Going Beyond Linear Mode Connectivity: The Layerwise Linear Feature Connectivity

Zhanpeng Zhou[1], Yongyi Yang[2], Xiaojiang Yang[1], Junchi Yan[1*], Wei Hu[2]

[1]Shanghai Jiao Tong University, [2]University of Michigan

*Corresponding author

Background: Linear Mode Connectivity

Linear Mode Connectivity (LMC)

Given dataset D and two modes $\boldsymbol{\theta}_A$, $\boldsymbol{\theta}_B$ that $\mathrm{Err}_D(\boldsymbol{\theta}_A) = \mathrm{Err}_D(\boldsymbol{\theta}_B)^*$, two mode $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ satisfy the *linear mode connectivity* if

$$\forall \alpha \in [0,1], \mathrm{Err}_D(\alpha\boldsymbol{\theta}_A + (1-\alpha)\boldsymbol{\theta}_B) \approx \mathrm{Err}_D(\boldsymbol{\theta}_A)$$

*$\mathrm{Err}_D(\boldsymbol{\theta})$ denotes the classification error of the network $f(\boldsymbol{\theta};\cdot)$ on the dataset D.

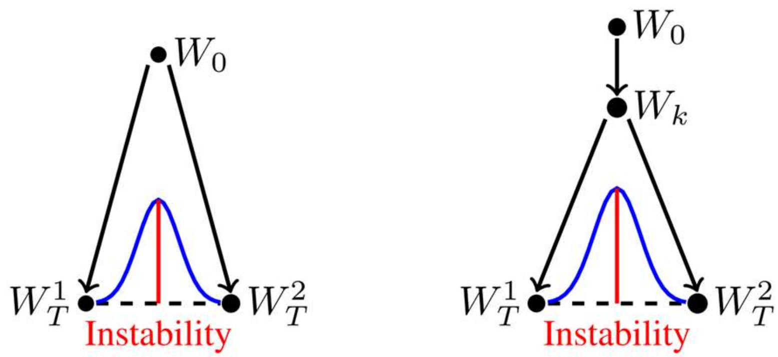


Fig. 1: Illustration of spawning method and LMC [1].

Frankle et al. [1] observed LMC for networks that are jointly trained for a short time before independent training (**spawning method**).

[1] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis.

Background: Permutation Method

Permutation Invariance.

Given an L-layer MLP f, we can permute the neurons of the MLP in each layer $\ell \in [L]$ without changing its functionality ($\pi = \{\boldsymbol{P}^{(\ell)}\}_{\ell\in[L]}$ are permutation matrices*):

$$f(\boldsymbol{\theta};\cdot) = f(\boldsymbol{\theta}';\cdot), \text{where } \boldsymbol{\theta} = \{\boldsymbol{W}^{(\ell)}\}_{\ell\in[\boldsymbol{L}]}, \boldsymbol{\theta}' = \{\boldsymbol{W}'^{(\ell)}\}_{\ell\in[\boldsymbol{L}]}$$

$$\forall \ell \in [L], \boldsymbol{W}'^{(\ell)} = \boldsymbol{P}^{(\ell)}\boldsymbol{W}^{(\ell)}, \boldsymbol{b}^{(\ell)} = \boldsymbol{P}^{(\ell)}\boldsymbol{b}^{(\ell)}, \boldsymbol{W}'^{(\ell+1)} = \boldsymbol{W}^{(\ell+1)}\boldsymbol{P}^{(\ell)}$$

*Note that $P^{(0)}$ and $P^{(L)}$ are all fixed to be identity matrix.

Independently trained networks can be *linearly connected* when considering *permutation invariance* (**permutation methods**)[2, 3].

[2] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks.
[3] Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries.

Background: Permutation Method

Ainsworth et al. [3] proposed *weight matching* and *activation matching* to achieve LMC:

$$\text{weight matching}^{*}: \min_{\pi} \sum_{\ell=1}^{L} \left\| \boldsymbol{W}_A^{(\ell)} - \boldsymbol{P}^{(\ell)} \boldsymbol{W}_B^{(\ell)} \boldsymbol{P}^{(\ell-1)^{\mathsf{T}}} \right\|_F^2$$

$$\text{Activation matching}^{*}: \min_{\pi} \sum_{\ell=1}^{L} \left\| \boldsymbol{H}_A^{(\ell)} - \boldsymbol{P}^{(\ell)} \boldsymbol{H}_B^{(\ell)} \right\|_F^2$$

*We denote ℓ-th layer feature as $\boldsymbol{H}^{(\ell)}$ over the dataset D. Subscript $\{A, B\}$ corresponds to modes $\boldsymbol{\theta}_A, \boldsymbol{\theta}_B$.

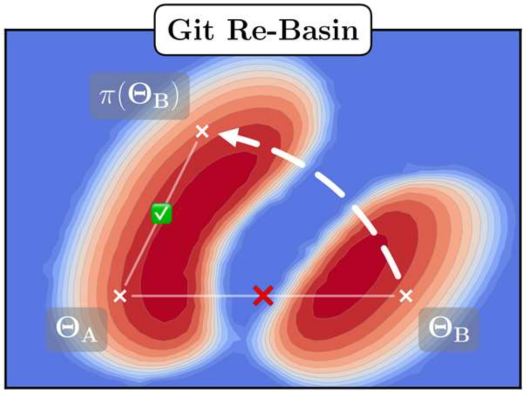
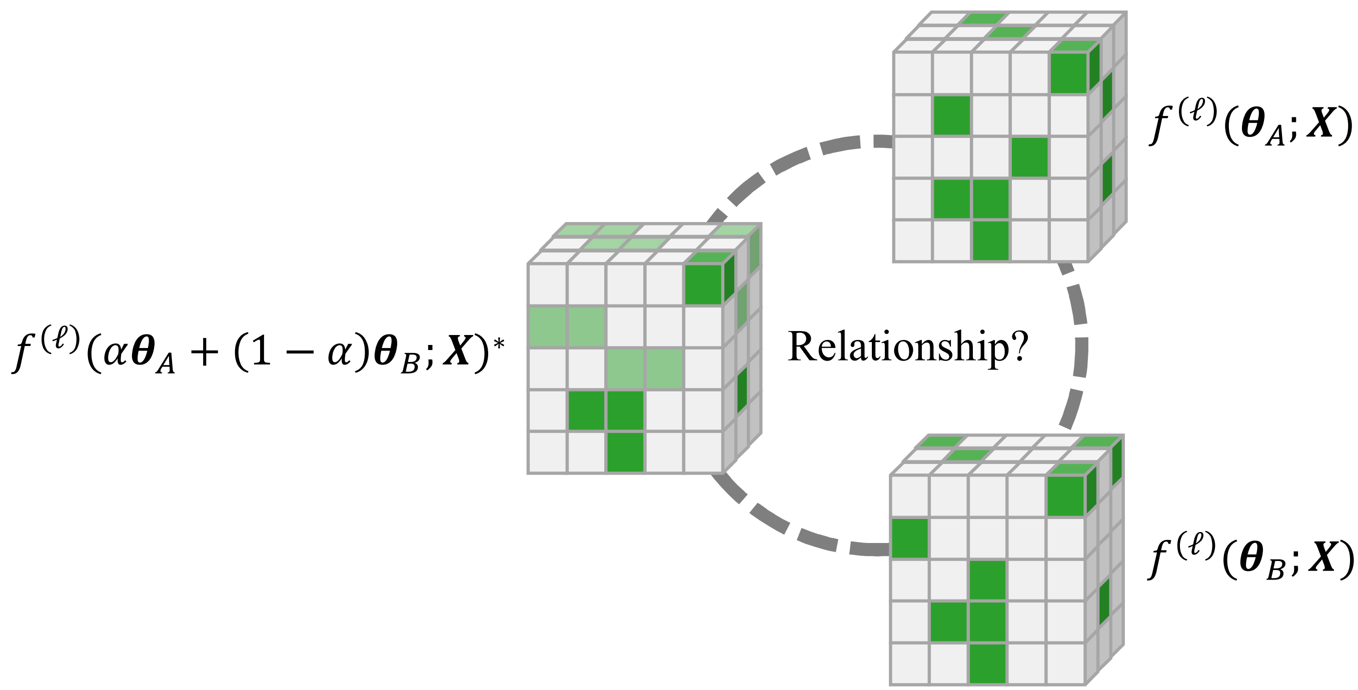


Fig. 2: Illustration of permutation [2].

[3] Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries.

Motivation



what happens to the internal features when we linearly interpolate the weights of two trained networks?

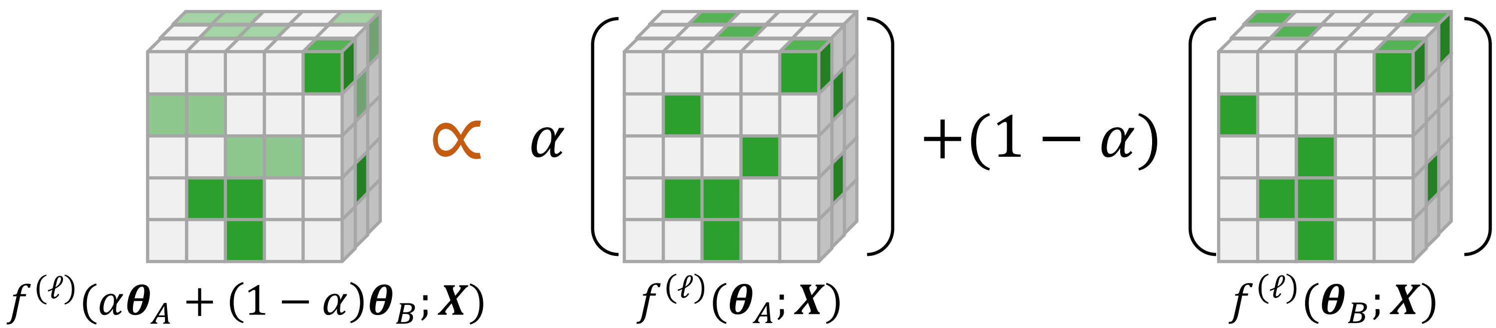
$^*f^{(\ell)}(\boldsymbol{\theta})$ denotes ℓ-th layer feature of the network $f(\boldsymbol{\theta};\cdot)$ over the dataset D.

Layerwise Linear Feature Connectivity

Layerwise Linear Feature Connectivity (LLFC)

Given dataset D and two modes $\boldsymbol{\theta}_A$, $\boldsymbol{\theta}_B$ of an L-layer neural network f, the modes $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ are *layerwise linearly feature connected* if:

$$\forall \ell \in [L], \forall \alpha \in [0,1], \exists c > 0, s.t., cf^{(\ell)}(\alpha\boldsymbol{\theta}_A + (1-\alpha)\boldsymbol{\theta}_B) = \alpha f^{(\ell)}(\boldsymbol{\theta}_A) + (1-\alpha) f^{(\ell)}(\boldsymbol{\theta}_B).$$



Layerwise Linear Feature Connectivity

LLFC always co-occurs with LMC in practice

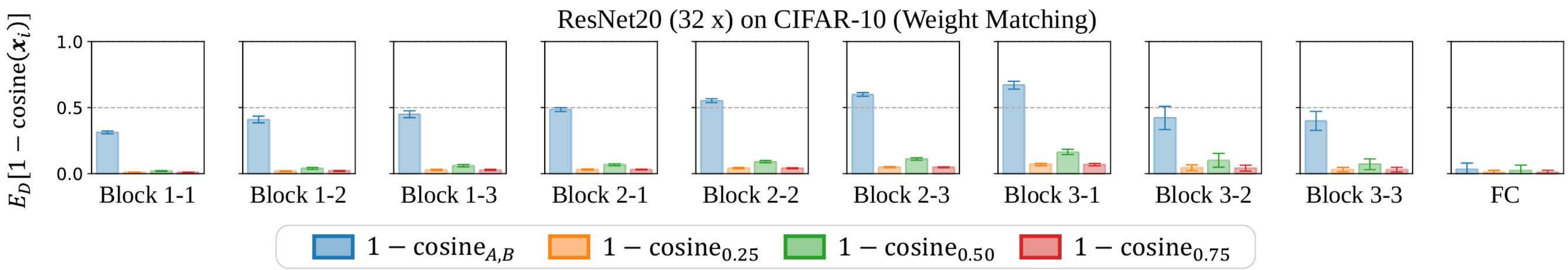


Fig. 3: Comparison of $E_D[1 - \text{cosine}_\alpha(\boldsymbol{x}_i)]^*$ and $E_D[1 - \text{cosine}_{A,B}(\boldsymbol{x}_i)]^*$, $\alpha \in \{.25, .5, .75\}$.

Lemma (LLFC implies LMC)

Two modes $\boldsymbol{\theta}_A$, $\boldsymbol{\theta}_B$ satisfy LLFC over dataset D and $\max\{\text{Err}_D(\boldsymbol{\theta}_A), \text{Err}_D(\boldsymbol{\theta}_B)\} \leq \epsilon$

$$\forall \alpha \in [0, 1], \text{Err}_D(\alpha\boldsymbol{\theta}_A + (1-\alpha)\boldsymbol{\theta}_B) \leq 2\epsilon.$$

$^*\text{cosine}_\alpha(\boldsymbol{x}_i) = \cos\langle f^{(\ell)}(\alpha\theta_A + (1-\alpha)\theta_B; \boldsymbol{x}_i), \alpha f^{(\ell)}(\theta_A; \boldsymbol{x}_i) + (1-\alpha) f^{(\ell)}(\theta_B; \boldsymbol{x}_i)\rangle$ and $\text{cosine}_{A,B}(\boldsymbol{x}_i) = \cos\langle f^{(\ell)}(\theta_A; \boldsymbol{x}_i), f^{(\ell)}(\theta_B; \boldsymbol{x}_i)\rangle$

Why LLFC Emerges?

Two simple conditions that leads to LLFC.

Condition I: Weak Additivity for ReLU Activations

Given dataset D, the modes $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ satisfy *weak additivity for ReLU activations* if

$$\forall \ell \in [L], \forall \alpha \in [0,1], \sigma\left(\alpha \widetilde{\boldsymbol{H}}_A^{(\ell)} + (1-\alpha)\widetilde{\boldsymbol{H}}_B^{(\ell)}\right) = \alpha\sigma\left(\widetilde{\boldsymbol{H}}_A^{(\ell)}\right) + (1-\alpha)\sigma\left(\widetilde{\boldsymbol{H}}_B^{(\ell)}\right).^*$$

*We denote ℓ-th layer pre-activations as $\widetilde{\boldsymbol{H}}^{(\ell)}$ over the dataset D and ReLU activation as $\sigma(\cdot)$.

Condition II: Commutativity

Given dataset D, the modes $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ satisfy *commutativity* if

$$\forall \ell \in [L], \boldsymbol{W}_A^{(\ell)}\boldsymbol{H}_A^{(\ell-1)} + \boldsymbol{W}_B^{(\ell)}\boldsymbol{H}_B^{(\ell-1)} = \boldsymbol{W}_B^{(\ell)}\boldsymbol{H}_A^{(\ell-1)} + \boldsymbol{W}_A^{(\ell)}\boldsymbol{H}_B^{(\ell-1)}.$$

Why LLFC Emerges?

Theorem (Condition I and II imply LLFC)
Given dataset D, if two modes $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ satisfy *weak additivity for ReLU activations* and *commutativity*, then

$$\forall \ell \in [L], \forall \alpha \in [0,1], f^{(\ell)}(\alpha\boldsymbol{\theta}_A + (1-\alpha)\boldsymbol{\theta}_B) = \alpha f^{(\ell)}(\boldsymbol{\theta}_A) + (1-\alpha) f^{(\ell)}(\boldsymbol{\theta}_B).^{*}$$

Weak additivity for ReLU activations and *commutativity* are verified empirical for modes that satisfy LMC/LLFC.

Justification of Permutation Method

Given a mode $\boldsymbol{\theta}_A$ and a permuted mode $\boldsymbol{\theta}'_B = \pi(\boldsymbol{\theta}_B)$ that satisfy LLFC, the *commutativity* is satisfied:

$$\forall \ell \in [L], \boldsymbol{W}_A^{(\ell)} \boldsymbol{H}_A^{(\ell-1)} + \boldsymbol{W'}_B^{(\ell)} \boldsymbol{H'}_B^{(\ell-1)} = \boldsymbol{W'}_B^{(\ell)} \boldsymbol{H}_A^{(\ell-1)} + \boldsymbol{W}_A^{(\ell)} \boldsymbol{H'}_B^{(\ell-1)} \quad (1)$$

Rewritten as:

$$\forall \ell \in [L], \left(\boldsymbol{W}_A^{(\ell)} - \boldsymbol{P}^{(\ell)} \boldsymbol{W}_B^{(\ell)} \boldsymbol{P}^{(\ell-1)^{\top}}\right)\left(\boldsymbol{H}_A^{(\ell-1)} - \boldsymbol{P}^{(\ell)} \boldsymbol{H}_B^{(\ell-1)}\right) = 0 \quad (2)$$

Connection to permutation methods

$$\text{weight matching:} \min_{\pi} \sum_{\ell=1}^{L} \left\| \boldsymbol{W}_A^{(\ell)} - \boldsymbol{P}^{(\ell)} \boldsymbol{W}_B^{(\ell)} \boldsymbol{P}^{(\ell-1)^{\top}} \right\|_F^2$$

$$\text{Activation matching:} \min_{\pi} \sum_{\ell=1}^{L} \left\| \boldsymbol{H}_A^{(\ell)} - \boldsymbol{P}^{(\ell)} \boldsymbol{H}_B^{(\ell)} \right\|_F^2$$

The two objectives correspond to the two factors of above equation.

Conclusion

Conclusion

- Identify Layerwise Linear Feature Connectivity (LLFC)
- Investigate the underlying contributing factors to LLFC
- Obtain novel insights into permutation methods

Future Directions

- Feature averaging methods
- Find a permutation directly enforcing the commutativity property
- Going Beyond Neural Network Feature Similarity: The Network Feature Complexity and Its Interpretation Using Category Theory