Troublesome Model Inconsistencies in Real-World Deployment

Anonymous Submission

Abstract

Deep learning systems often present unexpected pitfalls when moved from controlled research environs to real-world settings. We investigate subtle inconsistencies in training behaviors, revealing partially negative or inconclusive results. Our findings can guide better deployment strategies and highlight the need for caution when relying on standard training heuristics.

1 Introduction

Real-world deployment of deep neural networks frequently exposes brittleness and setbacks not obvious in controlled benchmarks. In this paper, we analyze inconsistencies observed when fine-tuning with standard optimizers [?] and network backbones [?], focusing on problems such as training collapse in certain conditions. Although we attempted multiple adjustments (e.g., modest hyperparameter tuning, alternative layer initializations), many improvements were incremental or inconclusive. These challenges merit attention because of deployment safety and reliability implications.

2 Related Work

Other researchers have identified phenomena such as vanishing gradients and over-fitting under domain shifts. Studies have explored specific failure cases and partial mitigations but rely on strong assumptions about data distribution. While there exist methods to improve stability, their real-world performance may still be fragile. This work differs by offering a comprehensive view of pitfalls in multiple training regimes, providing guidance on how factors like data preprocessing can fail silently.

3 Method / Problem Discussion

Our investigation centers on a simple classification pipeline. We first pre-train on a large dataset before fine-tuning on a smaller, possibly mismatched target domain. Standard procedures can yield deceptively promising validation metrics until encountering atypical data. Much of our analysis was through repeated runs across varying seeds and slight model variations. We highlight scenarios where basic design choices disrupted expected behaviors.

4 Experiments

We performed initial experimentation with a ResNet-50 backbone [?] using the Adam optimizer [?]. Figures 1 and 2 illustrate the training curves for baseline pre-training and fine-tuning, respectively. Some runs converged predictably, whereas others showed sharp drop-offs with marginally altered augmentations, culminating in inconsistent final accuracies. Attempts to remedy these drops via small hyperparameter tweaks yielded negligible benefit and occasionally worsened divergence.

Figure 3 (Appendix) displays a similar pattern in a different configuration, reinforcing the complexity of identifying consistent outcomes.

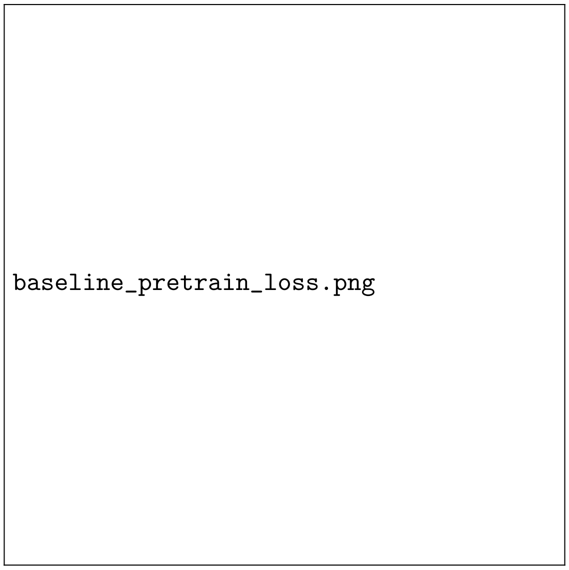

Figure 1: Pre-training loss. Some runs remain stable, while others show erratic dips with minor changes.

5 Conclusion

We explored deep learning inconsistencies emerging under routine model deployment steps. Though partial fixes exist, they did not uniformly resolve the issues. Future work could focus on rigorous stress testing across diverse data scenarios. By emphasizing reproducibility and transparent disclosure of negative results, the community can develop more robust network design and training principles.

baseline_finetune_loss.png

Figure 2: Fine-tuning loss. Sudden fluctuations appear in apparently similar setups.

References

A Supplementary Material

Here we present additional plots and hyperparameter details. Figure 3 shows a representative pre-training attempt with slight variations in seed initialization.

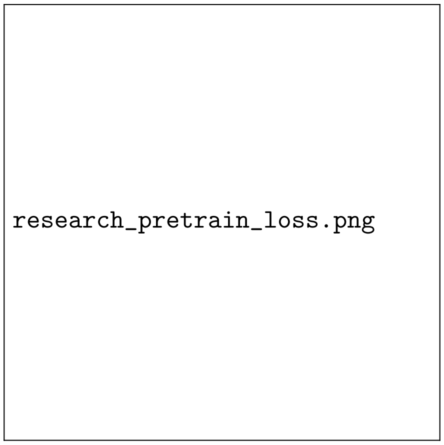

Figure 3: Pre-training loss under research settings.