University of Amsterdam
AMLAB Amsterdam Machine Learning Lab

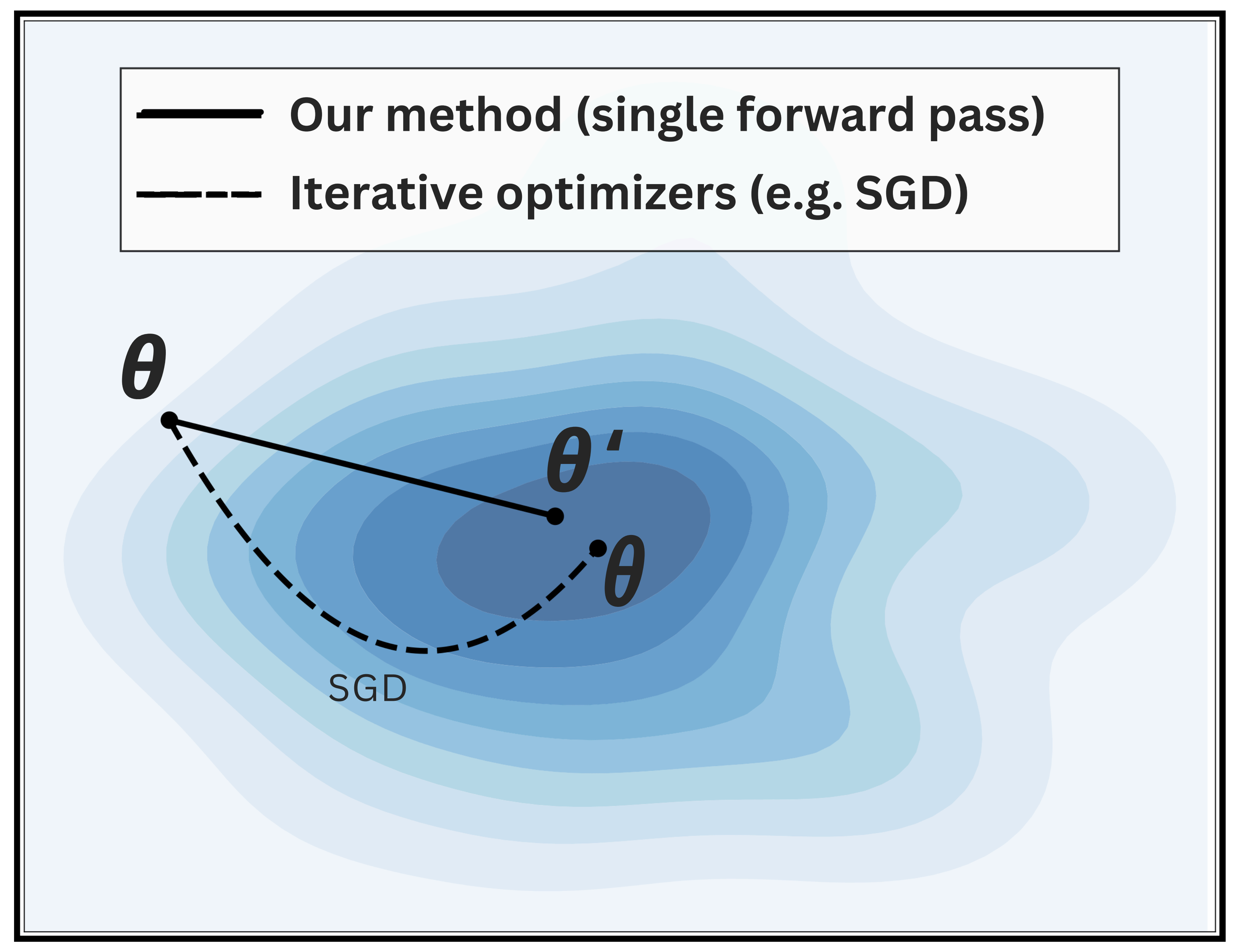
Symmetry-Aware Fully-Amortized Optimization with Scale Equivariant Graph Metanetworks

Bart Kuipers, Freek Byrman, Daniel Uyterlinde, Alejandro Garcia-Castellanos

SCAN ME

Introduction & Motivation

- Iterative optimizers are slow and expensive as models become increasingly larger.
- Amortized optimization accelerates the solution of related optimization problems by learning mappings that exploit shared structure across problem instances.
- Scale Equivariant Graph MetaNetworks (Kalogeropoulos et al., 2024) allow us to exploit symmetries found in NNs, to learn this mapping.



Conceptual idea of our fully-amortized meta-optimizer for a low-dimensional cost function

$$f_\phi(G,\boldsymbol{\theta}) \approx f^*(G,\boldsymbol{\theta}) \in \operatorname{argmin}_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}} \mathcal{C}(\boldsymbol{\theta}' \mid \boldsymbol{\theta}, G, \mathcal{D})$$

Our Contributions

- We leverage a symmetry-aware graph metanetwork **for single-shot finetuning a model's weights.**
- We prove that the **gauge freedom induced by scaling symmetries is strictly smaller in CNNs than in MLPs**, helping to explain performance differences observed in ScaleGMN experiments.

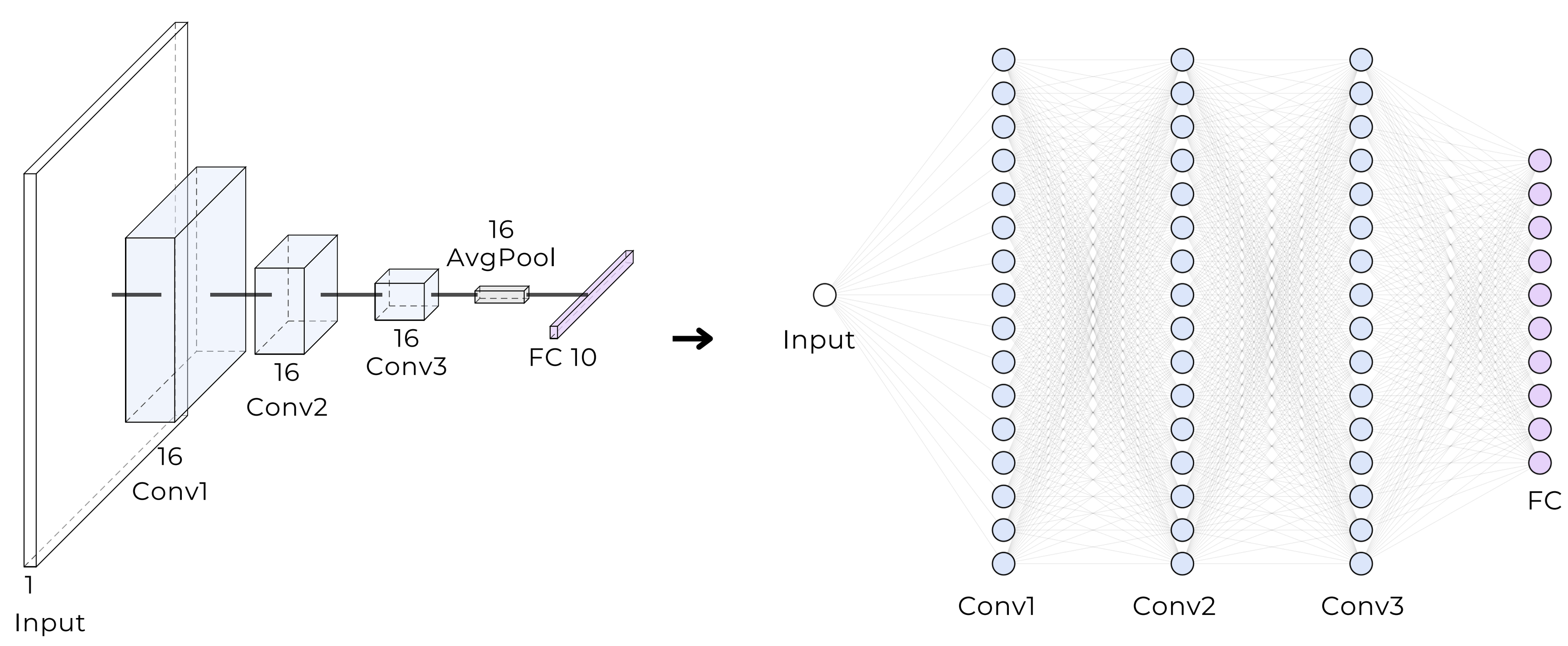
Scale Equivariant Graph Metanetworks

Metanetworks are models that take other NNs as input.

ScaleGMN is a graph metanetwork:

1. Maps a neural network to its **graph representation**: weights → edge features, biases → vertex features.
2. **Forward pass:** feature initialization, message passing, and feature updating.

ScaleGMN makes the components **equivariant to scaling and permutation symmetries.**



Mapping a CNN (left) to its corresponding graph structure (right)

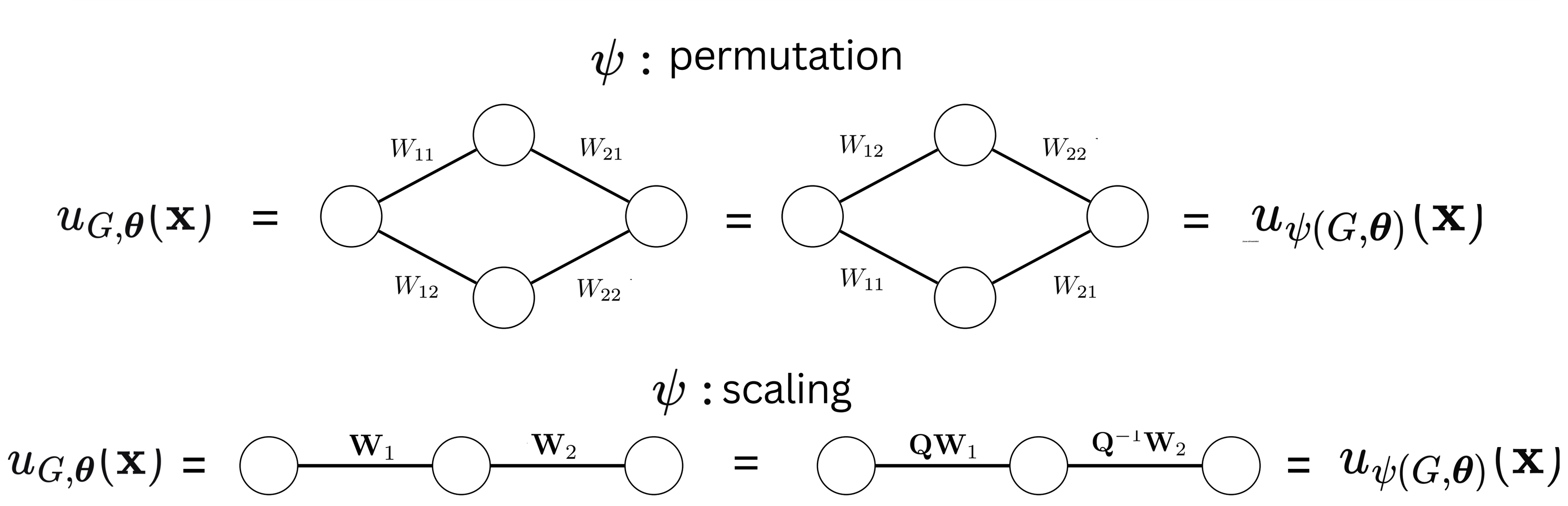
Network Symmetries

Neural network **symmetries** are transformations:

$$\psi : \mathcal{G} \times \boldsymbol{\Theta} \to \mathcal{G} \times \boldsymbol{\Theta}$$

That preserves network function:

$$u_{G,\boldsymbol{\theta}}(\mathbf{x}) = u_{\psi(G,\boldsymbol{\theta})}(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{X}, \quad \forall (G,\boldsymbol{\theta}) \in \mathcal{G} \times \boldsymbol{\Theta}$$



Methodology

We introduce an **amortized meta-optimization** model using the ScaleGMN framework:

- ScaleGMN maps input parameters to a new set of parameters:

$$\hat{f}_\phi : \mathcal{G} \times \boldsymbol{\Theta} \to \boldsymbol{\Theta}$$

- **Train a ScaleGMN** on a collection of pre-trained networks with:

$$\mathcal{L}(\phi; \boldsymbol{\theta}, \mathcal{B}) = \lambda \cdot \left\| \hat{f}_\phi(G,\boldsymbol{\theta}) \right\|_1 + \frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x},y)\in\mathcal{B}} \mathcal{L}_{\mathrm{CE}}\left(u_{G,\hat{f}_\phi(G,\boldsymbol{\theta})}(\mathbf{x}), y\right)$$

- Evaluate on **MLPs** and **CNNs**, activations **Tanh** and **ReLU**, with **cross-entropy** ($\lambda = 0$) and **sparsity** ($\lambda > 0$) objectives.
- Ablation: Breaking scale equivariance.

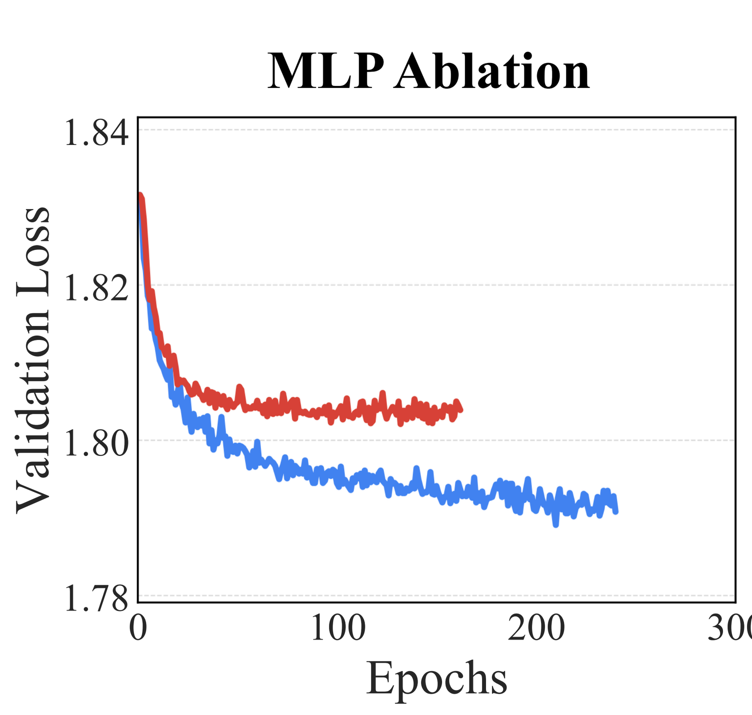
Results

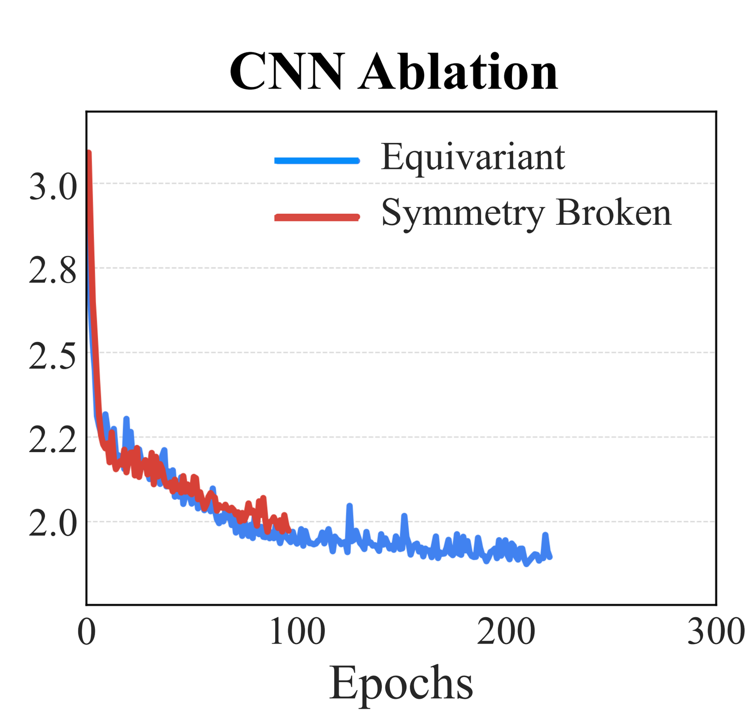
Ablation:

- Incorporating scaling symmetries **improves accuracy** and smoothens training.
- This effect is more apparent in MLPs than in CNNs, due to their **differences in guage freedom** *(full proof in paper)*.

Amortized Optimization:

- The ScaleGMN framework can be used for **single-shot optimization** *(see additional results on MLP architectures and other activation functions in the full paper)*.





Amortized Optimization:

Method	Cross-entropy ($\lambda = 0$) Avg Acc (%)	Cross-entropy ($\lambda = 0$) Max Acc (%)	Cross-entropy ($\lambda = 0$) Time (s)	L1-regularized ($\lambda > 0$) Avg Acc (%)	L1-regularized ($\lambda > 0$) Test Loss (CE+L1)	L1-regularized ($\lambda > 0$) Time (s)	L1-regularized ($\lambda > 0$) Sparsity (%)
CNN Architectures (Tanh)							
Initial performance	37.9	48.6	-	37.9	3.233	-	0
Metanetworks							
ScaleGMN-B	50.3	55.1	0.055	37.4	1.901	0.055	87.6
GMN-B (sym. broken)†	51.5	56.1	0.054	35.3	1.982	0.054	85.3
Iterative optimizer							
SGD (25 epochs)	41.0	50.6	59.5	39.8	2.651	82.5	35.9
SGD (50 epochs)	41.6	50.6	119	37.9	2.474	165	50.7
SGD (100 epochs)	43.5	51.1	238	40.4	2.189	330	64.4
SGD (150 epochs)	44.5	51.4	357	41.6	2.070	495	70.9