#### The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks

Rahim Entezari, Hanie Sedghi, Olga Saukh, Behnam Neyshabur





#### **Motivation**

#### Form an ensemble model

- In output space
- In weight space



#### **Motivation**

Form an ensemble model

- In output space
- In weight space (Embedded ML)



#### **Motivation**

Ensemble by weight averaging

requirements:

- 1. Functionally diverse solutions
- 2. Residing in one basin



#### Linear Mode Connectivity

Functionally different solutions:



Image credit: Mitchell Wortsman

#### **Linear Mode Connectivity**

Functionally different solutions:

Weight space averaging fails



Image credit: Mitchell Wortsman

#### **Linear Mode Connectivity**

Same basin:

When part of training trajectory is shared, the solutions are **linearly mode connected** (Frankle et al., 2019).



Image credit: Mitchell Wortsman

#### Stochastic Weight Averaging

Same basin:

e.g. SWA (Izmailov et al., 2018)



#### Question

Is there any way to make different solutions in one basin?

#### Conjecture

A, B, C, and D are minimas in different basins with barriers between pairs.



#### Conjecture

Taking permutations into account, there is likely no barrier in the linear interpolation between SGD solutions.



## Part 1: Observations over loss landscape shape

#### Barrier

$$B(\theta_1, \theta_2) = \sup_{\alpha} \left[ \left[ \mathcal{L}(\alpha \theta_1 + (1 - \alpha) \theta_2) \right] - \left[ \alpha \mathcal{L}(\theta_1) + (1 - \alpha) \mathcal{L}(\theta_2) \right] \right]$$

#### Effect of Width on barrier size

As the width increases, the barrier first increases and then decreases



#### **Deep Double Descent in Barrier**



#### Effect of Width on barrier size

As the width increases, the barrier first increases and then decreases



#### Effect of Width on barrier size:

Barrier saturates at high level in deeper models



#### Effect of **Depth** on barrier size

Low barrier when number of layers are low

Fast and significant barrier increase as more layers are added



#### Effect of **Depth** on barrier size

Low barrier when number of layers are low

Fast and significant barrier increase as more layers are added



#### Effect of **Depth** on barrier size



#### Effect of Task Complexity on barrier size

(architecture, task) has lower barrier if

the test error is lower



#### Effect of Task Complexity on barrier size

Effect of depth is stronger than (architecture, task) which leads to high barrier values for deep nets



# Part 2: evidences to support conjecture

#### Conjecture: recall

 $m_1, m_2$  are trained and converged



#### Conjecture: recall

 $\rm m^{}_{1}, \rm m^{}_{2}$  are trained and converged

There exists a permutation applied to  $m_2$ , making  $m_1$  and  $m_2$ ' Linearly Mode Connected.



#### Permutation



#### **Real World**



We train networks by running SGD with different random seeds and different initialization.

#### **Our Model**



Different final networks are obtained by applying random permutations to the same SGD solution.



• Our model satisfies the conjecture



- Our model satisfies the conjecture
- We show that Real world ~ Our model

Similar loss barrier between real world and our model **BEFORE** permutation search



Similar loss barrier between real world and our model **BEFORE** permutation search, across all datasets, architectures, width, and depth









#### Permutation Search: Simulated Annealing

Algorithm 1 Simulated Annealing (SA) for Permutation Search

1: procedure 
$$SA(\{\theta_i\}, i = 1..n, n \ge 2)$$
   
2:  $\pi_i = \pi_0, \forall i = 1..n$   
3: for  $k = 0; k < k_{max}; k++$  do  
4:  $T \leftarrow temperature(\frac{k+1}{k_{max}})$   
5: Pick random candidate permutations  $\{\hat{\pi}_i\}, \forall i = 1..n$   
6: if  $\Psi(P(\theta_i, \hat{\pi}_i)) < \Psi(P(\theta_i, \pi_i))$  then   
7:  $\pi_i \leftarrow \hat{\pi}_i$   
return  $\{\pi_i\}$ 

Simulated Annealing: Performance



Simulated Annealing: Performance



Similar loss barrier between real world and our model **AFTER** permutation search









• We show that Real world ~ Our model

#### Takeaways

- One way to form ensembles is to weight average solutions
- Conjecture: we can make different SGD solutions in one basin using permutations
- Our theoretical results + extensive experiments fall short of refuting our bold conjecture.

## Thanks

Code: https://github.com/rahimentezari/PermutationInvariance



entezari@tugraz.at

## Improve Search Algorithm:

Functional Difference

$$\delta E_l^{opt} = \frac{1}{2} (\tilde{\mathbf{w}}_{l,i}^A - \tilde{\mathbf{w}}_{l,j}^B)^\top \cdot \left( (\tilde{\mathbf{H}}_{l,i}^A)^{-1} + (\tilde{\mathbf{H}}_{l,j}^B)^{-1} \right)^{-1} \cdot (\tilde{\mathbf{w}}_{l,i}^A - \tilde{\mathbf{w}}_{l,j}^B)$$

He, Xiaoxi, Zimu Zhou, and Lothar Thiele. "Multi-task zipping via layer-wise neuron sharing." *arXiv* preprint arXiv:1805.09791 (2018).

Improve Search Algorithm:

FD vs. SA



He, Xiaoxi, Zimu Zhou, and Lothar Thiele. "Multi-task zipping via layer-wise neuron sharing." *arXiv* preprint arXiv:1805.09791 (2018).