

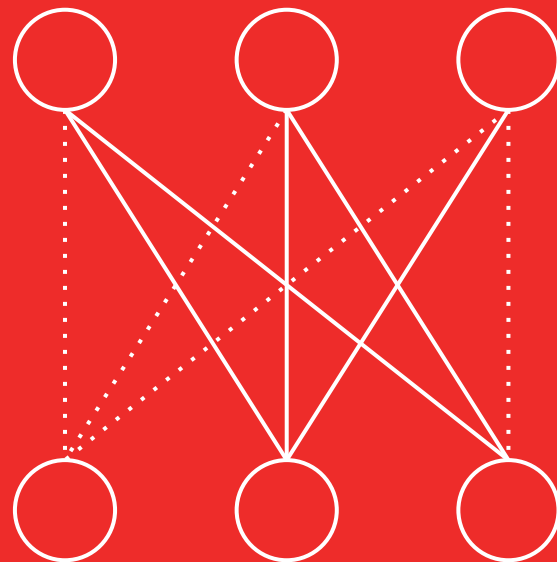
Sparse Training: Aligning Sparse Masks with Weight Symmetry

Yani Ioannou

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

Microsoft Research Cambridge

September 29th, 2025



UNIVERSITY OF
CALGARY

Sparse Training:

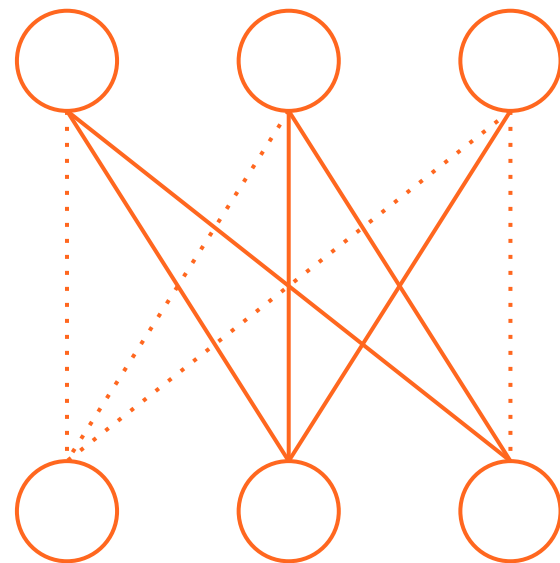
Aligning Sparse Masks with Weight Symmetry

1. Short Biography

2. Motivation

3. Background

4. Aligning Sparse Masks



Biography: Yani Ioannou

PhD, University of Cambridge, 2018

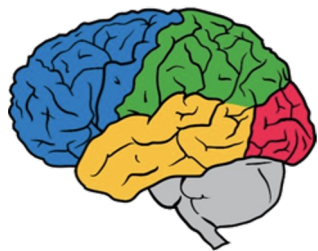
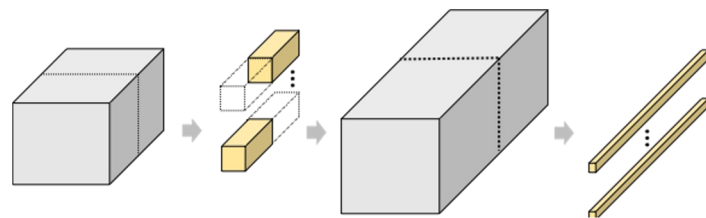
- **Prof. Roberto Cipolla** (Department of Engineering)
Dr. Antonio Criminisi (Microsoft Research)
- Microsoft Research PhD Scholarship

Training CNNs with Low-Rank Filters for Efficient Image Classification.

Yani Ioannou, Duncan Robertson, Jamie Shotton, Roberto Cipolla, Antonio Criminisi.
ICLR 2016

Deep Roots: Improving CNN Efficiency with Hierarchical Filter Groups

Yani Ioannou, Duncan Robertson, Roberto Cipolla, Antonio Criminisi.
CVPR 2017



2019 – 2020

**Google Brain (Toronto)
Visiting Researcher**

Gradient Flow in Sparse Neural Networks and How Lottery Tickets Win

Utku Evci, Yani Ioannou, Cem Keskin, Yann Dauphin
AAAI 2022 Oral Presentation



DYNAMIC SPARSE TRAINING WITH STRUCTURED SPARSITY

Mike Lasby¹, Anna Golubeva^{2,3}, Utku Evci⁴, Mihai Nica^{5,6}, Yani A. Ioannou¹

¹University of Calgary, ²Massachusetts Institute of Technology, ³IAIFI

⁴Google DeepMind, ⁵University of Guelph, ⁶Vector 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 $3.4\times/2.5\times$ on CPU for *online inference* and $1.7\times/13.0\times$ on GPU for inference with a batch size of 256 when compared to equivalent dense/unstructured (CSR) sparse layers, respectively.

1 INTRODUCTION

Dynamic Sparse Training (DST) methods such as RigL (Evci et al., 2021) are the state-of-the-art in sparse training methods for Deep Neural Networks (DNNs). DST methods typically learn *unstructured* masks resulting in 85–95% fewer weights than dense models, while maintaining dense-like generalization and typically outperforming masks found via pruning. Furthermore, sparse-to-sparse DST algorithms are capable of employing sparsity *both during training and inference*, unlike pruning and dense-to-sparse DST methods such as SR-STE (Zhou et al., 2021) which only exploit sparsity at inference time.

While models trained with DST methods are highly sparse and enable a large reduction in Floating Point Operations (FLOPs) in theory, realizing these speedups on hardware is challenging when the sparsity pattern is unstructured. Even considering recent advances in accelerating unstructured Sparse Neural Networks (SNNs) (Gale et al., 2020; Elsen et al., 2020; Ji & Chen, 2022), structured sparsity realizes much stronger acceleration on real-world hardware. On the other hand, structured sparse pruning often removes salient weights, resulting in worse generalization than comparable unstructured SNNs for the same sparsity level (Fig. 1a). Our work presents a best-of-both-worlds approach: we exploit the DST framework to learn *both* a highly-sparse *and* structured representation while maintaining generalization performance. In summary, our work makes the following contributions:



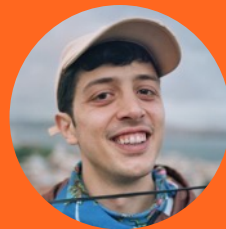
Michael Lasby

PhD Student
(Currently interning at Cerebras)



Anna Golubeva

Postdoctoral Fellow, MIT/IAIFI



Utku Evci

Research Scientist
Google DeepMind



Mihai Nica

Associate Professor
U. Guelph/Vector

● International Conference for Learning
Representations (ICLR) 2024

What's Left After Distillation? How Knowledge Transfer Impacts Fairness and Bias

Aida Mohammadshahi

aida.mohammadshahi@ucalgary.ca

Yani Ioannou

Department of Electrical and Software Engineering
Schulich School of Engineering, University of Calgary
Calgary, AB, Canada

yani.ioannou@ucalgary.ca

Reviewed on OpenReview: <https://openreview.net/forum?id=zBj46Y2fjN>

Abstract

Knowledge Distillation is a commonly used Deep Neural Network (DNN) compression method, which often maintains overall generalization performance. However, we show that even for balanced image classification datasets, such as CIFAR-100, Tiny ImageNet and ImageNet, as many as 41% of the classes are statistically significantly affected by distillation when comparing class-wise accuracy (i.e. class bias) between a teacher/distilled student or distilled student/non-distilled student model. Changes in class bias are not necessarily an undesirable outcome when considered outside of the context of a model's usage. Using two common fairness metrics, Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD) on models trained with the CelebA, Trifeature, and HateXplain datasets, our results suggest that increasing the distillation temperature improves the distilled student model's fairness, and the distilled student fairness can even surpass the fairness of the teacher model at high temperatures. Additionally, we examine individual fairness, ensuring similar instances receive similar predictions. Our results confirm that higher temperatures also improve the distilled student model's individual fairness. This study highlights the uneven effects of distillation on certain classes and its potentially significant role in fairness, emphasizing that caution is warranted when using distilled models for sensitive application domains.

1 Introduction

DNNs require significant computational resources, resulting in large overheads in compute, memory, and energy. Decreasing this computational overhead is necessary for many real-world applications where these costs would otherwise be prohibitive, or even make their application infeasible — e.g. the deployment of DNNs on mobile phones or edge devices with limited resources (Chen et al., 2016; Cheng et al., 2018; Gupta and Agrawal, 2022; Menghani, 2023). To address this challenge, DNN model compression methods have been developed that reduce the size and complexity of DNNs while maintaining their generalization performance (Cheng et al., 2017). One such widely used model compression method is Knowledge Distillation (distillation) (Hinton et al., 2015). Distillation has found extensive application in both industry and academia across various domains of artificial intelligence, encompassing areas such as Natural Language Processing (NLP) (Jiao et al., 2019; Fu et al., 2021; Liu et al., 2020), speech recognition (Ng et al., 2018; Gao et al., 2019; Perez et al., 2020), and visual recognition (Yan et al., 2019; Dou et al., 2020; Chawla et al., 2021), specifically image classification (Zhu et al., 2019; Chen et al., 2019; Gou et al., 2021).

Distillation involves transferring knowledge from a complex model with superior performance (referred to as the *teacher*) to a simpler model (known as the *student*). In practice this allows the student model to achieve comparable or even better generalization than the teacher model, while using far fewer parameters (Hinton et al., 2015; Gou et al., 2021). Despite the widespread use of distillation, evaluation of the impact of distillation since its proposal by (Hinton et al., 2015) has overwhelmingly focused almost exclusively on the impact it has on generalization performance (Cho and Hariharan, 2019; Mirzadeh et al., 2020).



Aida Mohammadshahi

MSc (Defended Jan. 2025)

ML Developer @ AltaML

- Presented at NeurIPS WiML Workshop in Dec. 2024
- Published in Transactions in Machine Learning Research (TMLR), March 2025

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 Ioannou¹

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

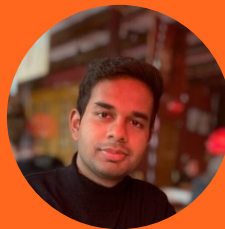
and practitioners alike. To overcome this issue, different model compression methods, such as pruning, quantization and knowledge distillation, have been proposed to reduce the model size at different phases of training or inference. Post-training model pruning ([Han et al., 2016](#)) has been shown 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 ([Evcil 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 with the optimization basin associated with the new random initialization. We illustrate our hypothesis in Figure 1.

To empirically validate our hypothesis, we obtain a sparse mask using Iterative Magnitude Pruning (IMP) ([Renda et al., 2020](#); [Han et al., 2015](#)) on model A (from Figure 1) and show that given a permutation that aligns the optimization basin of model A and a new random initialization, the mask can



Mohammed Adnan
PhD Student



Dr. Ekansh Sharma
PhD Graduate
University of Toronto



Rohan Jain
MSc Graduate



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

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

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

Muslims in ML (MusiML) Workshop

2:30-3:00 p.m. (Lightning Talk) *

4:30-5:00 p.m. (Internal Poster Session), 6:30-8:00 p.m. (Joint Poster Session for Affinity Groups)

* A Closer Look at Sparse Training in Deep Reinforcement Learning.

Muhammad Athar Ganale, Vincent Michalski, Samira Ebrahimi Kahou, Yani Ioannou.

This paper explores sparse training in DRL, highlighting methods to improve dynamic sparse training performance at high sparsity, underscoring the need for DRL-specific strategies.

Long-Tail Learning with Language Model Guided Curricula.

Mohammed Adnan, Rahul Krishnan, Yani Ioannou.

Improving performance on long-tail classes by leveraging LLMs to build curricula.

Women in Machine Learning (WiML) Workshop

6:30 p.m - 8:00 p.m (Joint Poster Session for Affinity Groups)

Learning to Reweight Examples in Backdoor Defense.

Yufan Feng, Benjamin Tan, Yani Ioannou.

We extend the online sample reweighting method from robust learning to the context of backdoor defense.

What's Left After Distillation? How Knowledge Transfer Impacts Fairness and Bias.

Aida Mohammadshahi, Yani Ioannou.

We explore the impact of knowledge distillation temperature on fairness for language and image classification models.

Friday, December 13th, 2024

Main Conference

4:30-7:30 p.m. (Poster Session)

Navigating Extremes: Dynamic Sparsity in Large Output Spaces

Nasibullah Nasibullah, Erik Schultheis, Mike Lasby, Yani Ioannou, Rohit Babbar.

Investigates Dynamic Sparse Training for large output spaces. Leveraging semi-structured sparsity, intermediate layers, and auxiliary loss, it enables end-to-end training with millions of labels.

Poster Location: East Exhibit Hall A-C #2004

Saturday, December 14th, 2024

UniReps: Unifying Representations in Neural Models

4:30-7:30 p.m. (Poster Session)

Winning Tickets from Random Initialization: Aligning Masks for Sparse Training.

Rohan Jain, Mohammed Adnan, Ekansh Sharma, Yani Ioannou.

Lottery Tickets can't be trained from random init. We show that permuting the mask to align with the new initialization's optimization basin results in a mask that better approaches LTH generalization.



- **Neural Information Processing Systems (NeurIPS) 2024 in Vancouver**
- **5 different works being presented by 6 CML students across main conference and workshops**

Monday, July 14th, 2025

4th Muslims in ML (MusiML) Workshop

West Meeting Room 211-214, 3:00 - 4:00 p.m. (Poster Session)

Does Compression Exacerbate Large Language Models' Social Bias?

Muhammad Athar Ganale · Mohammed Adnan · Arfa Raja · Shaina Roza · Yani Ioannou

Women in Machine Learning (WiML) Symposium

West Meeting Room 211-214, 3:00 - 4:00 p.m. (Poster Session)

Backdooring VLMs via Concept-Driven Triggers

Yufan Feng · Weimin Lyu · Yuxin Wang · Benjamin Tan · Yani Ioannou

Tuesday, July 15th, 2025

Main Conference

East Exhibition Hall A-B, 11:00 a.m. - 1:30 p.m. (Poster Session)

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

Mohammed Adnan · Rohan Jain · Ekansh Sharma · Rahul G. Krishnan · Yani Ioannou

Poster Location: East Exhibition Hall A-B #E-2106

Friday, July 18th, 2025

3rd Workshop on High-dimensional Learning Dynamics (HiLD)

West Meeting Room 118-120, 4:45 - 5:30 p.m. (Poster Session)

Understanding Normalization Layers for Sparse Training

Mohammed Adnan · Ekansh Sharma · Rahul G. Krishnan · Yani Ioannou

Saturday, July 19th, 2025

3rd Workshop on Efficient Systems for Foundation Models (ES-FoMo III)

East Exhibition Hall A, 1:00 - 2:30 p.m. (Poster Session)

SD²: Self-Distilled Sparse Drafters

Mike Lasby · Nish Sinnadurai · Valavan Manohararajah · Sean Lie · Yani Ioannou · Vithursan Thangarasa

Data in Generative Models Workshop: The Bad, the Ugly, and the Greats (DIG-BUGS)

West Ballroom A, 3:00 - 3:45 p.m. (Poster Session)

Backdooring VLMs via Concept-Driven Triggers

Yufan Feng · Weimin Lyu · Yuxin Wang · Benjamin Tan · Yani Ioannou

Workshop on Technical AI Governance

West Meeting Room 109-110, 3:00 - 4:00 p.m. (Poster Session)

Exploring Functional Similarities of Backdoored Models

Yufan Feng · Benjamin Tan · Yani Ioannou



- **International Machine Learning Conference (ICML) 2025 in Vancouver in July**
- **Five students from CML Lab presenting 6 different works across 6 workshops and the main conference**

Sparse Training:

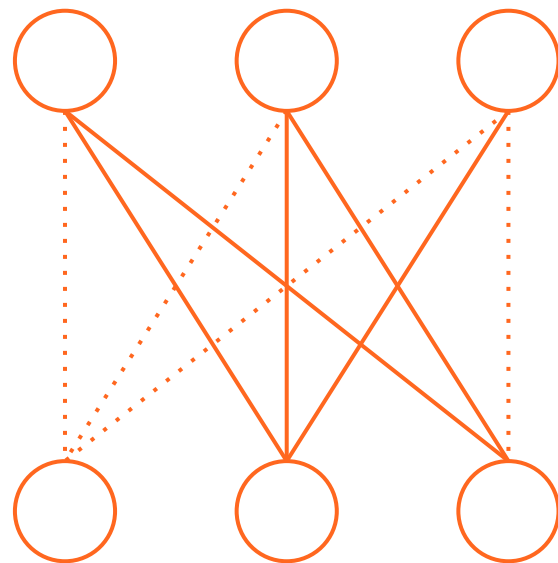
Aligning Sparse Masks with Weight Symmetry

1. Short Biography

2. Motivation

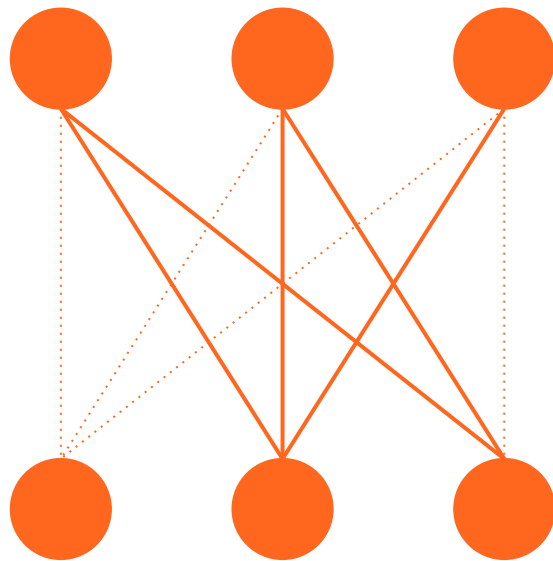
3. Background

4. Aligning Sparse Masks



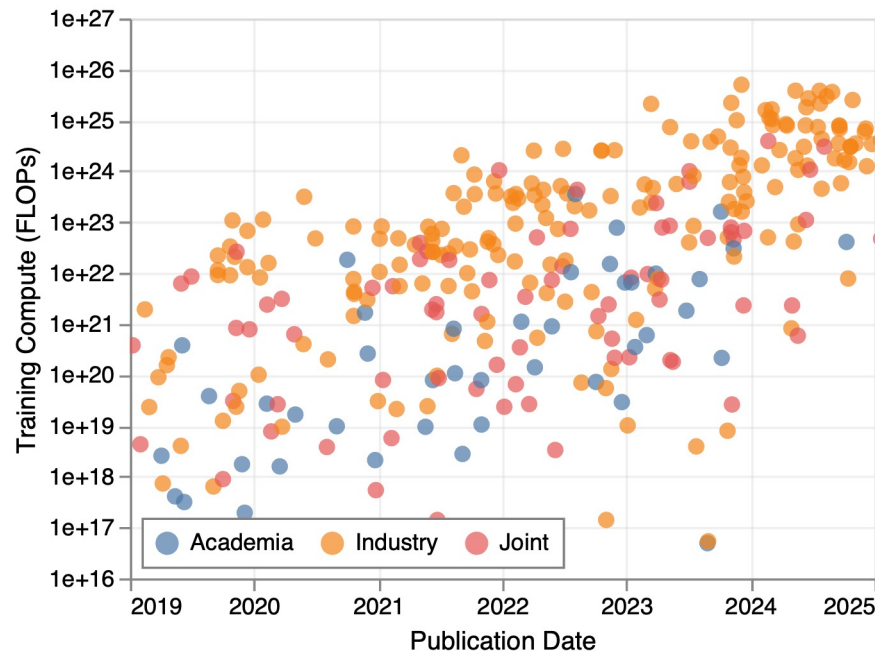
Why Sparse Neural Networks?

- 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

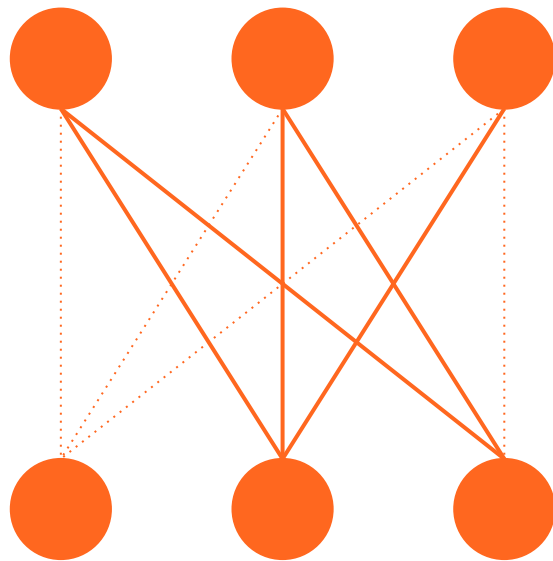
- State of the art models are becoming **exponentially more expensive to train**
- AI Research is **less accessible**
- Inference cost is increasingly important, sparse training shows promise in **learning better masks for inference** than pruning



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

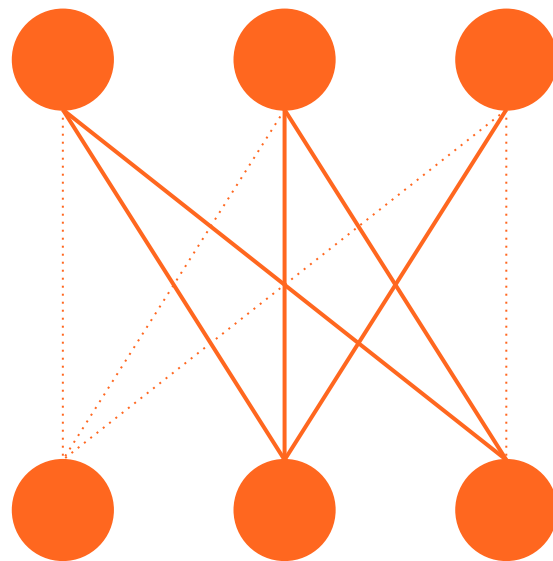
Motivation: Learning NN Structure

- 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

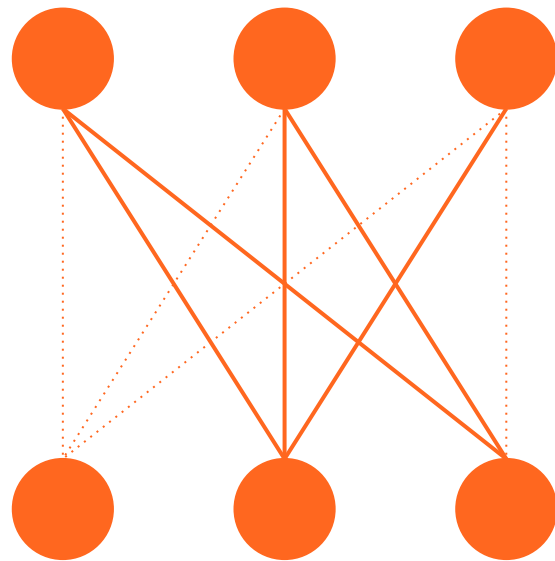
- Training NNs from random initialization is unreasonably effective... but **not always**
- Much of the “deep learning” progress can be attributed to improved NN training:
 - Initialization, normalization, residual connections, etc.
- **Sparse training breaks NN training**
- Understanding sparse training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

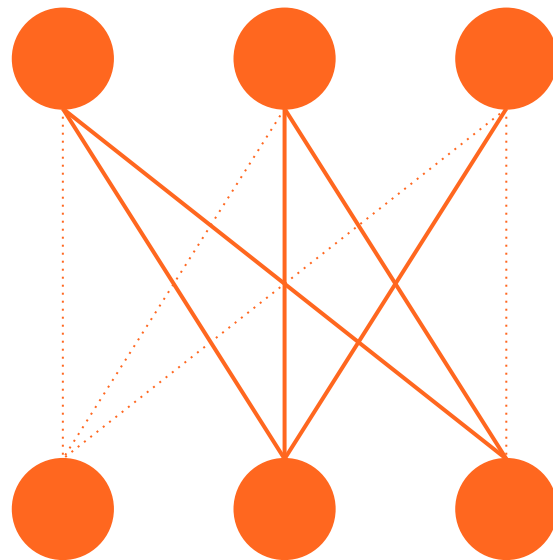
What we have been doing:

- **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^{*12} Rohan Jain^{*1} Ekansh Sharma³² Rahul G. Krishnan³² Yani Ioannou¹

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 with the new random initialization. We show that this is possible and associated with the new random initialization in our hypothesis in Figure 1.

To empirically validate our hypothesis, we use Iterative Magnitude Pruning (Han et al., 2015) on model A that given a permutation that aligns model A and a new random initialization. The sparse model will be trained to closer match the ge



Mohammed Adnan

PhD Student



Dr. Ekansh Sharma

PhD Graduate

University of Toronto



Rohan Jain

MSc Graduate



Dr. Rahul Krishnan

Assistant Professor

U. Toronto/Vector Institute

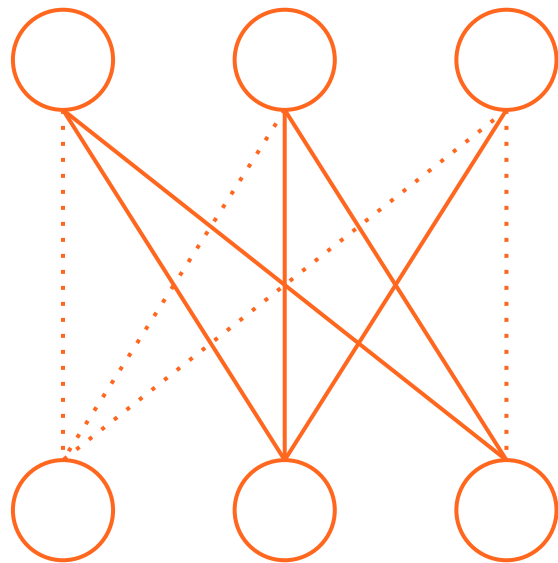
● Presented at International Conference
on Machine Learning (ICML) 2025

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

Sparse Training:

Aligning Sparse Masks with Weight Symmetry

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

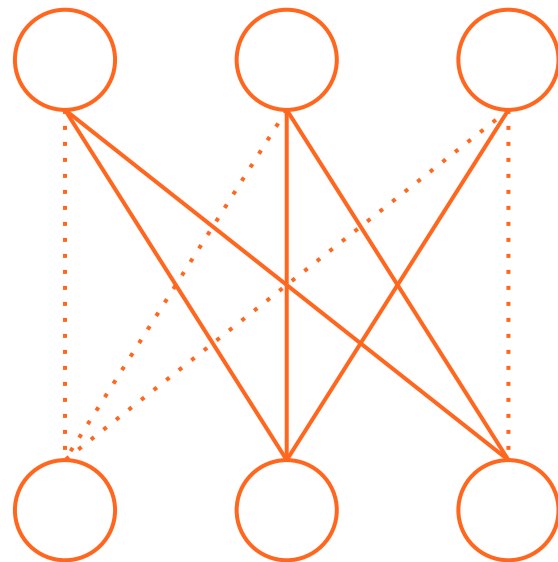


3. Background

i. Weight Symmetry

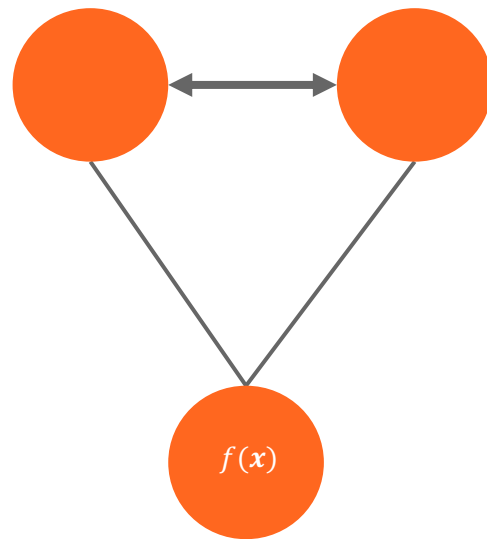
ii. Sparse Training Problem

iii. Lottery Ticket Hypothesis



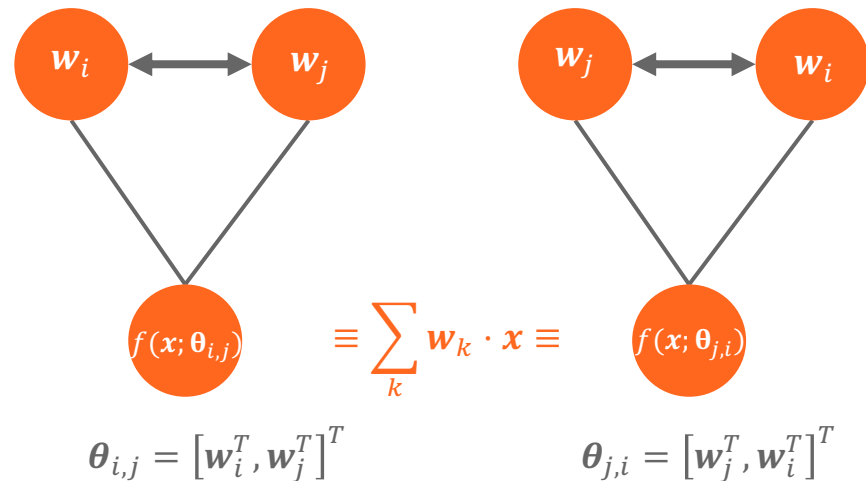
Weight Symmetry: Foundations

- 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*



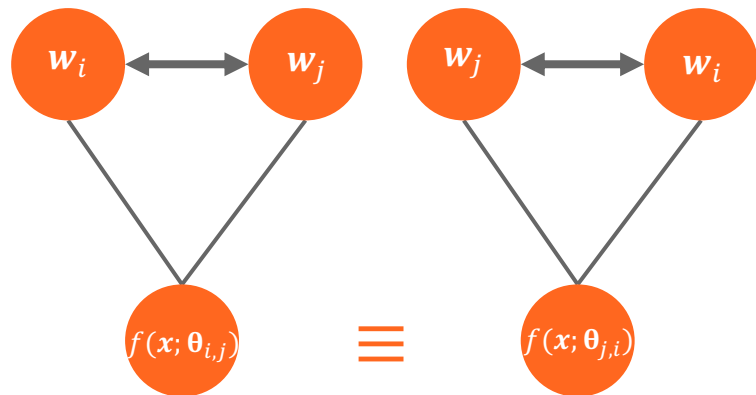
Weight Symmetry: Foundations

- 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})



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...

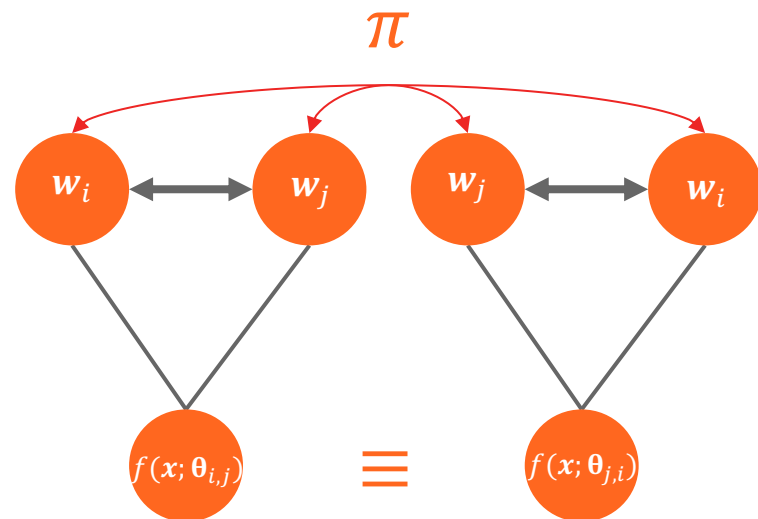


¹Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. *The role of permutation invariance in linear mode connectivity of neural networks*. ICLR 2022.

²Samuel 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 matching¹
 - Linear Assignment Problem (LAP) per layer
 - Maximizes correlation of weights/activations
 - Best results empirically for **very wide NNs**
- Activation matching more robust in general²



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

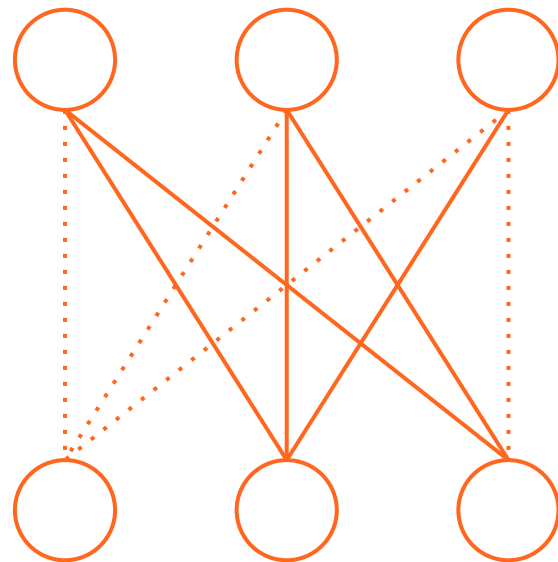
²Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. *Repair: Renormalizing permuted activations for interpolation repair*. ICLR 2023.

3. Background

i. Weight Symmetry

ii. Sparse Training Problem

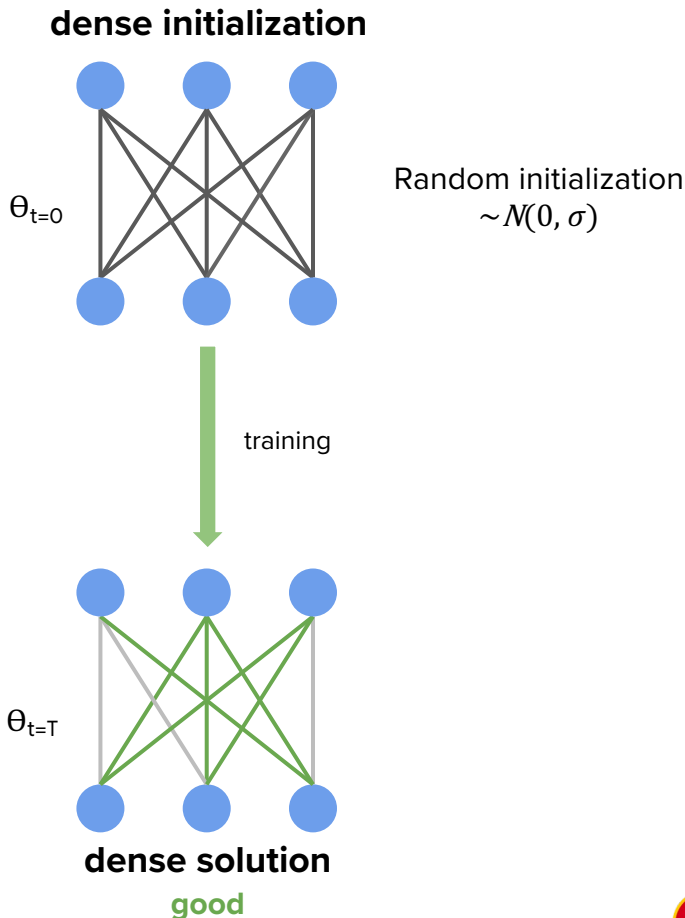
iii. Lottery Ticket Hypothesis



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**

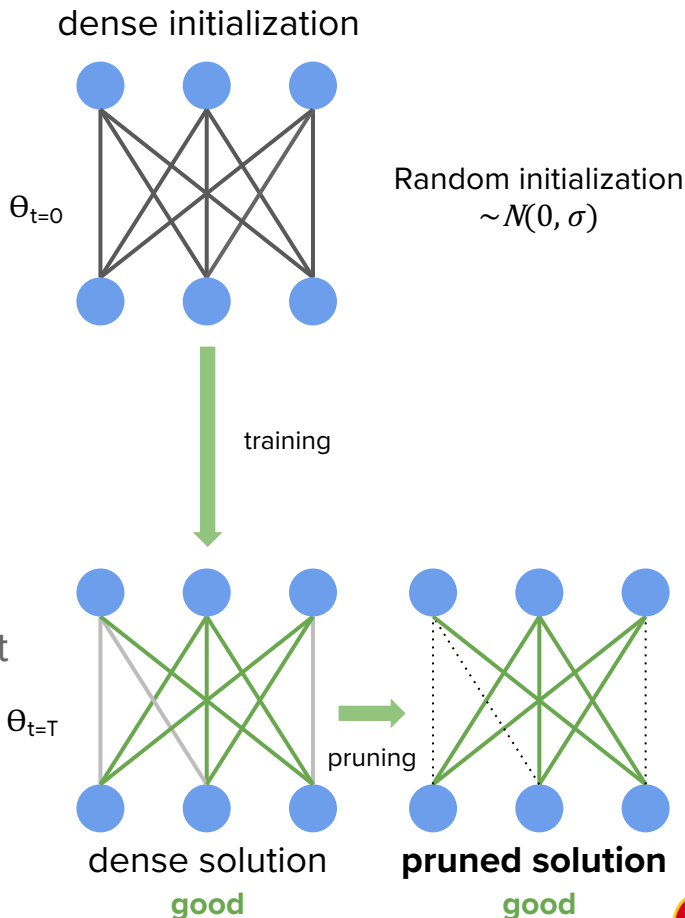
— High saliency weight
— Low saliency weight
..... Masked weight



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

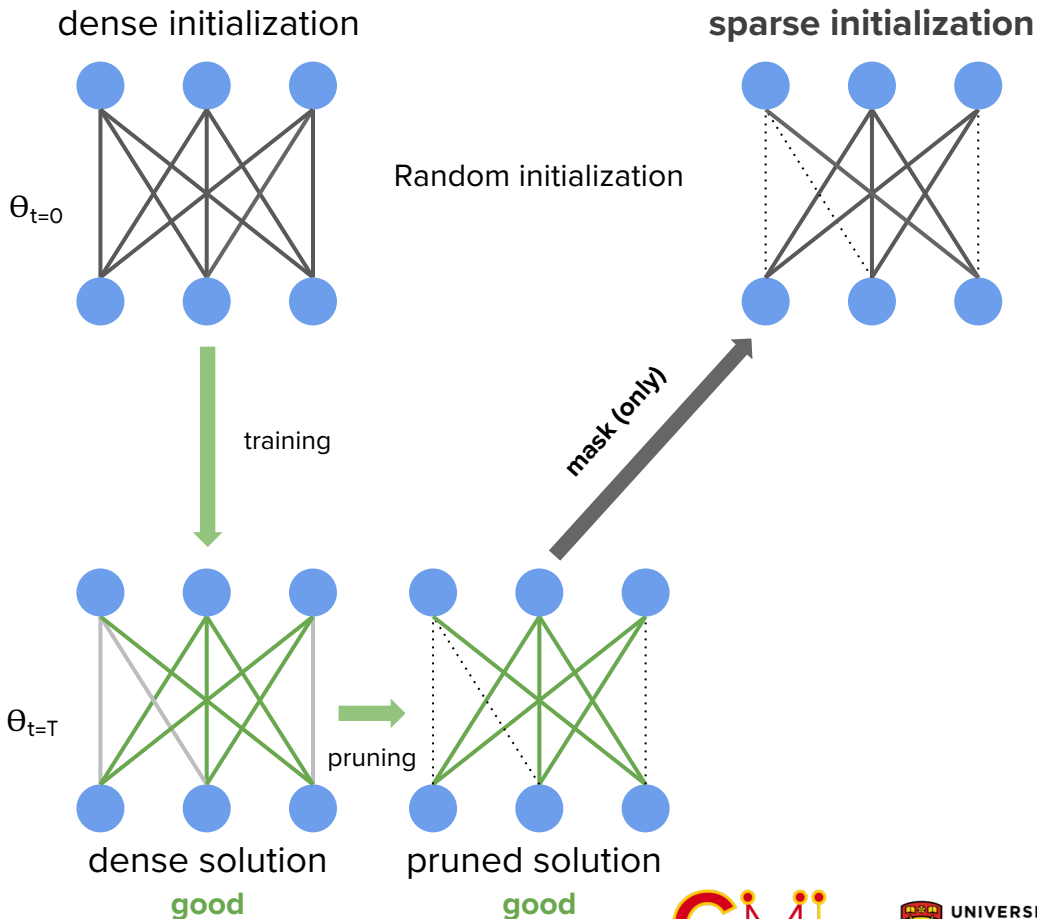
— High saliency weight
— Low saliency weight
..... Masked weight



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?**

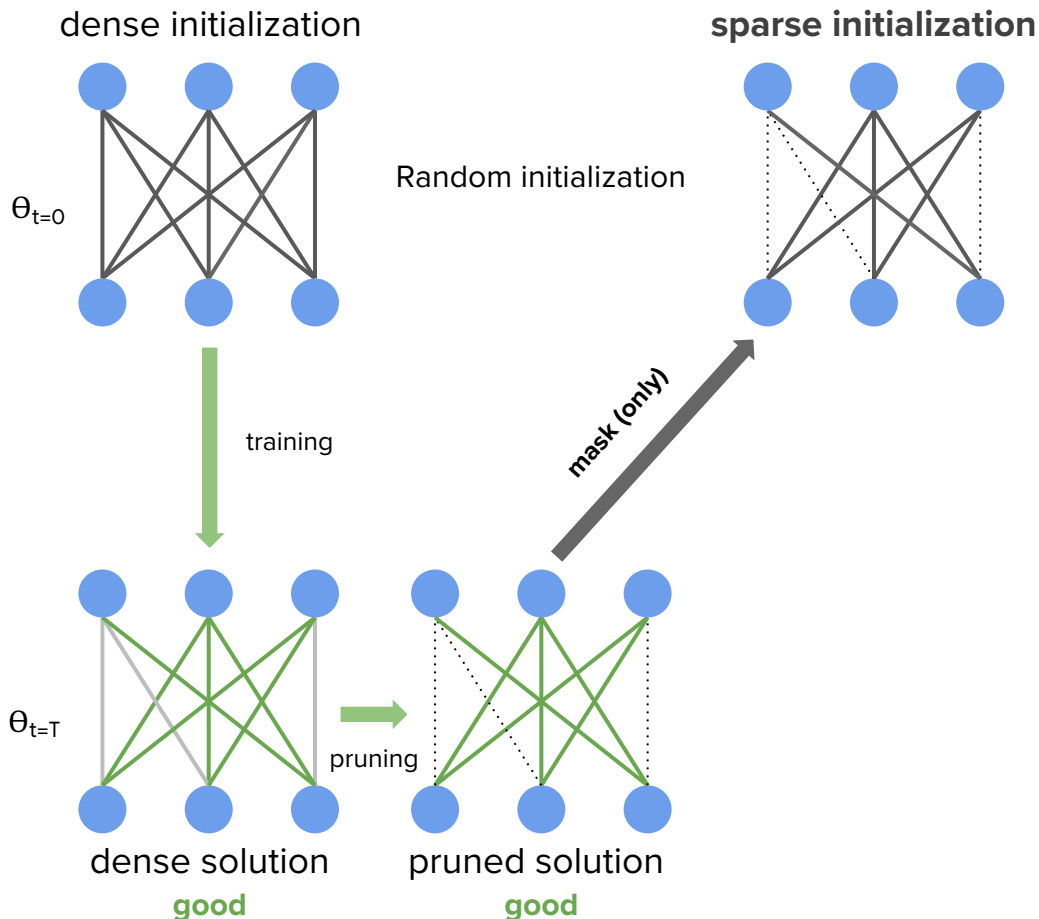
— High saliency weight
— Low saliency weight
..... Masked weight



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**...

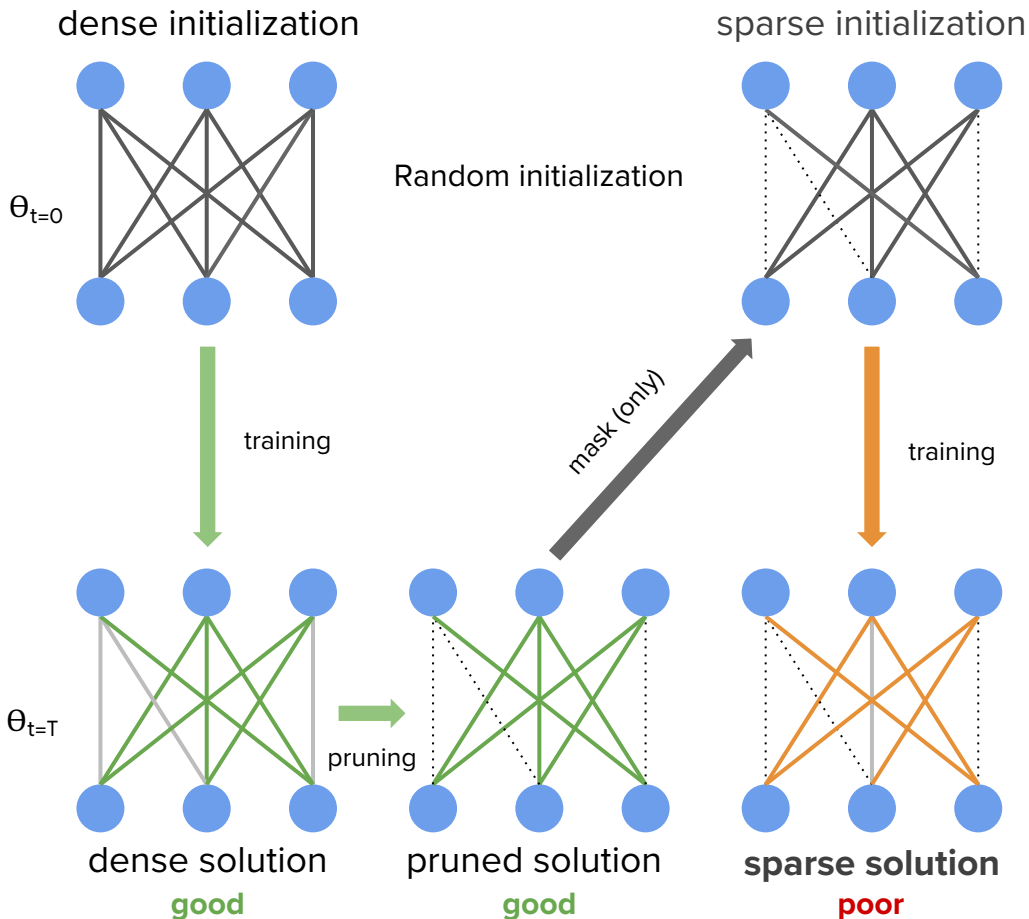
— High saliency weight
— Low saliency weight
..... Masked weight



Sparse Training Problem

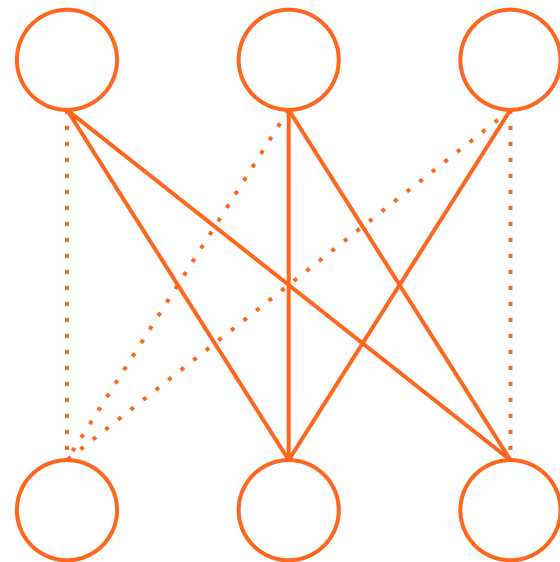
- The sparsely trained model (sparse solution) **doesn't generalize as well** as the original dense solution or pruned solution!

—— High saliency weight
—— Low saliency weight
..... Masked weight



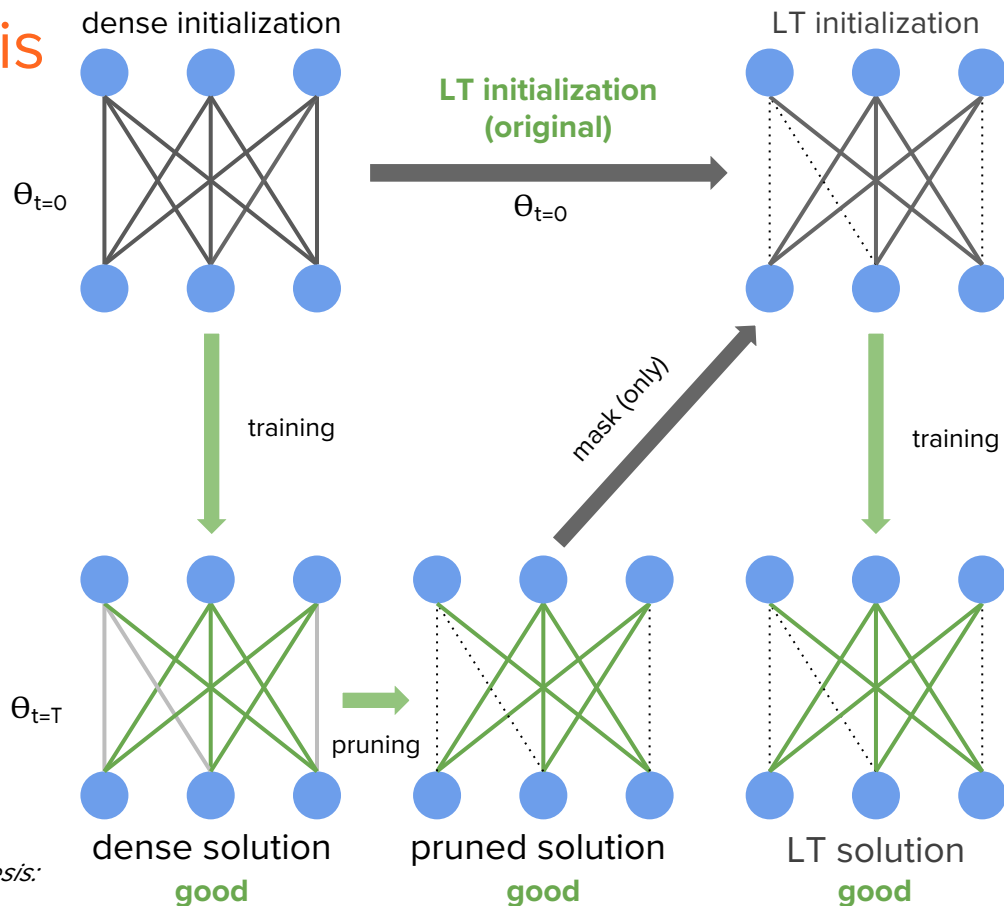
3. Background

- i. Weight Symmetry
- ii. Sparse Training Problem
- iii. Lottery Ticket Hypothesis**



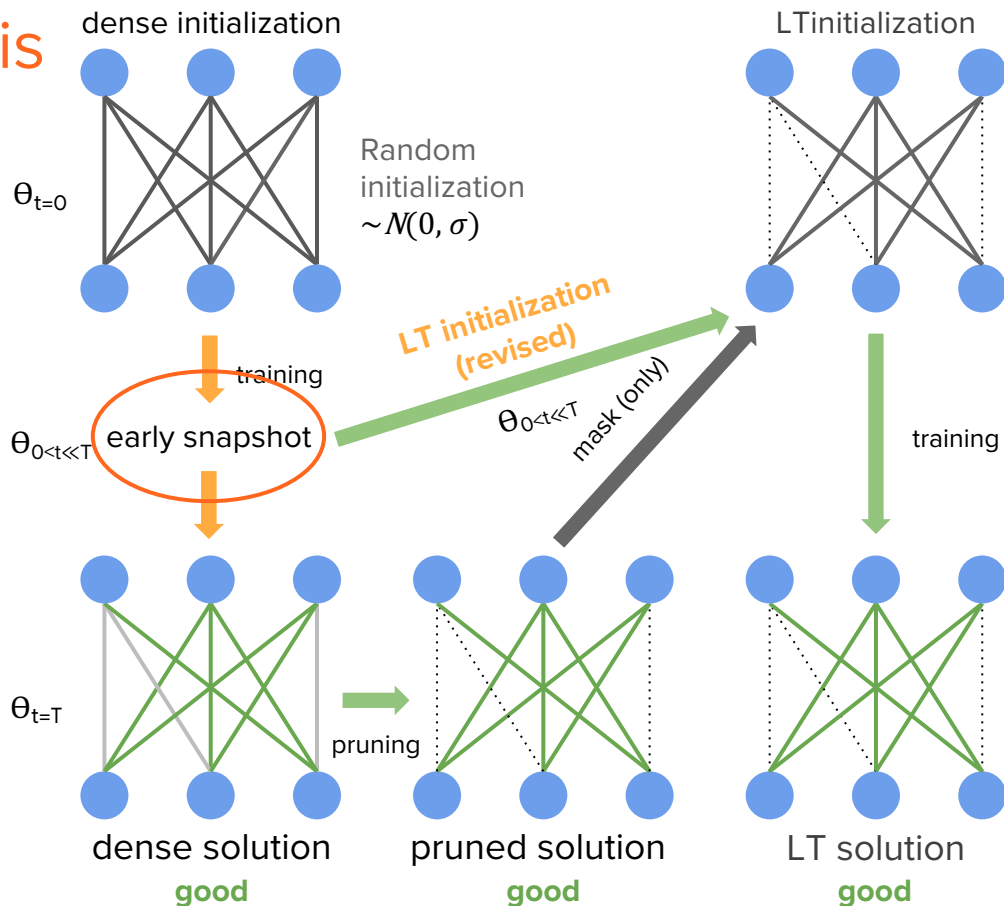
Lottery Ticket Hypothesis

- 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



Lottery Ticket Hypothesis (revised)

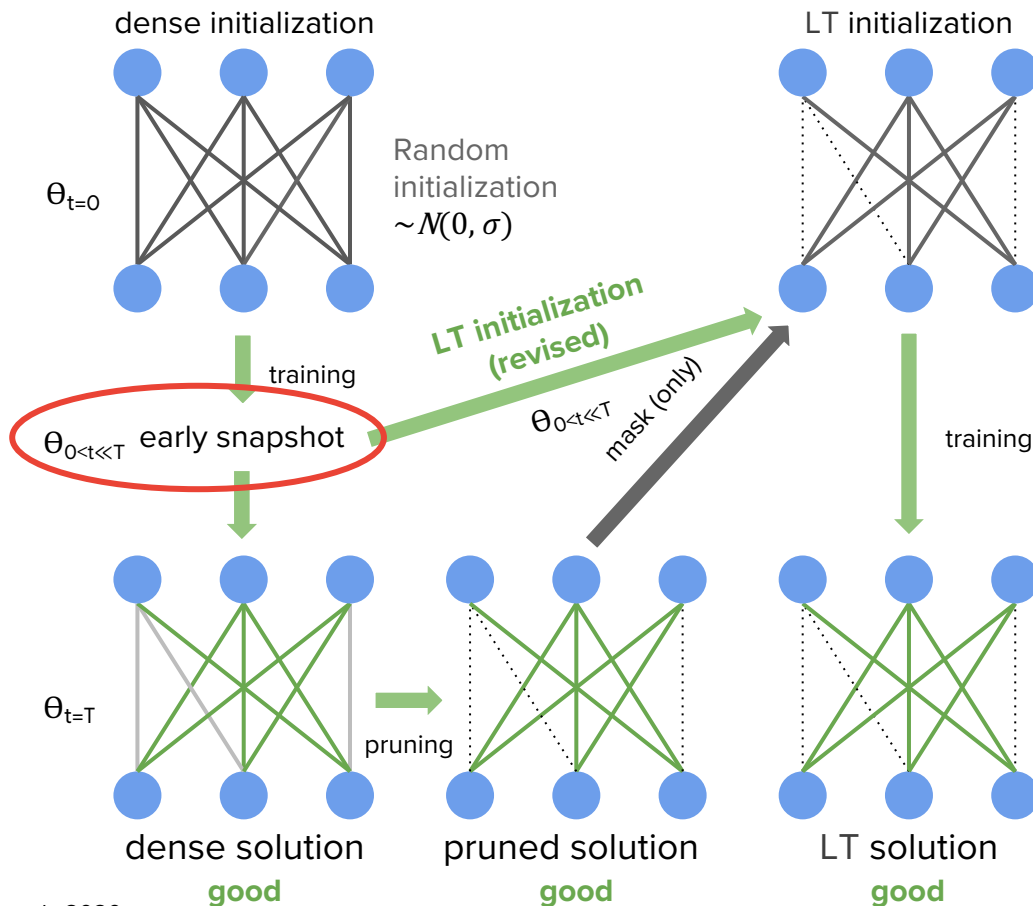
- This initialization was the original initialization the dense (pruned) model was trained from
- **LT initialization in general** is weights from early training¹
- This is **very expensive** to find



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

Lottery Tickets

- How random is the LT initialization?

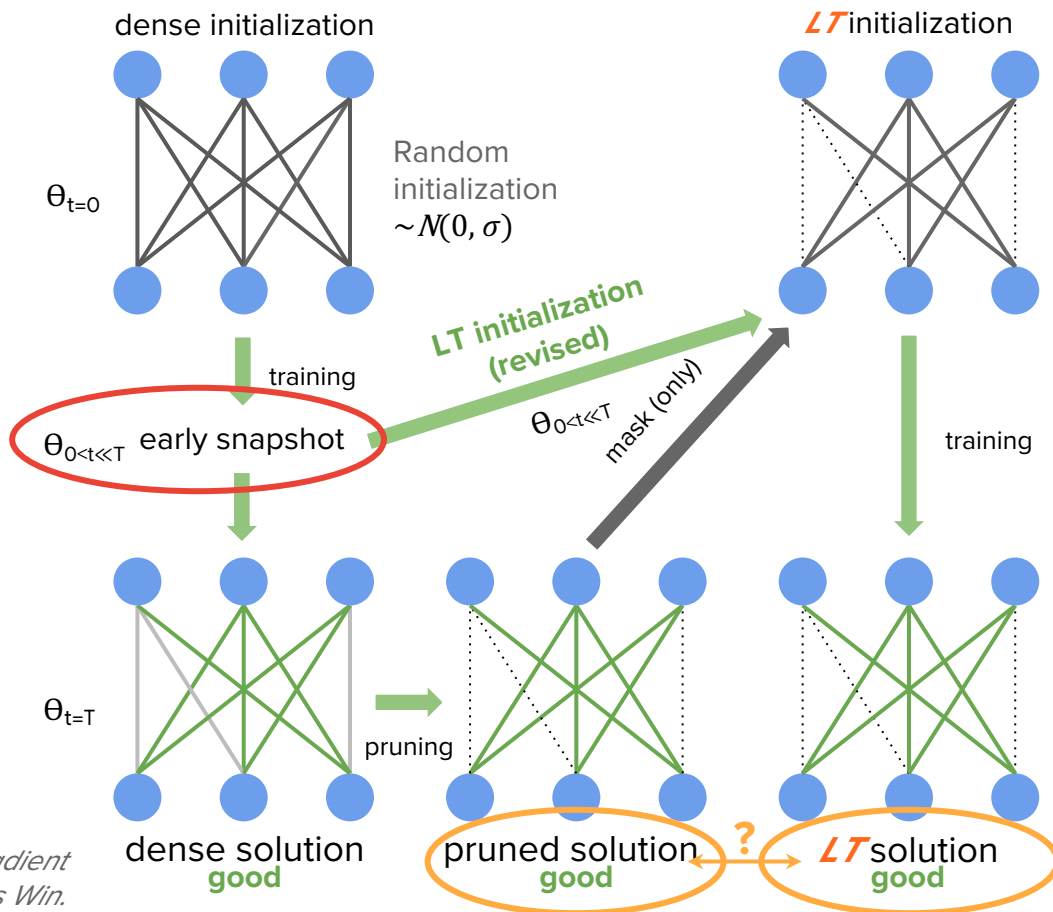


¹[Stabilizing the Lottery Ticket Hypothesis](#), Frankle et al., 2019

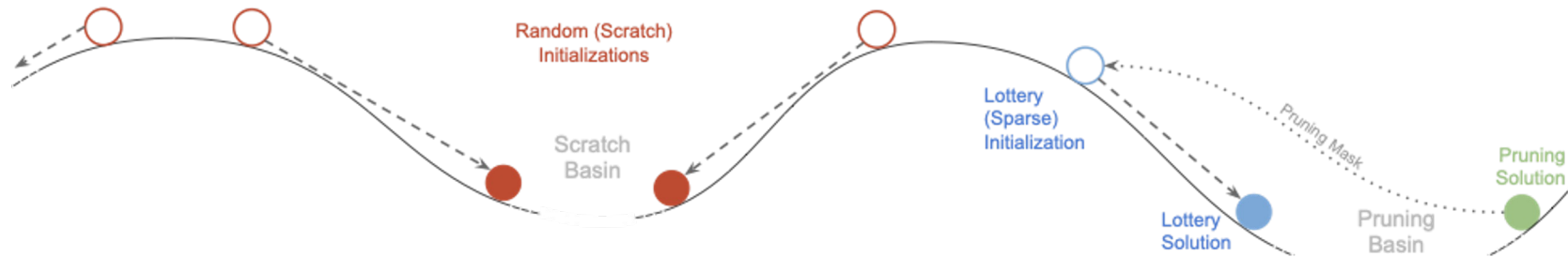
²[Linear Mode Connectivity and the Lottery Ticket Hypothesis](#), Frankle et al., 2020

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 basin¹



¹Utku 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

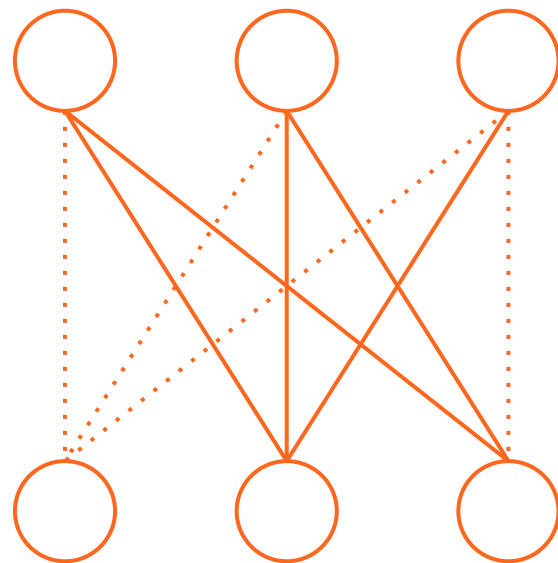
LTs appear to re-learn the pruned solution they are derived from

Sparse Training:

Aligning Sparse Masks with Weight Symmetry

1. Short Biography
2. Motivation
3. Background

4. Aligning Sparse Masks



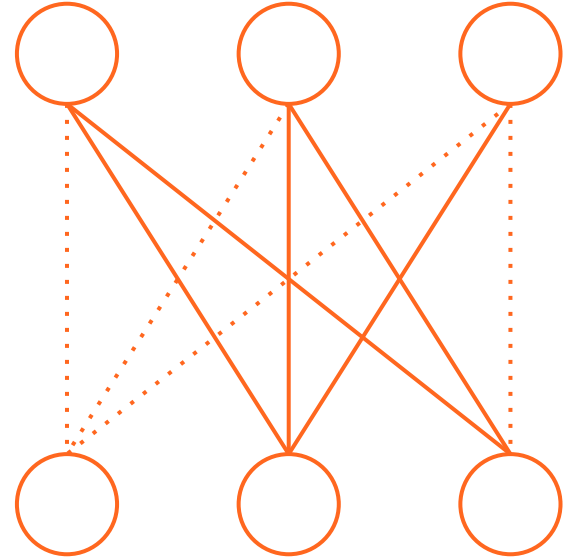
4. Aligning Sparse Masks

i. Hypothesis

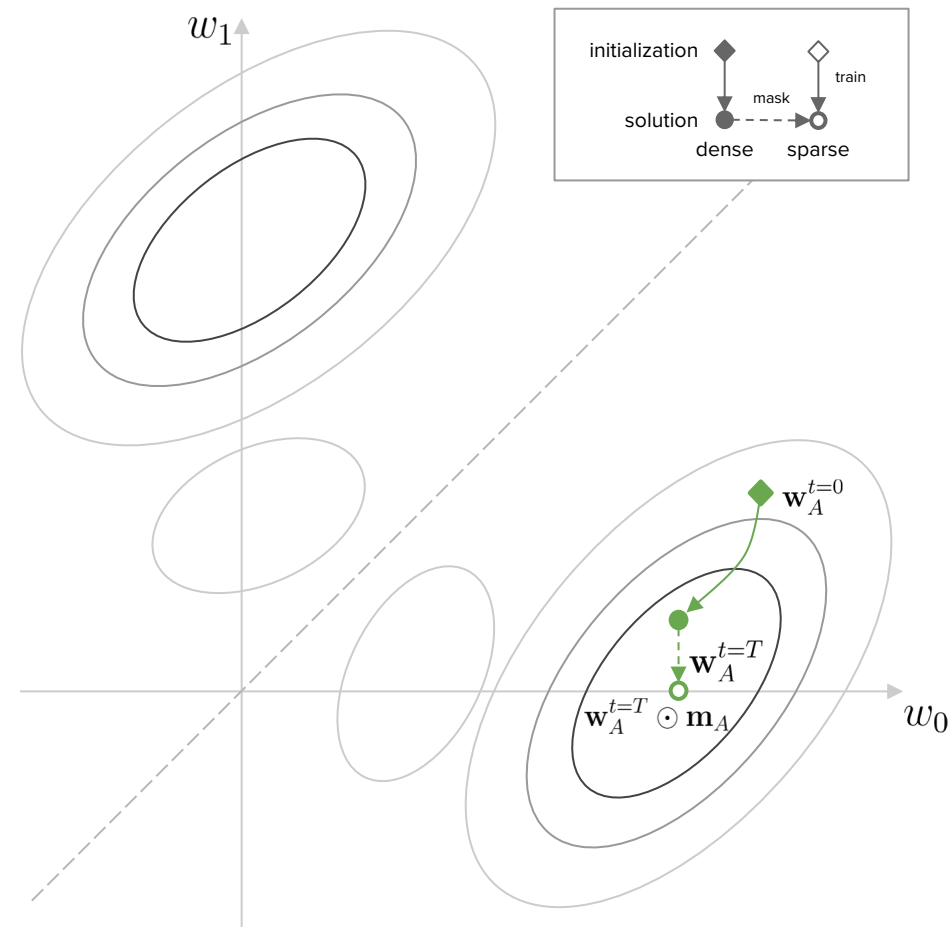
ii. Experimental Methodology

iii. Results

iv. Analysis



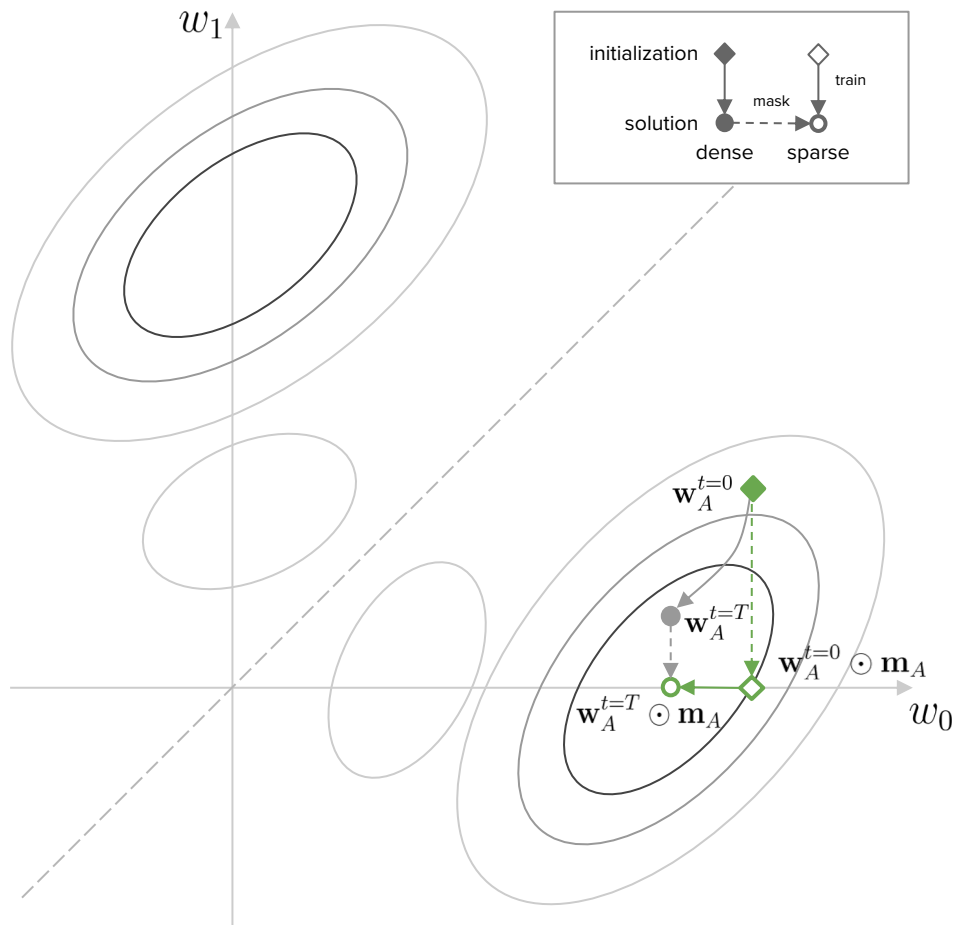
Pruning Loss Landscape



- 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.

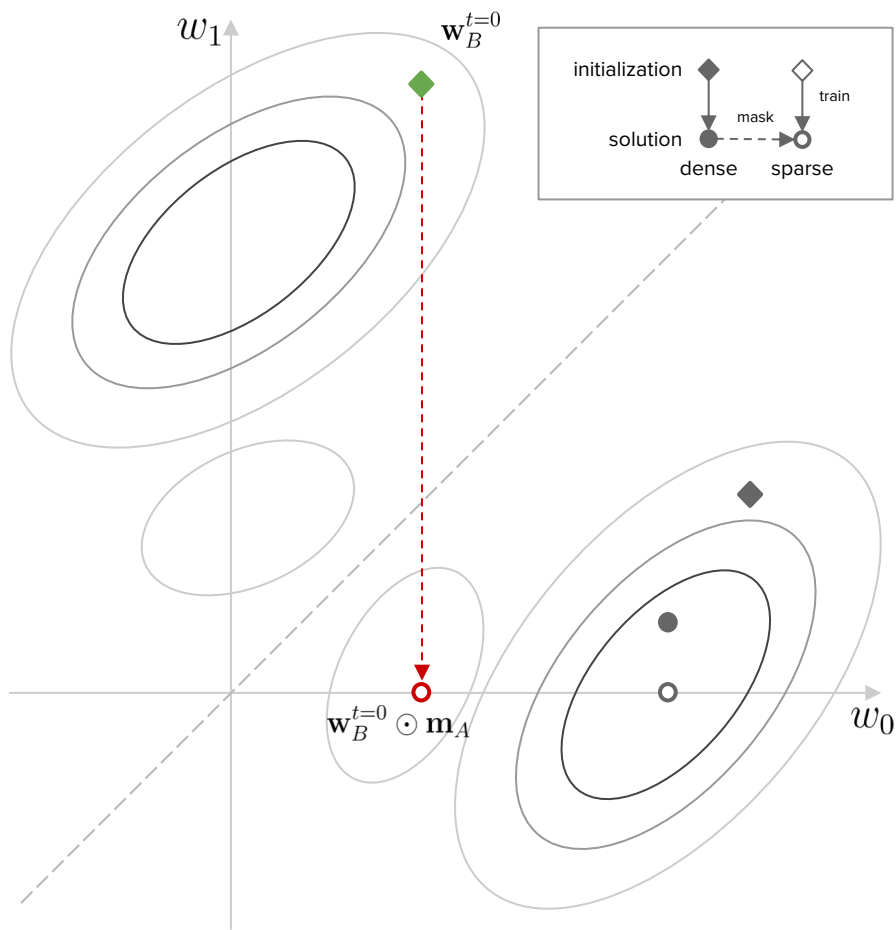
LTH Loss Landscape



- 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.

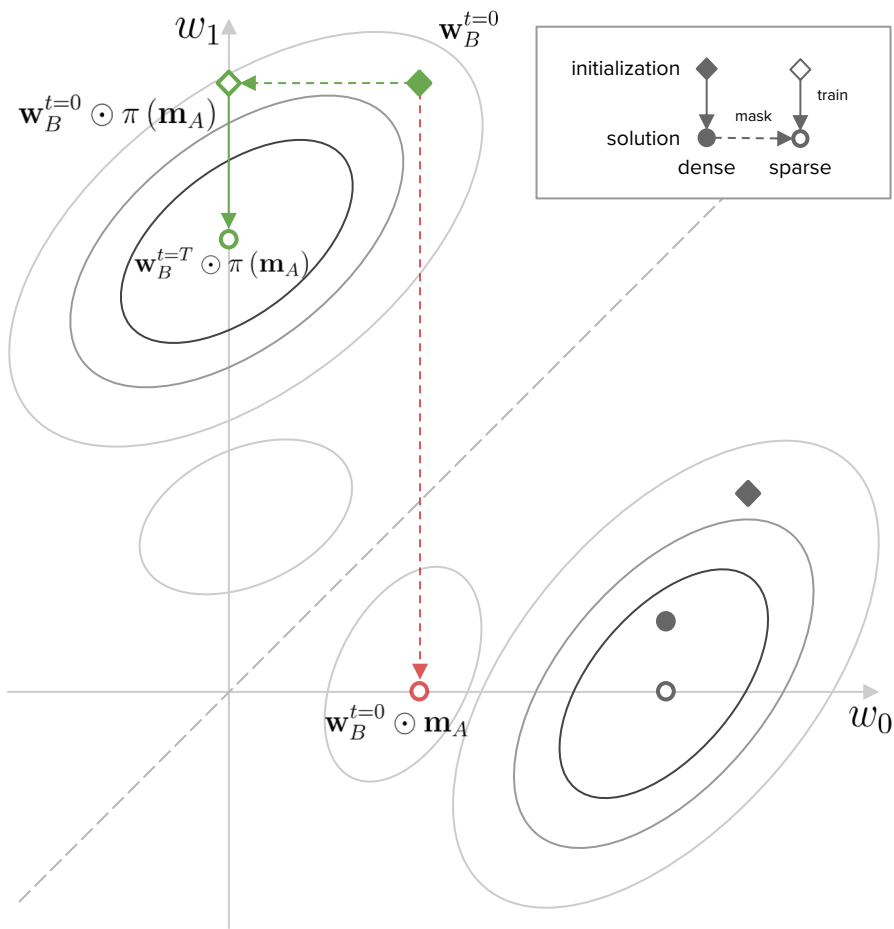
Sparse Loss Landscape



- 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.

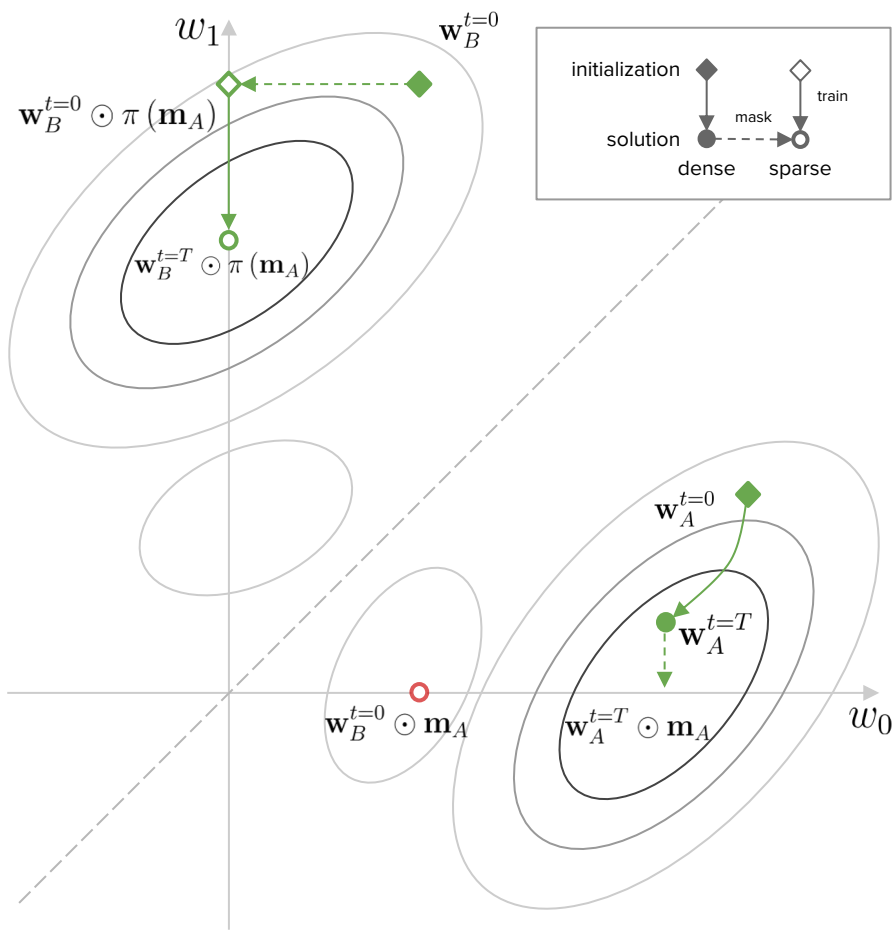
Our Hypothesis



- 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



- Recall¹: the basins of $w_A^{t=T}$ and $w_B^{t=T}$ are related by a permutation π :

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

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

$$\pi(m_A) = m_B$$

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

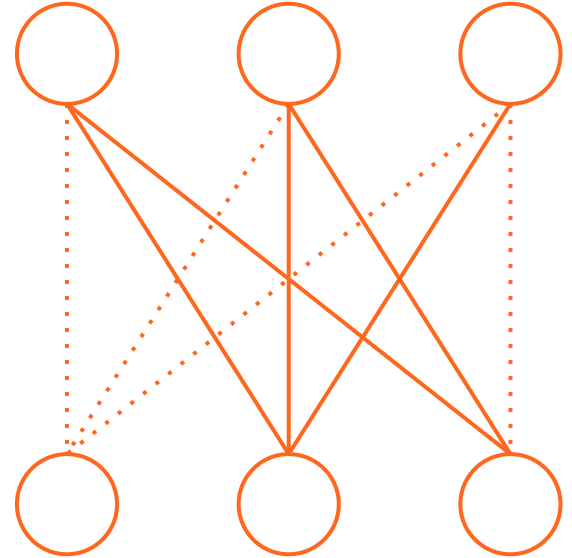
4. Aligning Sparse Masks

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

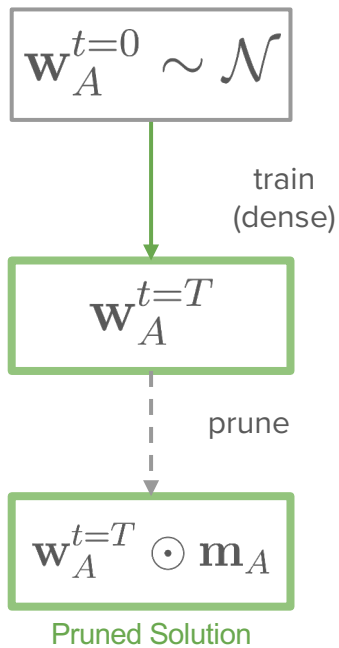
Dense Training
& Pruning

train

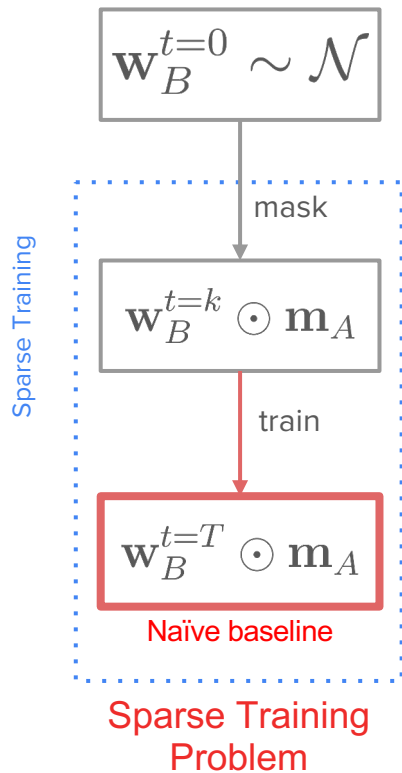
mask

match

.....

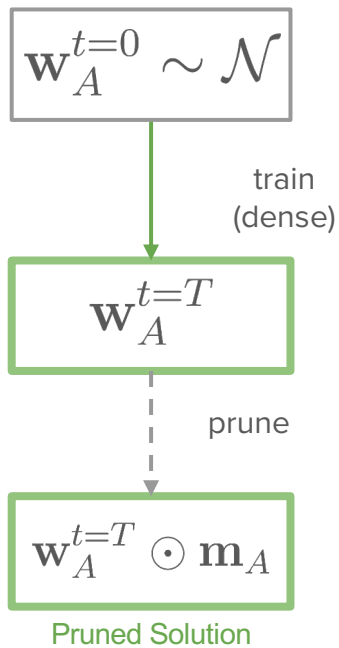


Dense Training
& Pruning



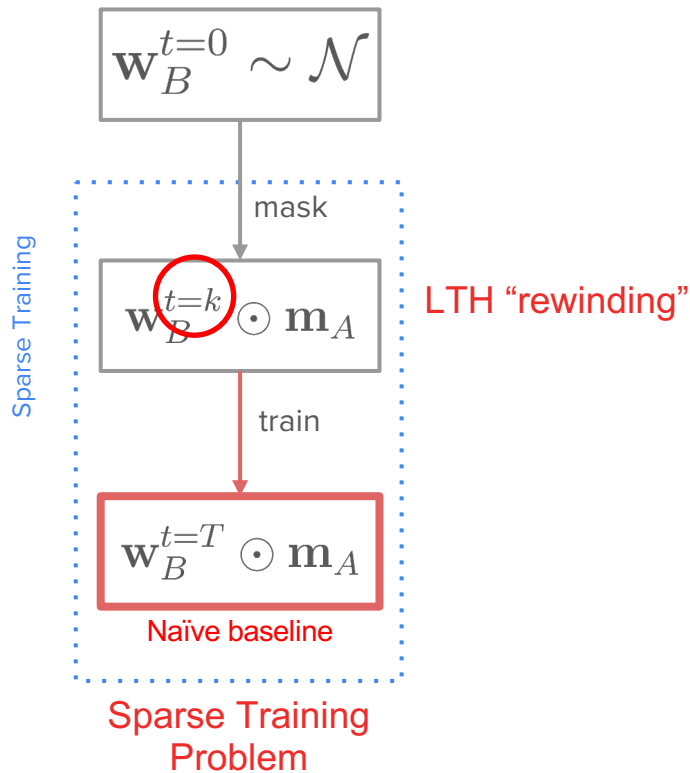
train
mask
match

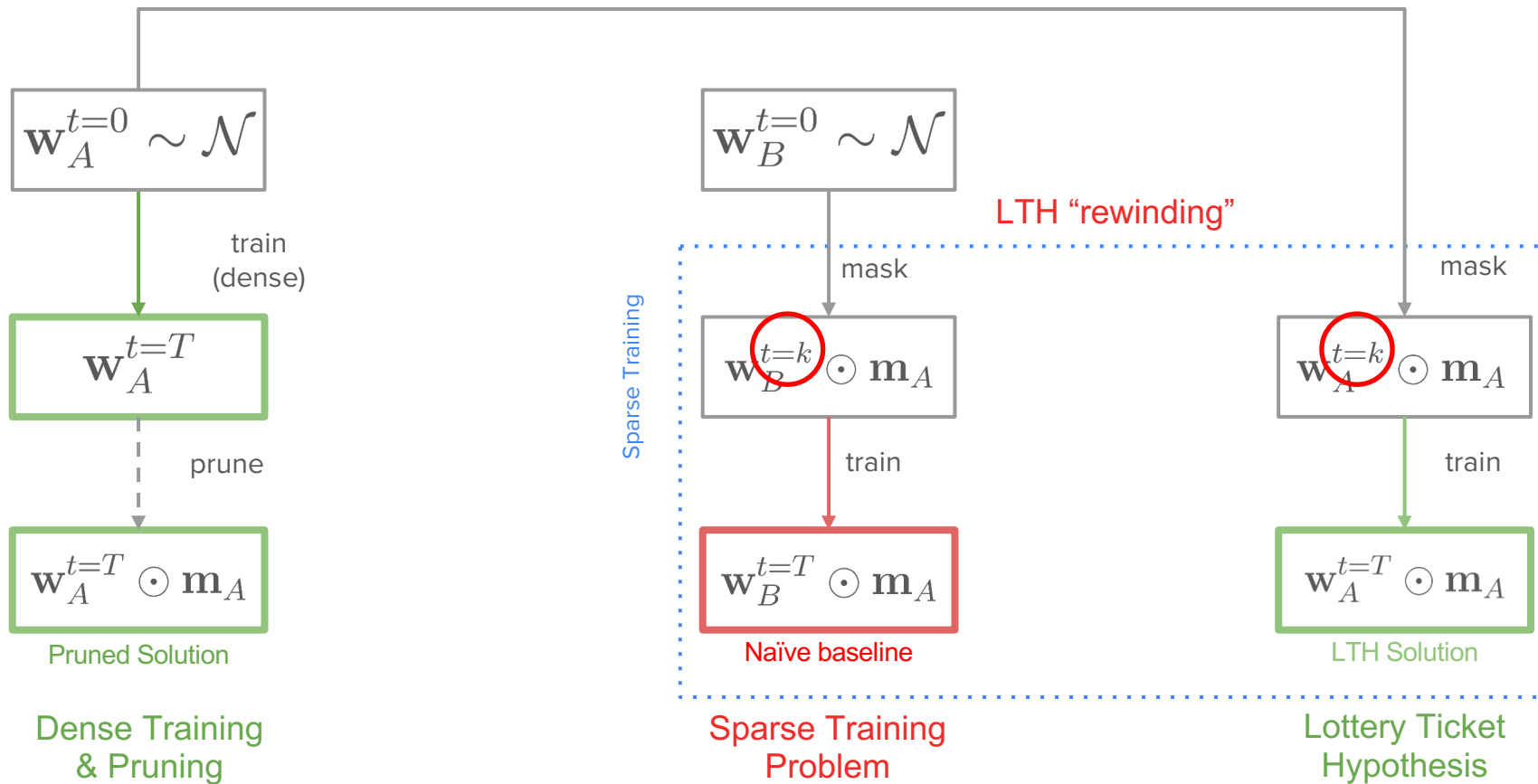




Dense Training
& Pruning

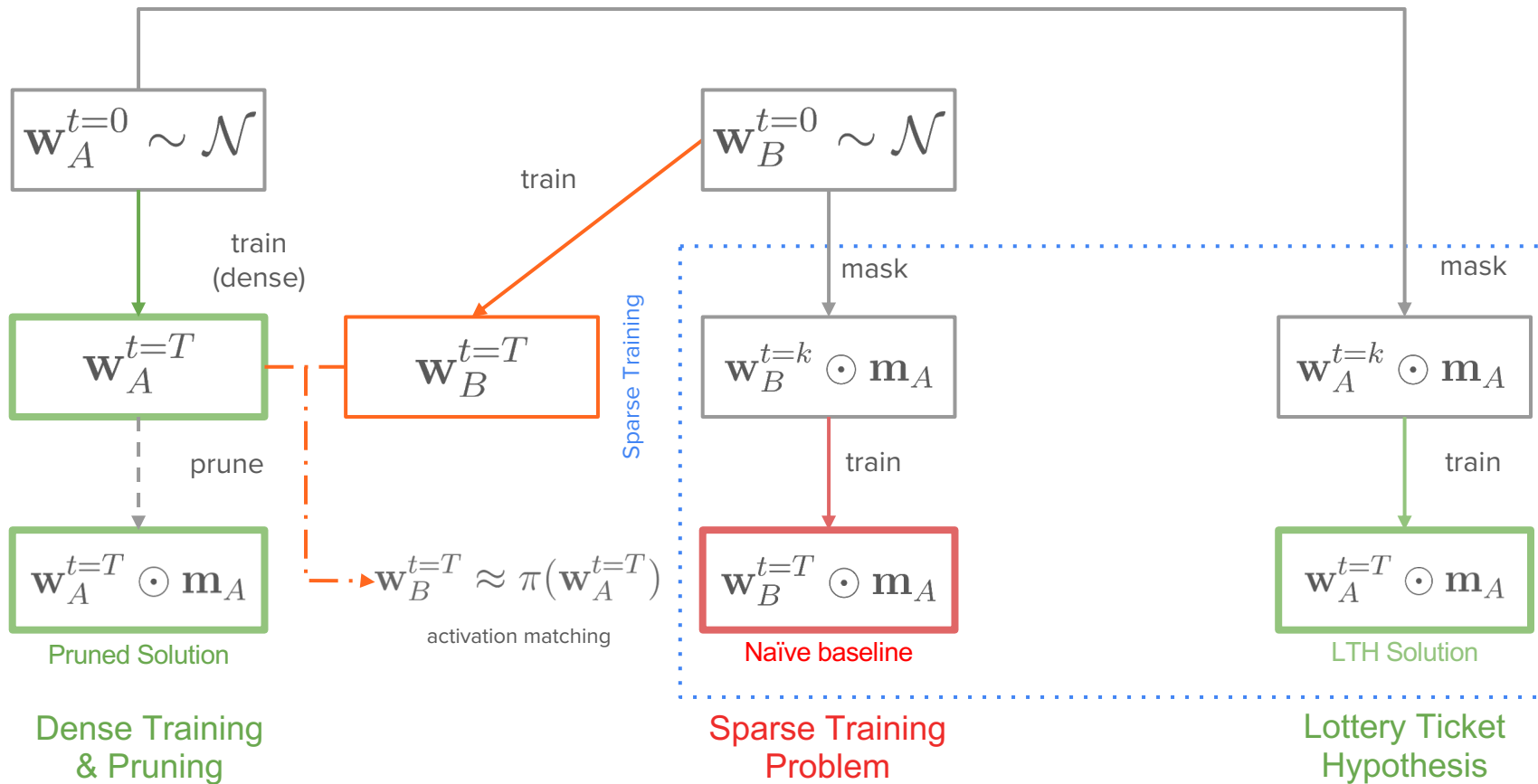
train
mask
match

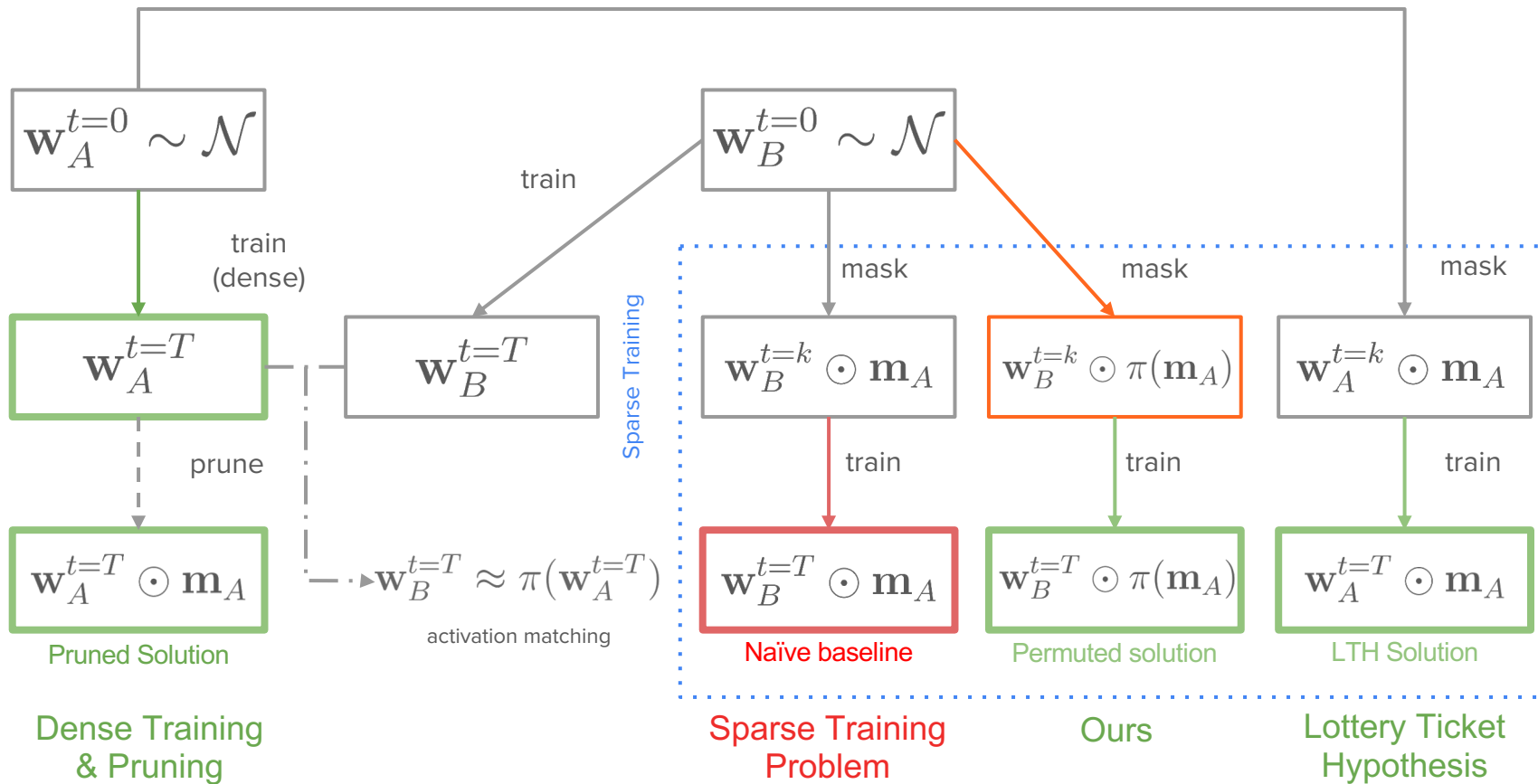


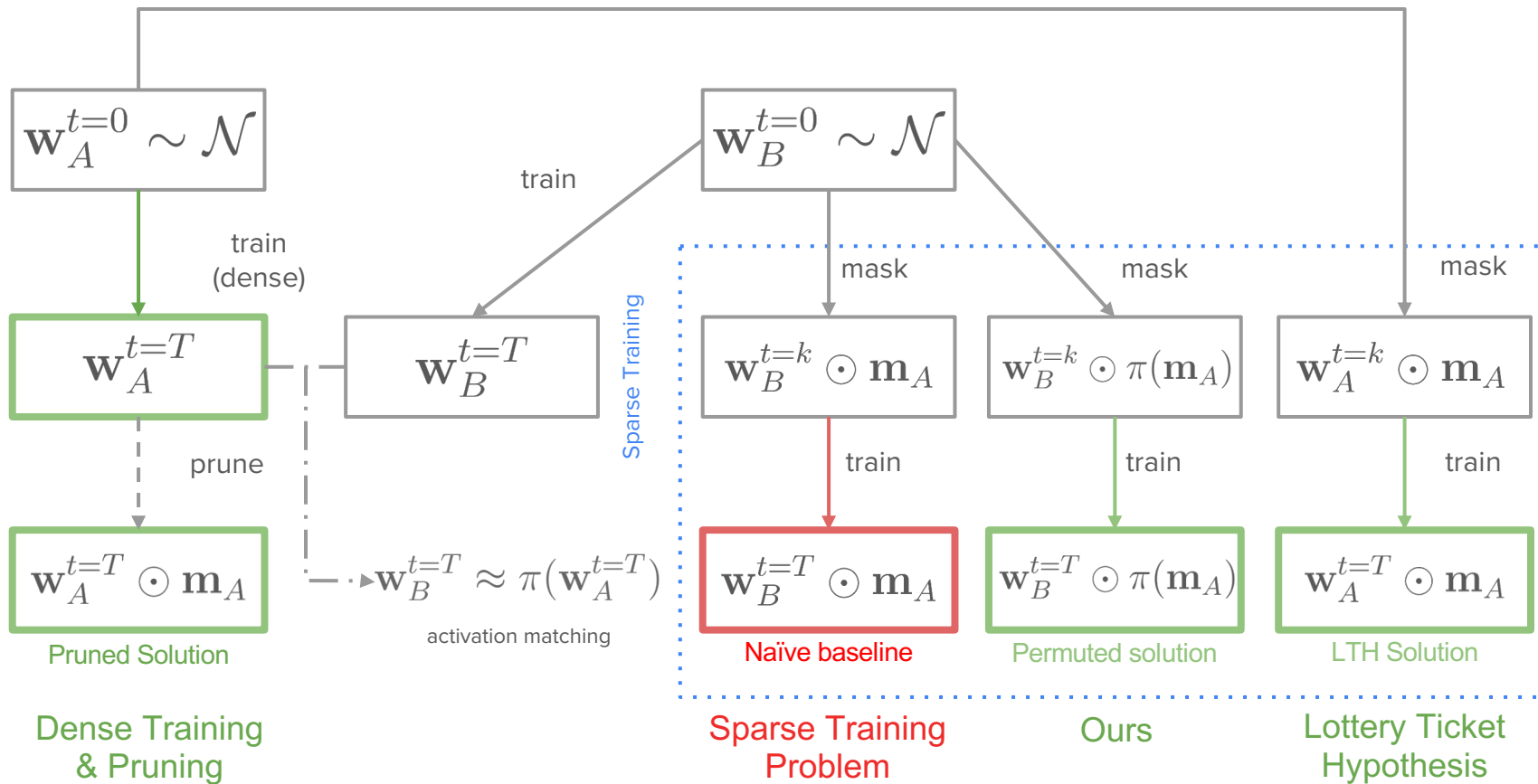


train
mask
match



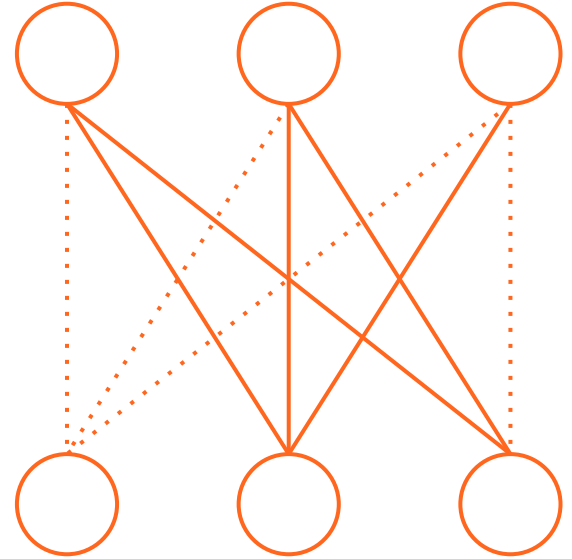


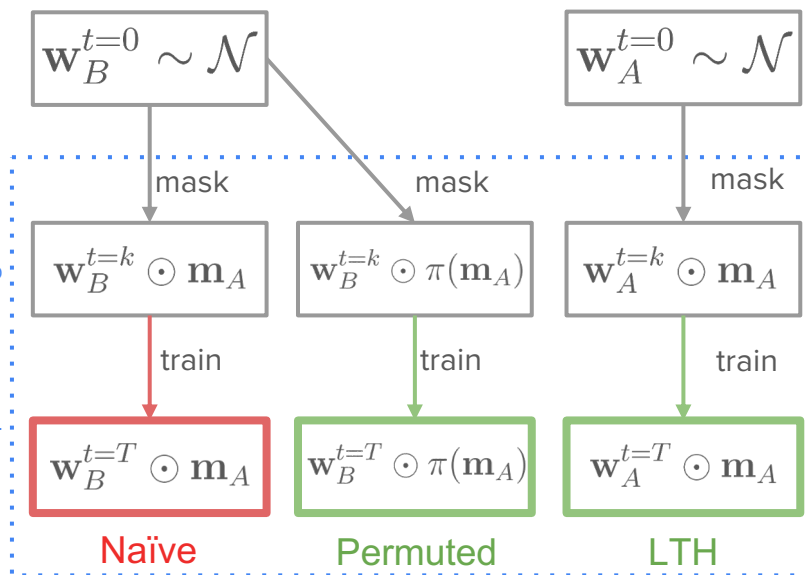




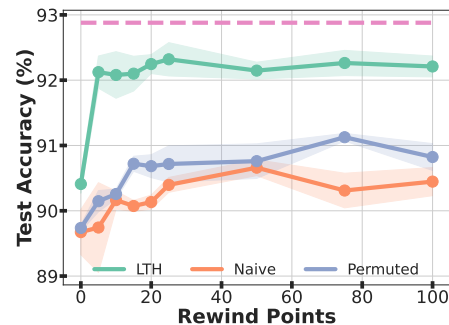
4. Aligning Sparse Masks

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

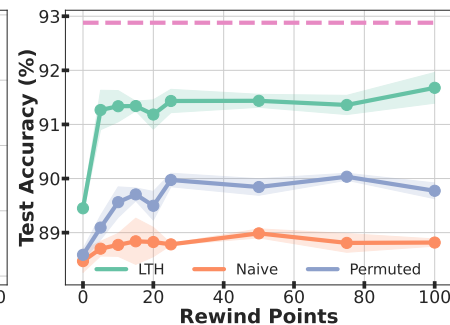




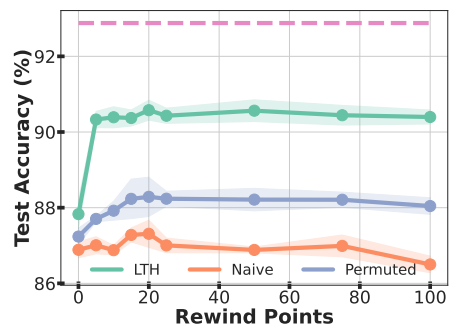
ResNet20 x {Width 1} on CIFAR-10



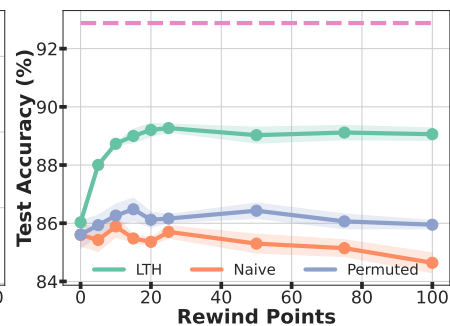
(a) sparsity = 0.80



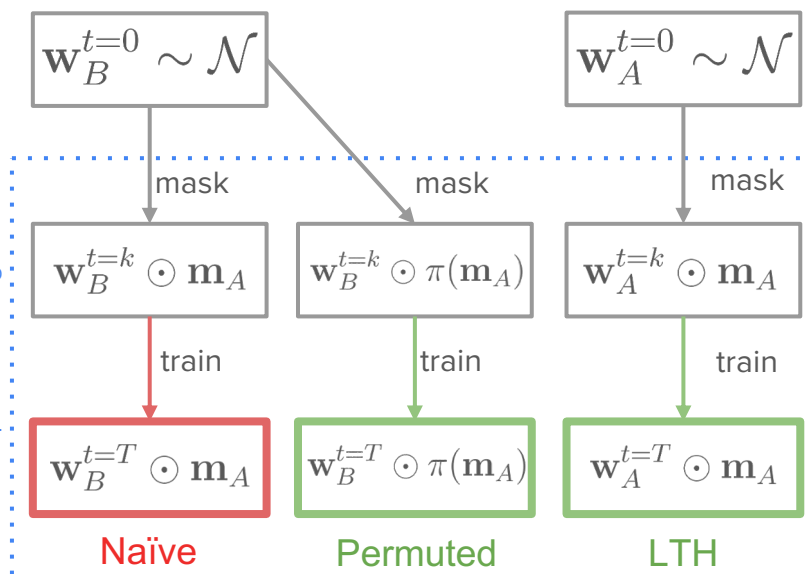
(b) sparsity = 0.90



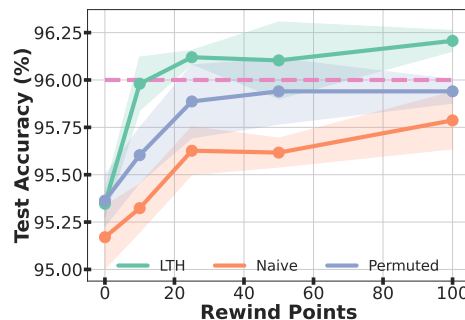
(c) sparsity = 0.95



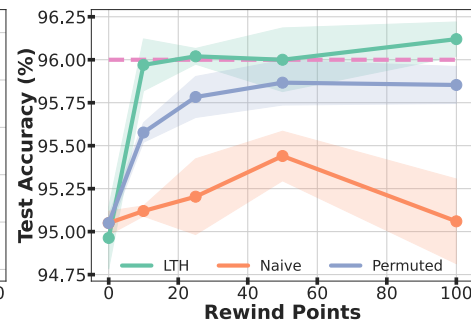
(d) sparsity = 0.97



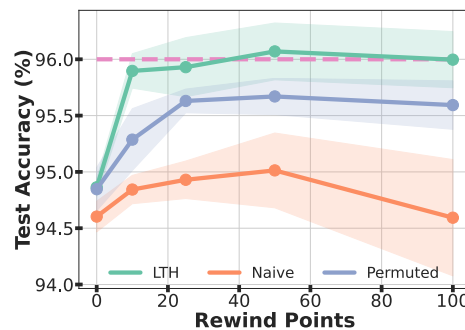
ResNet20 x {Width 8} on CIFAR-10



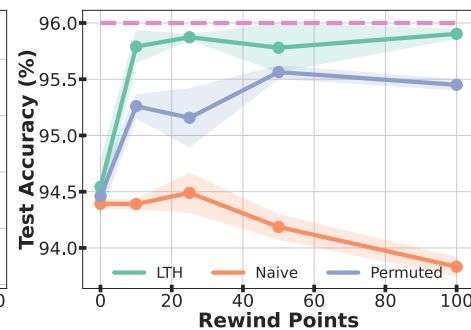
(a) sparsity = 0.80



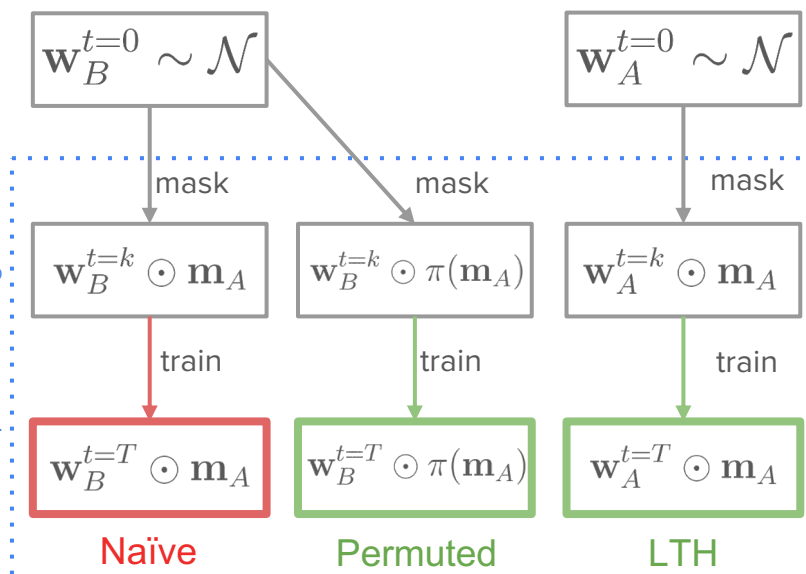
(b) sparsity = 0.90



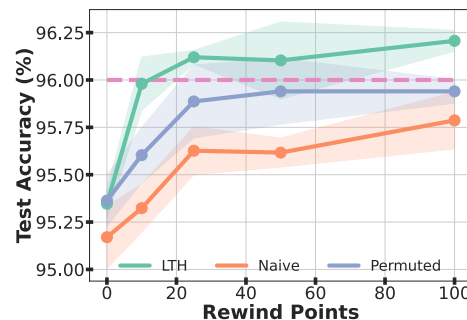
(c) sparsity = 0.95



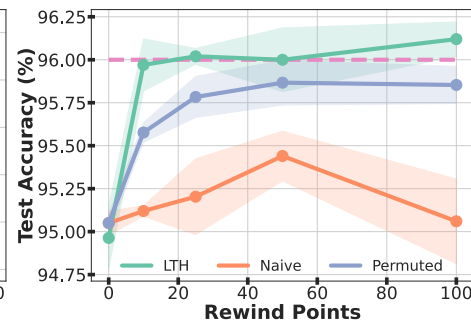
(d) sparsity = 0.97



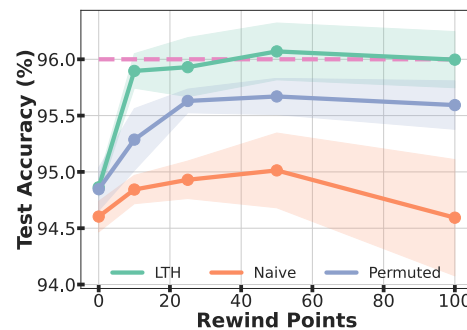
ResNet20 x {Width 8} on CIFAR-10



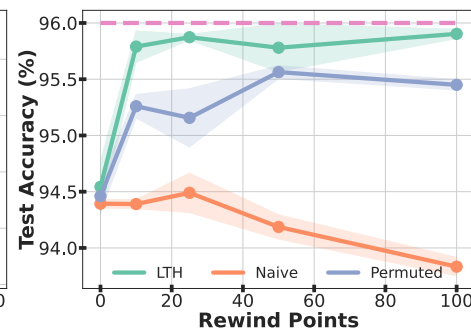
(a) sparsity = 0.80



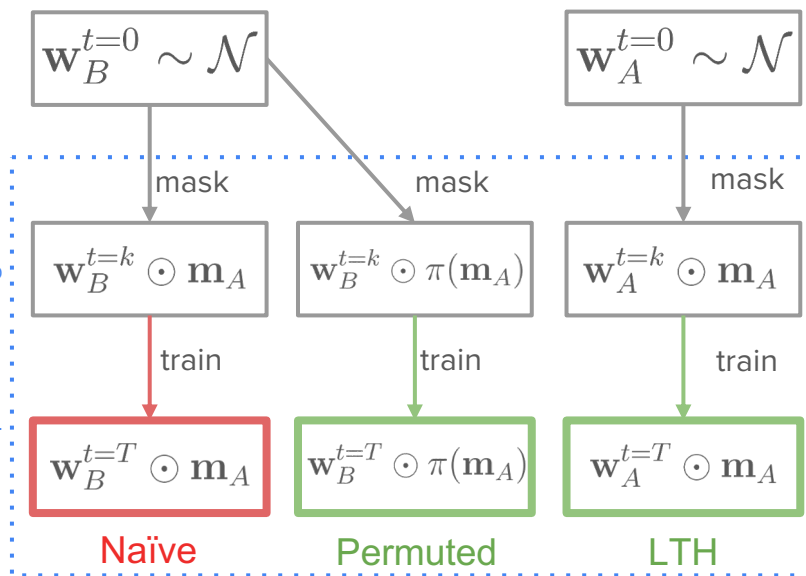
(b) sparsity = 0.90



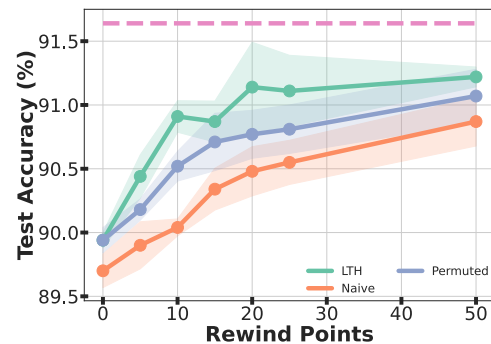
(c) sparsity = 0.95



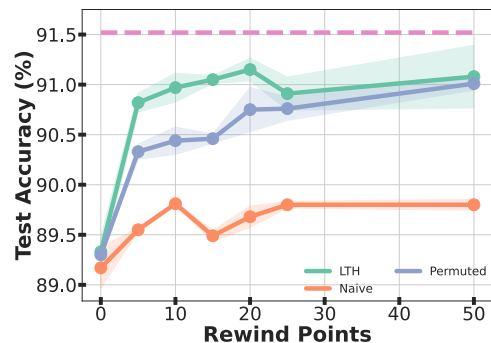
(d) sparsity = 0.97



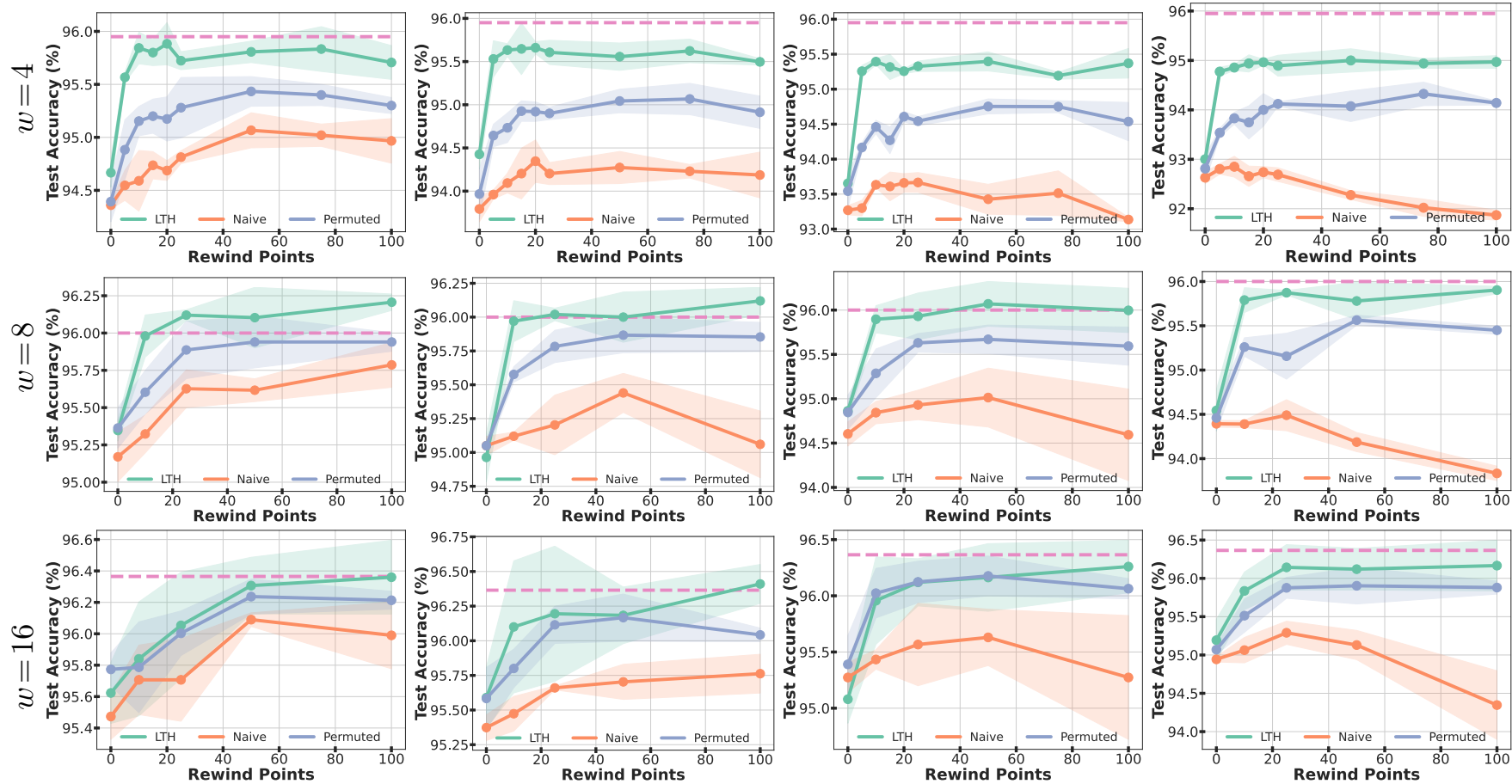
VGG11 x {Width 1} on CIFAR-10



(a) sparsity = 0.80



(b) sparsity = 0.90



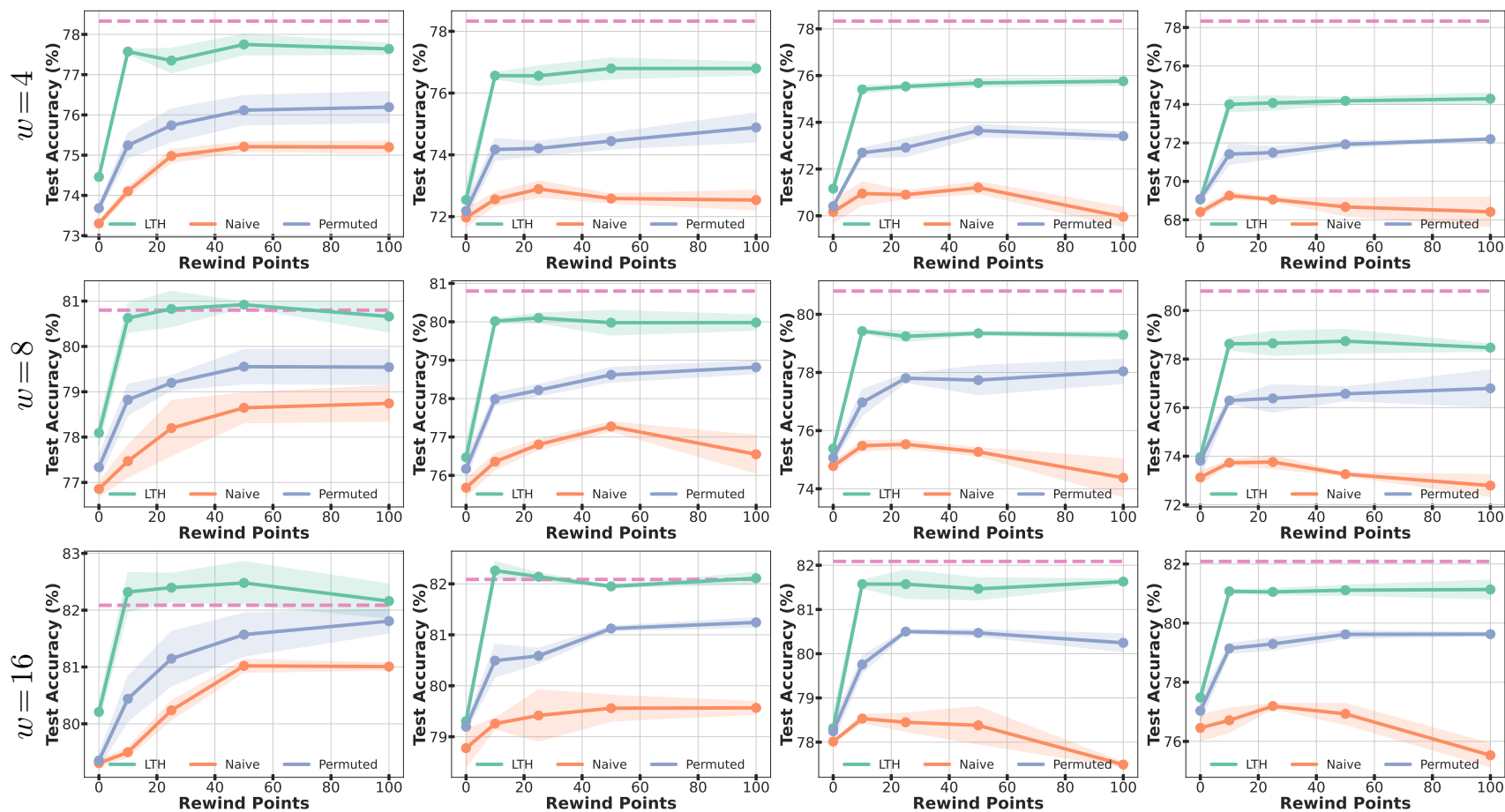
(a) sparsity = 0.80

(b) sparsity = 0.90

(c) sparsity = 0.95

(d) sparsity = 0.97

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



(a) sparsity = 0.80

(b) sparsity = 0.90

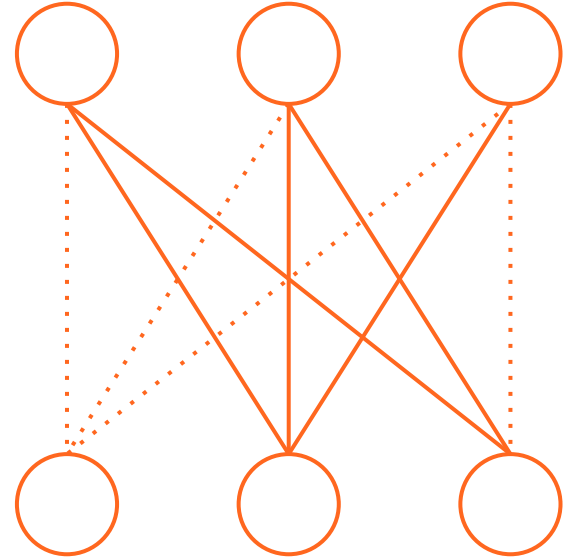
(c) sparsity = 0.95

(d) sparsity = 0.97

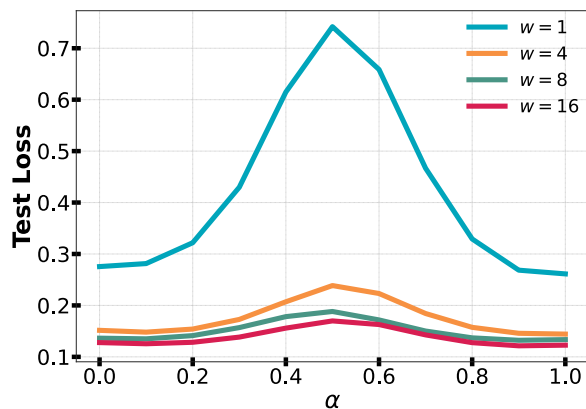
ResNet20 x {4,8,16} on CIFAR-100

4. Aligning Sparse Masks

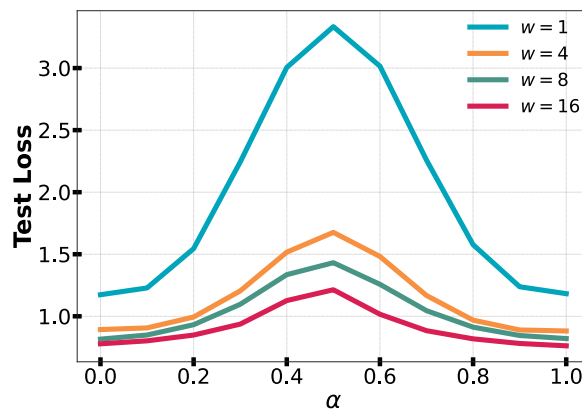
- i. Hypothesis
- ii. Experimental Methodology
- iii. Results
- iv. Analysis**



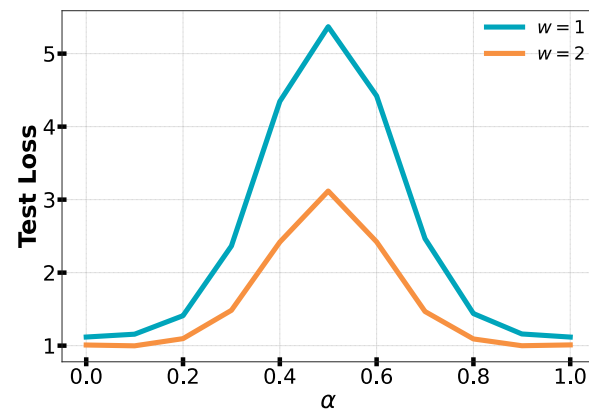
Effect of Model Width Multiplier



(a) ResNet20 $\times \{w\}$ / CIFAR-10



(b) ResNet20 $\times \{w\}$ / CIFAR-100



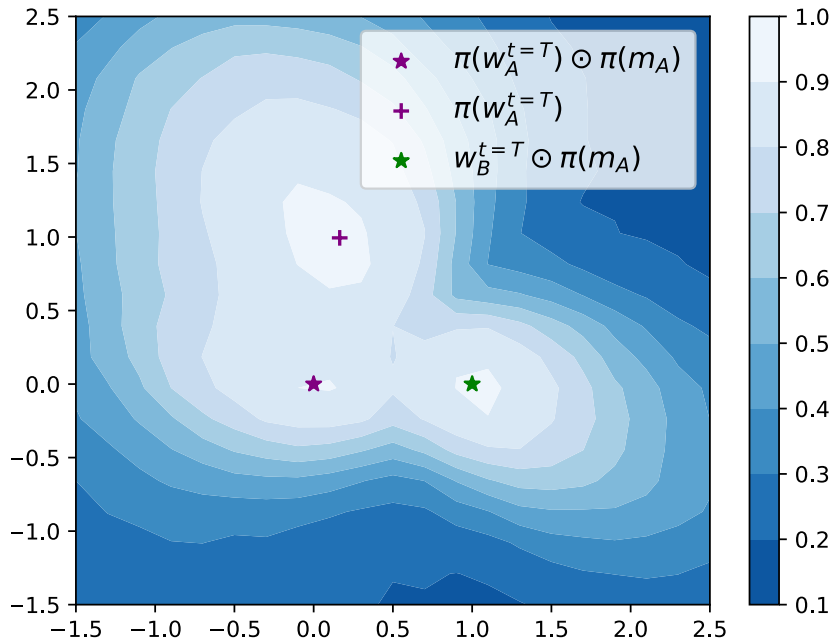
(c) ResNet50 $\times \{w\}$ / ImageNet

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

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: $\pi(w_A^{t=T}) \odot \pi(m_A)$
 - Soln. B masked by permuted mask: $w_B^{t=T} \odot \pi(m_A)$
- 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. (%)	Disagreement	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	91.15 ± 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	75.35 ± 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

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.

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



Défense nationale National Defence



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**