A Glimpse of Negative Results: Hidden Overfitting and Subtle Relational Gains

An Ambitious AI Researcher

Abstract

Real-world deep learning systems frequently exhibit unexpected failures, such as severe overfitting and hidden brittleness to data shifts. This paper presents negative and partially successful results from experiments aimed at improving model generalization across relational tasks. These lessons highlight important pitfalls for practical deployment of sequence and graph-based deep models in uncertain, ever-changing environments.

1 Introduction

Deep learning research often reports high-profile successes, but crucial insights are also gleaned from experiments that do not go as planned [**?**]. Failures such as unrecoverable overfitting in large-batch training and inconsistent performance across complex tasks reflect real challenges that practitioners face. This paper explores these pitfalls and partial improvements in a relational learning context. Our contributions revolve around two major findings: first, how batch-size scaling can induce overfitting and fragile test performance; second, how integrating relational inductive biases yields moderate but significant improvements on complex tasks.

2 Related Work

Prior research has documented erratic behavior in large-batch training and potential overfitting [**?**]. These concerns appear in both supervised [**?**] and self-supervised contexts. Relational architectures, such as graph neural networks, have shown promise for capturing structured dependencies [**?**], yet partial failures remain underexplored in the literature.

3 Method / Problem Discussion

We investigate a multi-task setup incorporating standard supervised objectives alongside relational constraints. Our pipeline adapts a baseline deep model to work with a graph-based relational module. The goal is to see if relational features mitigate the overfitting vulnerability we detected in baseline experiments.

4 Experiments

All experiments were implemented in a controlled environment to isolate overfitting dynamics. The baseline suffered from severe performance degradation with large batch sizes. The proposed relational approach improved training stability, yet only partially bridged the gap between training and validation metrics.

baseline_loss_curves.png

baseline_test_compwa_summary.png

Figure 1: (a) Training & validation loss curves show significant divergence at larger batch sizes, (b) Corresponding Test CompWA performance often drops sharply beyond a moderate batch size.

Figure 1 reveals how large batches facilitate faster initial training but then lead to sharply diminished results on unseen test data. The improved relational method, however, remains somewhat robust to these issues.

Figure 2 highlights the partial gains of integrating relational inductive biases. Although the final performance is improved, the model exhibits variance under domain shifts, which suggests the need for further refinement.

5 Conclusion

We presented negative and partially successful results examining overfitting tied to batch-size scaling and highlighting the incremental benefits of relational inductive biases. Our findings underscore the importance of thorough empirical checks before deploying large-batch training or complex relational architectures in production settings. Future work could investigate adaptive regularization strategies that address these pitfalls and enhance reliability across varied data distributions.

References

A Appendix

This appendix contains hyperparameters, logging details, and additional clarifications on the relational module. All removed figures showed minor variations on training metrics, thus omitted for brevity.

research_sprbench_metrics.png

research_test_cplxwa_summary.png

Figure 2: Relational architecture (left) shows moderate improvements in key metrics, while final CplxWA (right) indicates better stability across expectations.