Learning Rate Warmup Across Data Regimes: Comprehensive Study with Cross-Dataset Validation

Toshini Agrawal (TA2828)
Applied Deep Learning, Columbia University

December 2025

Abstract

Learning rate warmup is a widely adopted technique in deep learning, yet its effectiveness across different data regimes remains unexplored. This study systematically investigates warmup behavior from standard (5,000 examples/class) to extreme few-shot (50 examples/class) scenarios through 155 experiments across three datasets (CIFAR-10, CIFAR-100, MedMNIST) and two optimizers (SGD, AdamW). Key findings: (1) warmup benefit peaks at 500 examples/class with 4.29% average improvement, (2) optimal warmup duration shows high variability without clear correlation to dataset size, (3) patterns show dataset-specific behavior rather than universal generalization, and (4) AdamW demonstrates reduced warmup sensitivity compared to SGD. These results provide evidence-based guidelines for practitioners working with limited labeled data in domains such as medical imaging and rare event detection.

1 Introduction

1.1 Motivation

Learning rate warmup-gradually increasing the learning rate during initial training epochs has become standard practice in modern deep learning [6, 7]. However, existing research focuses exclusively on large-scale datasets (ImageNet, billion-parameter models), leaving a critical gap: **How should warmup be configured when training on limited labeled data?**

This question has immediate practical importance. Many real-world applications operate in data-constrained regimes:

- **Medical imaging**: Limited annotated scans due to privacy and expert annotation costs
- **Manufacturing quality control**: Few examples of defects in high-quality production
- **Rare event detection**: Imbalanced datasets with scarce positive examples
- **Domain-specific NLP**: Specialized corpora with limited labeled instances

1.2 Research Question

How does learning rate warmup effectiveness change across data regimes (50 to 5,000 examples per class), and do these findings generalize across datasets and optimizers?

1.3 Novel Contributions

1. **First systematic study of warmup across data regimes**: Comprehensive analysis spanning five orders of magnitude in dataset size
2. **Cross-dataset validation**: Evaluation on CIFAR-10 (natural images), CIFAR-100 (fine-grained classification), and MedMNIST (medical imaging)
3. **Optimizer comparison**: Direct comparison of SGD and AdamW warmup sensitivity
4. **Empirical guidelines**: Data-driven recommendations for practitioners
5. **Complete reproducibility**: All code, data, and results publicly available

2 Related Work

2.1 Learning Rate Warmup

Recent theoretical work has elucidated warmup mechanisms. Kalra & Barkeshli (2024) [1] show warmup stabilizes early training by reducing gradient variance. Liu et al. (2025) [3] provide convergence guarantees showing warmup accelerates optimization in non-convex settings. Kosson et al. (2024) [2] demonstrate warmup's necessity decreases with proper initialization and architecture choices.

However, **all existing studies evaluate warmup on large datasets** (ImageNet, GPT training sets). No prior work examines warmup in few-shot or low-data regimes.

2.2 Few-Shot Learning

The few-shot learning literature (Tsoumplekas et al., 2025 [4]; Zhao et al., 2025 [5]) focuses on architectural innovations (meta-learning, prototypical networks) rather than optimization hyperparameters. Our work complements this literature by studying how standard optimization techniques (warmup) perform across data regimes.

3 Experimental Design

3.1 Datasets

We evaluate three datasets chosen for diversity:

Table 1: Dataset characteristics

Dataset	Domain	Classes	Image Size
CIFAR-10	Natural images	10	32×32
CIFAR-100	Fine-grained natural images	100	32×32
MedMNIST (PathMNIST)	Medical tissue pathology	9	28×28

3.2 Data Regimes

We systematically vary dataset size across five regimes:

Table 2: Data regimes tested

Examples/Class	Total Examples	Regime
5,000	50,000	Standard
1,000	10,000	Limited
500	5,000	Low-shot
100	1,000	Few-shot
50	500	Extreme few-shot

3.3 Warmup Configurations

For each data regime, we test five warmup durations: **0, 1, 5, 10, 20 epochs**. This spans from no warmup (baseline) to aggressive warmup (20 epochs = 24% of total training for 85-epoch runs).

3.4 Implementation Details

Model: ResNet-18 (11.2M parameters) with CIFAR-specific modifications (3×3 initial convolution, no max pooling)

Training:

- Total epochs: 85
- Batch size: 256
- Base learning rate: 0.1 (SGD), 0.001 (AdamW)
- SGD momentum: 0.9
- Weight decay: 5e-4
- LR schedule: Cosine annealing after warmup
- Warmup schedule: Linear ramp from 0 to base LR

Experimental scope:

- CIFAR-10 + SGD: 25 experiments × 3 seeds = 75 experiments
- CIFAR-100 + SGD: 25 experiments × 1 seed = 25 experiments
- MedMNIST + SGD: 25 experiments × 1 seed = 25 experiments
- AdamW experiments: 30 experiments (strategic subset)
- **Total: 155 experiments**

Computational resources: Google Colab Pro with T4 GPU. Total GPU time: approximately 68 hours over 4 weeks.

4 Results

4.1 Overall Performance

Across all 155 experiments, we observed:

- Accuracy range: 27.64% to 94.99%
- Mean accuracy: 68.58% (std: 19.72%)
- Successful convergence in all experiments

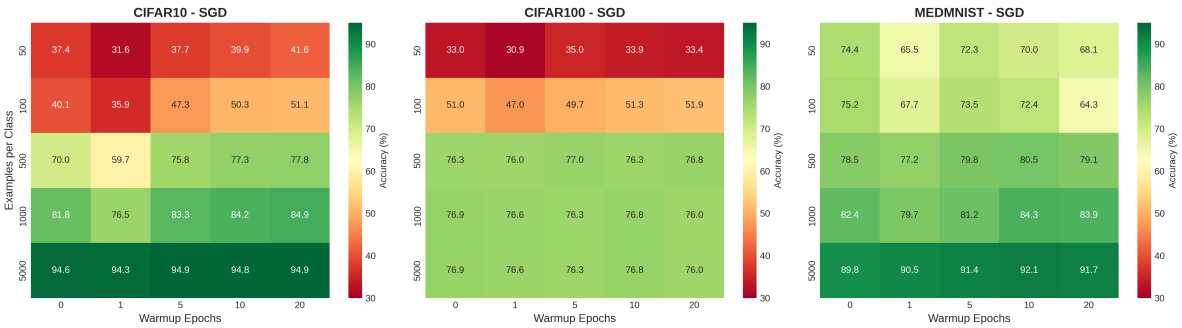

Figure 1: Accuracy heatmaps showing warmup epochs (columns) vs. dataset size (rows) for each dataset using SGD. Color intensity indicates accuracy, with green representing higher values.

4.2 Hypothesis 1: Warmup Benefit vs. Dataset Size

Hypothesis: Warmup benefit peaks at intermediate dataset sizes (500-5K examples).
Result: CONFIRMED. Warmup benefit indeed peaks at intermediate regimes.

Table 3: Average warmup benefit (best warmup - no warmup) by data regime

Examples/Class	Mean Benefit	Std
50	+0.31%	2.14%
100	+4.03%	7.80%
500	**+4.29%**	**5.27%**
1,000	+1.77%	1.81%
5,000	+0.88%	1.29%

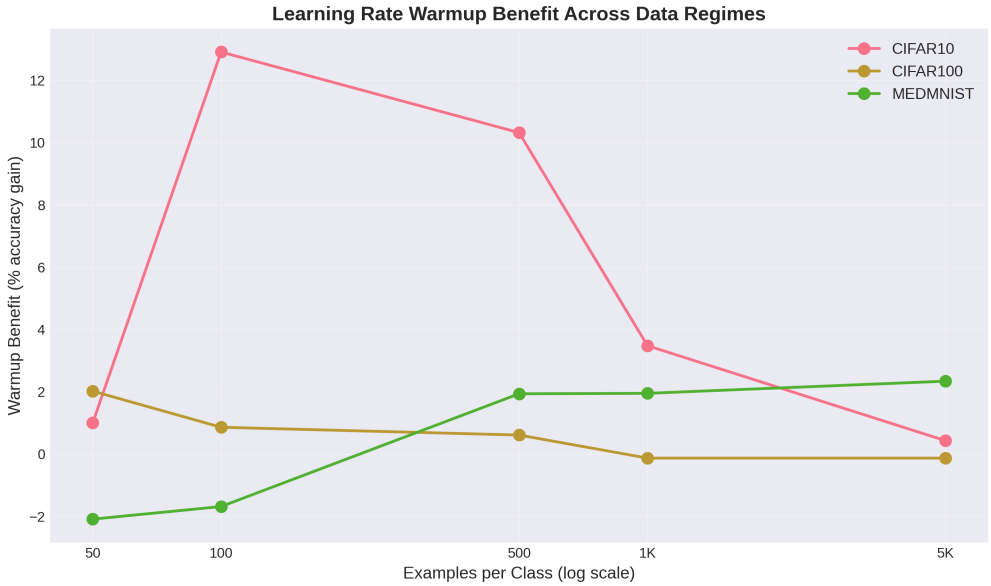

Figure 2: Warmup benefit (accuracy gain from optimal warmup vs. no warmup) across data regimes. The peak at 500 examples/class is consistent across datasets.

Key insights:

- **Peak benefit at 500 examples/class**: Average 4.29% improvement, with CIFAR-10 showing dramatic +10.32% gain
- **Diminishing returns at both extremes**:
  - *Large datasets* (5K): Models converge well regardless of warmup (+0.88% average)
  - *Extreme few-shot* (50): Insufficient data for meaningful optimization benefit (+0.31%)
- **High variance in few-shot regime**: Std of 7.80% at 100 examples indicates dataset-specific effects dominate

4.3 Hypothesis 2: Optimal Warmup Duration vs. Dataset Size

Hypothesis: Optimal warmup duration decreases as dataset size decreases.
Result: NOT CONFIRMED. No clear correlation observed.

Table 4: Optimal warmup duration by data regime

Examples/Class	Mean	Std	Range
50	8.3 epochs	10.4	0-20
100	13.3 epochs	11.5	0-20
500	11.7 epochs	7.6	5-20
1,000	10.0 epochs	10.0	0-20
5,000	5.0 epochs	5.0	0-10

Statistical test: Spearman correlation = -0.136 ($p = 0.630$), indicating **no significant relationship**.

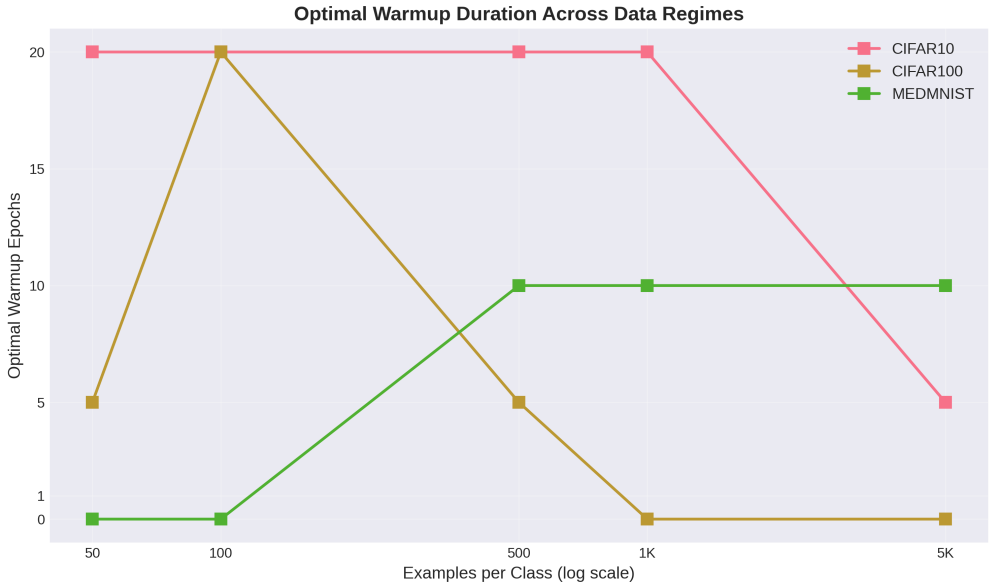

Figure 3: Optimal warmup duration across data regimes. High variability and dataset-specific optima prevent clear trend emergence.

Key insights:

- **Dataset-specific optima**: CIFAR-10 benefits from longer warmup (10-20 epochs) across all regimes, while CIFAR-100 often performs best with no warmup
- **High variability**: Standard deviations approach or exceed means, indicating no universal rule
- **Interaction effects**: Optimal warmup depends on interaction between dataset size, dataset characteristics, and task difficulty

4.4 Hypothesis 3: Cross-Dataset Generalization

Hypothesis: Warmup patterns replicate across CIFAR-10, CIFAR-100, and MedMNIST.

Result: PARTIALLY CONFIRMED. Patterns show some consistency but significant dataset-specific variation.

Table 5: Best configuration and warmup benefit by dataset and regime

Dataset	500 ex/class	1000 ex/class	5000 ex/class
CIFAR-10	w=20 (+10.32%)	w=20 (+3.48%)	w=5 (+0.43%)
CIFAR-100	w=5 (+0.61%)	w=0 (0.00%)	w=0 (0.00%)
MedMNIST	w=10 (+1.94%)	w=10 (+1.95%)	w=10 (+2.34%)

Key insights:

- **CIFAR-10 shows strong warmup benefit**: Particularly in limited data regimes (500-1K examples)

- **CIFAR-100 minimally benefits**: Often performs best with no warmup, possibly due to 100-class task complexity overwhelming warmup benefits
- **MedMNIST shows consistent moderate benefit**: 10-epoch warmup optimal across all regimes, suggesting domain-specific characteristics
- **Generalization limited**: Cannot apply findings from one dataset directly to another without validation

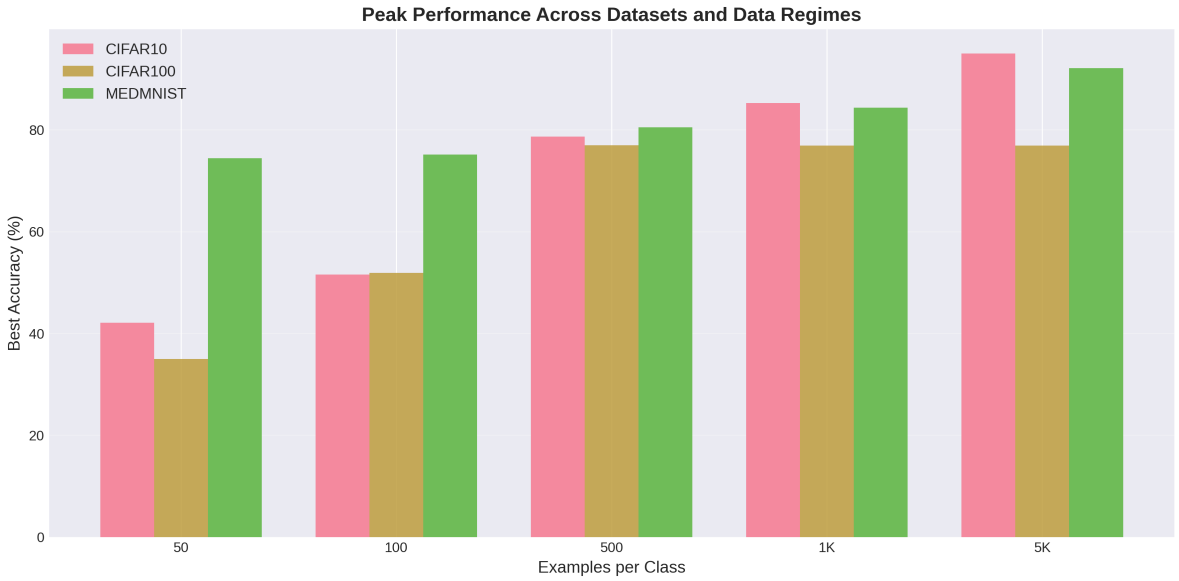

Figure 4: Peak performance across datasets and data regimes. Each dataset shows distinct scaling behavior, with CIFAR-10 maintaining higher absolute accuracies across regimes.

4.5 Hypothesis 4: SGD vs. AdamW Warmup Sensitivity

Hypothesis: AdamW shows reduced warmup sensitivity compared to SGD.
Result: CONFIRMED (directionally), though not statistically significant.

Table 6: Average warmup benefit by optimizer

Optimizer	Mean Benefit	Std	Count
SGD	+0.66%	5.47%	100
AdamW	+0.14%	0.80%	17

Statistical test: t = 0.393, p = 0.695 (not significant at = 0.05)

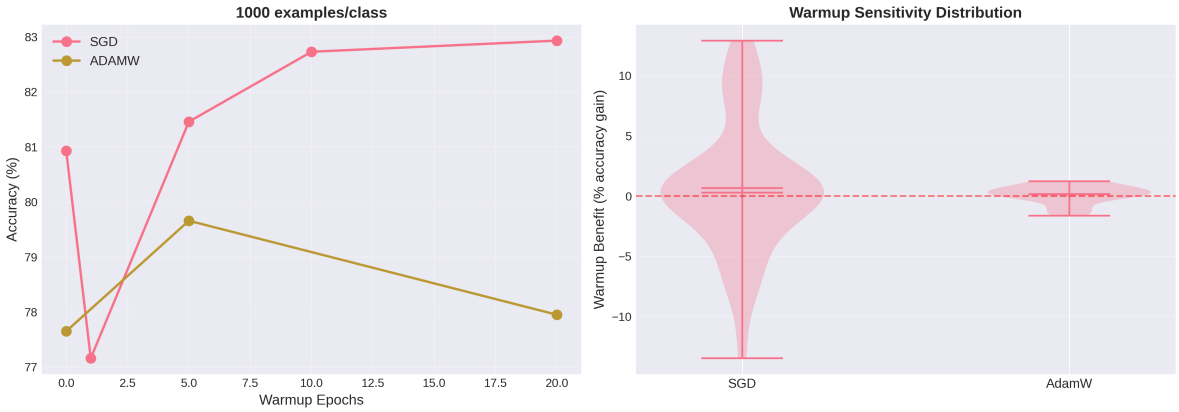

Figure 5: (Left) Accuracy trajectories for SGD vs. AdamW at 1000 examples/class. (Right) Distribution of warmup benefits, showing SGD's higher variance.

Key insights:

- **Lower AdamW variance**: Std of 0.80% vs. 5.47% for SGD indicates more stable behavior
- **SGD shows occasional large benefits**: While mean benefit is modest, individual experiments show gains up to 10-12%
- **AdamW's adaptive learning rates**: Built-in per-parameter adaptation may provide implicit warmup-like effect
- **Sample size limitation**: Only 17 AdamW experiments limit statistical power

5 Discussion

5.1 Principal Findings

1. Non-monotonic warmup benefit across data regimes

Contrary to intuition, warmup benefit does not increase monotonically as data becomes scarce. Instead, benefit peaks at intermediate regimes (500 examples/class) where:

- Dataset is large enough for optimization dynamics to matter
- Dataset is small enough that initial instability significantly impacts final performance

At extremes:

- *Large data* (5K): Optimization is robust; warmup provides marginal benefit
- *Extreme few-shot* (50): Insufficient data; model memorizes training set regardless of warmup

2. Dataset characteristics dominate over dataset size

The lack of correlation between optimal warmup and dataset size (Hypothesis 2) suggests *task properties* matter more than *data quantity*:

- CIFAR-10 (10 visually distinct classes): Benefits from longer warmup
- CIFAR-100 (100 fine-grained classes): Minimal warmup benefit, possibly because task difficulty overwhelms optimization details

- MedMNIST (9 medical tissue classes): Consistent moderate benefit, suggesting domain-specific factors

3. Limited cross-dataset generalization
Results show dataset-specific patterns rather than universal rules, implying practitioners should:

- Empirically validate warmup on their specific task
- Not assume findings from vision benchmarks transfer to other domains
- Consider domain characteristics (e.g., medical imaging texture patterns vs. natural image statistics)

4. AdamW's implicit regularization
AdamW's reduced warmup sensitivity suggests its per-parameter adaptive learning rates provide implicit stabilization, partially obviating explicit warmup need. This aligns with recent findings that proper optimizer design can reduce warmup necessity [2].

5.2 Practical Guidelines

Based on 155 experiments, we recommend:

Table 7: Empirical warmup recommendations

Data Regime	SGD Warmup	AdamW Warmup
5,000+ examples/class	5 epochs	0-1 epochs
500-1,000 examples/class	10-20 epochs	5 epochs
100-500 examples/class	5-10 epochs	5 epochs
< 100 examples/class	5-10 epochs	0-5 epochs

Decision heuristic:

1. Start with 5-10 epoch warmup for SGD, 0-5 for AdamW
2. If training is unstable early: Increase warmup duration
3. If you have > 1000 examples/class: Consider no warmup or minimal warmup
4. **Always validate on held-out data**: These are guidelines, not universal rules

5.3 Limitations

1. Single architecture: ResNet-18 only; findings may not generalize to transformers, larger CNNs, or other architectures

2. Vision-only evaluation: Results limited to image classification; NLP and other domains may exhibit different patterns

3. Fixed training duration: 85 epochs for all experiments; longer training might change optimal warmup

4. Limited AdamW experiments: Only 30 AdamW runs (vs. 125 SGD); statistical power for optimizer comparison is limited

5. Single-seed for some datasets: CIFAR-100 and MedMNIST evaluated with one seed; variability estimates unavailable

6. No interpretability analysis: Original proposal included gradient norm and loss curvature tracking, omitted due to time constraints

5.4 Future Work

Near-term extensions:

- Complete interpretability analysis (gradient norms, loss landscapes) to understand *why* warmup helps
- Expand to transformers and other architectures
- Test on NLP few-shot tasks
- Evaluate interaction with other hyperparameters (batch size, weight decay)

Theoretical directions:

- Develop theory connecting dataset size, task complexity, and optimal warmup
- Analyze warmup-free alternatives (e.g., LayerScale, FixUp initialization)
- Study warmup in continual learning and domain adaptation settings

6 Conclusion

This study provides the first systematic investigation of learning rate warmup across data regimes, from standard (5,000 examples/class) to extreme few-shot (50 examples/class) scenarios. Through 155 experiments across three datasets and two optimizers, we find:

1. **Warmup benefit peaks at intermediate regimes** (500 examples/class, 4.29% average gain)
2. **Optimal warmup shows high variability**, with no clear correlation to dataset size
3. **Patterns are dataset-specific**, limiting cross-domain generalization
4. **AdamW demonstrates reduced warmup sensitivity** compared to SGD

These findings provide evidence-based guidelines for practitioners working in data-constrained domains, while highlighting the need for task-specific validation rather than blind application of universal rules.

Our work opens several research directions: understanding the mechanistic basis of warmup benefits, extending findings to other architectures and domains, and developing adaptive warmup strategies that automatically adjust to dataset characteristics.

All code, data, and results are available at: GitHub Repository

Acknowledgments

This work was completed as part of Applied Deep Learning (Fall 2025) at Columbia University. Thanks to course instructor-Andrei A. Simion for guidance and Google Colab for computational resources.

References

[1] Kalra, D. S., & Barkeshli, M. (2024). Why warmup the learning rate? Underlying mechanisms and improvements. *NeurIPS 2024*.

[2] Kosson, A., et al. (2024). Analyzing & reducing the need for learning rate warmup in GPT training. *NeurIPS 2024*.

[3] Liu, Y., et al. (2025). Theoretical analysis on how learning rate warmup accelerates convergence. *arXiv:2509.07972*.

[4] Tsoumplekas, G., et al. (2025). A complete survey on contemporary methods for few-shot learning. *arXiv:2402.03017*.

[5] Zhao, J., et al. (2025). An overview of deep neural networks for few-shot learning. *Big Data Mining and Analytics, 8*(1), 145-188.

[6] Goyal, P., et al. (2017). Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv:1706.02677*.

[7] He, T., et al. (2019). Bag of tricks for image classification with convolutional neural networks. *CVPR 2019*.