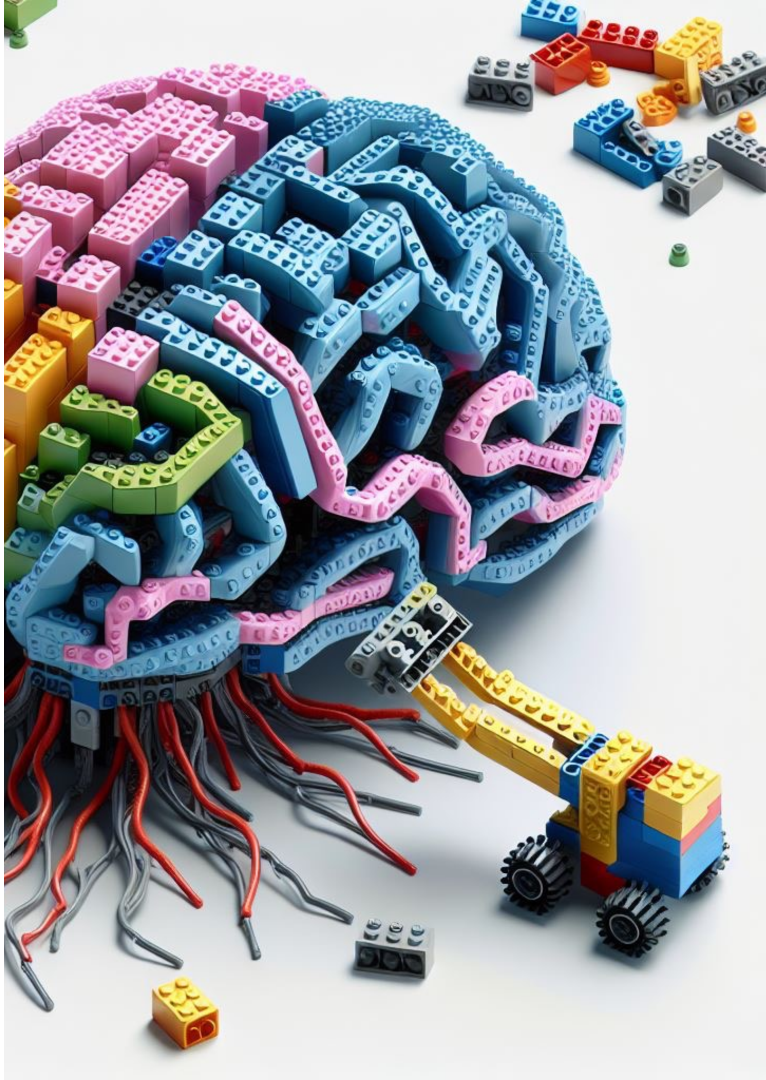
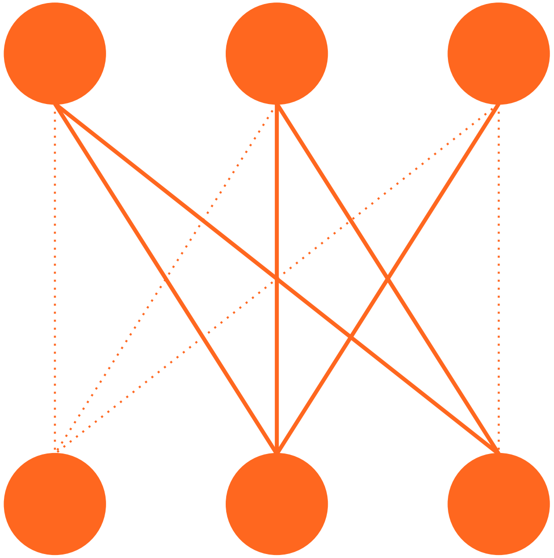


Image by Microsoft Co-pilot Designer

Calgary ML Lab Research

What we have been doing:

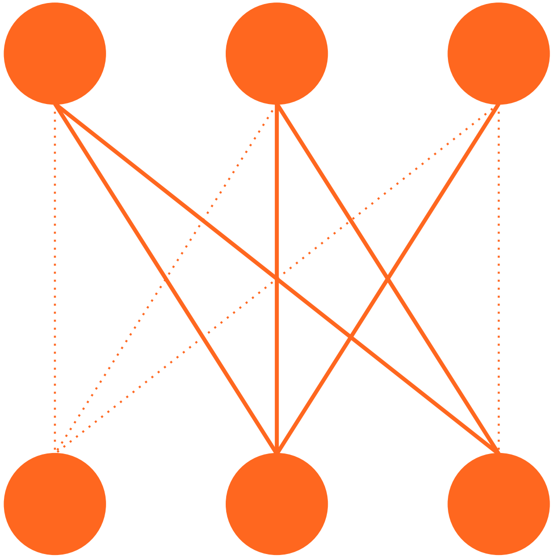
- Understanding why sparse training is difficult
  - **Winning Tickets from Random Initialization: Aligning Masks for Sparse Training.** Rohan Jain, Mohammed Adnan, Ekansh Sharma, and Yani Ioannou. 2nd Workshop on Unifying Representations in Neural Models (UniReps), NeurIPS 2024 Workshops, Vancouver, BC, Canada.
    - Come visit us at NeurIPS Dec 10-15th, 2024!
  - **Gradient Flow in Sparse Neural Networks and How Lottery Tickets Win.** Utku Evci, Yani A. Ioannou, Cem Keskin, and Yann Dauphin. In Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI) 2022, Vancouver, BC, Canada.



Calgary ML Lab Research

What we have been doing:

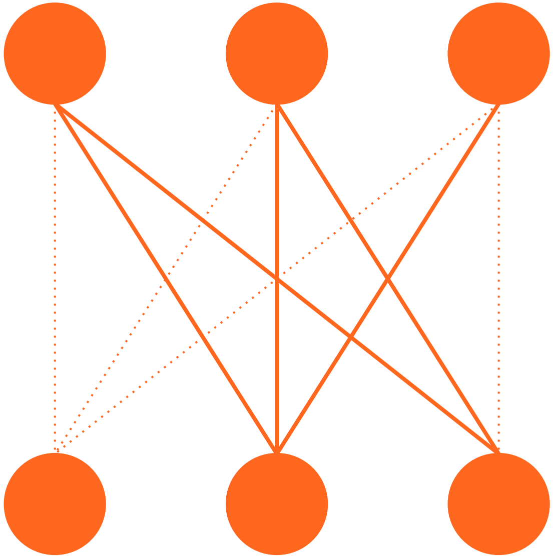
- Making SOTA sparse training methods more practical
  - **Dynamic Sparse Training with Structured Sparsity.** Mike Lasby, Anna Golubeva, Utku Evci, Mihai Nica, and Yani Ioannou. In International Conference on Learning Representations (ICLR), Vienna, Austria 2024.
  - **Navigating Extremes: Dynamic Sparsity in Large Output Spaces.** Nasib Ullah, Erik Schultheis, Mike Lasby, Yani Ioannou, and Rohit Babbar. In 38th Annual Conference Neural Information Processing Systems (NeurIPS) 2024, Vancouver, BC, Canada 2024.
    - Come visit us at NeurIPS Dec 10-15th, 2024!



Calgary ML Lab Research

What I will talk about today:

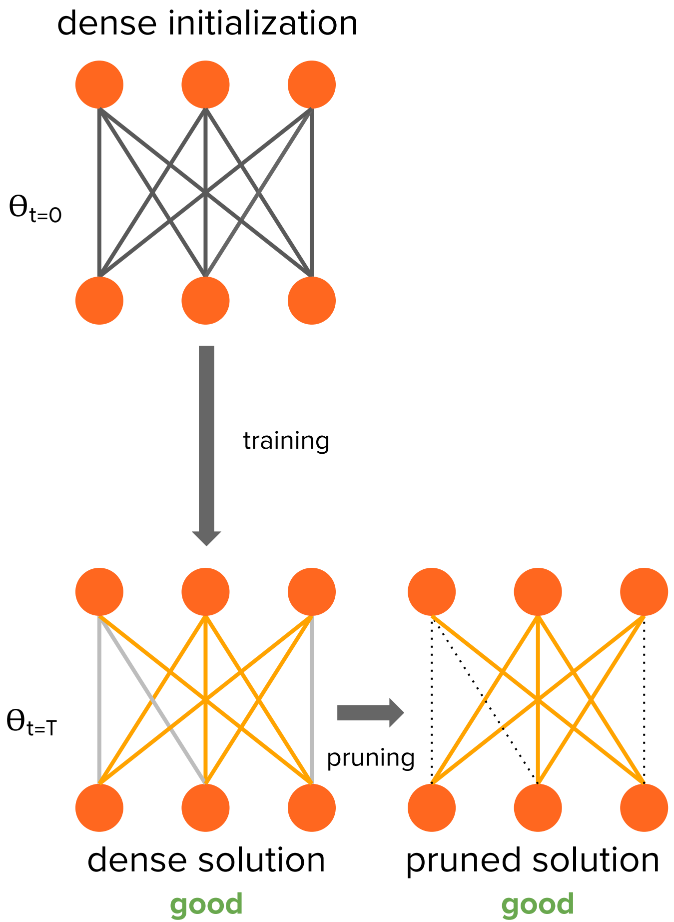
- **Dynamic Sparse Training with Structured Sparsity.** Mike Lasby, Anna Golubeva, Utku Evci, Mihai Nica, and Yani Ioannou. In International Conference on Learning Representations (ICLR), Vienna, Austria 2024.
- **But what is Dynamic Sparse Training? ...**



Dynamic Sparse Training

Pruning Dense Models

- Pruning is an effective method to find a sparse mask for a dense models
- Sparse masks of 85 – 95% with similar generalization!
- However, **we still need to train a dense model…**



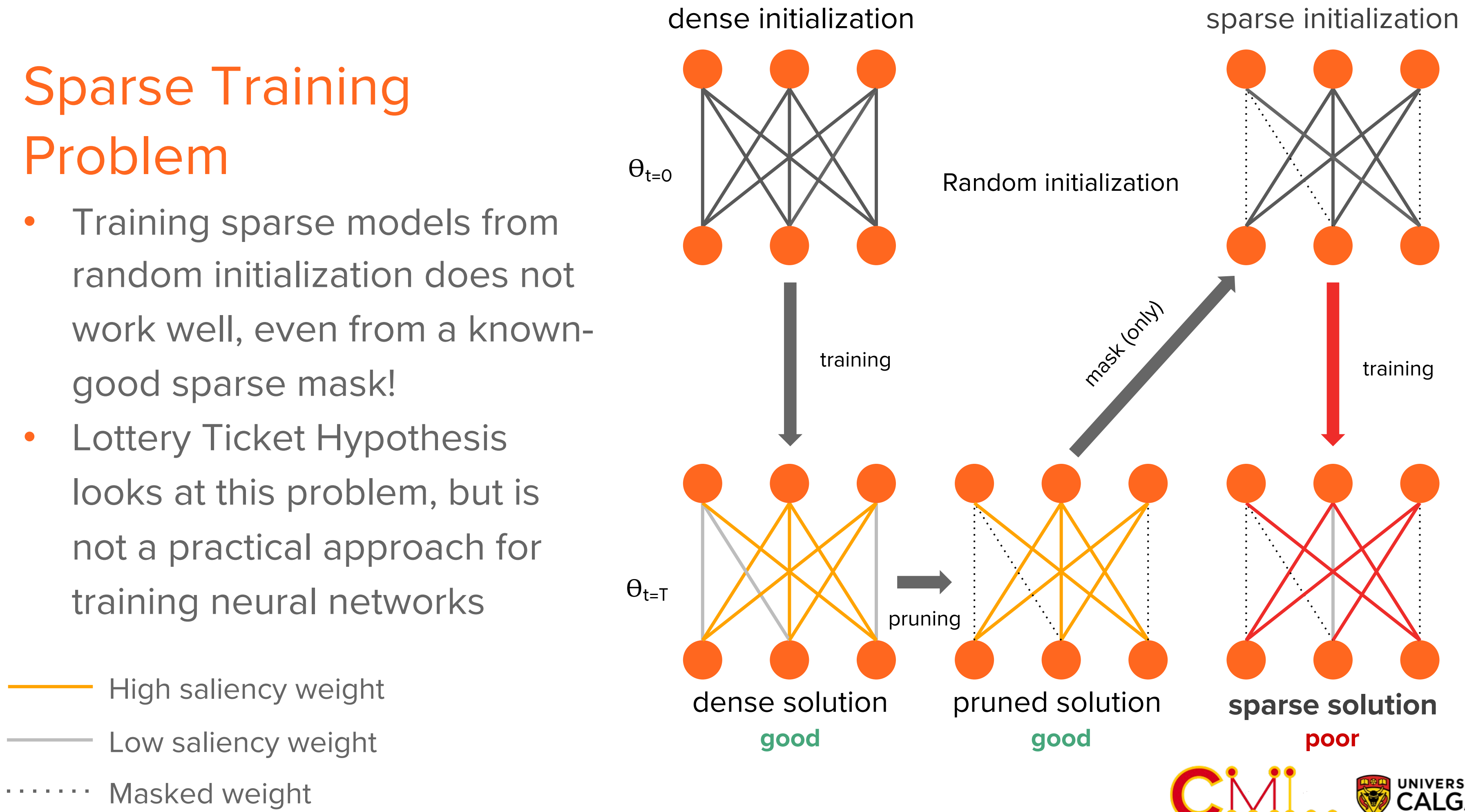
High saliency weight

Low saliency weight

Masked weight

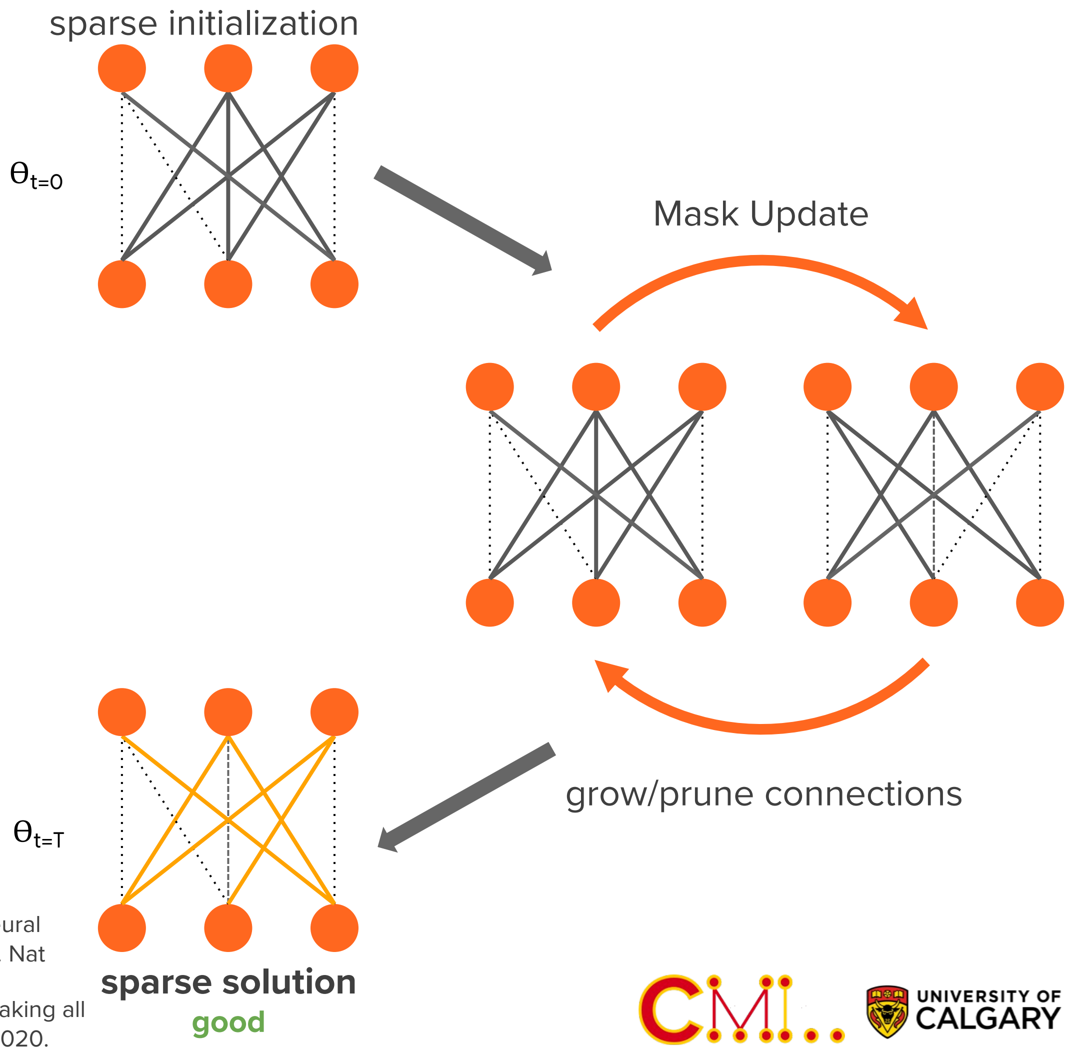
Sparse Training Problem

- Training sparse models from random initialization does not work well, even from a known-good sparse mask!
- Lottery Ticket Hypothesis looks at this problem, but is not a practical approach for training neural networks



Dynamic Sparse Training

- Dynamic Sparse Training (DST), e.g. Sparse Evolutional Training[1] (SET) and Rigging the Lottery Ticket[2] (RigL), are alternatives
- DST trains **sparse-to-sparse**: i.e. from sparse initialization to sparse solution
- Achieves **similar generalization to dense training!**



1. Mocanu, D.C., Mocanu, E., Stone, P. et al. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. Nat Commun 9, 2383 (2018).
2. Evci, U., Gale, T., Menick, J., Castro, P. S., Elsen, E. Rigging the lottery: Making all tickets winners, International Conference on Machine Learning (ICML), 2020.

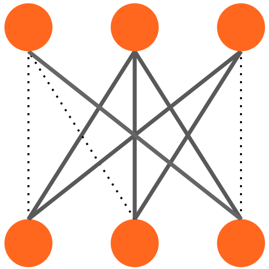
Dynamic Sparse Training

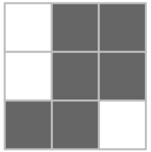
So why aren't we all using DST for deep neural network training or inference?

- **Uses unstructured sparse weight matrices**: hard to accelerate in practice on GPU/CPU

sparse initialization

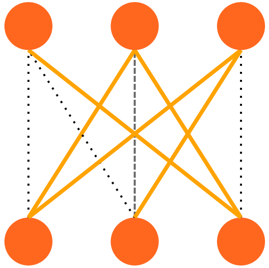
$\theta_{t=0}$





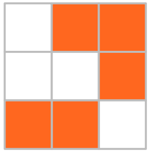
unstructured sparse
initialization

DST
training



$\theta_{t=T}$

sparse solution
good

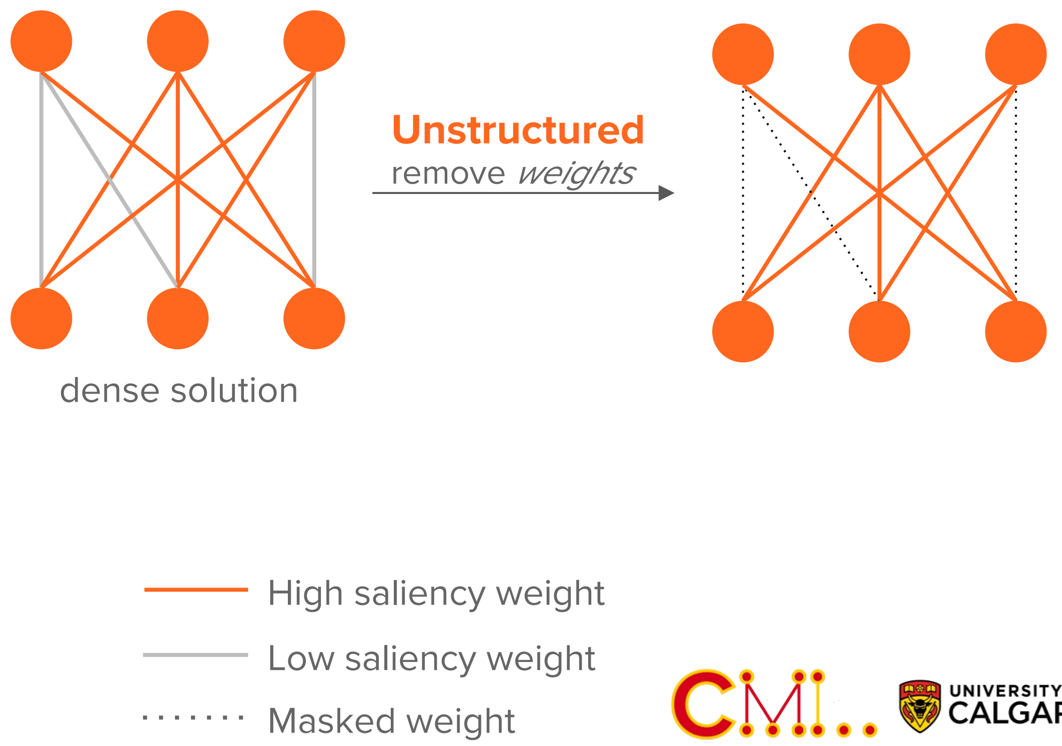


unstructured sparse
solution

1. Mocanu, D.C., Mocanu, E., Stone, P. et al. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. Nat Commun 9, 2383 (2018).
2. Evci, U., Gale, T., Menick, J., Castro, P. S., Elsen, E. Rigging the lottery: Making all tickets winners, International Conference on Machine Learning (ICML), 2020.

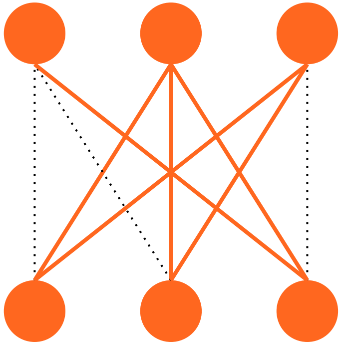
Background:
Unstructured v.s.
Structured Sparsity

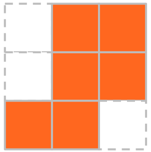
Sparsity/Pruning Categories



Unstructured Pruning

(remove weights)





sparse
weight matrix

High saliency weight

Low saliency weight

Masked weight

Pro:

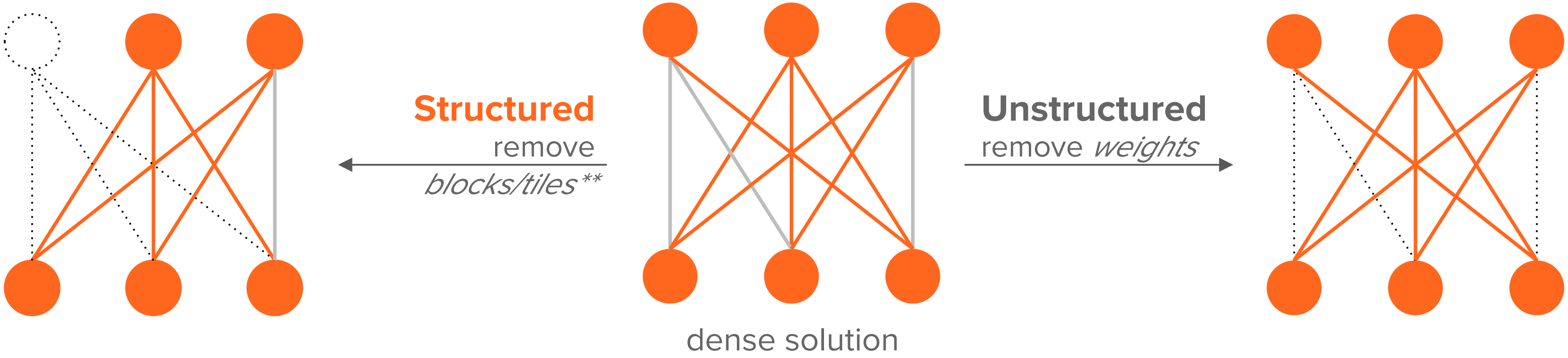
- Good generalization at **very high sparsity** (even 85-95%)
- Fewer **theoretical** FLOPS

Con:

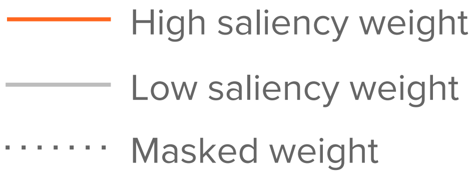
- Poorly supported by acceleration libraries/hardware
- Theoretical speedups not realized on real-world hardware

UNIVERSITY OF CALGARY

Sparsity/Pruning Categories

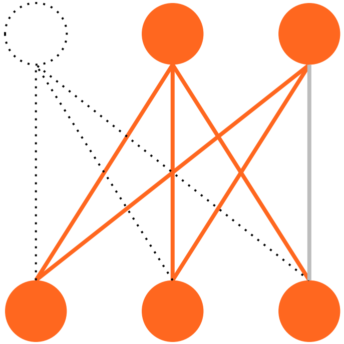


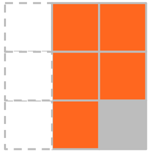
** Removes tiles or blocks of contiguous weights



Structured Pruning

(removing blocks/tiles)





smaller dense
weight matrix

High saliency weight
Low saliency weight
Masked weight

Pro:

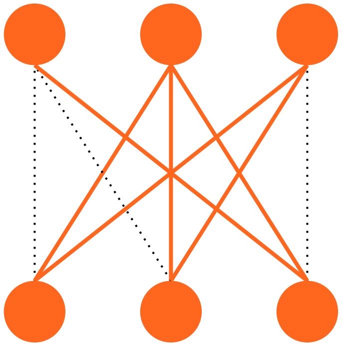
- Better supported by acceleration libraries (BLAS) / hardware **(faster in practice!)**
- It’s effectively a smaller dense model if removing neurons

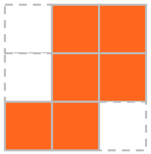
Con:

- Poor generalization at **very high sparsity**

N:M Structured Pruning

(keep N weights in contiguous blocks of size M)





N:M (2:3) weight matrix

High saliency weight

Low saliency weight

Masked weight

Pro:

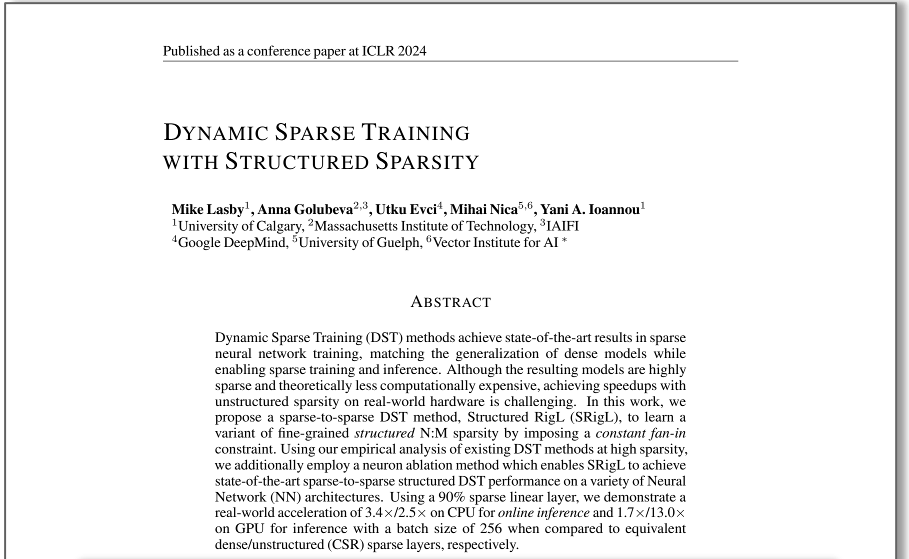
- Better supported by acceleration libraries / hardware than unstructured **(e.g. 2:4 on Nvidia Ampere)**
- Good generalization, similar to unstructured (depending on N:M)

Con:

- Not as fast to accelerate as block sparsity

Can DST Learn structured Sparse Models?

Dynamic Sparse Training with Structured Sparsity.
Mike Lasby, Anna Golubeva, Utku Evci, Mihai Nica, Yani Ioannou
International Conference on Learning Representations (ICLR) 2024

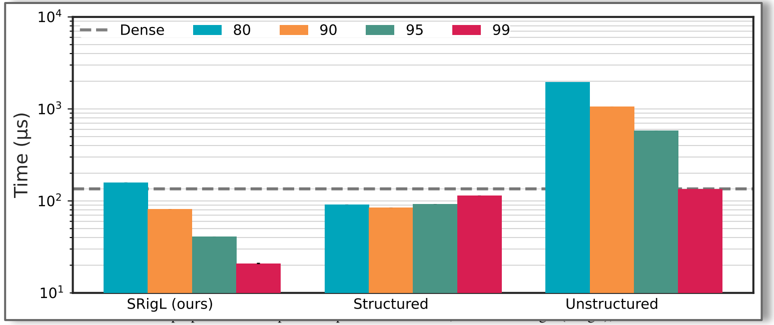

Published as a conference paper at ICLR 2024

DYNAMIC SPARSE TRAINING WITH STRUCTURED SPARSITY

Mike Lasby[1], Anna Golubeva[2,3], Utku Evci[4], Mihai Nica[5,6], Yani A. Ioannou[1]
[1]University of Calgary, [2]Massachusetts Institute of Technology, [3]IAIFI
[4]Google DeepMind, [5]University of Guelph, [6]Vector Institute for AI *

ABSTRACT

Dynamic Sparse Training (DST) methods achieve state-of-the-art results in sparse neural network training, matching the generalization of dense models while enabling sparse training and inference. Although the resulting models are highly sparse and theoretically less computationally expensive, achieving speedups with unstructured sparsity on real-world hardware is challenging. In this work, we propose a sparse-to-sparse DST method, Structured RigL (SRigL), to learn a variant of fine-grained *structured* N:M sparsity by imposing a *constant fan-in* constraint. Using our empirical analysis of existing DST methods at high sparsity, we additionally employ a neuron ablation method which enables SRigL to achieve state-of-the-art sparse-to-sparse structured DST performance on a variety of Neural Network (NN) architectures. Using a 90% sparse linear layer, we demonstrate a real-world acceleration of 3.4×/2.5× on CPU for *online inference* and 1.7×/13.0× on GPU for inference with a batch size of 256 when compared to equivalent dense/unstructured (CSR) sparse layers, respectively.

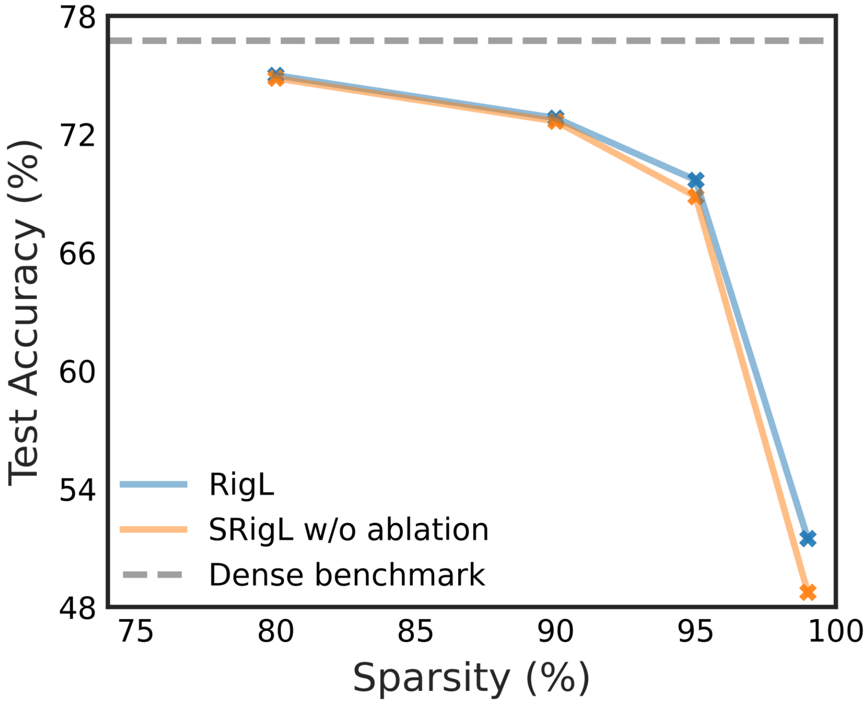


Mike Lasby
Anna Golubeva
Utku Evci
Mihai Nica

- Used DST to learn a variant of N:M structured sparsity
- Does not limit generalization performance
- Show GPU acceleration with the learned models (**1.7x** @90%)

Structured DST: Initial Results

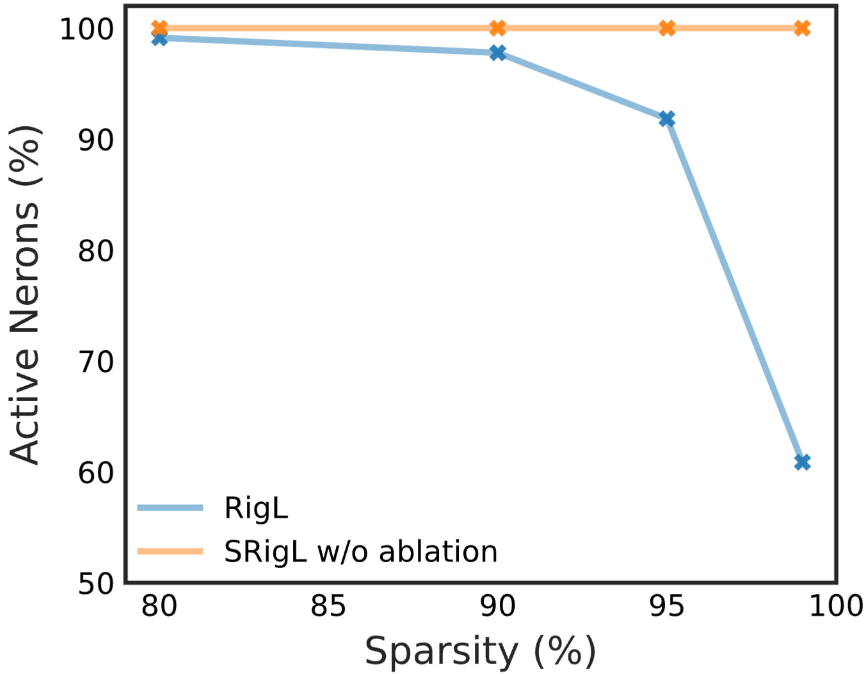
- Enforced structured sparsity in existing state-of-the-art DST method (RigL)
- We saw similar generalization with as unstructured RigL **up to 90% sparsity**
- At high sparsity (>= 90%) we found constant fan-in did not match RigL results...



ImageNet / ResNet50

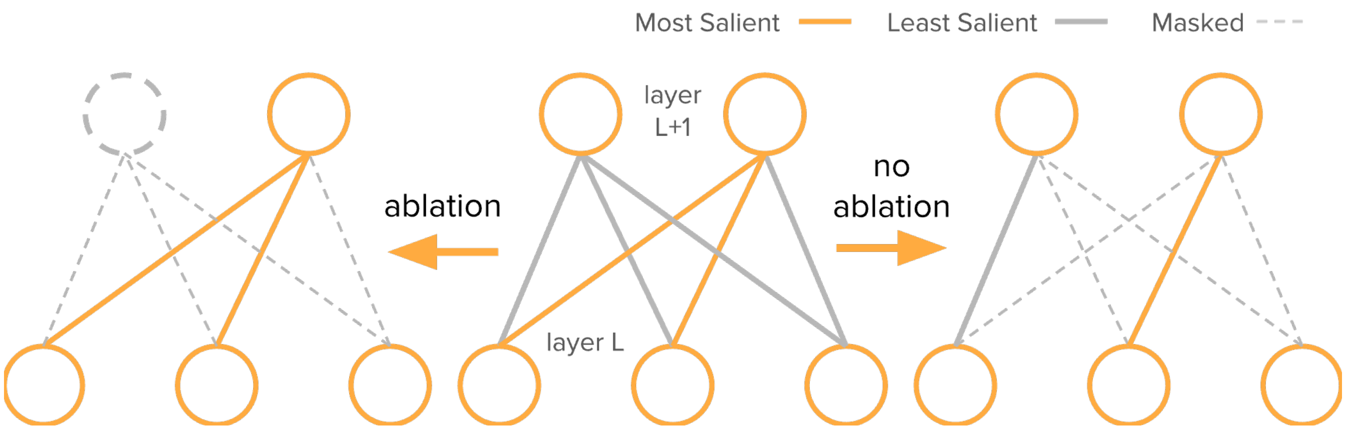
Neuron Ablation in DST Methods

- We investigated what unstructured RigL learned at high sparsity (>90%)
- We found **RigL ablates many neurons**, i.e. it removes *whole neurons…*



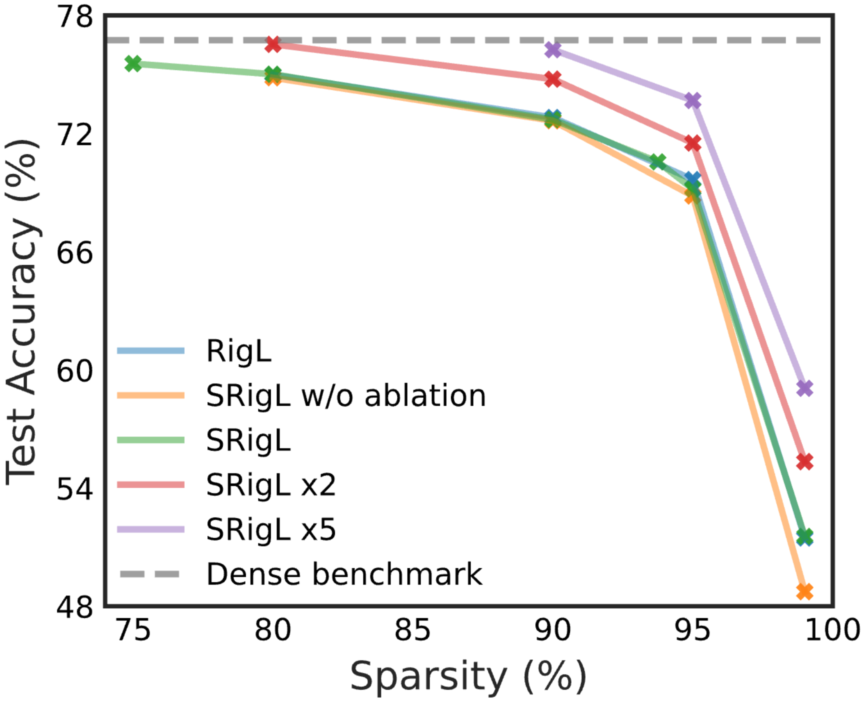
ImageNet / ResNet50

Neuron Ablation



Neuron Ablation in SRigL

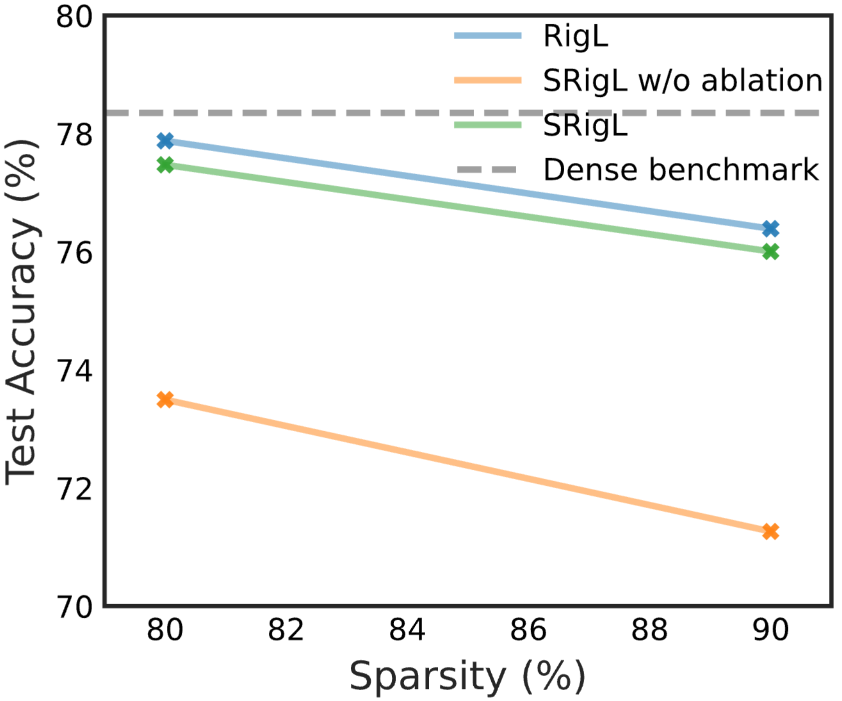
- Effectively RigL at high sparsity **learns to reduce the width of layers!**
- Our naïve structured sparsity constraint prohibited ablation
- At high sparsity, **ablation is required** to maintain generalization
- Extended training of structured RigL w/ablation **matches dense training baseline**, even at 90% sparsity (like unstructured RigL)!



ImageNet / ResNet50

Neuron Ablation in DST Methods

- Our findings also applied to **Transformer** MLP layers in Vision Transformers (ViT)
- In fact, neuron ablation is **even more effective** with ViT compared to convolutional models!



ImageNet / ResNet50

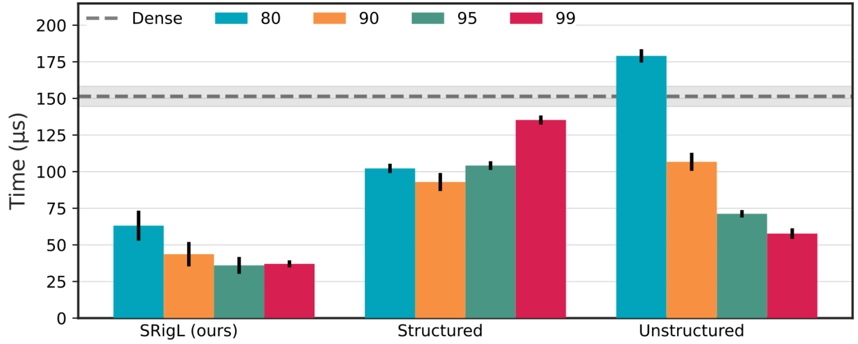
Benchmarks

- Vision Transformer MLP layers
- CPU (top): **3.4x faster than dense inference at 90% sparsity**
- GPU (bottom): **1.7x faster than dense at 90% sparsity**

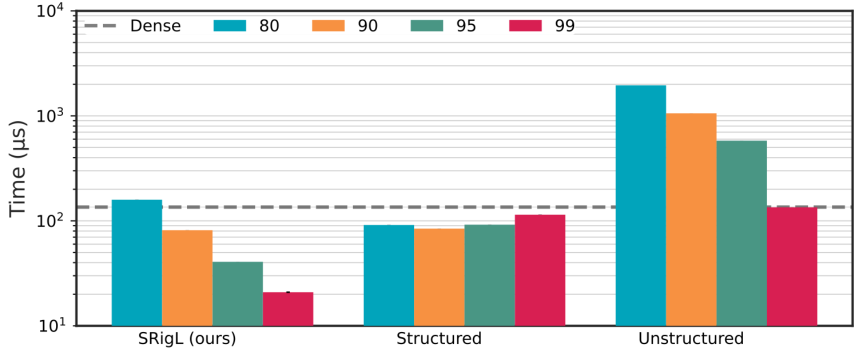
Table 4: Top-1 test accuracy of ViT-B/16 trained on ImageNet with or w/o neuron ablation

	RigL	SRigL	
sparsity (%)†		w/o	w/ ablation
80	**77.9**	73.5	77.5
90	**76.4**	71.3	76.0
0	*dense ViT-B/16*: 78.35		

†Sparsity level set for all modules *except multi-headed attention input projections*, which remain dense. See Appendix D.3 for more details.



(a) CPU online inference



(b) GPU inference with batch size of 256

Figure 4: Comparing real-world timings for a fully-connected layer extracted from a ViT-B/16 model trained with SRigL when compressed using the condensed representation learned by SRigL to structured (i.e. the same layer accelerated using only the ablated neurons without exploiting the fine-grained sparsity), and unstructured (i.e. Compressed Sparse Row (CSR)) representations. The median over a minimum of 5 runs is shown, while the error bars show the std. dev. ***Note:*** *the increased timings for the 95 & 99% sparse structured representations is due to SRigL ablating relatively fewer neurons at these sparsities compared to 80 and 90%.* **(a) CPU wall-clock timings for online inference** on an Intel Xeon W-2145. For online (single input) inference, our condensed representation at 90% is *3.4× faster than dense* and *2.5 × faster than unstructured sparsity*. See Appendix I. **(b) GPU wall-clock timings for inference with a batch size of 256** on an NVIDIA Titan V. At 90% sparsity, our condensed representation is *1.7× faster than dense* and *13.0× faster than unstructured (CSR)* sparse layers. Note y-axis is log-scaled.

Conclusion

- Dynamic Sparse Training methods learn **NN structure during training**, and **learn representations as well as dense training** for vision CNNs/Transformers
- We show that DST methods can also learn **hardware-aware** sparsity patterns, to be easier to accelerate on real-world hardware (GPUs/CPUs)
  - Can improve real-world inference costs of CNNs/Transformers
  - Much of the progress in deep learning is by finding AI methods/models that can take better advantage of our current hardware – e.g. Transformers, AlexNet
- We show that DST methods already learn how to remove whole neurons during training, i.e. they **learn to reduce the width of the model** when it's advantageous!

UNIVERSITY OF CALGARY

Future Directions

- Dynamic Sparse Training methods are **slow to converge**, taking up to 5x more iterations to train than dense methods
  - Some of our current work shows promise in improving this
  - Does not affect inference costs!
- DST methods have shown promise **in learning better architectures for novel data domains**
  - Currently co-supervising a student using DST to improve electricity price forecasting
  - **Hoping to work with more domain experts/application domains to know if we can learn better structures for other domains!**

Questions?

www.calgaryml.com

Yani Ioannou
yani.ioannou@ucalgary.ca

DeepMind

NSERC
CRSNG

Meta

Google

CML

www.calgaryml.com

MIT

VECTOR
INSTITUTE

amazon | science

ALBERTA
INNOVATES

CML

UNIVERSITY OF
CALGARY