Novelty Assessment Report

Paper: Understanding the Mechanisms of Fast Hyperparameter Transfer

PDF URL: https://openreview.net/pdf?id=Q7mLKxQ8qk

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

The growing scale of deep learning models has rendered exhaustive hyperparameter (HP) optimization prohibitively expensive. A promising solution is the use of scale-aware HPs, which can enable direct transfer of optimal settings from small-scale grid searches to large models with minimal performance loss. Such approaches are useful when the optimal settings converge "fast" enough with scale. While approaches like the Maximal Update Parameterization (μP) have empirically displayed fast transfer when scaling model width, a deeper conceptual understanding of the mechanisms that enable this is still missing. Our work establishes a systematic conceptual framework for analyzing fast HP transfer across different synthetic and practical scenarios. In synthetic settings, we present various quantitative examples where transfer either offers a provable computational advantage or fails even under μP. We then propose a key property that enables the fast transfer often observed in practice: through a novel decomposition of the optimization trajectory, we identify one component that rapidly converges with model width and determines the optimal HPs, and the other that continues to improve the loss with increased width but has negligible impact on HP choice. We conjecture that this decomposition elucidates the key mechanisms behind fast transfer and empirically validate it in practical settings such as LLM training.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Hyperparameter Transfer Across Neural Network Scales**

A total of **48 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Parametrization-Based Transfer Methods**
- **Optimizer-Specific Transfer**
- **Bayesian and Meta-Learning Transfer**
- **Application-Specific Transfer Methods**
- **Model Initialization and Warmstarting**
- **Theoretical Foundations and Analysis**

Complete Taxonomy Tree

- Hyperparameter Transfer Across Neural Network Scales Survey Taxonomy
- Parametrization-Based Transfer Methods
  - Maximal Update Parametrization (µP) and Extensions ★ (7 papers)
  - [0] Understanding the Mechanisms of Fast Hyperparameter Transfer (Anon et al., 2026) View paper
  - [1] Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer (Yang, 2022) View paper
  - [15] Tuning large neural networks via zero-shot hyperparameter transfer (Ge Yang, 2021) View paper
  - [18] Super Consistency of Neural Network Landscapes and Learning Rate Transfer (Thomas Hofmann, 2024) View paper
  - [29] Understanding Scaling Laws in Deep Neural Networks via Feature Learning Dynamics (Zihan Yao, 2025) View paper
  - [45] An Empirical Study of P Learning Rate Transfer (Lingle, 2024) View paper
  - [48] How Width Scaling Affects Neural Networks: Generalization, Optimal Hyperparameters, Feature Learning and Beyond (Haas, n.d.) View paper
  - Multi-Axis Parametrization Extensions (6 papers)
  - [7] Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit (Bordelon, 2023) View paper
  - [8] Completed Hyperparameter Transfer across Modules, Width, Depth, Batch and Duration (Bruno Mlodozeniec, 2025) View paper
  - [11] Time Transfer: On Optimal Learning Rate and Batch Size In The Infinite Data Limit (Filatov, 2024) View paper
  - [16] Scaling optimal LR across token horizons (Bjorck, 2024) View paper
  - [24] Power scheduler: A batch size and token number agnostic learning rate scheduler (Shen, 2024) View paper
  - Specialized Architecture Parametrizations (4 papers)
  - [17] nit Scaling: Simple and Scalable FP8 LLM Training (S Narayan, 2025) View paper
  - [26] Maximal Update Parametrization and Zero-Shot Hyperparameter Transfer for Fourier Neural Operators (Li, 2025) View paper
  - [30] Î¼-Parametrization for Mixture of Experts (Jan Malasnicki, 2025) View paper
- Optimizer-Specific Transfer
  - Matrix-Preconditioned and Second-Order Optimizers (4 papers)
  - [13] Dion: Distributed orthonormalized updates (Ahn, 2025) View paper
  - [19] Hyperparameter Transfer Enables Consistent Gains of Matrix-Preconditioned Optimizers Across Scales (Shikai Qiu, 2025) View paper
  - [25] Î¼ P2: Effective Sharpness Aware Minimization Requires Layerwise Perturbation Scaling (Haas, 2024) View paper
  - [35] Practical Efficiency of Muon for Pretraining (- -, 2025) View paper

    - Layerwise and Component-Specific Learning Rates (3 papers)
    - [21] Decoupled Relative Learning Rate Schedules (Jan Ludziejewski, 2025) View paper
    - [38] Function-Space Learning Rates (Milsom, 2025) View paper
    - [39] Robust Layerwise Scaling Rules by Proper Weight Decay Tuning (Fan Zhiyuan, 2025) View paper
    - Optimizer Comparison and Selection Studies (2 papers)
    - [36] Fantastic Pretraining Optimizers and Where to Find Them (Wen, 2025) View paper
    - [43] Scaling Exponents Across Parameterizations and Optimizers (Everett, 2024) View paper
- Bayesian and Meta-Learning Transfer
    - Multi-Task Bayesian Optimization (2 papers)
    - [2] Scalable hyperparameter transfer learning (Valerio Perrone, 2018) View paper
    - [47] Hyperparameter Transfer Learning with Adaptive Complexity (HorvÃ¡th, 2022) View paper
    - Sequential and Ordered Task Transfer (2 papers)
    - [22] Hyperparameter Transfer Across Developer Adjustments (Danny Stoll, 2020) View paper
    - [46] Obeying the Order: Introducing Ordered Transfer Hyperparameter Optimisation (Hellan, 2023) View paper
    - Neural Process-Based Transfer (1 papers)
    - [44] Transferable Neural Processes for Hyperparameter Optimization (Wei Ying, 2022) View paper
- Application-Specific Transfer Methods
    - Domain-Specific Applications (3 papers)
    - [4] Autone: Hyperparameter optimization for massive network embedding (Ke Tu, 2019) View paper
    - [5] Deep hyperparameter transfer learning for diabetic retinopathy classification (Mahesh S Patil, 2021) View paper
    - [40] Diffusing to the Top: Boost Graph Neural Networks with Minimal Hyperparameter Tuning (Lin, 2024) View paper
    - Search Algorithm and Infrastructure (6 papers)
    - [3] A new hyperparameters optimization method for convolutional neural networks (Hua Cui, 2019) View paper
    - [6] Hydro:{Surrogate-Based} hyperparameter tuning service in datacenters (Q Hu, 2023) View paper
    - [10] Tune as you scale: Hyperparameter optimization for compute efficient training (Fetterman, 2023) View paper
    - [14] Learning to Accelerate: Tuning Data Transfer Parameters (B Didrich, 2025) View paper
    - [20] Syne tune: A library for large scale hyperparameter tuning and reproducible research (D Salinas, 2022) View paper
    - [23] A novel hierarchical hyper-parameter search algorithm based on greedy strategy for wind turbine fault diagnosis (Yan Zhang, 2022) View paper
- Model Initialization and Warmstarting
    - Progressive Model Growth and Warmstarting (3 papers)
    - [9] Warmstarting for scaling language models (Mallik, 2024) View paper
    - [12] Longcat-flash technical report (Meituan LongCat Team, 2025) View paper
    - [34] A Comparative Survey: Reusing Small Pre-Trained Models for Efficient Large Model Training (Dhroov Pandey, 2024) View paper
    - Heterogeneous Model Transfer (2 papers)
    - [33] Heterogeneous Model Transfer between Different Neural Networks (Guangcong Wang, 2021) View paper
    - [42] Rethinking Binary Hyperparameters for Deep Transfer Learning (J. Plested, 2021) View paper
- Theoretical Foundations and Analysis (4 papers)
    - [27] Constructing large-scale low-latency network from small optimal networks (R. Mizuno, 2016) View paper
    - [28] Learning in Large Neural Networks (D. Anguita, 2025) View paper
    - [32] NC_SGOI: A Node Classification Method for a Streaming Graph Using Lightweight Variable Graph Neural Network (Zhipeng Sun, 2025) View paper
    - [41] Theoretical Foundations of Deep Learning: Optimization, Generalization, and Scaling (Ghosh, 2024) View paper

Narrative

Core task: Hyperparameter transfer across neural network scales. The field addresses how to reuse hyperparameters—such as learning rates, initialization schemes, and optimizer settings—when moving from small to large models, thereby reducing the cost of tuning at scale. The taxonomy organizes work into six main branches. Parametrization-Based Transfer Methods focus on reparametrizing network weights and learning rates so that optimal settings remain stable across widths or depths, with Maximal Update Parametrization (μP) and its extensions forming a prominent line of research. Optimizer-Specific Transfer examines how particular optimizers (e.g., Adam, SGD) behave under scaling and how their hyperparameters can be adapted. Bayesian and Meta-Learning Transfer leverages probabilistic models or learned transfer functions to predict good configurations. Application-Specific Transfer Methods tailor strategies to domains like vision or language modeling. Model Initialization and Warmstarting explores using smaller pretrained models to seed larger ones. Finally, Theoretical Foundations and Analysis provides scaling laws and convergence guarantees that underpin transfer strategies.

Several active themes emerge across these branches. One central question is whether transfer can be made nearly automatic—works like Tensor Programs V[1] and Zero Shot Transfer[15] aim for zero-shot or minimal-tuning regimes—versus methods that accept some residual search, as in CNN Hyperparameter Optimization[3] or Syne Tune[20]. Another contrast lies between width-centric parametrizations (μP-style approaches) and depth or layer-specific scaling rules. Fast Hyperparameter Transfer[0] sits within the Parametrization-Based Transfer branch, specifically under μP and Extensions, emphasizing efficient transfer with minimal retuning. It shares conceptual ground with Tensor Programs V[1] and Mu Learning Rate[45], which also exploit structured parametrizations to stabilize hyperparameters. Compared to Zero Shot Transfer[15], which targets immediate applicability, Fast Hyperparameter Transfer[0] may allow modest adjustments while still achieving strong cross-scale performance, positioning it as a practical middle ground in the parametrization-driven landscape.

Related Works in Same Category

The following **6 sibling papers** share the same taxonomy leaf node with the original paper:

1. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer

Authors: Yang, Greg, Greg Yang, Hu, Edward J., et al. (32 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

Hyperparameter (HP) tuning in deep learning is an expensive process, prohibitively so for neural networks (NNs) with billions of parameters. We show that, in the recently discovered Maximal Update Parametrization (muP), many optimal HPs remain stable even as model size changes. This leads to a new HP tuning paradigm we call muTransfer: parametrize the target model in muP, tune the HP indirectly on a smaller model, and zero-shot transfer them to the full-sized model, i.e., without directly tuning...

Relationship Analysis

Both papers belong to the Maximal Update Parametrization (µP) and Extensions category, focusing on hyperparameter transfer across neural network scales using µP-based parametrization methods. The original paper investigates the theoretical mechanisms underlying fast hyperparameter transfer by proposing a trajectory-based loss decomposition framework that explains why optimal hyperparameters converge faster than performance metrics under µP. The candidate paper (Tensor Programs V) is the foundational empirical work that introduced the µTransfer technique, demonstrating that µP enables zero-shot hyperparameter transfer from small to large models across various architectures (Transformers, ResNets) without providing the mechanistic explanations that the original paper develops.

2. Tuning large neural networks via zero-shot hyperparameter transfer

Authors: Ge Yang, J. Edward Hu, I. Babuschkin, Szymon Sidor, Xiaodong Liu, et al. (10 authors total) | **Year/Venue**: 2021 | **URL**: View paper

Abstract

â¦ While previous experiments scaled only width, here we will also scale depth, as discussed â¦ Leveraging the discovery of a feature learning neural network infinite-width limit, we â¦

Relationship Analysis

Both papers belong to the Maximal Update Parametrization (µP) and Extensions category, focusing on hyperparameter transfer across neural network scales using µP-based parametrization methods. The original paper investigates the theoretical mechanisms explaining why µP enables fast hyperparameter transfer, proposing a loss decomposition framework that separates width-stable and width-sensitive components to understand convergence rates. The candidate paper (Yang et al.) is the foundational work that introduced µP and empirically demonstrated zero-shot hyperparameter transfer (µTransfer) from small to large models, establishing the practical effectiveness of the approach without providing the deeper mechanistic analysis that the original paper develops.

3. Super Consistency of Neural Network Landscapes and Learning Rate Transfer

Authors: Thomas Hofmann, Alexandru Meterez, Lorenzo Noci, Antonio Orvieto | **Year/Venue**: 2024 • Neural Information Processing Systems | **URL**: View paper

Abstract

Recently, there has been growing evidence that if the width and depth of a neural network are scaled toward the so-called rich feature learning limit (\mup and its depth extension), then some hyperparameters -- such as the learning rate -- exhibit transfer from small to very large models. From an optimization perspective, this phenomenon is puzzling, as it implies that the loss landscape is consistently similar across very different model sizes. In this work, we study the landscape through the l...

Relationship Analysis

Both papers belong to the µP and Extensions category, investigating how hyperparameters transfer across neural network scales under Maximal Update Parametrization. While the original paper focuses on understanding the mechanisms of fast hyperparameter transfer through trajectory decomposition and loss component analysis, the candidate paper examines landscape consistency through Hessian spectral properties (particularly sharpness) and connects this to feature learning dynamics. The key difference is that the original paper provides a trajectory-based decomposition framework to explain when and why transfer works, whereas the candidate paper characterizes transfer through the lens of landscape geometry and its relationship to the NTK regime.

4. Understanding Scaling Laws in Deep Neural Networks via Feature Learning Dynamics

Authors: Zihan Yao, Ruoyu Wu, Tianxiang Gao | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The empirical success of deep learning is often attributed to scaling laws that predict consistent gains as model, data, and compute grow; however, large models can exhibit training instability and diminishing returns, suggesting that scaling laws describe what success looks like but not when and why scaling succeeds or fails. A central obstacle is the lack of a rigorous understanding of feature learning at large depth. While muP characterizes feature-learning dynamics in the infinite-width limi...

Relationship Analysis

Both papers belong to the Maximal Update Parametrization (µP) and Extensions category, focusing on scaling-aware hyperparameter transfer in neural networks. While the original paper investigates the mechanisms enabling fast hyperparameter transfer across width scales through trajectory decomposition and loss analysis, the candidate paper extends µP theory to the depth dimension by deriving Neural Feature Dynamics (NFD) as a forward-backward SDE system in the joint infinite-width and infinite-depth limit. The key distinction is that the original paper focuses on understanding why width-based µP transfer works efficiently in practice, whereas the candidate paper addresses the theoretical gap in depth scaling and proposes corrections for multi-layer residual blocks where depth-µP fails.

5. An Empirical Study of P Learning Rate Transfer

Authors: L Lingle | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ Empirically, ÂµP is also reported to enable zero-shot hyperparameter transfer from small proxy â¦ between 1/âD and 1/D can be treated as a constant multiplier in the model architecture. â¦

Relationship Analysis

Both papers belong to the Maximal Update Parametrization (µP) and Extensions category, focusing on hyperparameter transfer across neural network scales using µP-based approaches. The original paper develops a conceptual framework to understand the mechanisms enabling fast hyperparameter transfer under µP, proposing a loss decomposition into width-stable and width-sensitive components to explain why optimal hyperparameters converge faster than performance metrics. The candidate paper provides an extensive empirical study of µP learning rate transfer in transformers, investigating practical compatibility with various architectural choices (e.g., attention mechanisms, normalizations, optimizers) and demonstrating transfer up to 10B parameters, but does not develop theoretical mechanisms or loss decomposition frameworks for understanding transfer.

6. How Width Scaling Affects Neural Networks: Generalization, Optimal Hyperparameters, Feature Learning and Beyond

Authors: MSM Haas | **URL**: View paper

Abstract

â¦ (ÂµP) has been shown to induce hyperparameter transfer and improved generalization at large â¦ By preserving maximal stable feature learning in all layers, increasing model size in ÂµP is â¦

Relationship Analysis

Both papers belong to the Maximal Update Parametrization (µP) and Extensions category, focusing on hyperparameter transfer across neural network widths using µP-based scaling rules. The original paper investigates the mechanisms enabling fast hyperparameter transfer by decomposing optimization trajectories into width-stable and width-sensitive components, while the candidate paper is a dissertation that broadly studies width-scaling effects including generalization, optimal hyperparameters, and feature learning, with one chapter extending µP to Sharpness Aware Minimization (μP^2) and another analyzing standard parameterization under cross-entropy loss. The key difference is that the original paper provides a focused mechanistic analysis of why µP enables fast transfer through trajectory decomposition, whereas the candidate offers a comprehensive treatment of multiple width-scaling phenomena beyond the core µP framework.

Contributions Analysis

Overall novelty summary. The paper proposes a conceptual framework for understanding when and why hyperparameter transfer succeeds or fails across model widths, particularly under Maximal Update Parametrization (µP). It resides in the 'Maximal Update Parametrization (µP) and Extensions' leaf, which contains seven papers—a moderately populated research direction within the broader parametrization-based transfer landscape. This leaf focuses on width-scaling methods that preserve optimization dynamics, distinguishing itself from multi-axis extensions (depth, batch size) and optimizer-specific approaches. The work aims to move beyond empirical demonstrations of µP's effectiveness toward a principled understanding of the mechanisms enabling fast transfer.

The taxonomy reveals that parametrization-based methods form one of six major branches addressing hyperparameter transfer. Neighboring leaves include 'Multi-Axis Parametrization Extensions' (six papers handling depth and batch size jointly) and 'Specialized Architecture Parametrizations' (four papers for MoE, FNO, and FP8 training). The 'Optimizer-Specific Transfer' branch explores complementary strategies through preconditioned optimizers and layerwise learning rates, while 'Bayesian and Meta-Learning Transfer' pursues probabilistic surrogate models. The paper's focus on decomposing optimization trajectories to isolate width-stable components connects conceptually to theoretical foundations but remains grounded in the parametrization paradigm rather than optimizer design or meta-learning.

Among fifteen candidates examined, the framework contribution encountered one potentially refutable prior work out of ten candidates reviewed, while the loss decomposition (three candidates) and synthetic examples (two candidates) showed no clear refutations. The limited search scope—top-K semantic matches plus citation expansion—means these statistics reflect a targeted sample rather than exhaustive coverage. The framework contribution appears most exposed to overlap, likely because conceptual analyses of µP's mechanisms have been explored in sibling papers. The decomposition and synthetic examples may offer more distinctive technical angles, though the small candidate pools (two to three papers each) constrain confidence in their novelty assessment.

Based on the examined literature, the work occupies a moderately crowded research direction with established parametrization methods but contributes theoretical depth to understanding transfer mechanisms. The analysis covers a focused set of candidates rather than the full field, so conclusions about novelty remain provisional. The decomposition approach and synthetic counterexamples may represent the most original contributions, while the overarching framework builds incrementally on existing µP scholarship within a well-defined but active research area.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Conceptual framework for analyzing fast hyperparameter transfer

Description: The authors develop a formal framework to analyze when hyperparameter transfer is computationally useful, defining fast transfer as occurring when optimal hyperparameters converge faster than the evaluation metric. They connect this to computational efficiency through theorems showing when transfer strategies outperform direct tuning.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Be aware of overfitting by hyperparameter optimization!

URL: View paper

Brief Assessment

Overfitting Hyperparameter Optimization[53] focuses on demonstrating that hyperparameter optimization can lead to overfitting and that pre-set hyperparameters may perform similarly. It does not develop a formal framework analyzing when hyperparameter transfer is computationally useful or define fast transfer through convergence rate theorems as the original paper does.

2. Calibrated Dataset Condensation for Faster Hyperparameter Search

URL: View paper

Brief Assessment

Calibrated Dataset Condensation[54] focuses on dataset condensation for hyperparameter search by matching validation-performance rankings, not on analyzing when hyperparameter transfer across model scales is computationally efficient through convergence rate analysis.

3. Tune as you scale: Hyperparameter optimization for compute efficient training

URL: View paper

Brief Assessment

Tune as Scale[10] focuses on cost-aware Bayesian optimization for hyperparameter tuning across model scales, but does not develop a formal framework analyzing when hyperparameter transfer is computationally useful through convergence rate theorems as the original paper does.

4. Scalable hyperparameter transfer learning

URL: View paper

Brief Assessment

Scalable Hyperparameter Transfer[2] focuses on multi-task Bayesian optimization with neural network feature learning for transfer across different datasets/tasks, not on analyzing convergence rates or computational efficiency conditions for hyperparameter transfer as defined in the original paper's framework.

5. Dion: Distributed orthonormalized updates

URL: View paper

Brief Assessment

Dion[13] focuses on distributed orthonormalized optimizer updates for efficient large-scale training, not on analyzing hyperparameter transfer mechanisms or computational efficiency of transfer strategies across model scales.

6. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer

URL: View paper

Prior Art Analysis

Tensor Programs V[1] demonstrates that the conceptual framework for analyzing when hyperparameter transfer is computationally useful was already established in their work. They formally define the conditions under which optimal hyperparameters remain stable across model scales and connect this to computational efficiency through their µTransfer paradigm. The candidate paper provides explicit theoretical foundations (Theorems 13-14) showing when transfer strategies outperform direct tuning, and empirically validates these conditions across multiple architectures including transformers and ResNets. This prior work establishes the same core concepts that the original paper claims as novel contributions.

Evidence

Evidence 1 - **Rationale**: Tensor Programs V[1] explicitly describes their framework for HP transfer and its connection to computational efficiency, demonstrating prior establishment of the conceptual framework claimed as novel by the original paper. - **Original**: we formally define fast hp transfer in terms of convergence rates of optimal hps and the loss, and provide a direct connection to the 'usefulness' of transfer when performing compute-optimal grid search. - **Candidate**: this reduces the tuning problem of an (arbitrarily) large model to that of a (fixed-sized) small model. our overall procedure, which we call µtransfer, is summarized in algorithm 1 and fig. 2, and the hps we cover are summarized in tables 1 and 2.

Evidence 2 - **Rationale**: Both papers connect HP transfer to computational efficiency and performance advantages. Tensor Programs V[1] provides the foundational framework showing when transfer outperforms direct tuning, which the original paper builds upon. - **Original**: theorem 2 (informal). suppose an ~n-α and bn ~n-β. given a compute budget of fflops, let pdir(f) and ptr(f) be the compute-optimal performance under the direct and transfer strategies, respectively. then as f→∞ , ptr(f) ~pdir(f) iff β = α/2 and ptr(f) ≪pdir(f) iff β >α/2. - **Candidate**: better performance: µtransfer is not just about predicting how the optimal learning rate scales in sp. in general, we expect the µtransferred model to outperform its sp counterpart with learning rate optimally tuned. for example, this is the case in fig. 1 with the width-8192 transformer. we discuss...

Evidence 3 - **Rationale**: Tensor Programs V[1] demonstrates the computational efficiency framework through concrete examples showing massive speedup in HP tuning, establishing the practical connection between HP convergence and computational usefulness that the original paper claims as novel. - **Original**: we provide insight into the puzzle of fast hp transfer by first developing a framework for reasoning about hp transfer in section 2. then in section 3 we formally define fast hp transfer in terms of convergence rates of optimal hps and the loss - **Candidate**: speedup: it provides massive speedup to the tuning of large models. for example, we are able to outperform published numbers of (350m) bert-large [11] purely by zero-shot hp transfer, with tuning cost approximately equal to 1 bert-large pretraining. likewise, we outperform the published numbers of t...

7. Practical Efficiency of Muon for Pretraining

URL: View paper

Brief Assessment

Muon Pretraining[35] focuses on practical efficiency of the Muon optimizer and hyperparameter transfer via mup scaling, not on developing a formal framework for analyzing when hyperparameter transfer is computationally useful through convergence rate theorems.

8. Scaling Laws for Fine-Grained Mixture of Experts

URL: View paper

Brief Assessment

Fine Grained MoE[55] focuses on scaling laws for mixture of experts models with a granularity hyperparameter, not on hyperparameter transfer mechanisms or computational efficiency of transfer strategies across model scales.

9. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit

URL: View paper

Brief Assessment

Depthwise Transfer ResNets[7] focuses on hyperparameter transfer across both width and depth in residual networks using 1/√depth scaling, while the original paper analyzes fast transfer specifically across width using µP parameterization with a trajectory decomposition framework. The candidate does not address the original's core mechanism of top-k loss decomposition for explaining fast transfer.

10. Two-step hyperparameter optimization method: Accelerating hyperparameter search by using a fraction of a training dataset

URL: View paper

Brief Assessment

Two Step Optimization[52] focuses on reducing computational cost by evaluating hyperparameters on data subsets, not on analyzing transfer across model scales or defining convergence rate conditions for computational efficiency.

Contribution 2: Loss decomposition into width-stable and width-sensitive components

Description: The authors introduce a novel trajectory-level loss decomposition that separates the linearized loss change into top-k components (which remain width-invariant and determine optimal hyperparameters) and residual components (which improve with width but minimally affect hyperparameter choice). This decomposition provides a mechanistic explanation for fast transfer.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. The empirical impact of neural parameter symmetries, or lack thereof

URL: View paper

Brief Assessment

Parameter Symmetries[49] focuses on removing parameter space symmetries in neural networks through architectural modifications (w-asymmetric and σ-asymmetric networks), not on loss decomposition or hyperparameter transfer mechanisms across network widths.

2. How Width Scaling Affects Neural Networks: Generalization, Optimal Hyperparameters, Feature Learning and Beyond

URL: View paper

Brief Assessment

Width Scaling Effects[48] focuses on width-dependent parameterizations (μP, μP^2) and their effects on hyperparameter transfer, feature learning, and generalization. It does not present a trajectory-level loss decomposition into top-k and residual components as described in the original paper.

3. Invariant polynomials and machine learning

URL: View paper

Brief Assessment

Invariant Polynomials[50] focuses on using invariant polynomial generators in machine learning for particle physics, not on neural network width scaling or hyperparameter transfer mechanisms. The paper addresses symmetry-preserving representations rather than optimization trajectory decomposition across model widths.

Contribution 3: Synthetic examples demonstrating conditions for fast transfer

Description: The authors provide concrete synthetic examples including random features regression (where fast transfer provably occurs) and two-layer ReLU networks (where transfer can be slow even under maximal update parameterization), illustrating that fast transfer depends on structural properties of the training process rather than being guaranteed by parameterization alone.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Tuning large neural networks via zero-shot hyperparameter transfer

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

2. Sparse maximal update parameterization: A holistic approach to sparse training dynamics

URL: View paper

Brief Assessment

Sparse Maximal Update[51] focuses on sparse neural network training dynamics and hyperparameter transfer across sparsity levels, not on random features regression or conditions for fast transfer in the sense studied by the original paper.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Understanding the Mechanisms of Fast Hyperparameter Transfer View paper
- [1] Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer View paper
- [2] Scalable hyperparameter transfer learning View paper
- [3] A new hyperparameters optimization method for convolutional neural networks View paper
- [4] Autone: Hyperparameter optimization for massive network embedding View paper
- [5] Deep hyperparameter transfer learning for diabetic retinopathy classification View paper
- [6] Hydro:{Surrogate-Based} hyperparameter tuning service in datacenters View paper
- [7] Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit View paper
- [8] Completed Hyperparameter Transfer across Modules, Width, Depth, Batch and Duration View paper
- [9] Warmstarting for scaling language models View paper
- [10] Tune as you scale: Hyperparameter optimization for compute efficient training View paper
- [11] Time Transfer: On Optimal Learning Rate and Batch Size In The Infinite Data Limit View paper
- [12] Longcat-flash technical report View paper
- [13] Dion: Distributed orthonormalized updates View paper
- [14] Learning to Accelerate: Tuning Data Transfer Parameters View paper
- [15] Tuning large neural networks via zero-shot hyperparameter transfer View paper
- [16] Scaling optimal LR across token horizons View paper
- [17] nit Scaling: Simple and Scalable FP8 LLM Training View paper
- [18] Super Consistency of Neural Network Landscapes and Learning Rate Transfer View paper
- [19] Hyperparameter Transfer Enables Consistent Gains of Matrix-Preconditioned Optimizers Across Scales View paper
- [20] Syne tune: A library for large scale hyperparameter tuning and reproducible research View paper
- [21] Decoupled Relative Learning Rate Schedules View paper
- [22] Hyperparameter Transfer Across Developer Adjustments View paper
- [23] A novel hierarchical hyper-parameter search algorithm based on greedy strategy for wind turbine fault diagnosis View paper
- [24] Power scheduler: A batch size and token number agnostic learning rate scheduler View paper
- [25] Î¼ P2: Effective Sharpness Aware Minimization Requires Layerwise Perturbation Scaling View paper
- [26] Maximal Update Parametrization and Zero-Shot Hyperparameter Transfer for Fourier Neural Operators View paper
- [27] Constructing large-scale low-latency network from small optimal networks View paper
- [28] Learning in Large Neural Networks View paper
- [29] Understanding Scaling Laws in Deep Neural Networks via Feature Learning Dynamics View paper
- [30] Î¼-Parametrization for Mixture of Experts View paper
- [31] μ-Parametrization for Mixture of Experts View paper
- [32] NC_SGOI: A Node Classification Method for a Streaming Graph Using Lightweight Variable Graph Neural Network View paper
- [33] Heterogeneous Model Transfer between Different Neural Networks View paper
- [34] A Comparative Survey: Reusing Small Pre-Trained Models for Efficient Large Model Training View paper
- [35] Practical Efficiency of Muon for Pretraining View paper
- [36] Fantastic Pretraining Optimizers and Where to Find Them View paper
- [37] Scaling Optimal LR Across Token Horizon View paper
- [38] Function-Space Learning Rates View paper
- [39] Robust Layerwise Scaling Rules by Proper Weight Decay Tuning View paper

- [40] Diffusing to the Top: Boost Graph Neural Networks with Minimal Hyperparameter Tuning View paper
- [41] Theoretical Foundations of Deep Learning: Optimization, Generalization, and Scaling View paper
- [42] Rethinking Binary Hyperparameters for Deep Transfer Learning View paper
- [43] Scaling Exponents Across Parameterizations and Optimizers View paper
- [44] Transferable Neural Processes for Hyperparameter Optimization View paper
- [45] An Empirical Study of P Learning Rate Transfer View paper
- [46] Obeying the Order: Introducing Ordered Transfer Hyperparameter Optimisation View paper
- [47] Hyperparameter Transfer Learning with Adaptive Complexity View paper
- [48] How Width Scaling Affects Neural Networks: Generalization, Optimal Hyperparameters, Feature Learning and Beyond View paper
- [49] The empirical impact of neural parameter symmetries, or lack thereof View paper
- [50] Invariant polynomials and machine learning View paper
- [51] Sparse maximal update parameterization: A holistic approach to sparse training dynamics View paper
- [52] Two-step hyperparameter optimization method: Accelerating hyperparameter search by using a fraction of a training dataset View paper
- [53] Be aware of overfitting by hyperparameter optimization! View paper
- [54] Calibrated Dataset Condensation for Faster Hyperparameter Search View paper
- [55] Scaling Laws for Fine-Grained Mixture of Experts View paper