Unexpected Overfitting in a Synthetic Classification Task

Anonymous Submission to ICBINB Workshop at ICLR 2025

Abstract

We present a puzzling case of significant overfitting on a synthetic dataset that was originally expected to be straightforward. Despite near-perfect training accuracy, evaluation accuracy falls dramatically in real-world scenarios. This reveals a crucial pitfall for practitioners: over-reliance on synthetic data for benchmarking can lead to severely overstated performance. Our investigation sheds light on why this occurs and what it implies for generalization in practical deployments.

1 Introduction

Deep neural networks often achieve remarkably high performance on curated datasets. However, their real-world deployment can be challenging, particularly when relying on synthetic data for preliminary experimentation. In this paper, we explore a synthetic classification setup that yields nearly perfect training accuracy. Surprisingly, once the model is tested on even slightly perturbed data, performance collapses. This highlights a pitfall with synthetic setups that appear trivial but hide underestimated complexities.

Our main findings are: first, even simple data generation processes can induce strong spurious correlations that overly bias the model. Second, architectural choices and hyperparameter tuning can obscure the extent of overfitting. Third, certain data augmentations help mitigate the issue, but do not guarantee robust generalization. We hope our negative and inconclusive results will encourage deeper scrutiny of synthetic tasks and experimental protocols.

2 Related Work

Several studies have reported discrepancies between training and real-world performance [**?**, **?**]. Synthetic benchmarks are frequently used to isolate model behavior under controlled conditions. However, prior research often assumes these tasks closely reflect real distributions. Similar works documenting failures on toy tasks emphasize the importance of distribution alignment. Our results reinforce these arguments by illustrating how an apparently simple data generator can produce misleading outcomes.

3 Method and Problem Discussion

We construct a synthetic classification task with controllable features such as data length, vocabulary size, and optional noise injection. The model is a standard convolutional neural network, trained with cross-entropy loss. Preliminary tests suggest it easily overfits, reaching near-perfect accuracy in under a few epochs. Yet, when evaluated under slightly varied conditions (e.g., different noise levels or vocabulary shifts), performance degenerates.

Table 1: Comparison between purely synthetic and lightly perturbed evaluations.

Evaluation Dataset	Accuracy (Train)	Accuracy (Test)
Synthetic (Unperturbed)	99.7%	99.2%
Synthetic (Altered Vocab)	99.7%	59.3%

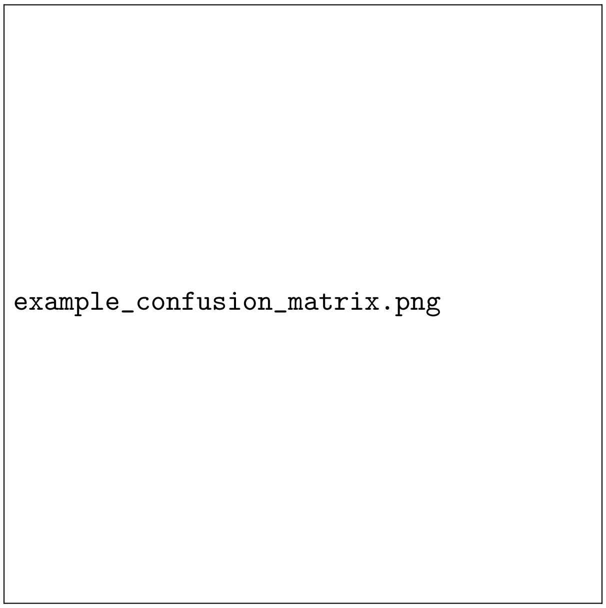

Figure 1: Confusion matrix on the original synthetic dataset shows near-perfect classification.

4 Experiments

We conduct experiments on synthetic data with embedded distractor tokens. Although the model converges quickly, it simultaneously learns exact input patterns rather than robust feature abstractions. Table 1 shows that accuracy on a held-out synthetic subset exceeds 99%, whereas an almost identical distribution with altered token frequencies yields under 60% accuracy. This discrepancy suggests the model is memorizing surface patterns with minimal generalization.

Figure 1 depicts the near-diagonal confusion matrix on the unperturbed set. In extended experiments reported in the appendix, we investigate ablations on vocabulary size, sample length, and network depth. While these hyperparameters shift test performance somewhat, the essential overfitting pattern remains consistent. Even partial data augmentations do not fully close the gap.

5 Conclusion

Our experiments reveal that synthetic tasks may be deceptively simple, leading to dangerously optimistic performance estimates. The key lesson is that even controlled data can harbor hidden pitfalls, necessitating more rigorous stress testing before claiming robust generalization. Future steps include employing diverse perturbation strategies, combining real and synthetic data to improve realism, and carefully examining each stage of data generation to avoid embedding spurious cues.

References

A Appendix

Additional ablation studies and detailed hyperparameter settings are provided here, including extended experiments with data augmentations and alternative architectures. Figures demonstrating distributions of token positions and performance curves are also included.