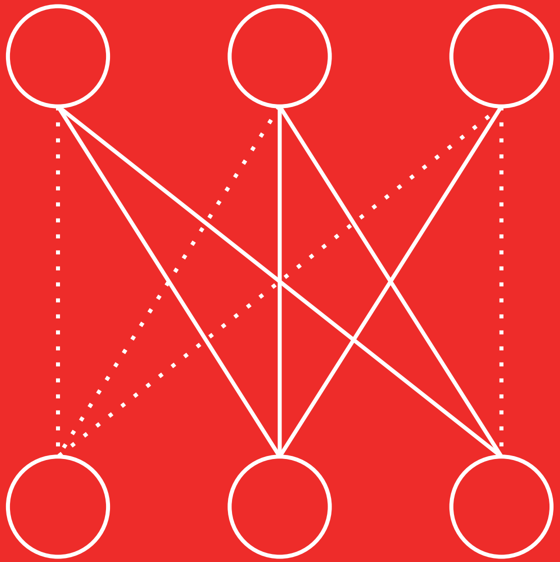
Sparse Training: Aligning Sparse Masks with Weight Symmetry



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

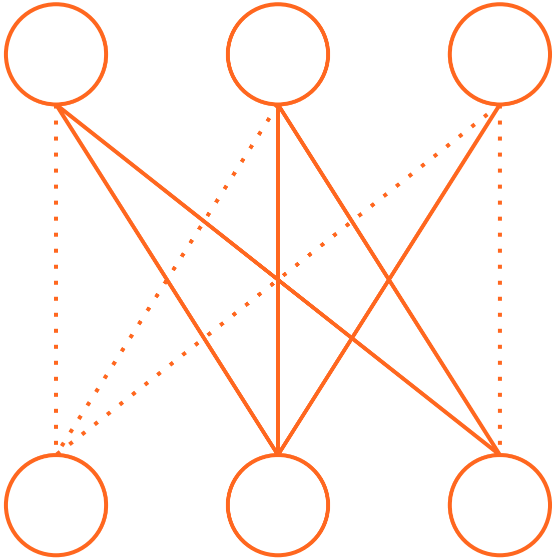
University of Bath

September 30th, 2025

UNIVERSITY OF CALGARY

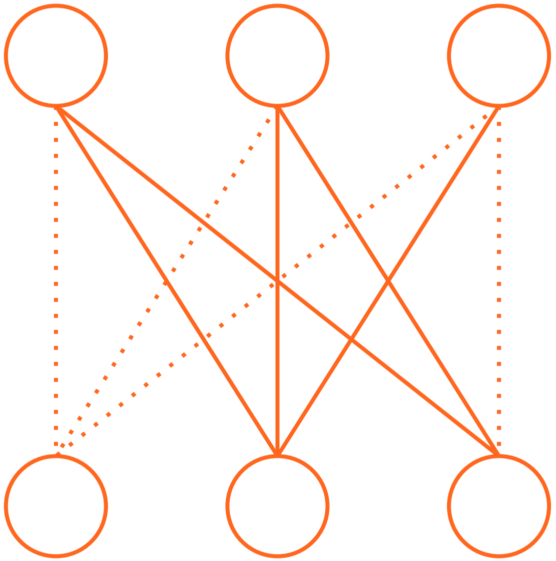
Sparse Training: Aligning Sparse Masks with Weight Symmetry

1. **Short Biography**
2. Motivation
3. Background
4. Aligning Sparse Masks




UNIVERSITY OF
CAMBRIDGE
Microsoft
VECTOR
INSTITUTE
NASA
GY69 UES
WAYVE
Google

Sparse Training:
Aligning Sparse Masks with Weight Symmetry

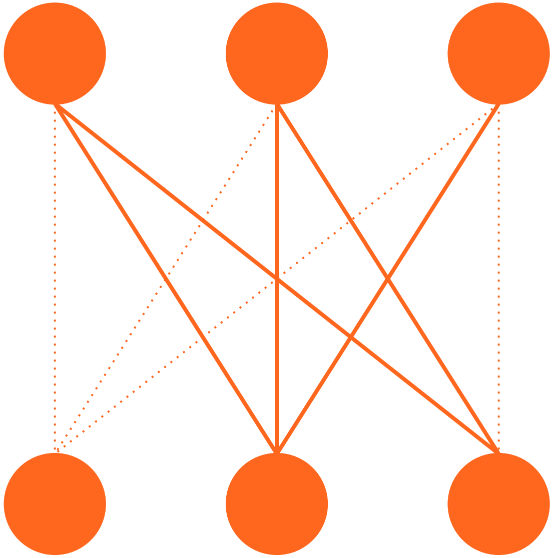


1. Short Biography
2. **Motivation**
3. Background
4. Aligning Sparse Masks

Why Sparse Neural Networks?

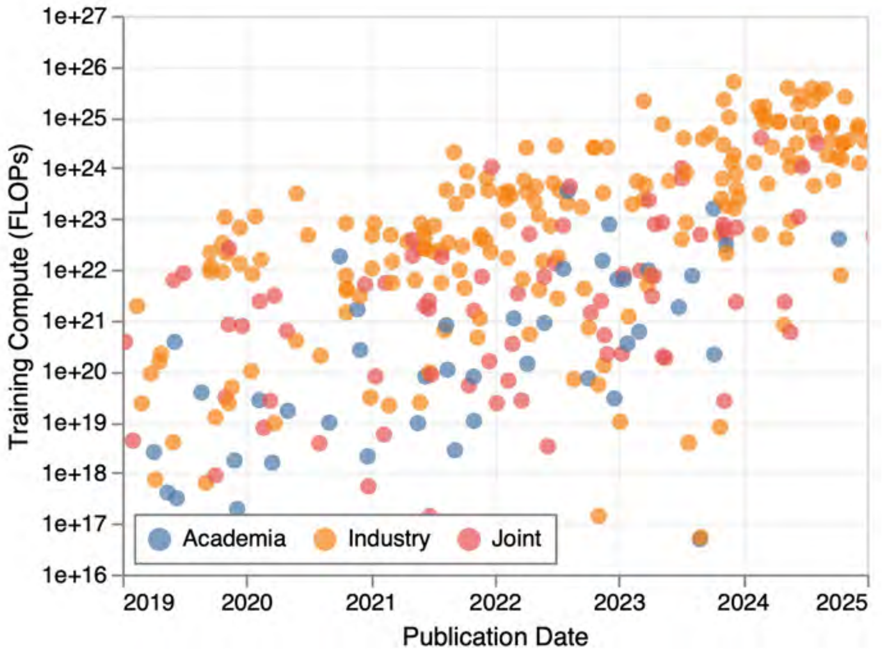
- We will focus on **weight sparsity**, but there are other forms of sparsity (e.g. activation)

- Reducing the **cost of NN training and inference**
- **Learning NN structure** from data
- Understanding & **improving NN training**



Motivation: Efficiency

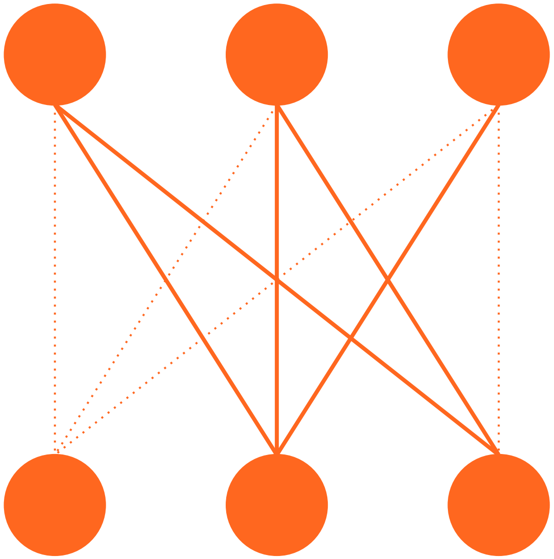
- State of the art models are becoming **exponentially more expensive to train**
- AI Research is **less accessible**
- Inference cost is increasingly important, sparse training shows promise in **learning better masks for inference** than pruning



Training Cost (FLOPS) for State-of-the-Art ML Models
(data Epoch AI)

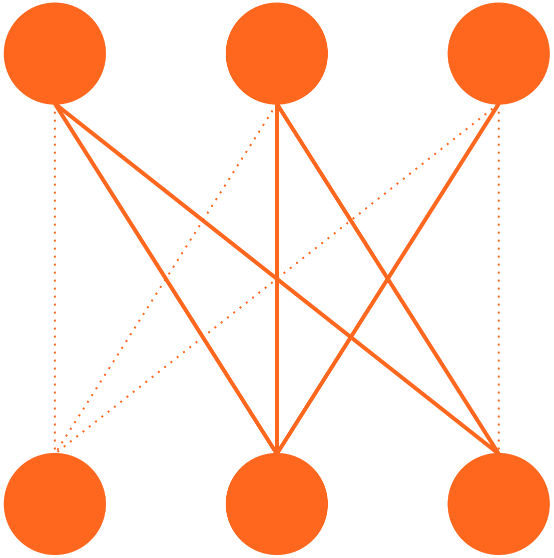
Motivation: Learning NN Structure

- In practice we rarely use fully-connected NNs for **learning representations (features)...**
- Instead, we must use our **domain knowledge** to change **the structure** of the model
  - CNNs, Transformers, Graph NNs, ...
- These are technically sparse neural networks also, but are hand-designed, **not learned**
- **Can we learn NN structure & inductive biases from data?**



Motivation: Understanding Learning

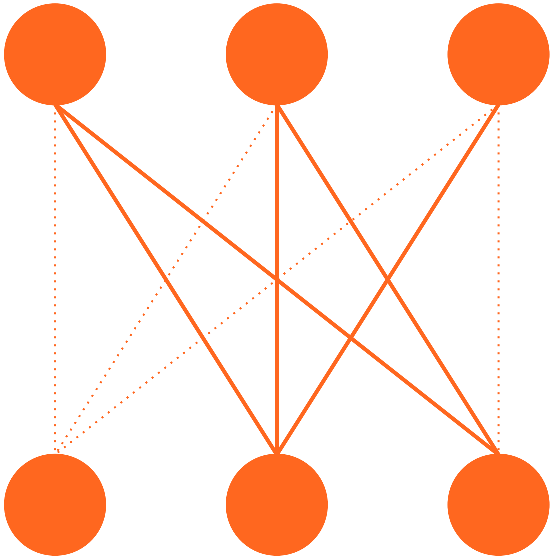
- Training NNs from random initialization is unreasonably effective… but **not always**
- Much of the “deep learning” progress can be attributed to improved NN training:
  - Initialization, normalization, residual connections, etc.
- **Sparse training breaks NN training**
- Understanding sparse training could improve our fundamental understanding of training



Calgary ML Lab Research

What we have been doing:

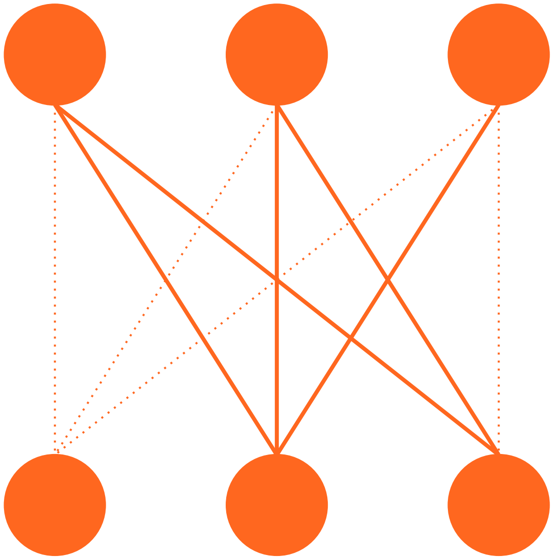
- Using SOTA sparse training methods **to accelerate practical** real-world problems
  - **Dynamic Sparse Training with Structured Sparsity.** Mike Lasby, Anna Golubeva, Utku Evci, Mihai Nica, and Yani Ioannou. In International Conference on Learning Representations (ICLR), Vienna, Austria 2024.
  - **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.



Calgary ML Lab Research

What we have been doing:

- **Understanding why sparse training is difficult**
  - **Sparse Training from Random Initialization: Aligning Lottery Ticket Masks using Weight Symmetry.** Mohammed Adnan, Rohan Jain, Ekansh Sharma, Rahul Krishnan, and Yani Ioannou. In Proceedings of Forty-second International Conference on Machine Learning (ICML) 2025, Vancouver, BC, Canada.
  - **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.



Sparse Training from Random Initialization: Aligning Lottery Ticket Masks using Weight Symmetry

Mohammed Adnan [*,1,2] **Rohan Jain** [*,1] **Ekansh Sharma** [3,2] **Rahul G. Krishnan** [3,2] **Yani Ioannou** [1]

arXiv:2505.05143v2 [cs.LG] 10 Jun 2025

Abstract

The Lottery Ticket Hypothesis (LTH) suggests there exists a sparse LTH mask and weights that achieve the same generalization performance as the dense model while using significantly fewer parameters. However, finding a LTH solution is computationally expensive, and a LTH's sparsity mask does not generalize to other random weight initializations. Recent work has suggested that neural networks trained from random initialization find solutions within the same basin modulo permutation, and proposes a method to align trained models within the same loss basin. We hypothesize that misalignment of basins is the reason why LTH masks do not generalize to new random initializations and propose permuting the LTH mask to align with the new optimization basin when performing sparse training from a different random init. We empirically show a significant increase in generalization when sparse training from random initialization with the permuted mask as compared to using the non-permuted LTH mask, on multiple datasets (CIFAR-10/100 & ImageNet) and models (VGG11 & ResNet20/50). Our codebase for reproducing the results is publicly available at here.

1. Introduction

In recent years, foundation models have achieved state-of-the-art results for different tasks. However, the exponential increase in the size of state-of-the-art models requires a similarly exponential increase in the memory and computational costs required to train, store and use these models — decreasing the accessibility of these models for researchers and practitioners alike. To overcome this issue, different model compression methods, such as pruning, quantization and knowledge distillation, have been proposed to reduce the model size at different phases of training or inference. Post-training model pruning (Han et al., 2016) has been shown to be effective in compressing the model size, and seminal works have demonstrated that large models can be pruned after training with minimal loss in accuracy (Gale et al., 2019; Han et al., 2015). While model pruning makes inference more efficient, it does not reduce the computational cost of training the model.

Motivated by the goal of training a sparse model from a random initialization, Frankle & Carbin (2019) demonstrated that training with a highly sparse mask is possible and proposed the Lottery Ticket Hypothesis (LTH) to identify sparse subnetworks that, when trained, can match the performance of a dense model. The key caveat is that a dense model must first be trained to find the sparse mask, which can *only* be used with the same random initialization that was used to train the dense model. Despite LTH seeing significant interest in the research community, LTH masks cannot be used to train from a new random initialization. Furthermore, it has been observed empirically that the LTH is impractical for finding a diverse set of solutions (Evci et al., 2022).

This posits our main research questions: *How can we train a LTH mask from a different random initialization while maintaining good generalization? Would doing so find a more diverse set of solutions than observed with the LTH itself?*

In this work, we try to understand why the LTH does not work for different random initializations from a weight-space symmetry perspective. Our hypothesis is that to reuse the LTH winning ticket mask with a different random initialization, the winning ticket mask obtained needs to be permuted such that it aligns w
associated with the new random i
our hypothesis in Figure 1.

To empirically validate our hypo
mask using Iterative Magnitude Pr
2020; Han et al., 2015) on model A
that given a permutation that alig
of model A and a new random in
be reused. The sparse model (wit
be trained to closer match the ge

[*]Equal contribution [1]Schulich School of Engineering, University of Calgary [2]Vector Institute for AI [3]Dept. of Computer Science, University of Toronto. Correspondence to: Mohammed Adnan <adnan.ahmad@ucalgary.ca>, Yani Ioannou <yani.ioannou@ucalgary.ca>.

Mohammed Adnan
PhD Student

Rohan Jain
MSc Graduate

Dr. Ekansh Sharma
PhD Graduate
University of Toronto

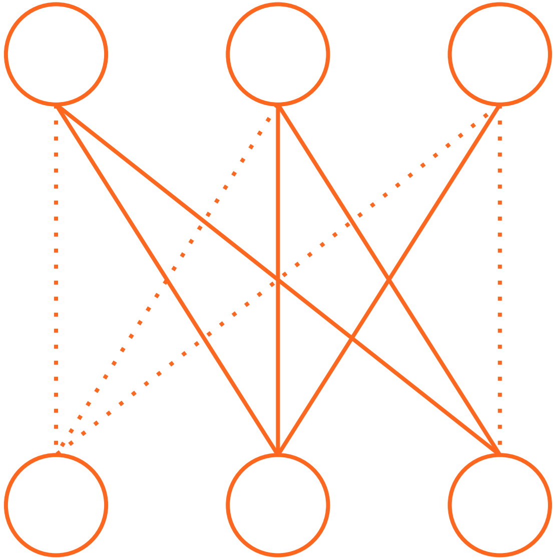
Dr. Rahul Krishnan
Assistant Professor
U. Toronto/Vector Institute

- Presented at International Conference on Machine Learning (ICML) 2025

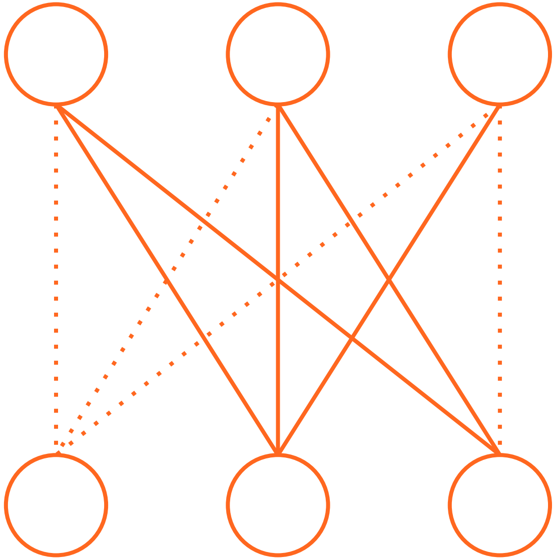
UNIVERSITY OF CALGARY

Sparse Training: Aligning Sparse Masks with Weight Symmetry

1. Short Biography
2. Motivation
3. **Background**
4. Aligning Sparse Masks



3. Background



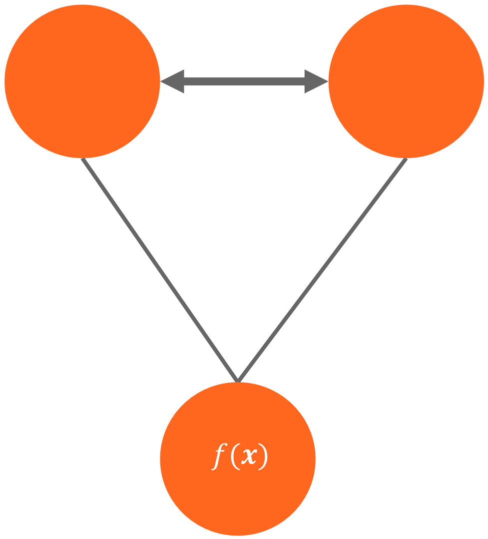
i. Weight Symmetry

ii. Sparse Training Problem

iii. Lottery Ticket Hypothesis

Weight Symmetry: Foundations

- NN layers are ***permutation invariant:*** the ordering of neurons is arbitrary
- Different permutations result in the same function, but **different parameterizations**
  - i.e. model is a different point in weight space
- NN are an example of what is generally known as a ***symmetric function***



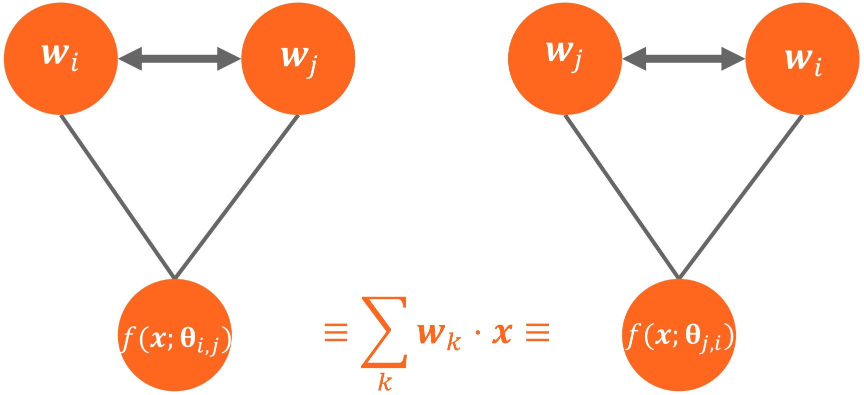
Robert Hecht-Nielsen. On The Algebraic Structure of Feed-forward Network Weight Spaces. Advanced Neural Computers. 1990.

Weight Symmetry: Foundations

- Different permutations result in same function, but **different weight parameterizations**

- For a NN with L layers, and layer width w, the **number of permutations** is:
$$(w!)^L$$

- NN permutations often number more than atoms in universe (10^{80})

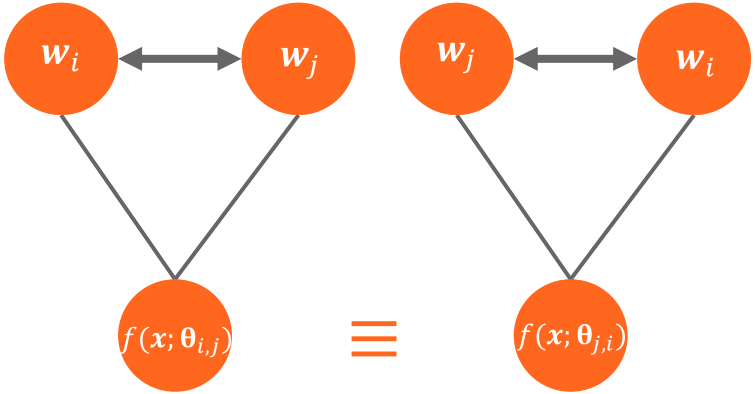


$\theta_{i,j} = [w_i^T, w_j^T]^T$ $\quad\quad$ $\theta_{j,i} = [w_j^T, w_i^T]^T$

Robert Hecht-Nielsen. On The Algebraic Structure of Feed-forward Network Weight Spaces. Advanced Neural Computers. 1990.

Weight Symmetry: Implications

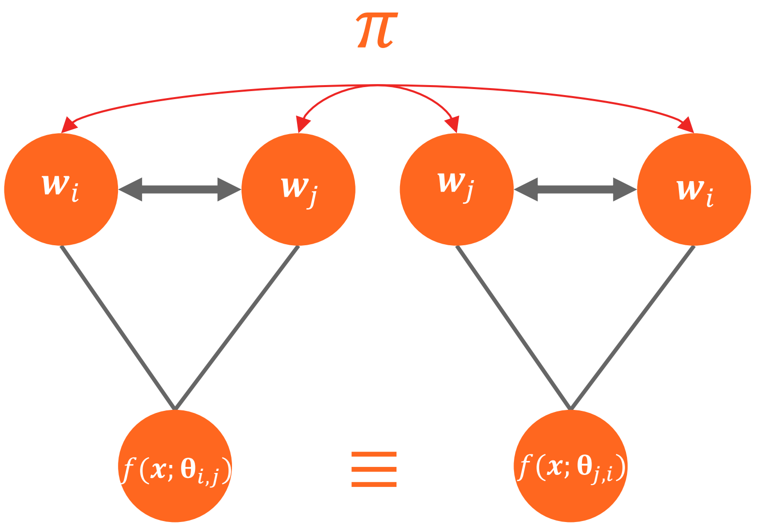
- No **unique** minima (or solutions) in weight space
- Why 1st-order optimization can find **good solutions** with **random init**[2]
- May exist only one "basin" modulo permutations[1,2], e.g. why random init. find similar solutions...



[1]Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. ICLR 2022.
[2]Samuel K. Ainsworth, Jonathan Hayase, Siddhartha Srinivasa. Git Re-Basin: Merging Models modulo Permutation Symmetries. ICLR 2023.

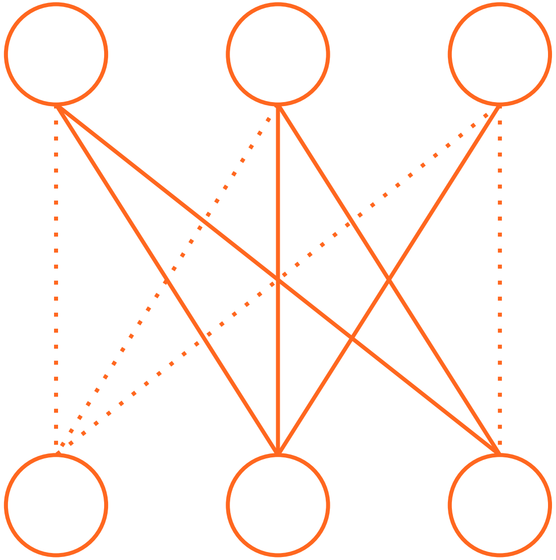
Permutation Alignment/Mapping

- Finding exact π for deep NN is NP Hard
- **Greedy approximation** w/ weight matching[1]
  - Linear Assignment Problem (LAP) per layer
  - Maximizes correlation of weights/activations
  - Best results empirically for **very wide NNs**
- Activation matching more robust in general[2]



[1]Samuel K. Ainsworth, Jonathan Hayase, Siddhartha Srinivasa. Git Re-Basin: Merging Models modulo Permutation Symmetries. ICLR 2023.
[2]Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. Repair: Renormalizing permuted activations for interpolation repair. ICLR 2023.

3. Background



i. Weight Symmetry

ii. Sparse Training Problem

iii. Lottery Ticket Hypothesis

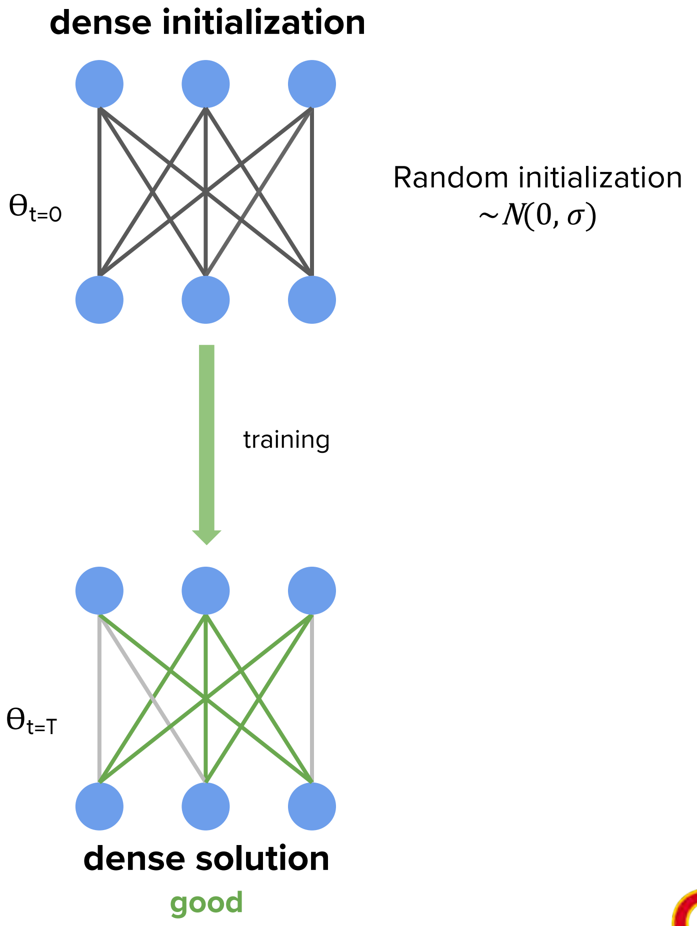
Standard NN Training

- Train a dense NN from a random initialization to find a **dense solution**
- This solution generalizes well — in fact similarly even for different random init.!
- **Recall: weight symmetry can explain this**

High saliency weight

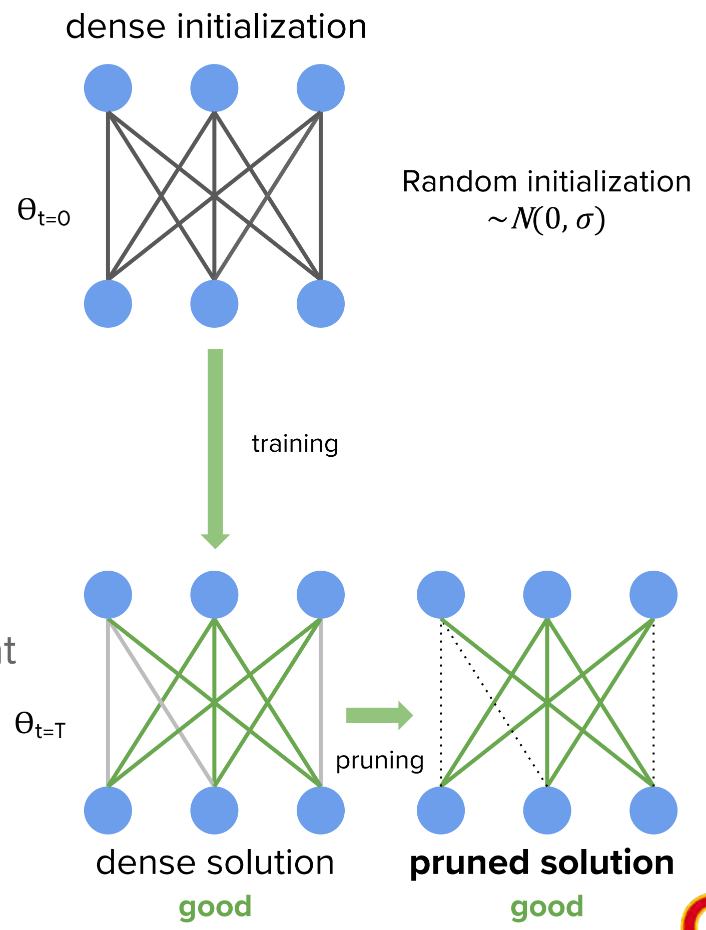
Low saliency weight

Masked weight



Unstructured Pruning

- Prune **low saliency weights**
  - Most commonly remove smallest magnitude weights
- "One-shot" pruning
  - Train and then prune once
- Iterative pruning
  - Train a bit, prune a bit, repeat several times



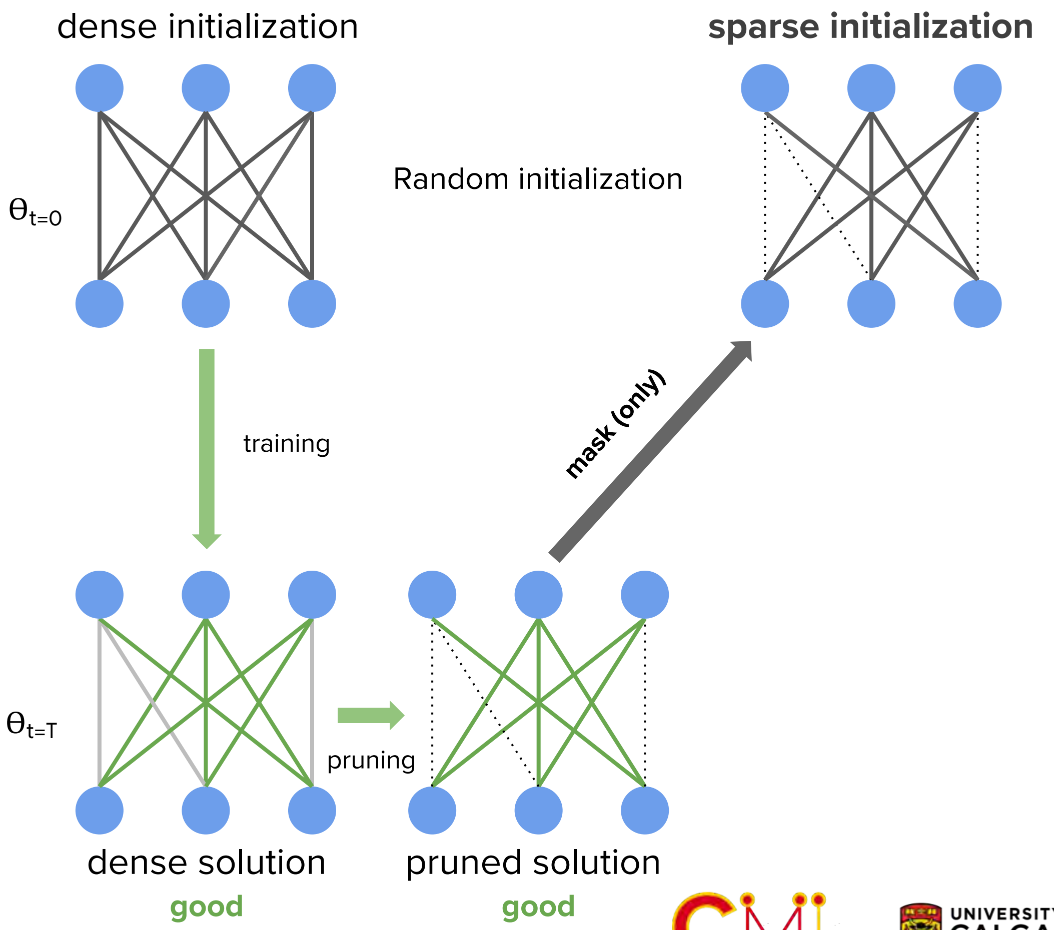
High saliency weight

Low saliency weight

Masked weight

Sparse Training?

- We know **we don't need ~85-95% of weights at inference**…
- Lots of methods to prune after training… but **can we train pruned NNs from random initialization?**

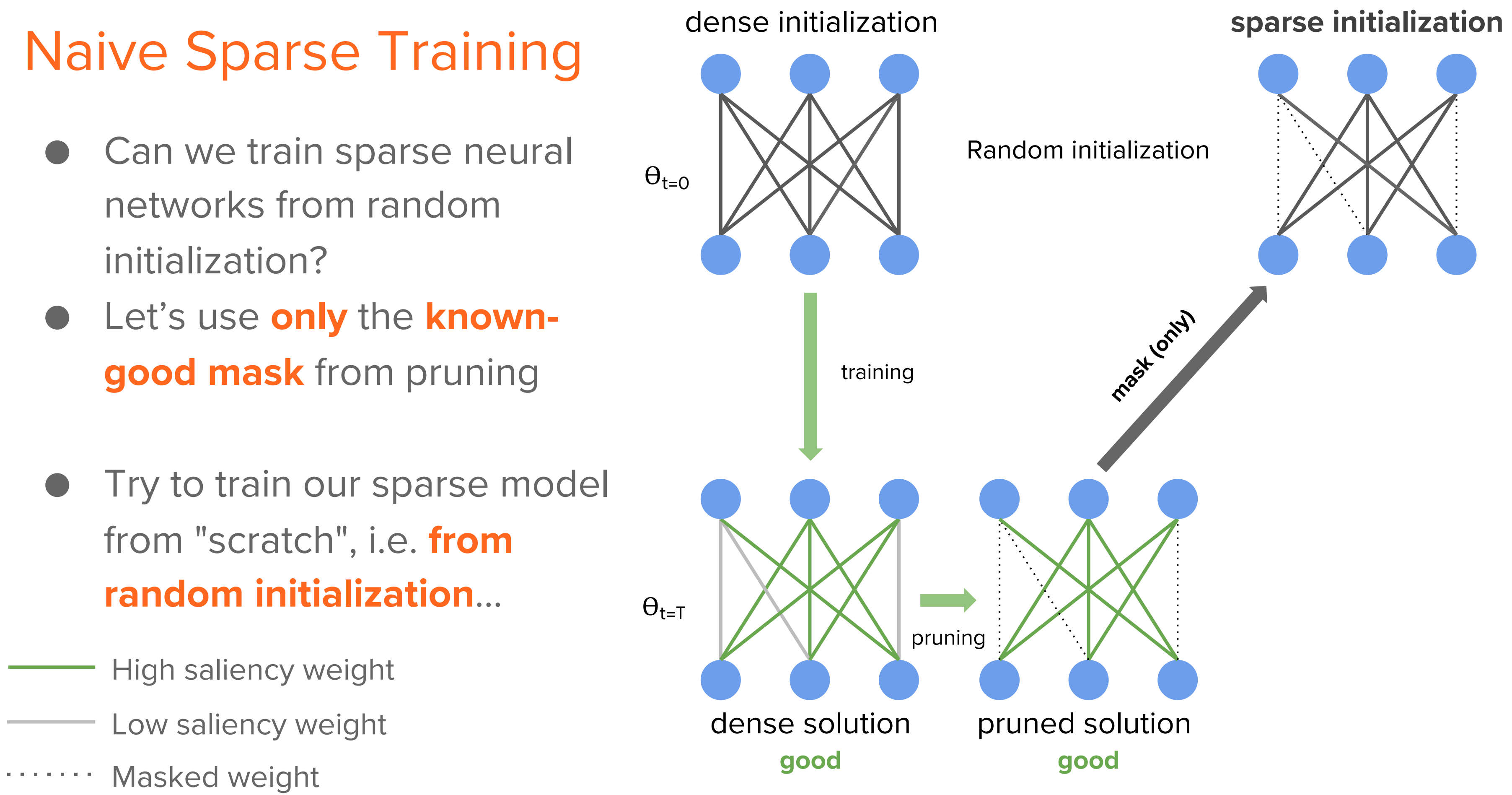


High saliency weight
Low saliency weight
Masked weight

Naive Sparse Training

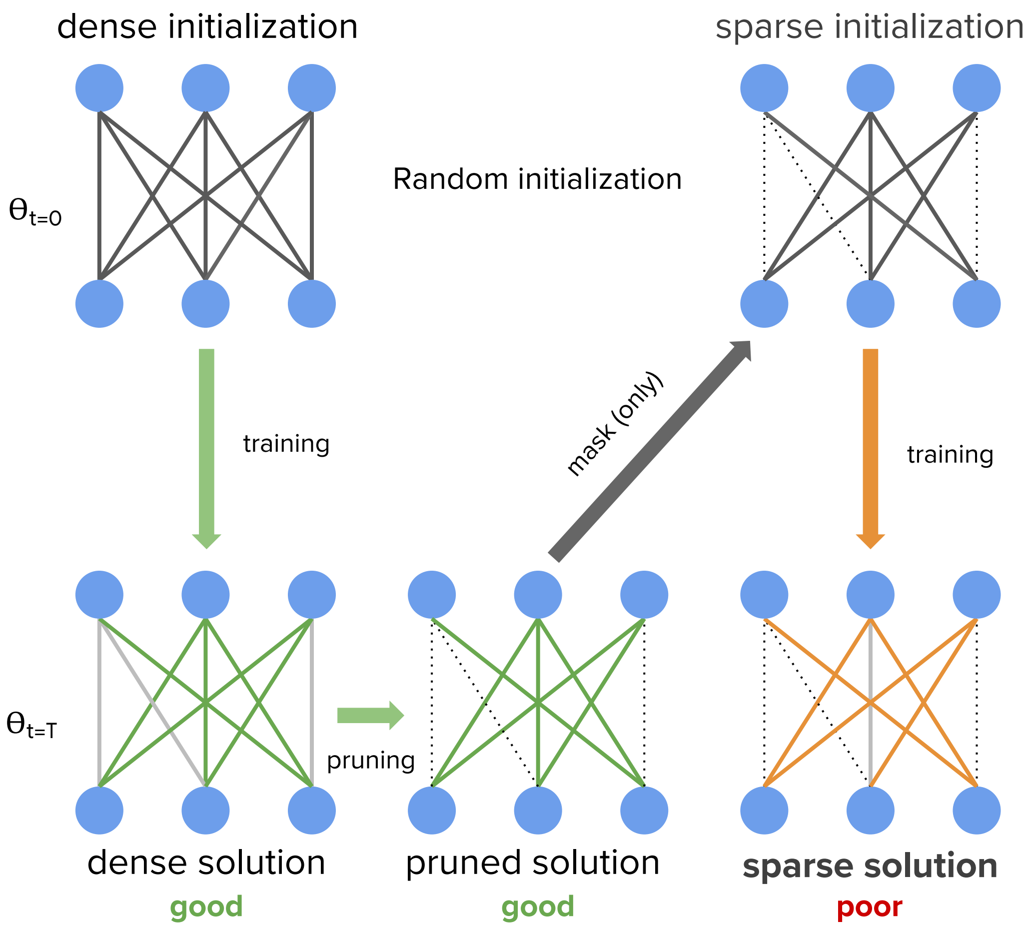
- Can we train sparse neural networks from random initialization?
- Let’s use **only** the **known-good mask** from pruning

- Try to train our sparse model from "scratch", i.e. **from random initialization**...



Sparse Training Problem

- The sparsely trained model (sparse solution) **doesn't generalize as well** as the original dense solution or pruned solution!

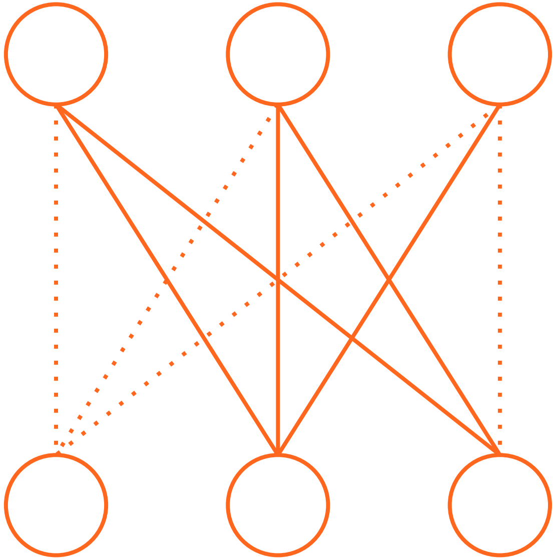


High saliency weight

Low saliency weight

Masked weight

3. Background



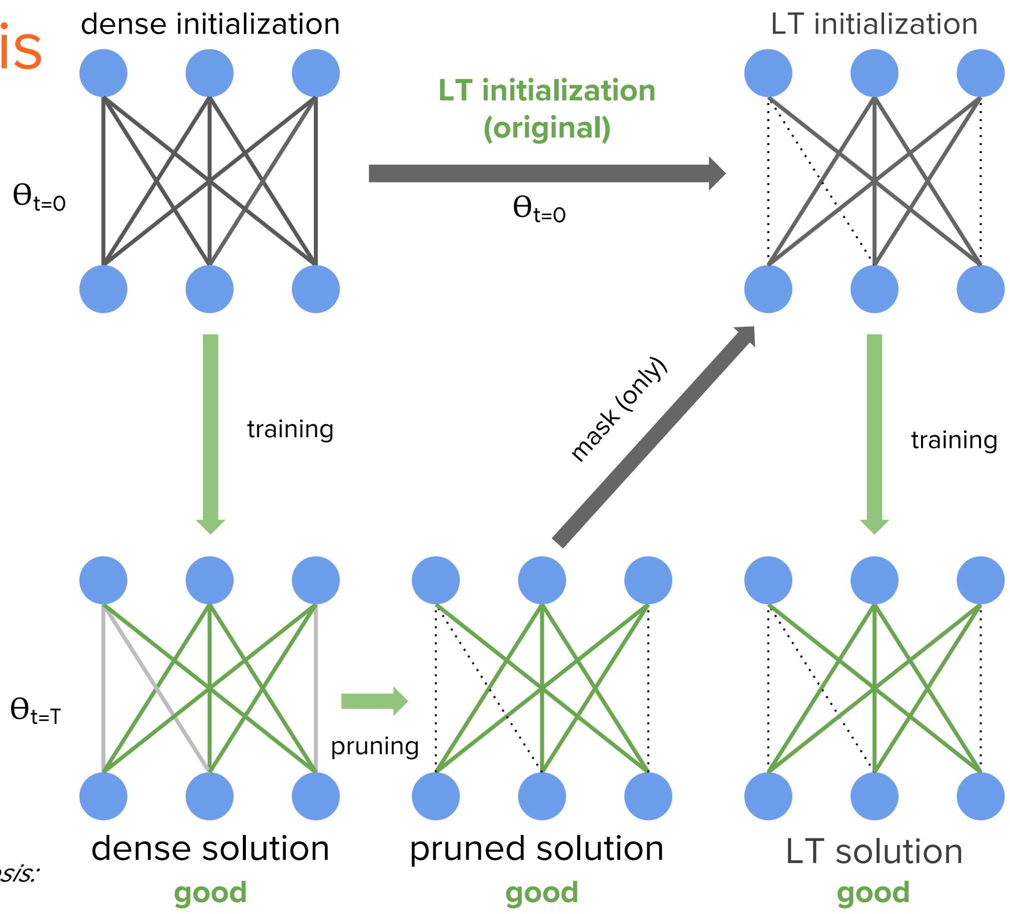
i. Weight Symmetry

ii. Sparse Training Problem

iii. Lottery Ticket Hypothesis

Lottery Ticket Hypothesis

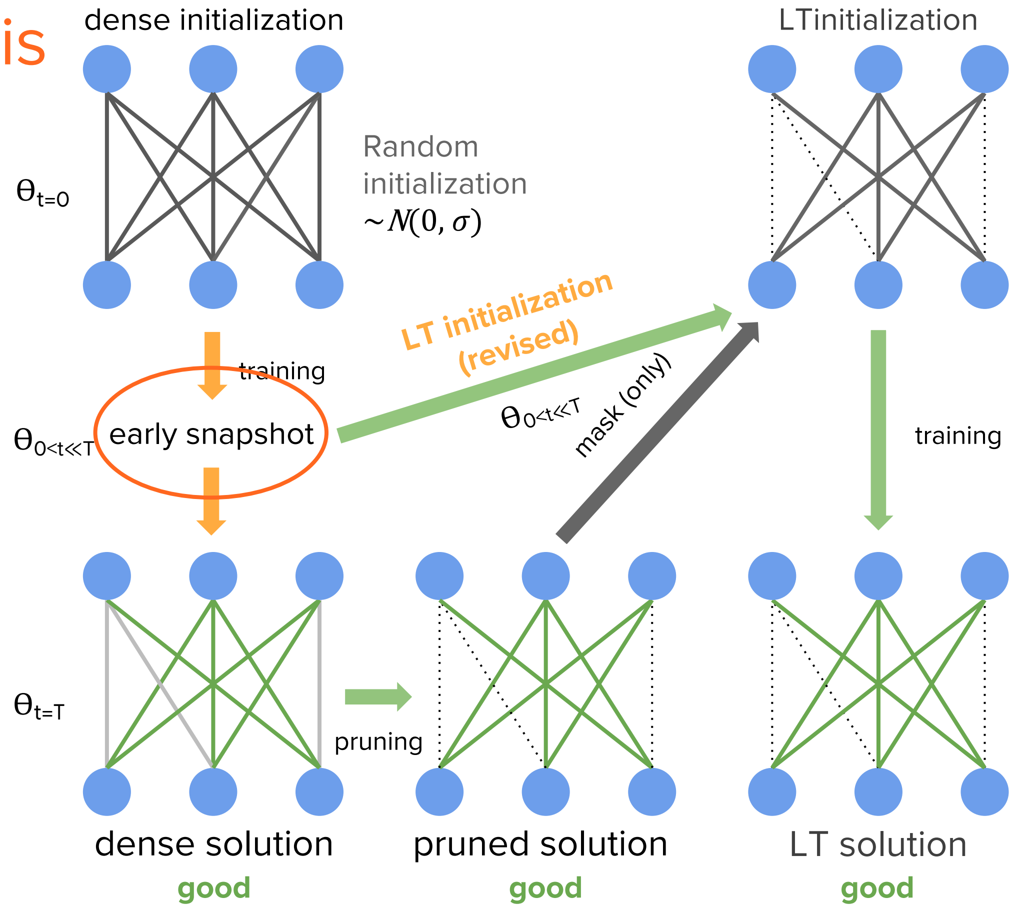
- An unstructured sparse NN, when trained from a **Lottery Ticket "initialization"** can **generalize well**
- This initialization was the **original initialization** the dense (pruned) model was trained from



Jonathan Frankle and Michael Carbin. *The Lottery Ticket Hypothesis: Training Pruned Neural Networks*. International Conference on Learning Representations (ICLR), 2019

Lottery Ticket Hypothesis (revised)

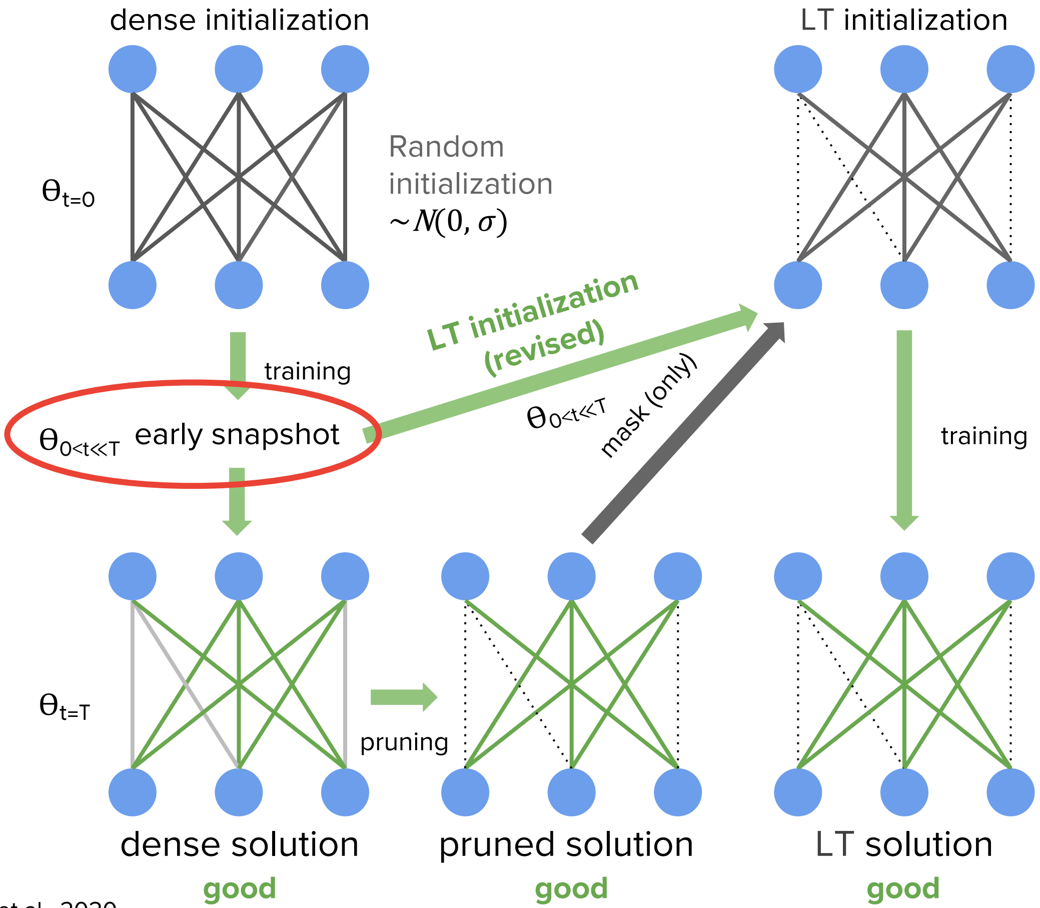
- This initialization was the original initialization the dense (pruned) model was trained from
- **LT initialization in general** is weights from early training[1]
- This is **very expensive** to find



[1] Linear Mode Connectivity and the Lottery Ticket Hypothesis. Frankle et al., 2020

Lottery Tickets

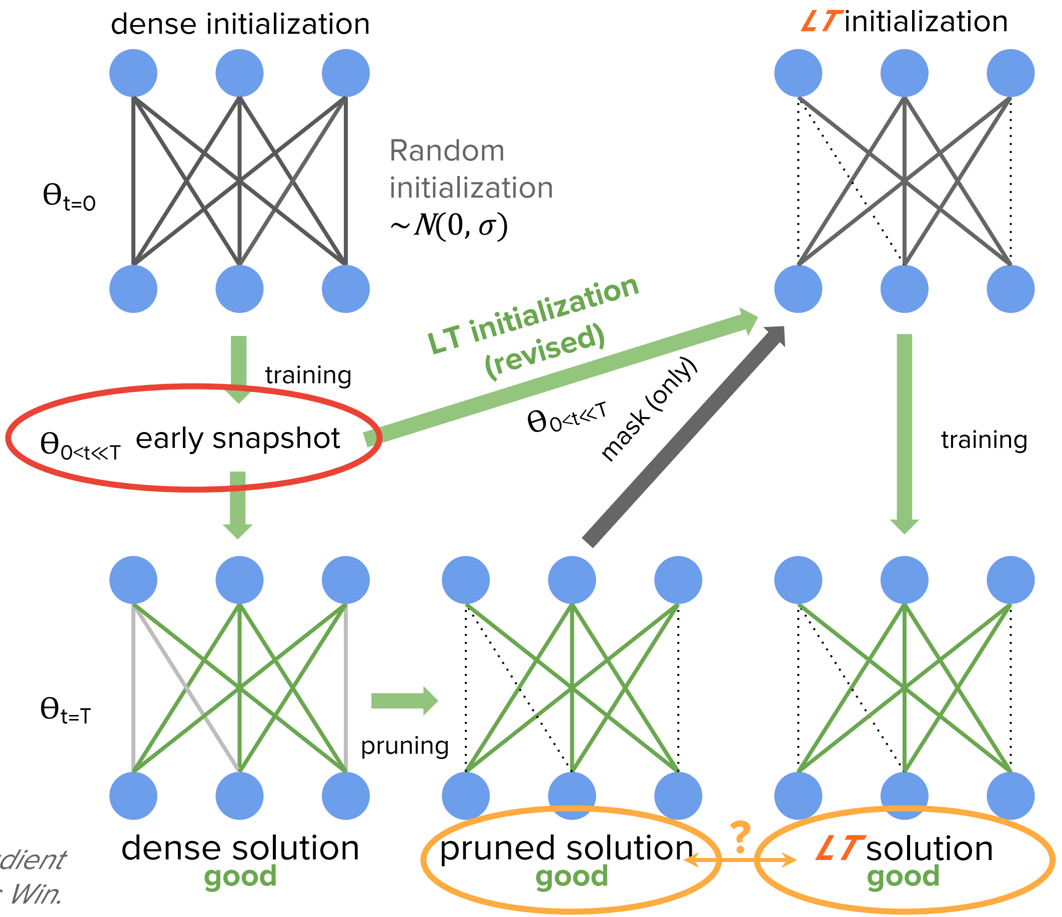
- How random is the LT initialization?



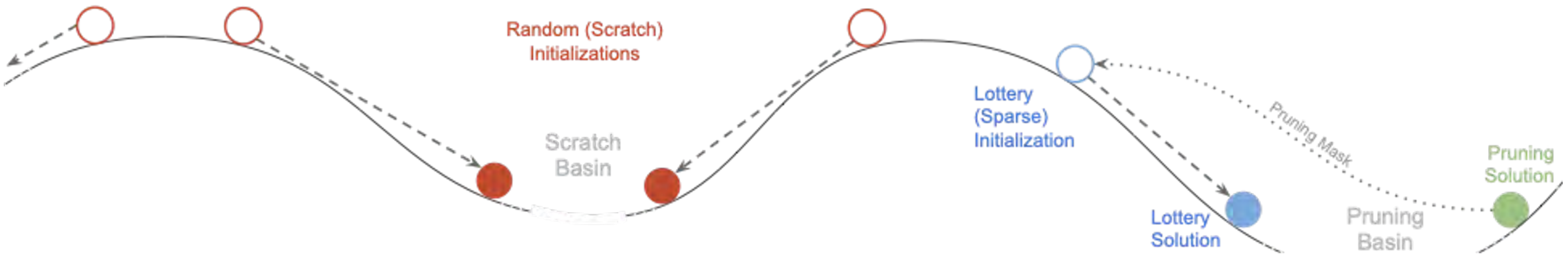
[1]Stabilizing the Lottery Ticket Hypothesis, Frankle et al., 2019
[2]Linear Mode Connectivity and the Lottery Ticket Hypothesis, Frankle et al., 2020

Lottery Tickets

- How “random” is the LTH “initialization”? **Not very...**
- **LTH doesn’t work with an arbitrary random init!**
- In previous work we showed LTs are re-learning extremely similar solutions within the same basin[1]



[1]Utku Evci, Yani Ioannou, Cem Keskin, Yann Dauphin. *Gradient Flow in Sparse Neural Networks and How Lottery Tickets Win.* AAAI 2022

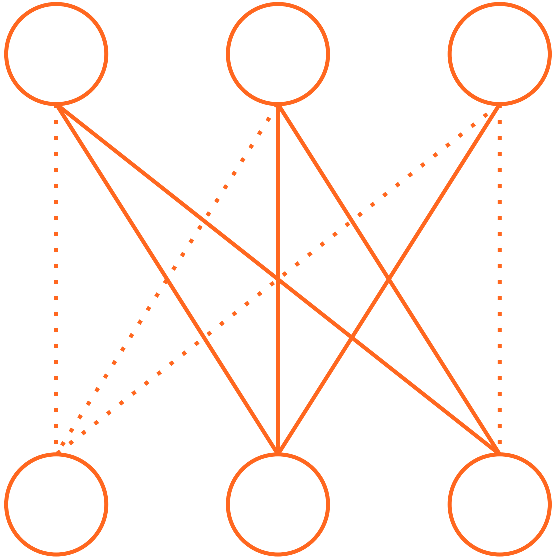


1. LT solution is close to the pruned solution
2. LT/pruned solution is the same basin of convergence
3. LT/pruned solution's learn very similar functions

LTs appear to re-learn the pruned solution they are derived from

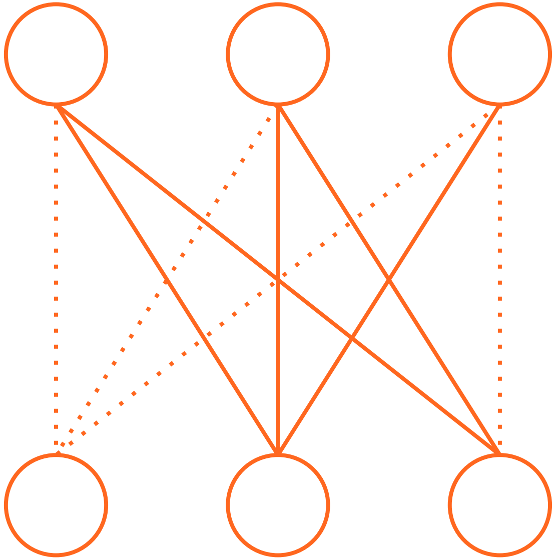
Utku Evci, Yani Ioannou, Cem Keskin, Yann Dauphin. *Gradient Flow in Sparse Neural Networks and How Lottery Tickets Win.* AAAI 2022

Sparse Training:
Aligning Sparse Masks with Weight Symmetry



1. Short Biography
2. Motivation
3. Background
4. **Aligning Sparse Masks**

4. Aligning Sparse Masks



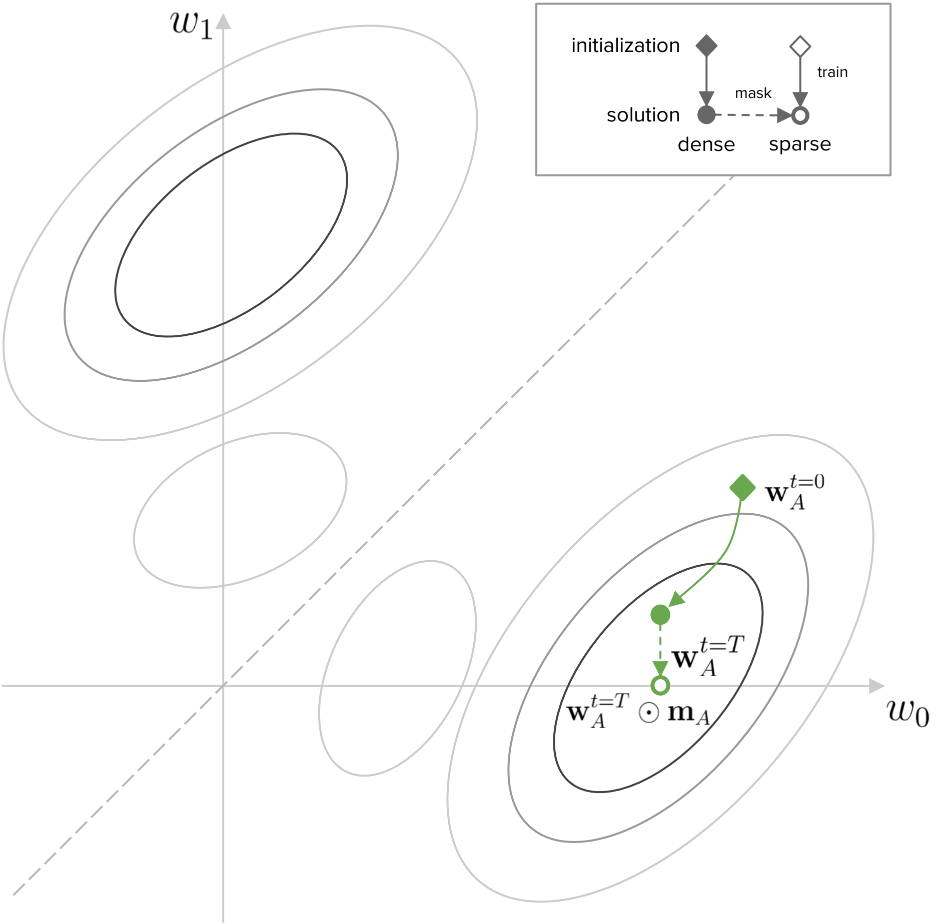
i. Hypothesis

ii. Experimental Methodology

iii. Results

iv. Analysis

Pruning Loss Landscape

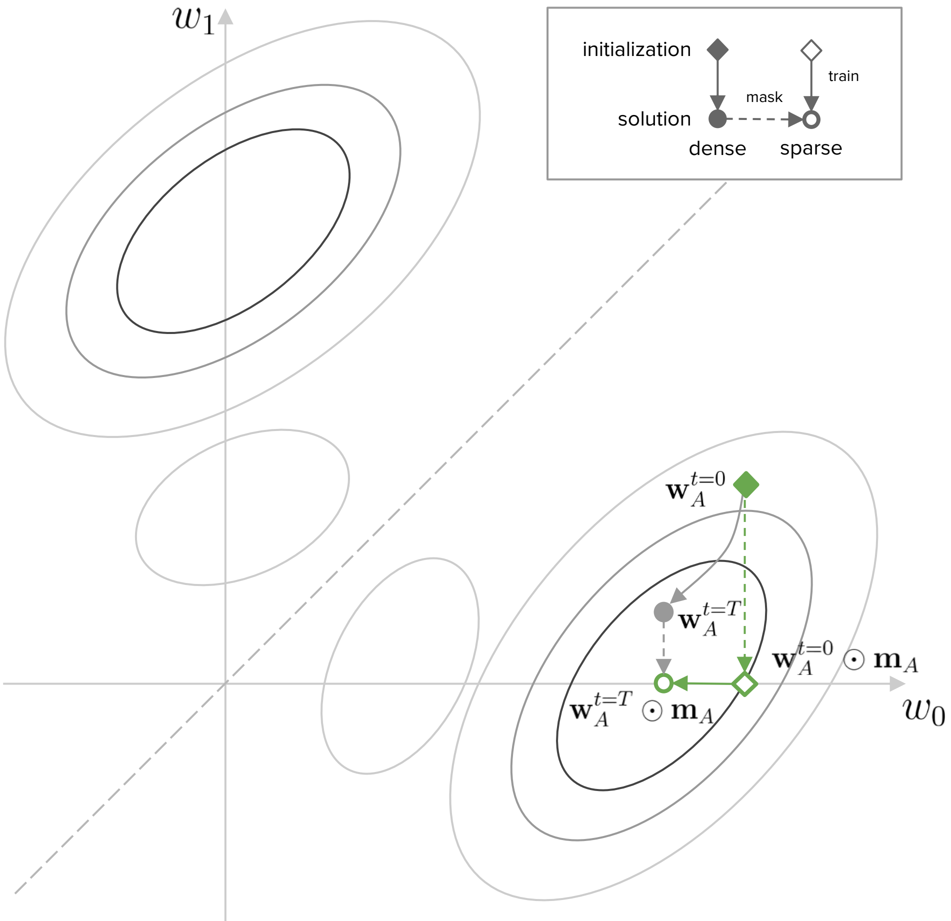


- loss landscape with two weights $\mathbf{w}_A = (w_0, w_1)$
- Train from $\mathbf{w}_A^{t=0}$ to soln. $\mathbf{w}_A^{t=T}$
- Prune $\mathbf{w}_A^{t=T}$ with $\mathbf{m}_A = (1, 0)$

Figure 7. A 2D loss landscape visualization of our method in the setting of a model with a single layer and two parameters on a single input scale.

UNIVERSITY OF CALGARY

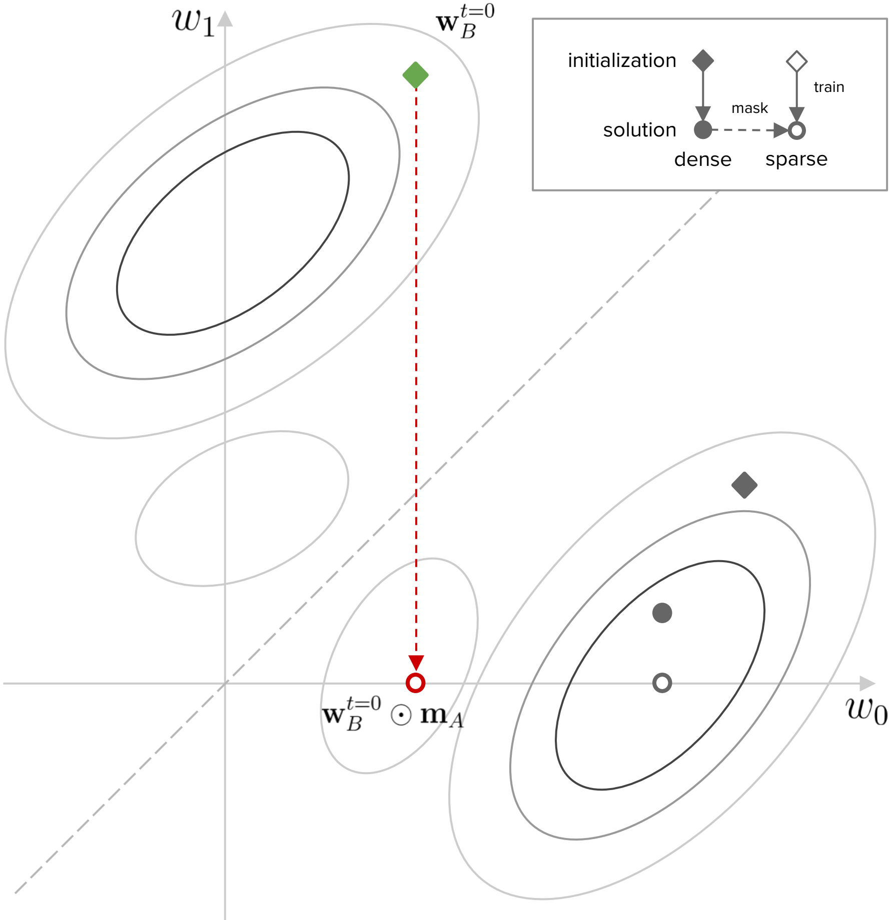
LTH Loss Landscape



- Project (prune) re-using $\mathbf{m}_A$
- Train from $\mathbf{w}_A^{t=0} \odot \mathbf{m}_A$
- End training at $\mathbf{w}_A^{t=T} \odot \mathbf{m}_A$

Figure 7. A 2D loss landscape visualization of our method in the setting of a model with a single layer and two parameters on a single input scale.

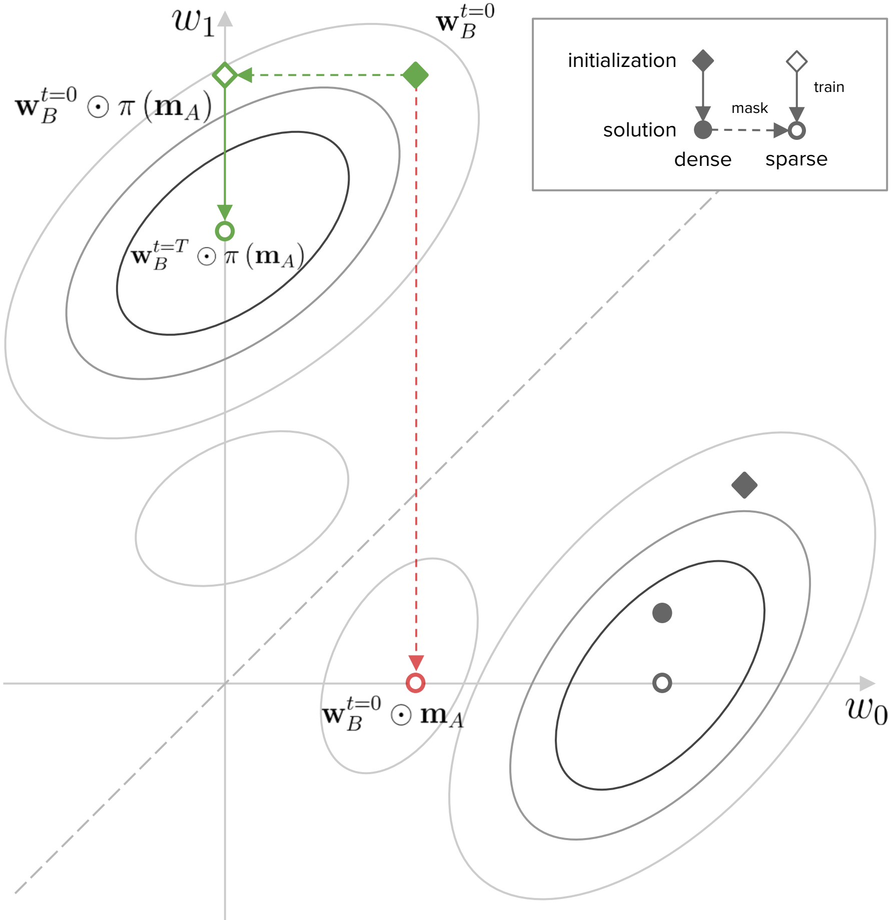
Sparse Loss Landscape



- Train w/ new random init. $\mathbf{w}_B^{t=0}$
- Re-using $\mathbf{m}_A$ is illustrated
  - This is clearly the wrong axis to project to from new initialization
  - Masked init falls outside basin
- Training from $\mathbf{w}_B^{t=0} \odot \mathbf{m}_A$ doesn't find good soln.

Figure 7. A 2D loss landscape visualization of our method in the setting of a model with a single layer and two parameters on a single input scale.

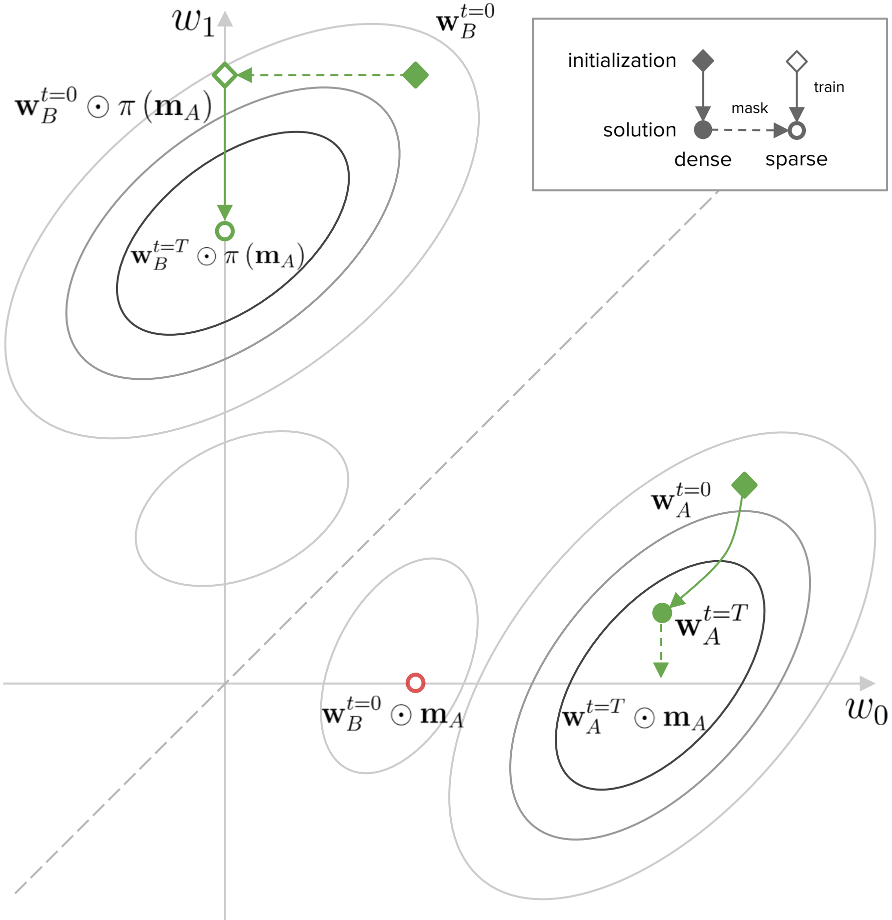
Our Hypothesis



- Hypothesis: Sparse training from random init does not work well because the mask is misaligned with the new basin of $w_B^{t=0}$
- **Can we adapt the mask m_A derived from $w_A^{t=0}$ for $w_B^{t=0}$?**

Figure 7. A 2D loss landscape visualization of our method in the setting of a model with a single layer and two parameters on a single input scale.

Our Hypothesis



- Recall[1]: the basins of $w_A^{t=T}$ and $w_B^{t=T}$ are related by a permutation π:

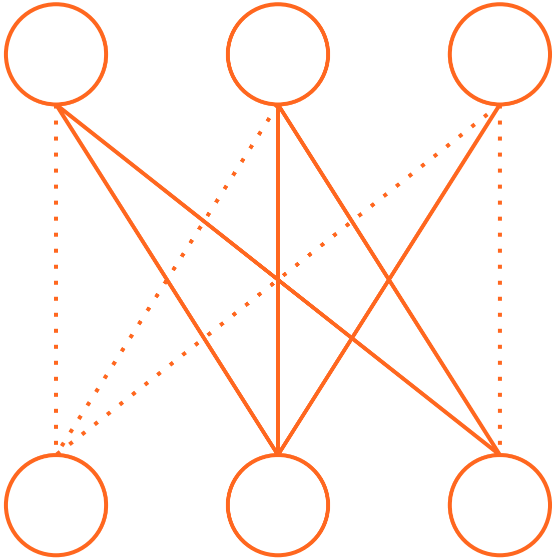
$$\pi\left(w_A^{t=T}\right) = w_B^{t=T}$$

- Are the masks for different basins also related by the **same permutation**?

$$\pi(\mathbf{m}_A) = \mathbf{m}_B$$

[1]Samuel K. Ainsworth, Jonathan Hayase, Siddhartha Srinivasa. Git Re-Basin: Merging Models modulo Permutation Symmetries. ICLR 2023.

4. Aligning Sparse Masks



i. Hypothesis

ii. Experimental Methodology

iii. Results

iv. Analysis

$$\mathbf{w}_A^{t=0} \sim \mathcal{N}$$

train
(dense)

$$\mathbf{w}_A^{t=T}$$

Dense Solution

Dense Training

train

mask

match

$$\mathbf{w}_A^{t=0} \sim \mathcal{N}$$

train (dense)

$$\mathbf{w}_A^{t=T}$$

prune

$$\mathbf{w}_A^{t=T} \odot \mathbf{m}_A$$

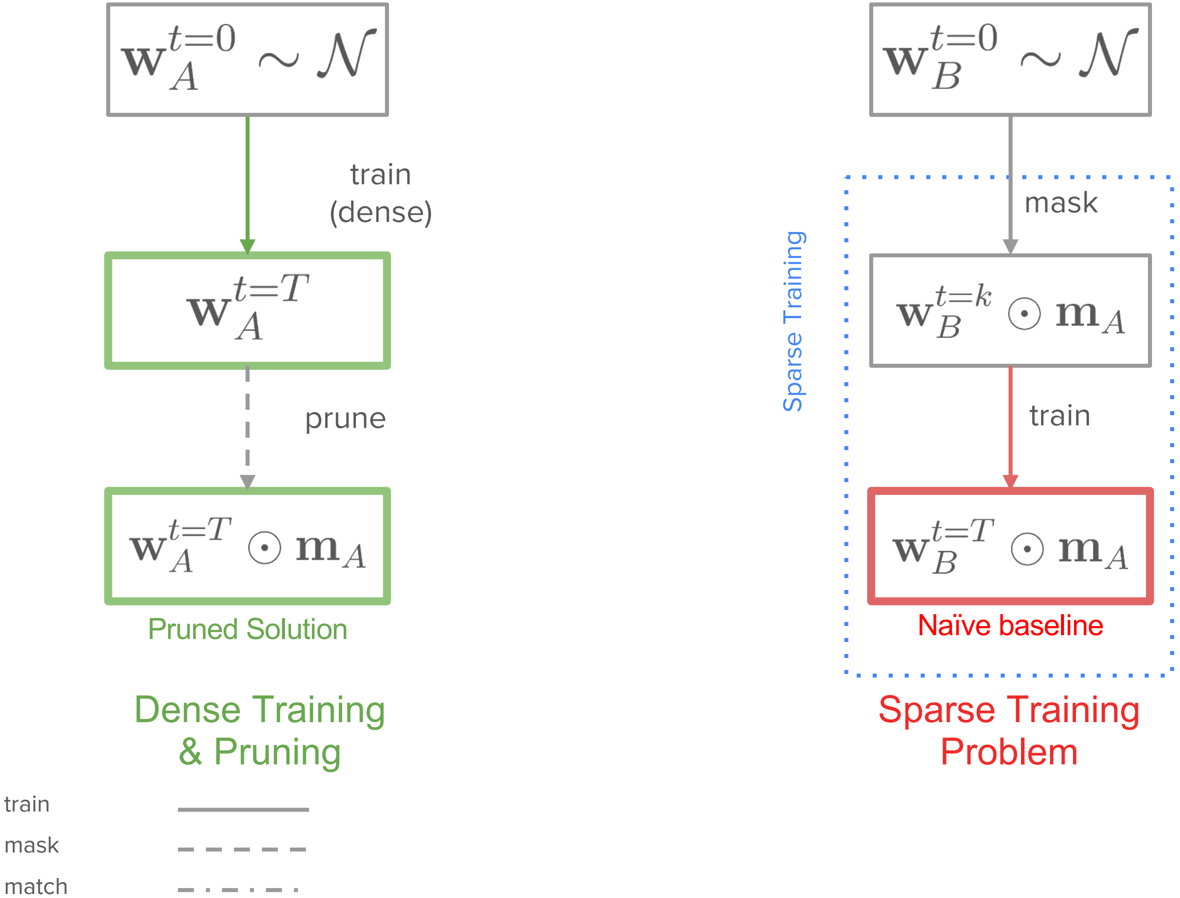
Pruned Solution

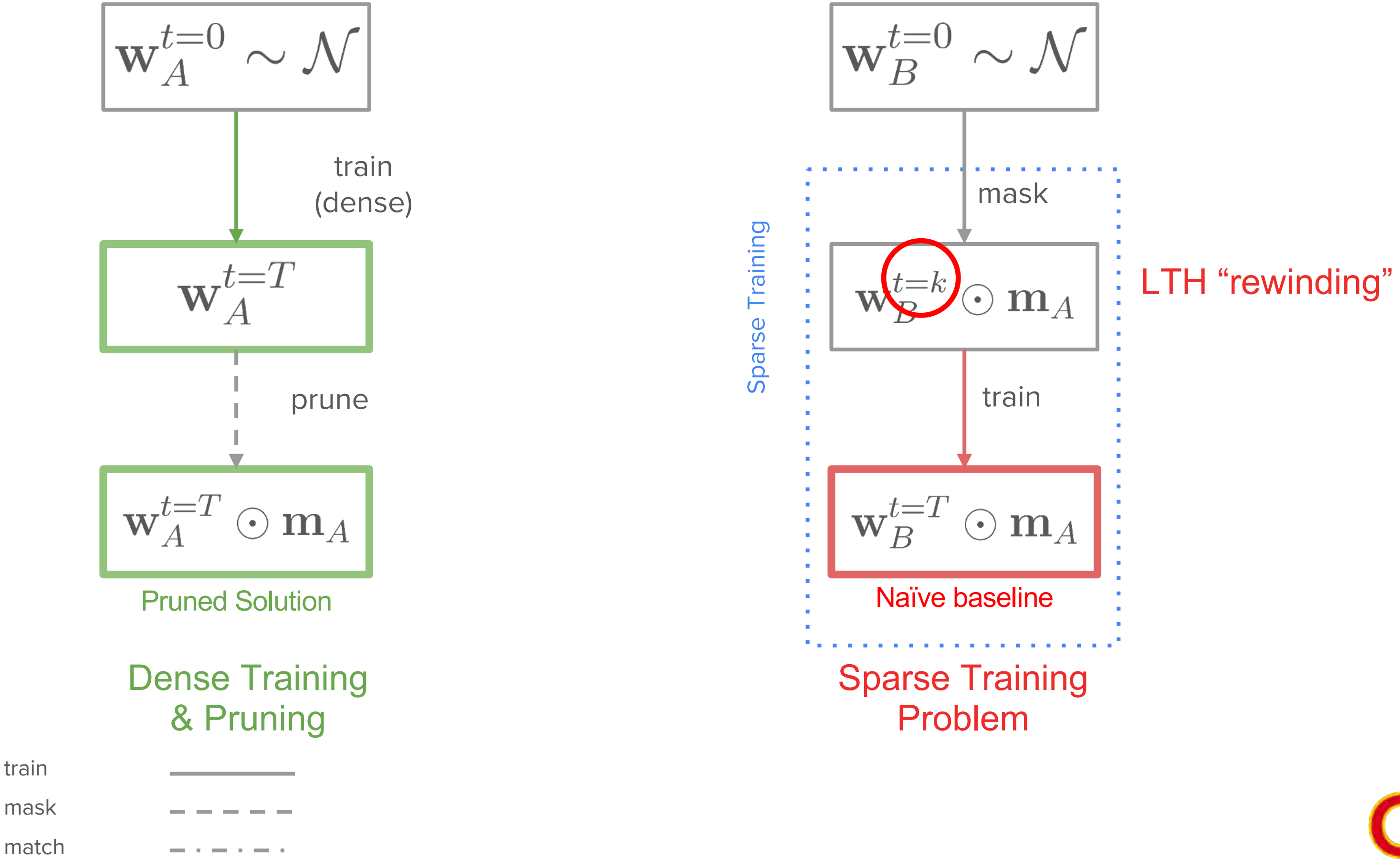
Dense Training & Pruning

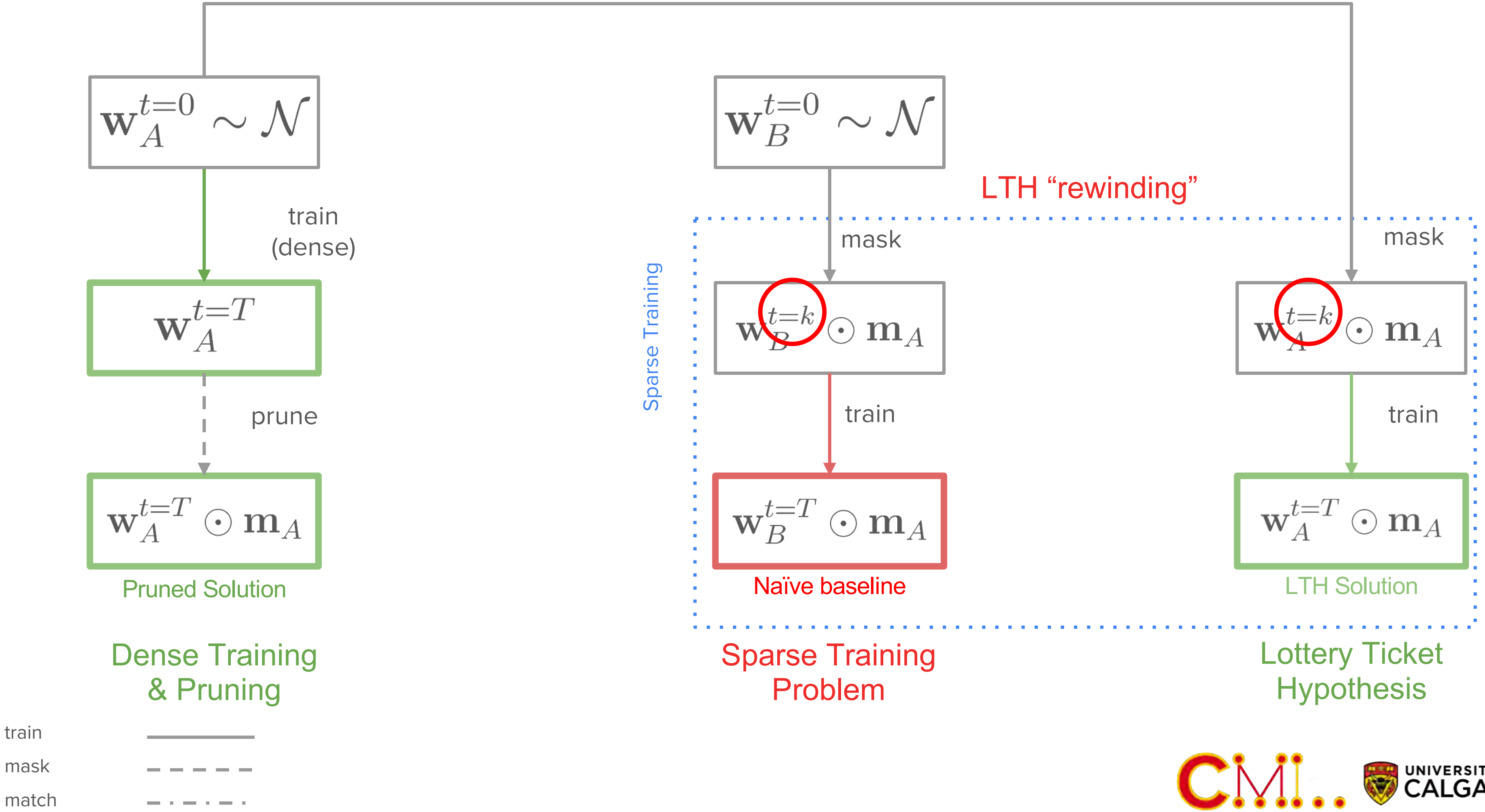
train

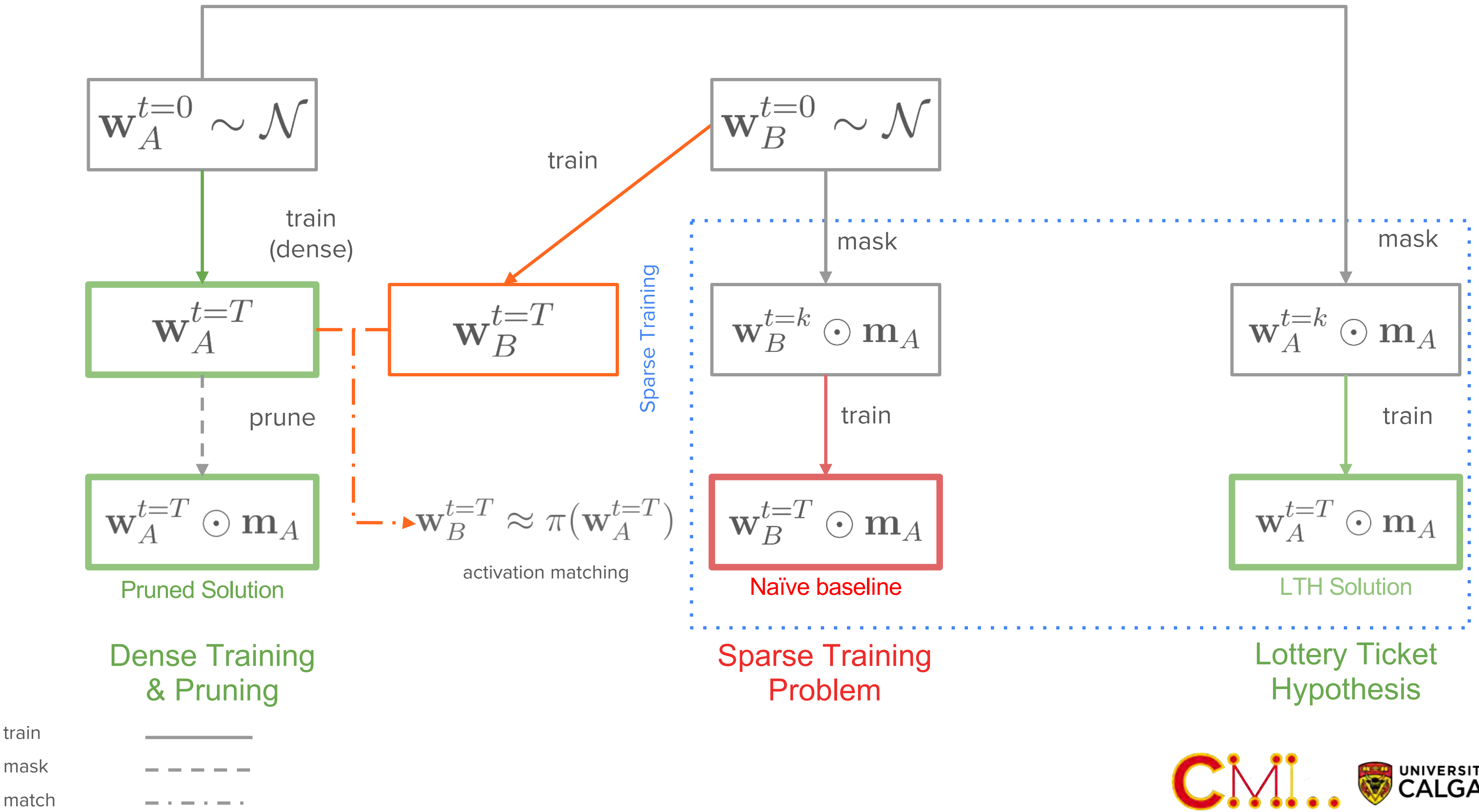
mask

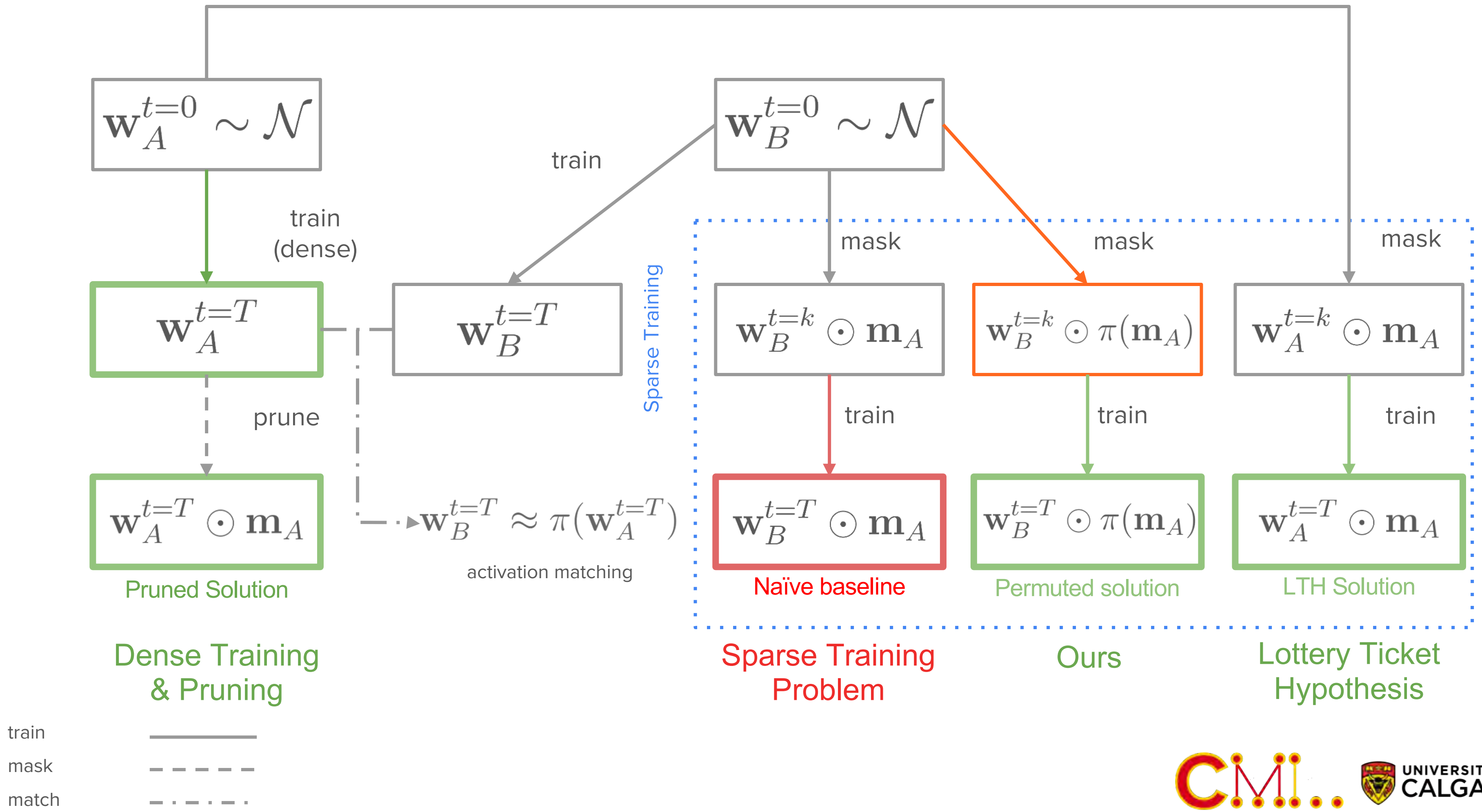
match


$\mathbf{w}_A^{t=0} \sim \mathcal{N}$
train (dense)
$\mathbf{w}_A^{t=T}$
prune
$\mathbf{w}_A^{t=T} \odot \mathbf{m}_A$
Pruned Solution
Dense Training & Pruning
$\mathbf{w}_B^{t=0} \sim \mathcal{N}$
Sparse Training
mask
$\mathbf{w}_B^{t=k} \odot \mathbf{m}_A$
train
$\mathbf{w}_B^{t=T} \odot \mathbf{m}_A$
Naïve baseline
Sparse Training Problem
train
mask
match

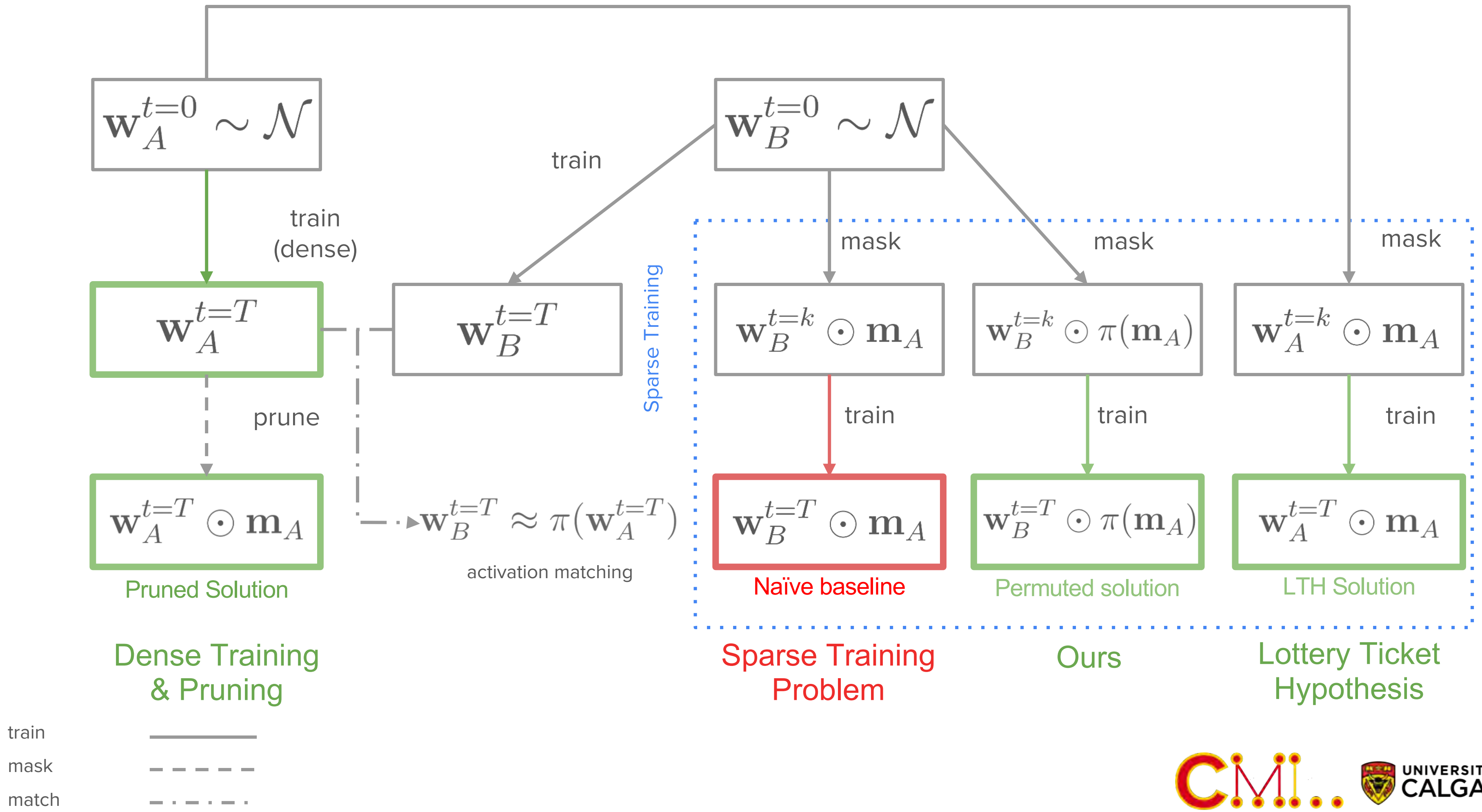

$\mathbf{w}_A^{t=0} \sim \mathcal{N}$
train
(dense)
$\mathbf{w}_A^{t=T}$
prune
$\mathbf{w}_A^{t=T} \odot \mathbf{m}_A$
Pruned Solution
Dense Training
& Pruning
$\mathbf{w}_B^{t=0} \sim \mathcal{N}$
mask
Sparse Training
$\mathbf{w}_B^{t=k} \odot \mathbf{m}_A$
LTH “rewinding”
train
$\mathbf{w}_B^{t=T} \odot \mathbf{m}_A$
Naïve baseline
Sparse Training
Problem
train
mask
match


$\mathbf{w}_A^{t=0} \sim \mathcal{N}$
train
(dense)
$\mathbf{w}_A^{t=T}$
prune
$\mathbf{w}_A^{t=T} \odot \mathbf{m}_A$
Pruned Solution
Dense Training
& Pruning
$\mathbf{w}_B^{t=0} \sim \mathcal{N}$
LTH “rewinding”
Sparse Training
mask
$\mathbf{w}_B^{t=k} \odot \mathbf{m}_A$
train
$\mathbf{w}_B^{t=T} \odot \mathbf{m}_A$
Naïve baseline
Sparse Training
Problem
mask
$\mathbf{w}_A^{t=k} \odot \mathbf{m}_A$
train
$\mathbf{w}_A^{t=T} \odot \mathbf{m}_A$
LTH Solution
Lottery Ticket
Hypothesis
train
mask
match


$\mathbf{w}_A^{t=0} \sim \mathcal{N}$
$\mathbf{w}_B^{t=0} \sim \mathcal{N}$
train
train
(dense)
mask
mask
$\mathbf{w}_A^{t=T}$
$\mathbf{w}_B^{t=T}$
Sparse Training
$\mathbf{w}_B^{t=k} \odot \mathbf{m}_A$
$\mathbf{w}_A^{t=k} \odot \mathbf{m}_A$
prune
train
train
$\mathbf{w}_A^{t=T} \odot \mathbf{m}_A$
$\mathbf{w}_B^{t=T} \approx \pi(\mathbf{w}_A^{t=T})$
$\mathbf{w}_B^{t=T} \odot \mathbf{m}_A$
$\mathbf{w}_A^{t=T} \odot \mathbf{m}_A$
activation matching
Pruned Solution
Naïve baseline
LTH Solution
Dense Training
& Pruning
Sparse Training
Problem
Lottery Ticket
Hypothesis
train
mask
match

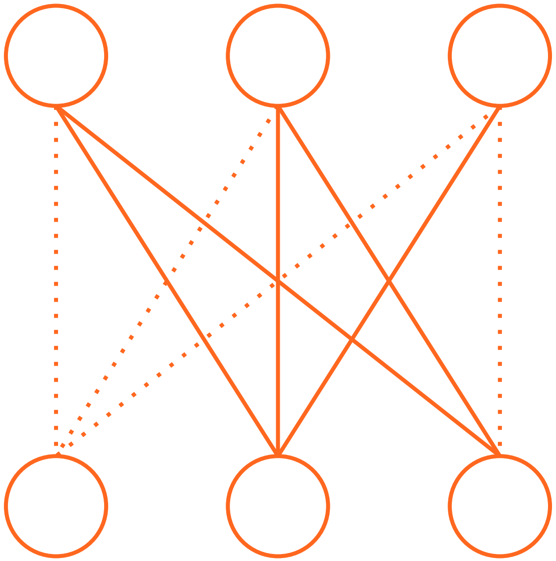


UNIVERSITY OF CALGARY



$\mathbf{w}_A^{t=0} \sim \mathcal{N}$
$\mathbf{w}_B^{t=0} \sim \mathcal{N}$
train
train (dense)
mask
mask
mask
$\mathbf{w}_A^{t=T}$
$\mathbf{w}_B^{t=T}$
Sparse Training
$\mathbf{w}_B^{t=k} \odot \mathbf{m}_A$
$\mathbf{w}_B^{t=k} \odot \pi(\mathbf{m}_A)$
$\mathbf{w}_A^{t=k} \odot \mathbf{m}_A$
prune
train
train
train
$\mathbf{w}_A^{t=T} \odot \mathbf{m}_A$
$\mathbf{w}_B^{t=T} \approx \pi(\mathbf{w}_A^{t=T})$
$\mathbf{w}_B^{t=T} \odot \mathbf{m}_A$
$\mathbf{w}_B^{t=T} \odot \pi(\mathbf{m}_A)$
$\mathbf{w}_A^{t=T} \odot \mathbf{m}_A$
Pruned Solution
activation matching
Naïve baseline
Permuted solution
LTH Solution
Dense Training & Pruning
Sparse Training Problem
Ours
Lottery Ticket Hypothesis
train
mask
match

4. Aligning Sparse Masks

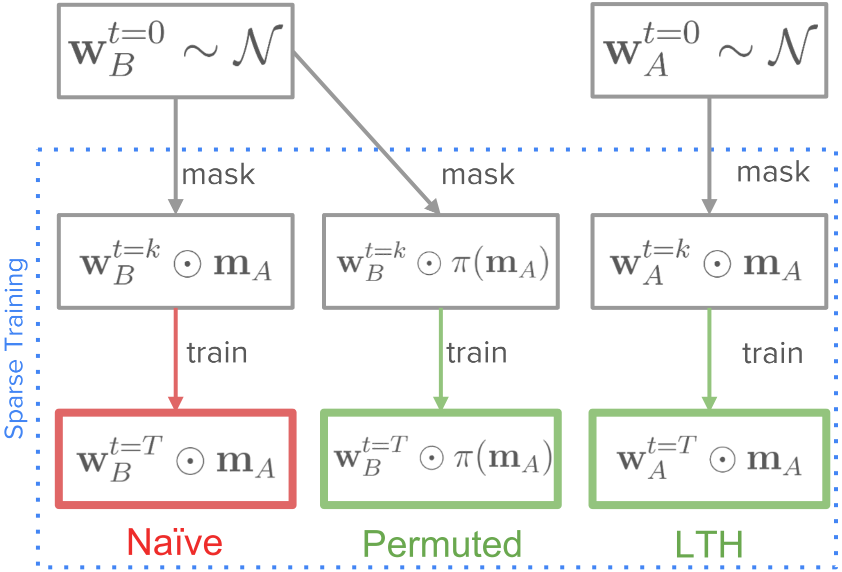


i. Hypothesis

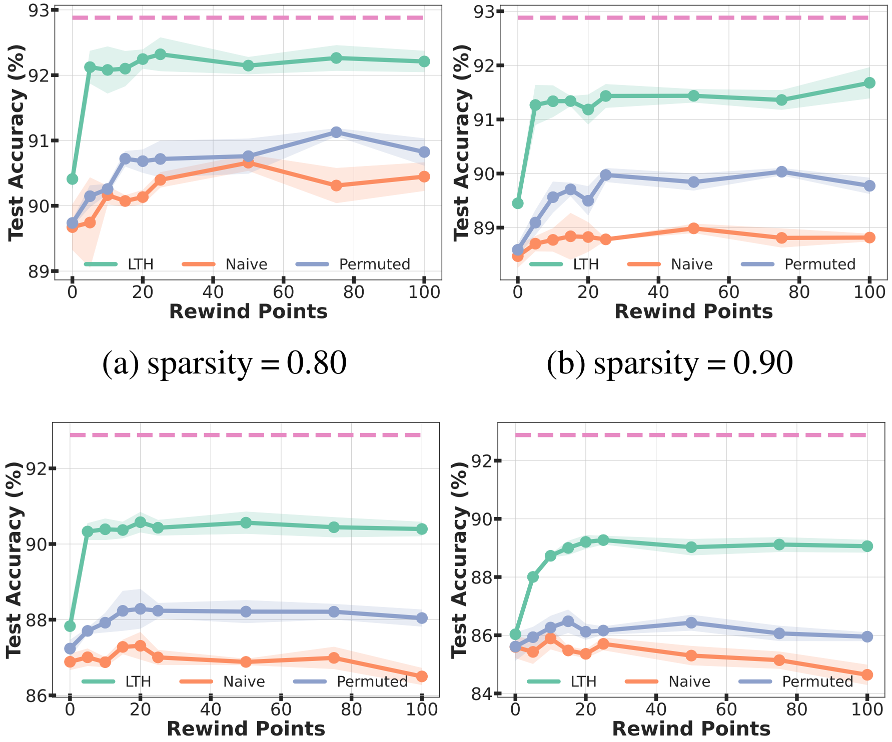
ii. Experimental Methodology

iii. Results

iv. Analysis



ResNet20 x {Width 1} on CIFAR-10

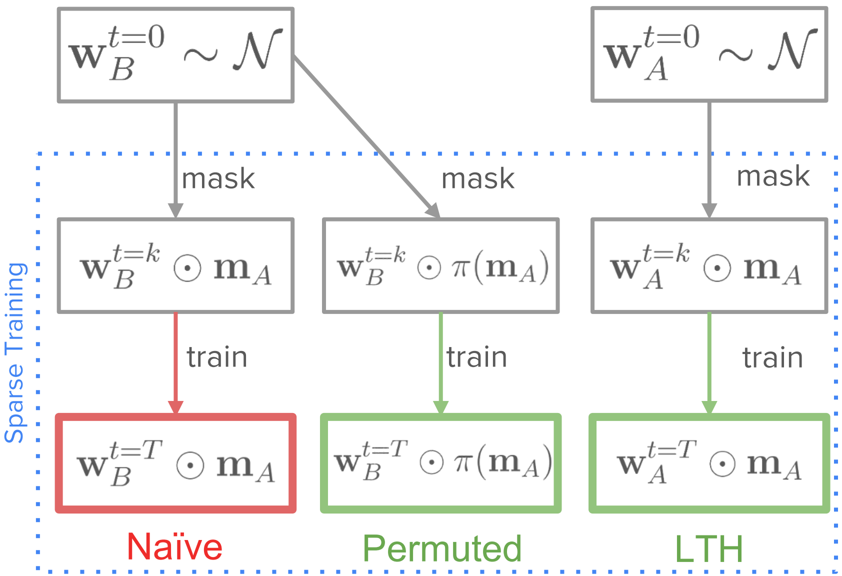


(a) sparsity = 0.80

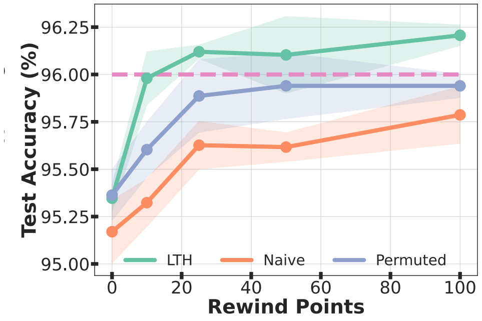
(b) sparsity = 0.90

(c) sparsity = 0.95

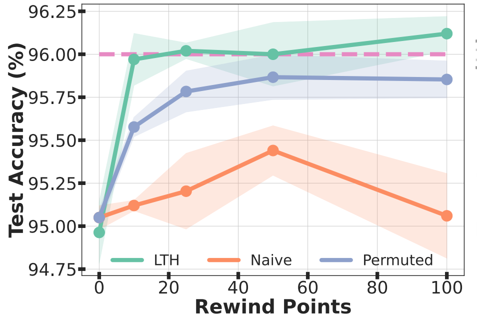
(d) sparsity = 0.97



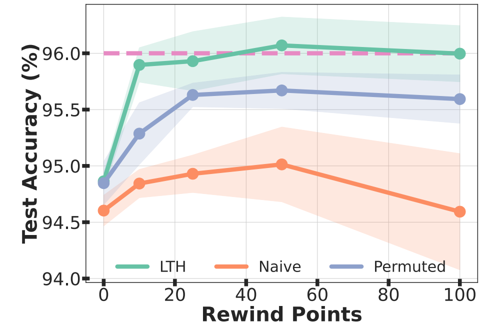
ResNet20 x {Width 8} on CIFAR-10



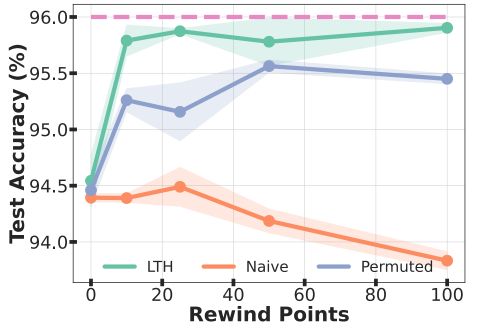
(a) sparsity = 0.80



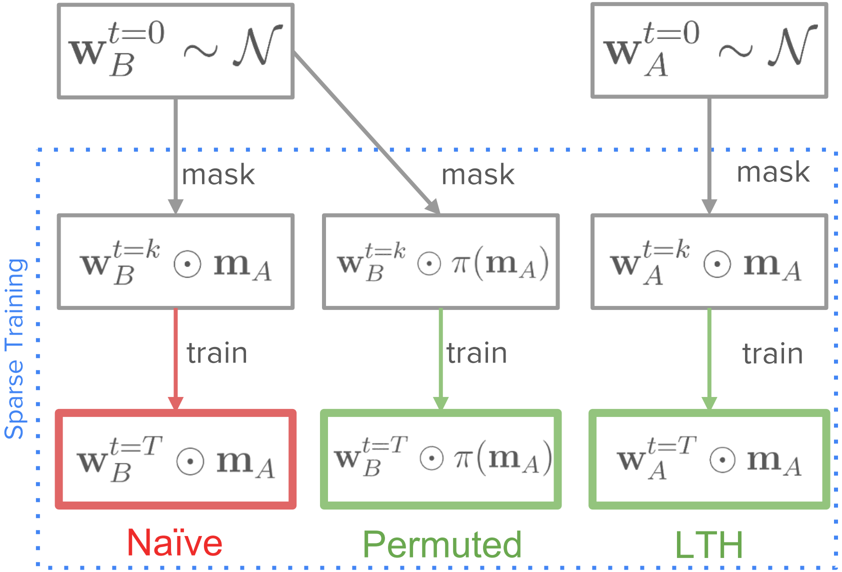
(b) sparsity = 0.90



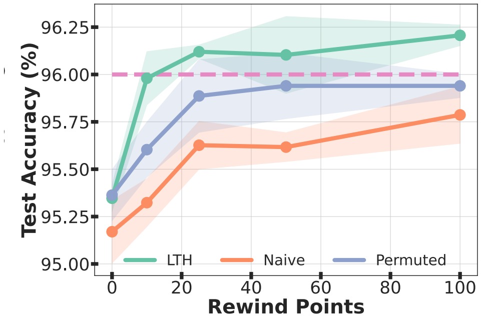
(c) sparsity = 0.95



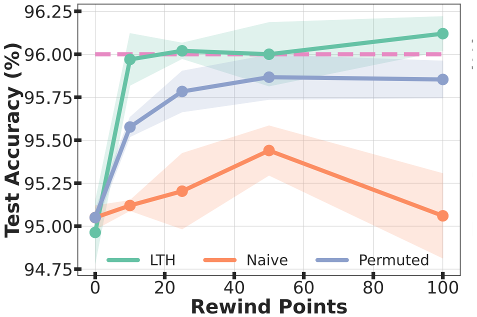
(d) sparsity = 0.97



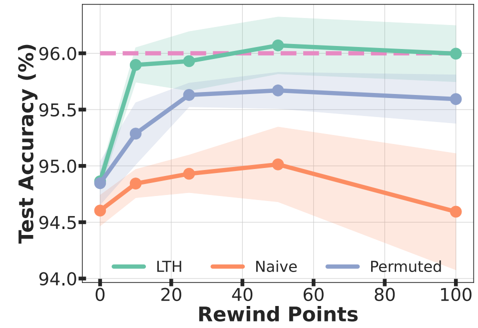
ResNet20 x {Width 8} on CIFAR-10



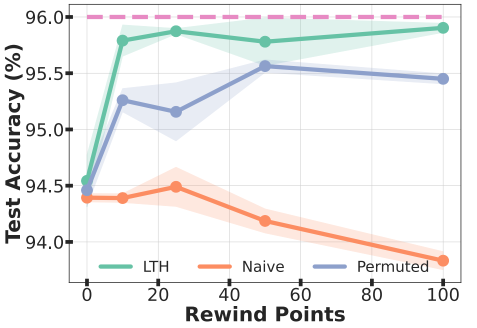
(a) sparsity = 0.80



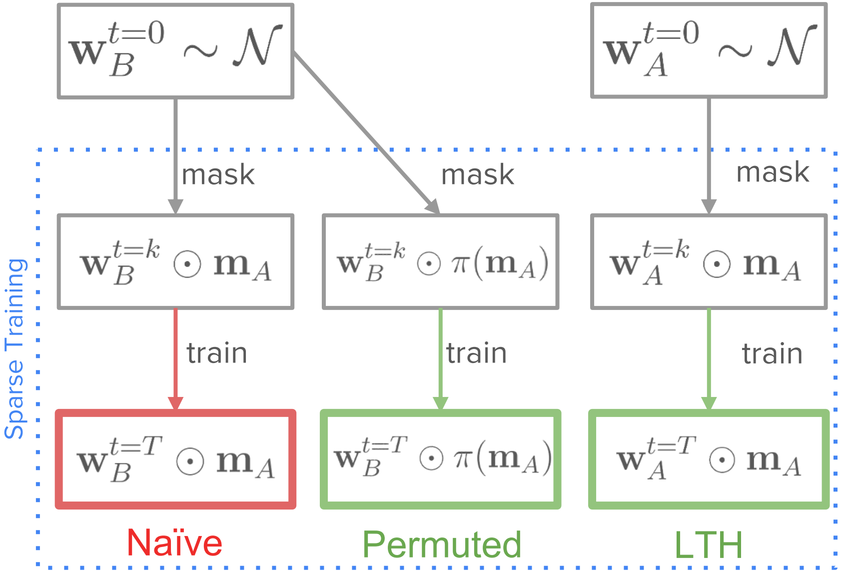
(b) sparsity = 0.90



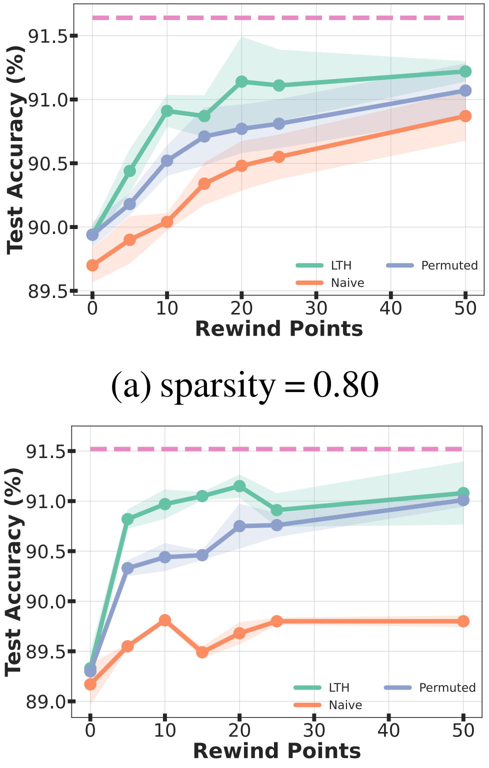
(c) sparsity = 0.95



(d) sparsity = 0.97



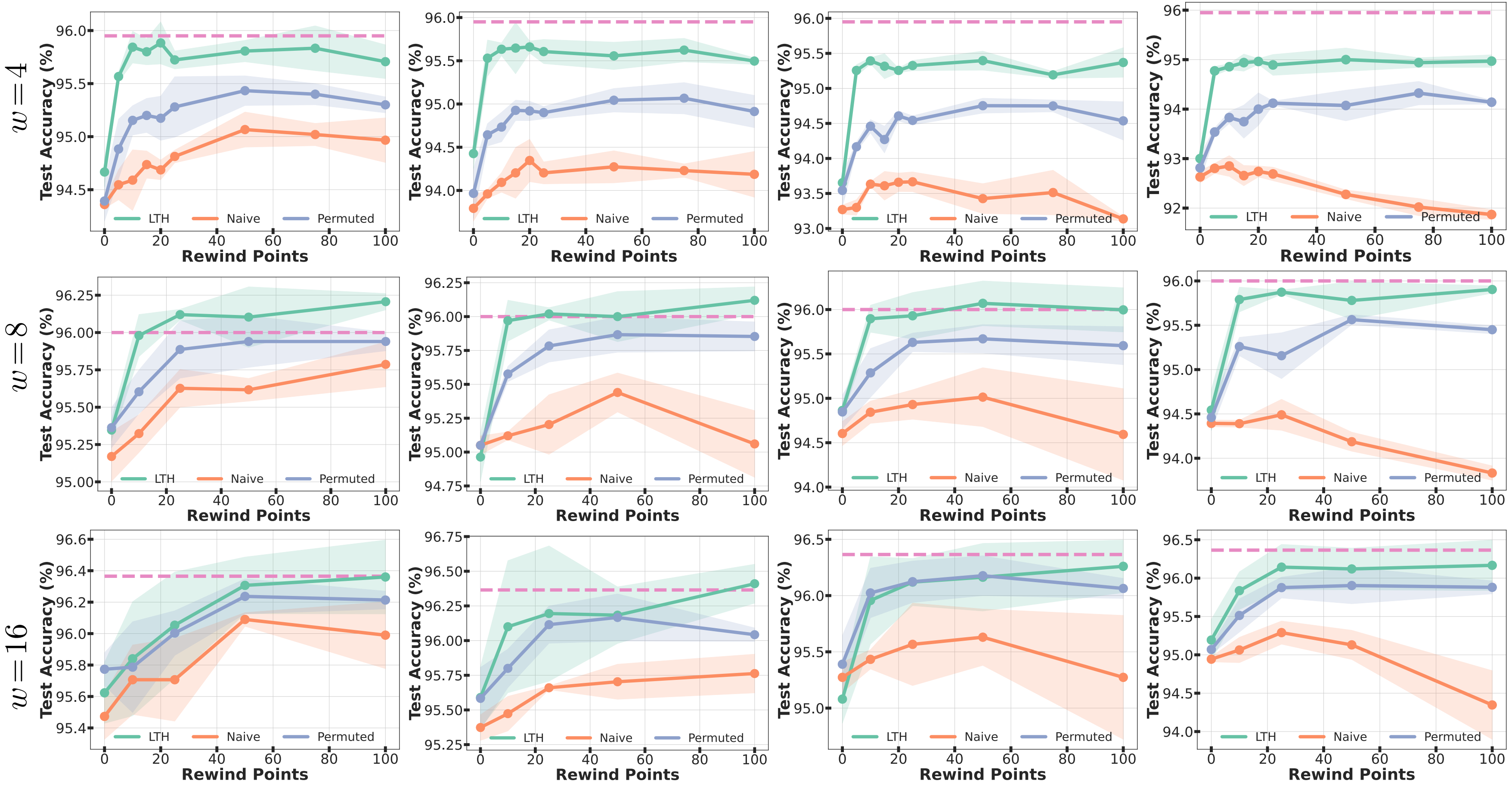
VGG11 x {Width 1} on CIFAR-10



(a) sparsity = 0.80

(b) sparsity = 0.90

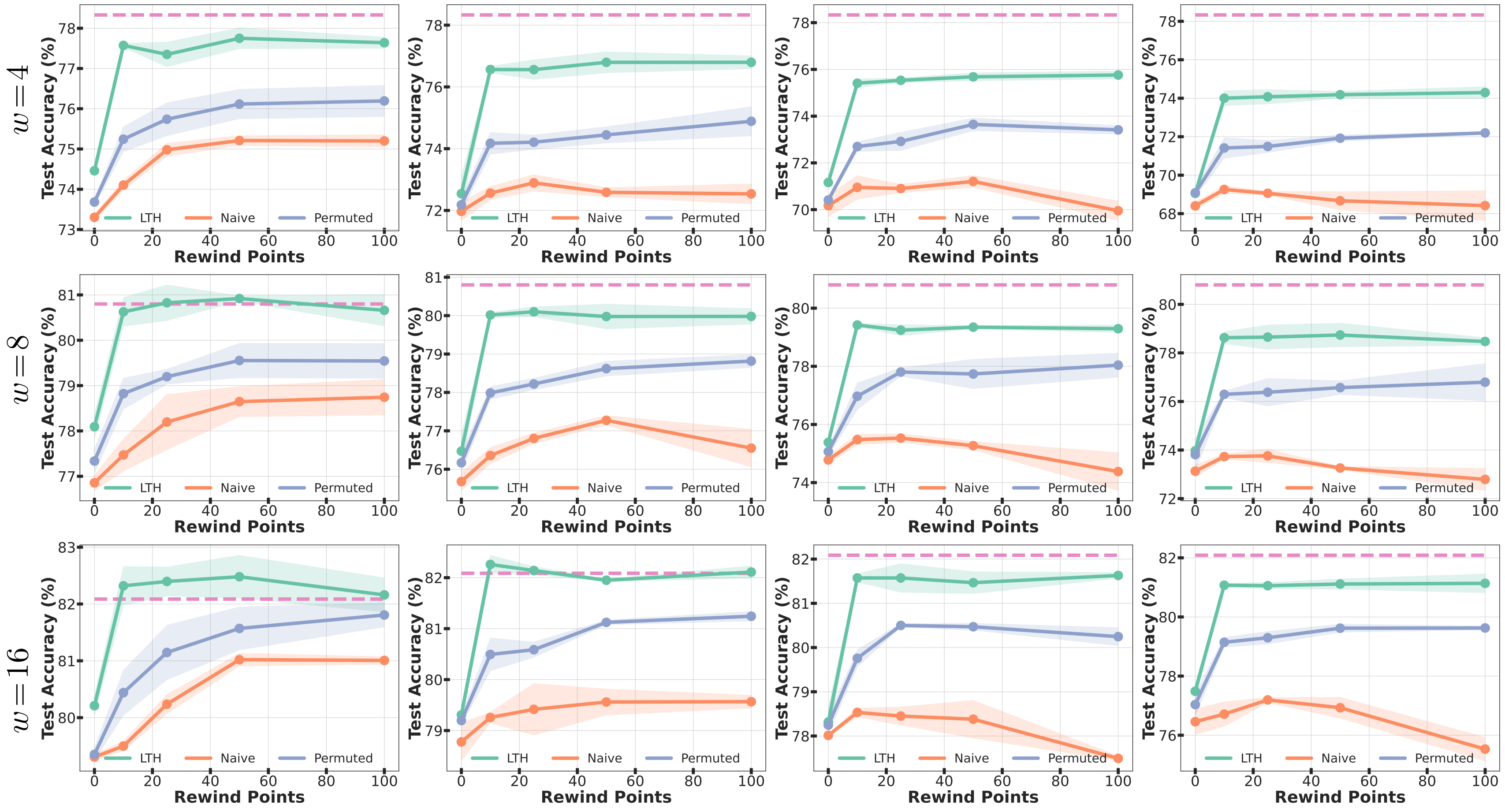
CMI UNIVERSITY OF CALGARY



(a) sparsity = 0.80 (b) sparsity = 0.90 (c) sparsity = 0.95 (d) sparsity = 0.97

ResNet20 x {4,8,16} on CIFAR-10

UNIVERSITY OF CALGARY

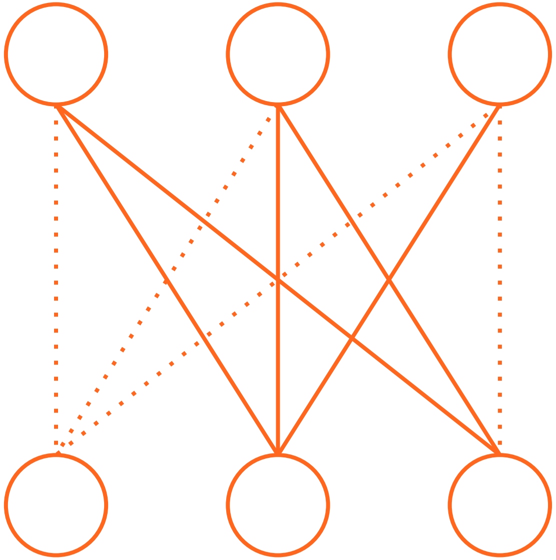


(a) sparsity = 0.80 (b) sparsity = 0.90 (c) sparsity = 0.95 (d) sparsity = 0.97

ResNet20 x {4,8,16} on CIFAR-100

UNIVERSITY OF CALGARY

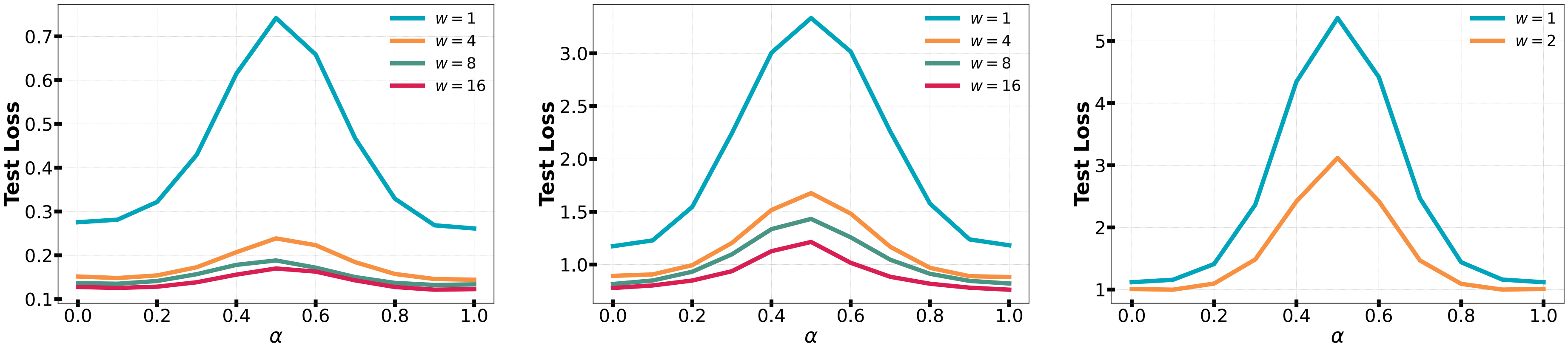
4. Aligning Sparse Masks



i. Hypothesis
ii. Experimental Methodology
iii. Results
iv. Analysis

CMI.. UNIVERSITY OF CALGARY

Effect of Model Width Multiplier



(a) ResNet20×$\{w\}$/CIFAR-10 (b) ResNet20×$\{w\}$/CIFAR-100 (c) ResNet50×$\{w\}$/ImageNet

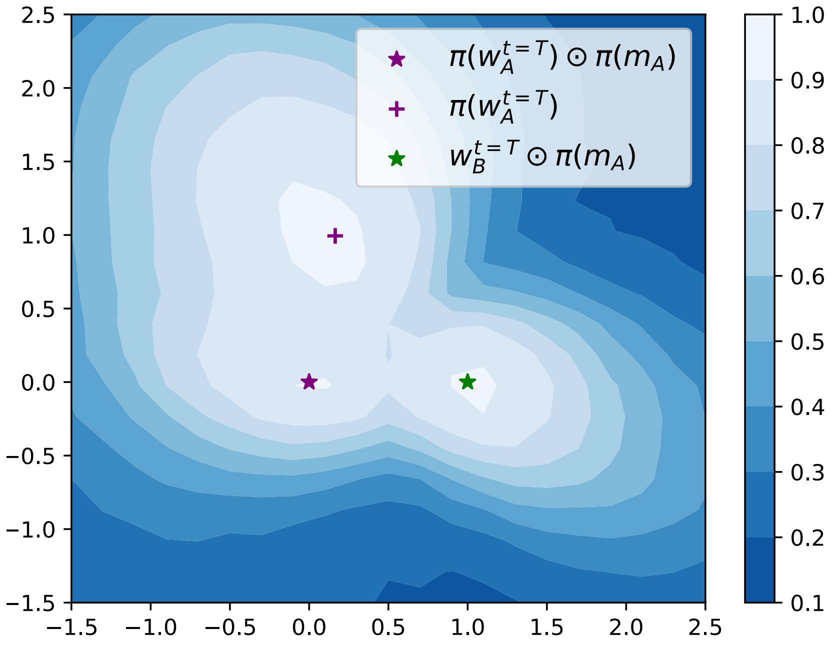
Demonstrating LMC by linearly interpolating between $\mathbf{w}_B^{t=T}$ and $\pi(\mathbf{w}_B^{t=T})$.

- Larger width exhibits better linear mode connectivity, i.e. lower loss barriers between $\mathbf{w}_B^{t=T}$ and $\pi(\mathbf{w}_B^{t=T})$
- As the width of the model increases, the approximate permutation matching algorithm is more accurate, reducing the loss barrier
- Our results are best when model width multiplier is high

UNIVERSITY OF CALGARY

Analysis of 0-1 Loss Basin of Solutions

- If our permutation matching is only approximate, are solutions in same basin?
- Here we analyze the 0-1 loss of three solutions, plotting their planar cross-section
  - Permuted dense soln. A: $\pi(w_A^{t=T})$
  - Permuted masked soln. A: $\pi(\mathbf{w}_A^{t=T})\odot\pi(\mathbf{m_A})$
  - Soln. B masked by permuted mask: $\mathbf{w}_B^{t=T}\odot\pi(\mathbf{m_A})$
- In same basin, but different modes



Functional Diversity

- Our previous work showed that the LTH relearns a highly similar solution
- Unlike LTH, we can reuse the LTH mask with different random initializations
- We do see improved function diversity over LTH, comparable to dense!
- More computationally efficient way to improve diversity than iterative magnitude pruning alone

Mask	Test Accuracy (%)	Ensemble Acc. (%)	Disagree-ment	KL	JS
		ResNet20×{1}/CIFAR-10			
none (dense)	92.76 ± 0.106	-	-	-	-
IMP	91.09 ± 0.041	93.25	0.093	0.352	0.130
LTH	$\mathbf{91.15 \pm 0.163}$	91.43	0.035	0.038	0.011
permuted	89.38 ± 0.170	**91.75**	0.107	**0.273**	**0.091**
naive	88.68 ± 0.205	91.07	**0.113**	0.271	0.089
		ResNet20×{4}/CIFAR-100			
none (dense)	78.37 ± 0.059	-	-	-	-
IMP	74.46 ± 0.321	79.27	0.259	1.005	0.372
LTH	$\mathbf{75.35 \pm 0.204}$	75.99	0.117	0.134	0.038
permuted	72.48 ± 0.356	**77.85**	0.278	0.918	0.327
naive	71.05 ± 0.366	76.15	**0.290**	**0.970**	**0.348**

[1]Utku Evci, Yani Ioannou, Cem Keskin, Yann Dauphin. *Gradient Flow in Sparse Neural Networks and How Lottery Tickets Win.* AAAI 2022

Conclusion

- The Lottery Ticket Hypothesis excited the community on the possibility of sparse training and sparse mask re-use, **but LTH is limited to re-learning the same soln**.
- **We explain the sparse training problem**: misalignment between a pruned mask and the loss basin of a new random initialization prevents effective re-use of sparse masks for training
- **We show how to re-use a mask to find new solutions:**
    - We can approximately permute an existing sparse mask for a new random initialization, although this is currently computationally expensive
    - We found the functional diversity of sparse training solutions to be comparable to dense training when using permuted masks.

UNIVERSITY OF CALGARY

Future Directions

- Improving the efficiency and/or efficacy of permutation alignment would make the method we propose more practical
- Explaining and/or avoiding weight ”rewinding”, i.e. checkpoints in LTH/sparse training
  - Notably Dynamic Sparse Training (DST) methods do not need this, but learn masks
- We see high function diversity with our method of sparse training:
  - Can we efficiently create ensembles using permutations of sparse masks?
  - Could help align weight sparse experts in MoEs for merging

NSERC
CRSNG

ALBERTA
INNOVATES

CMI..
www.calgaryml.com

Défense nationale
National Defence

amazon | science

Digital Research Alliance of Canada
Alliance de recherche numérique du Canada

Questions?

www.calgaryml.com

Yani Ioannou
yani.ioannou@ucalgary.ca

Sparse Training from Random Initialization: Aligning Lottery Ticket Masks using Weight Symmetry
Mohammed Adnan, Rohan Jain, Ekansh Sharma, Rahul Krishnan, and Yani Ioannou
In Forty-second International Conference on Machine Learning (ICML), Jul 2025

UNIVERSITY OF CALGARY