Hyperparameter Transfer in Graph Neural Networks

Gage DeZoort · Boris Hanin
Department of Operations Research & Financial Engineering, Princeton University

PRINCETON UNIVERSITY

Why hyperparameter transfer?

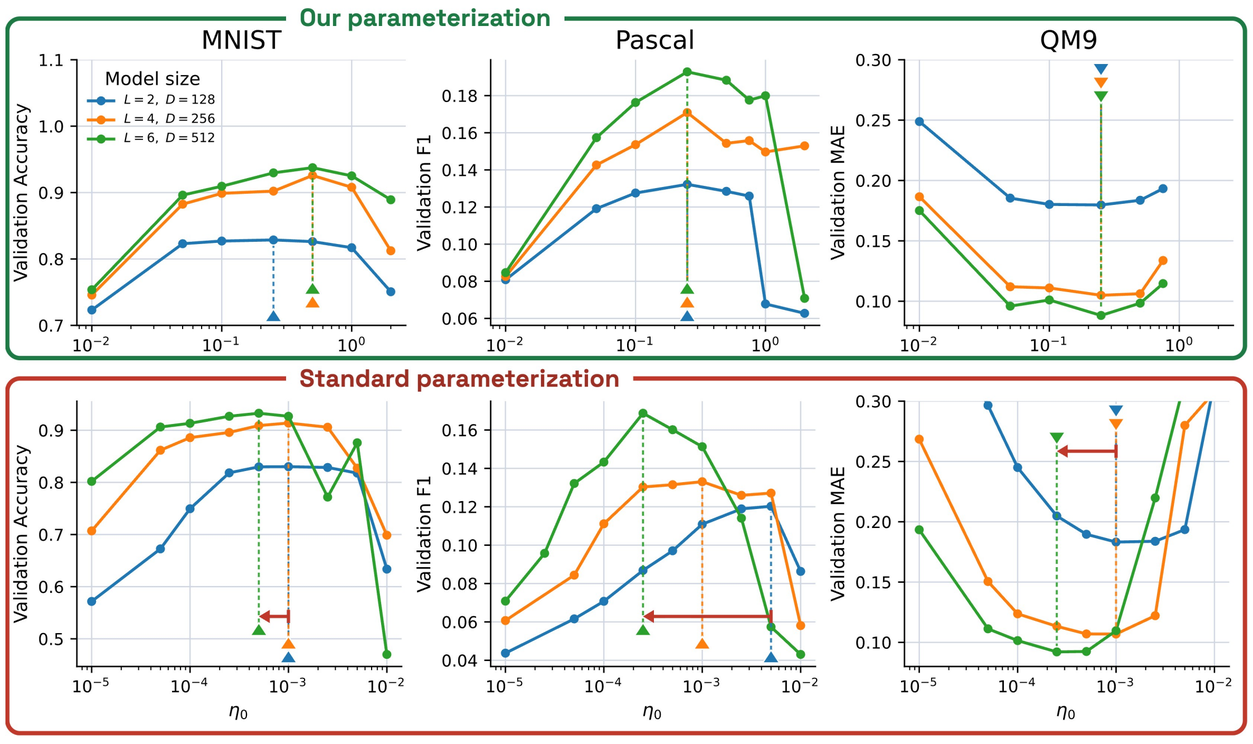
1. **Cheap proxy tuning** — tune small, scale large
   Tuning is prohibitively expensive at large **width D** and **depth L**; transfer infers near-optimal hyperparameters for large models from much smaller ones, where they are cheap to find.
2. **Limiting behavior** — training dynamics converge
   As model size increases, performance increases and training dynamics converge to a well-defined and fully non-linear limit.

One shared base learning rate across scale

With Adam, the learning rate **shrinks as width grows**

$$\mathrm{LR} = \frac{\textbf{Base LR}}{\sqrt{\textbf{Width}}}$$

...but the **base LR transfers**: tune it once, on a small model



Trained with Adam. **Our parameterization (top row): transfer holds;** the best η_0 range is consistent across scale. **Standard parameterization (bottom row): transfer breaks;** the best η_0 range drifts across scale.

1 Transfer parameterization for GNNs

GRAPH NOTATION

N nodes and M edges; node features $\mathbf{X} \in \mathbb{R}^{N \times n_0}$; adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$

MODEL AND INITIALIZATION

Encoder $\quad \mathbf{X}^{(0)} = \frac{1}{\sigma\sqrt{n_0}}\,\mathbf{X}\,\mathbf{W}^{(0)}$ $\quad$ weights $\sim \mathcal{N}(0, \sigma^2)$

Message passing $\quad \tilde{\mathbf{X}}^{(\ell+1)} = \mathbf{X}^{(\ell)} + \frac{1}{L}\,\mathrm{GNN}^{(\ell+1)}\big(\mathbf{X}^{(\ell)}, \mathbf{A}\big)$

MLP $\quad \mathbf{X}^{(\ell+1)} = \tilde{\mathbf{X}}^{(\ell+1)} + \frac{1}{L}\,\mathrm{MLP}^{(\ell+1)}\big(\tilde{\mathbf{X}}^{(\ell+1)}\big)$

Order-1 blocks neither grow nor shrink their inputs as D grows; for $\ell = 0, \ldots, L-1$

Decoder $\quad \mathbf{Z} = \frac{1}{\sigma D}\,\mathbf{X}^{(L)}\,\mathbf{W}^{(L+1)}$ $\quad$ weights $\sim \mathcal{N}(0, \sigma^2)$

$$\mathrm{GNN}(\mathbf{X}, \mathbf{A}) = \frac{1}{\sqrt{D}}\,\mathbf{P}\,\mathbf{X}\,\mathbf{W}$$

$\mathbf{P}$: message-passing operator built from $\mathbf{A}$; $W_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$

↓ OPTIMIZATION: PER-LAYER LEARNING RATES

	SGD	Adam
Init scale σ	$\sqrt{L}$	$1/\sqrt{D}$
Encoder, decoder $\eta^{(0)}, \eta^{(L+1)}$	$\eta_0 D \sigma^2$	$\eta_0 \sigma$
Residual $\eta^{(1\ldots L)}$	$\eta_0 D L$	$\eta_0 / \sqrt{D}$

One global learning rate: $\boldsymbol{\eta_0 DL}$ (SGD) or $\boldsymbol{\eta_0/\sqrt{D}}$ (Adam); only η_0 is tuned.

2 SGD + first-layer correction

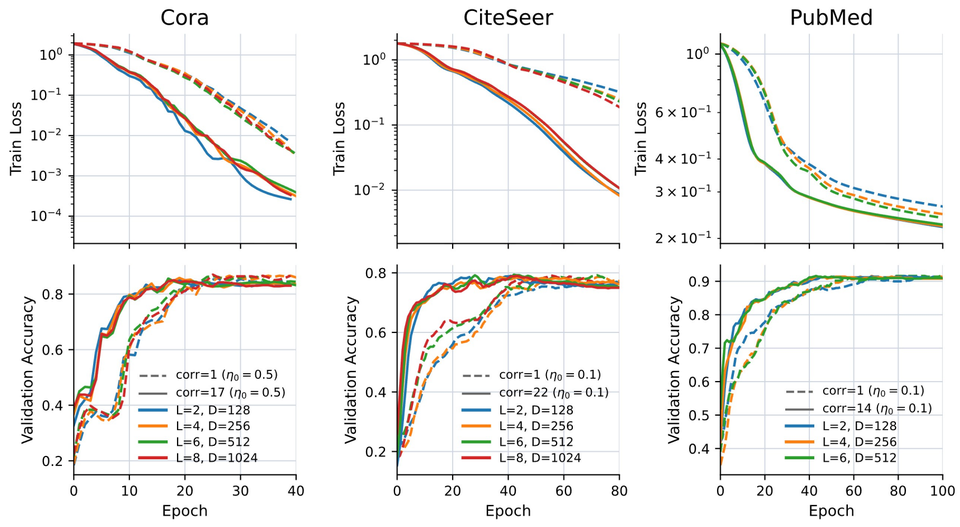
$$\eta_{\mathrm{SGD}}^{(0)} = \eta_0 D \sigma^2 \times \widehat{C}, \quad \widehat{C} = \operatorname*{mean}_{\alpha,\beta} \frac{n_0\sqrt{N_\beta}}{\big\|\bar{\mathbf{X}}_\alpha \mathbf{X}_\beta^\top\big\|_2}$$

One-step GD analysis, averaged over pairs of graphs α, β; $\bar{\mathbf{X}}_\alpha$: mean node feature of α

Estimated $\widehat{C}$ per dataset

Sparse bag-of-words features: $\widehat{C} \gg 1$, so the encoder rate must grow.

Cora	CiteSeer	PubMed
16.7	22.0	13.5

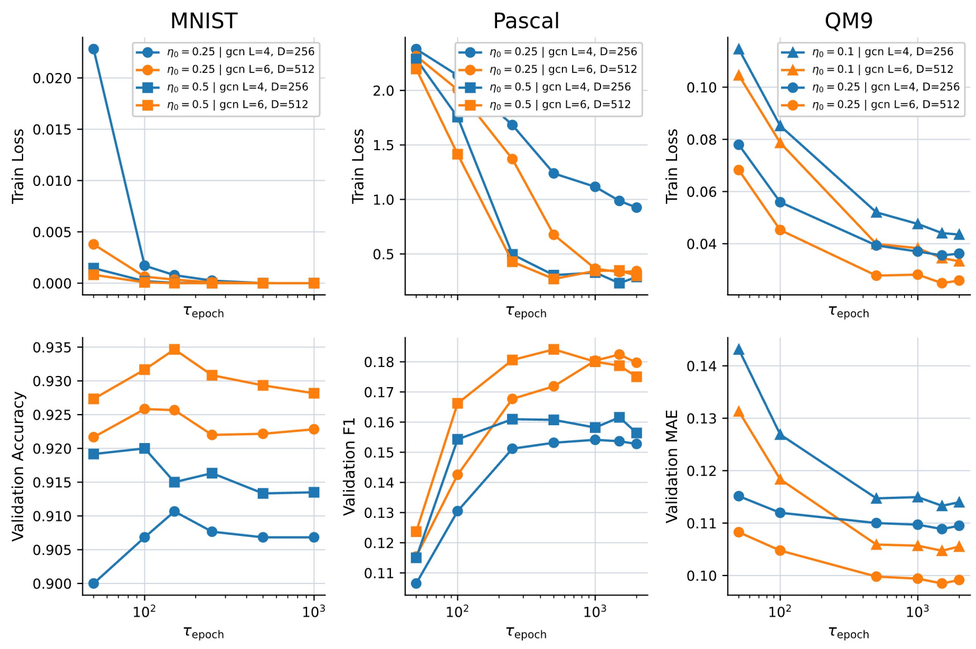


With (solid) vs. without (dashed) the correction at near-optimal η_0.

3 Adam & AdamW

Adam needs $\eta = \eta_0/\sqrt{D}$; AdamW is an EMA with timescale $(\lambda\eta)^{-1}$, so weight decay must scale with width:

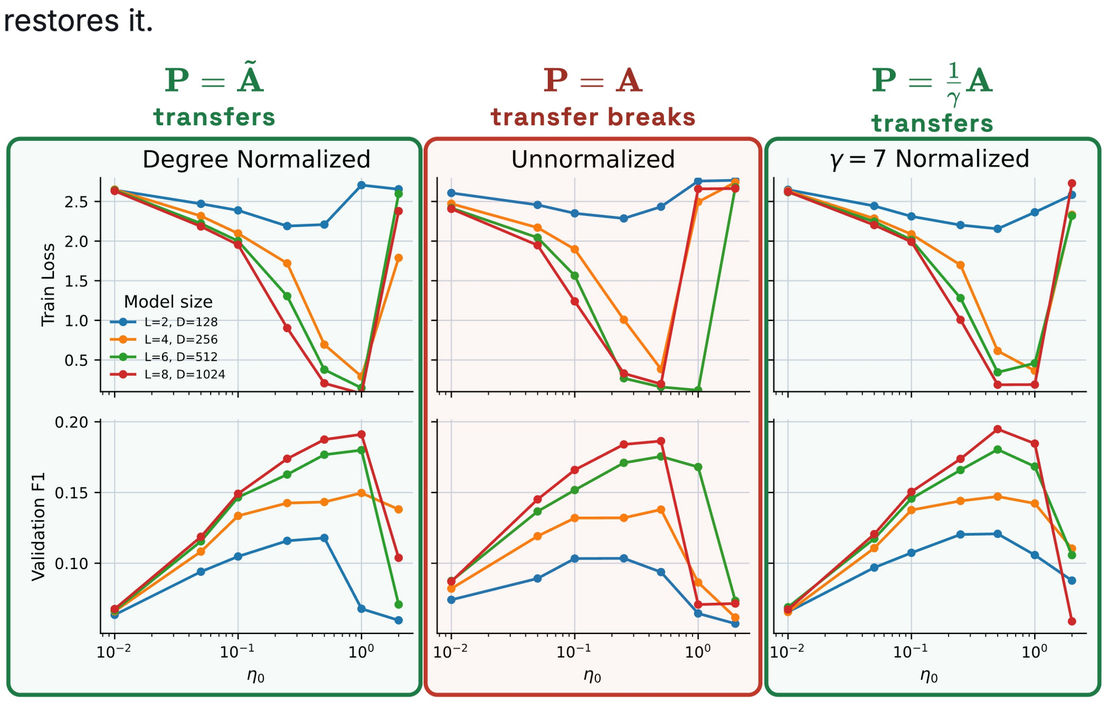
Weight decay $\quad \lambda = \lambda_0\sqrt{D}$ $\quad$ so $\tau_{\mathrm{epoch}} = B/(\lambda_0 \eta_0 N_{\mathrm{train}})$ transfers across scale



Weight decay transfer via τ_{epoch} with AdamW.

4 Normalize message passing

Message passing with the adjacency matrix shifts the activation scale by $\gamma = \|\mathbf{AX}\|_F / \|\mathbf{X}\|_F$; degree normalization ($\mathbf{P} = \tilde{\mathbf{A}}$) or dividing out γ restores it.



Adam, PascalVOC-SP, 250 epochs; here $\gamma = 7$.

Future directions

- Transfer for edge/hyperedge-aware GNNs
- GNN scaling laws, compared with language and vision
- Large-scale, heterogeneous, temporal and heterophilous graphs
- Benchmarking different aggregation operators
- Attention, graph positional encodings (e.g. Laplacian encodings, WIRE), deeper encoders/decoders

Gage DeZoort · jdezoort@princeton.edu
Paper: **arxiv.org/abs/2607.05017**
LoG 2026 · Northeastern University, Boston