Uncovering Real-World Overfitting Pitfalls in Deep Learning

Anonymous Submission

Abstract

We investigate overfitting pitfalls that arise in real-world deployment, focusing on hidden factors that degrade generalization. Our findings underscore the subtlety of certain design choices (e.g. label smoothing) and highlight how partial remedies can fail under non-ideal conditions. These observations matter for practical deep learning pipelines, where discrepancies between expectation and reality can squander model performance.

1 Introduction

Overfitting remains a core challenge in deep learning [**?**]. In practice, small changes in data distribution, labeling practices, or hyperparameters can expose severe flaws. We show how attempts to mitigate overfitting (e.g. label smoothing) may only partially succeed. These findings stem from a series of controlled experiments, revealing pitfalls that are easily overlooked in real-world scenarios.

2 Related Work

Prior studies on overfitting [**?**] often focus on synthetic benchmarks, sometimes neglecting subtle real-world issues. Label smoothing techniques [**?**] have been proposed to combat overconfidence, but the benefits can be context-dependent. Other ablation analyses [**?**] show that seemingly minor adjustments to regularization can cause unexpected optimization behaviors. Our work extends these insights by illustrating additional pitfalls during deployment.

3 Method / Problem Discussion

We consider a classification task where data exhibits distribution shifts. Our baseline model uses standard cross-entropy. Key variants include label smoothing and dropout. Despite careful tuning, performance often deteriorates under domain shifts. We track training/validation curves, confusion matrices, and downstream performance metrics to pinpoint persistent shortcomings.

Figure 1: Baseline performance exhibits steep overfitting after early epochs.

4 Experiments

Experiments were conducted on a real-world dataset where the baseline achieves initial gains but overfits rapidly. We illustrate these trends in Figures 1-3. Label smoothing mildly reduces the gap between training and validation metrics, though overfitting remains an issue.

Negative or inconclusive outcomes emerged: confusion matrices revealed misclassification spikes for minority classes in new domains. These analyses, included in the Appendix, confirm that typical regularization strategies do not always suffice.

5 Conclusion

We presented evidence of overfitting pitfalls in realistic settings. Our experiments confirm that partial solutions (label smoothing, dropout) can help but are not panaceas. Future work might explore novel forms of adaptive regularization that address the nuanced distributional shifts common in practical pipelines.

Figure 2: Modified approach with partial mitigation. Gains are noticeable but not robust.

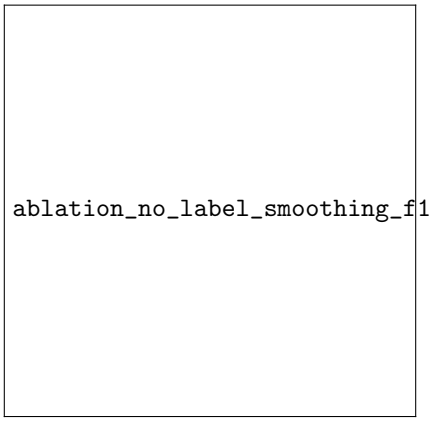

Figure 3: Ablation shows label smoothing improves F1 but gaps persist in shifted conditions.

References

A Appendix

Additional confusion matrices, ablation details, and extended plots (including per-seed training accuracy) are provided here for completeness.