When Validation Deceives: Surprising Shortfalls in a Color-Shape Model

Anonymous Submission

Abstract

We investigate a color-shape matching model whose validation performance suggests strong generalization. However, comprehensive tests reveal significant pitfalls and overfitting. Our analysis demonstrates how high validation results may be misleading for real-world deployment, offering insights to the community on avoiding such pitfalls.

1 Introduction

Deep neural networks often exhibit promising validation metrics that fail to translate into comparable gains in real-world settings. We explore a color-shape dataset targeted at object classification, where the model reliably learns to match color and shape in controlled training and validation splits. Nevertheless, the model struggles to handle subtle variations or domain shifts. This discrepancy highlights risks in over-reliance on validation metrics. Our key contributions include revealing: (1) how seemingly strong validation performance can obscure subpar generalization, and (2) lessons from ablation studies that underscore the fragility of color-shape alignment mechanisms.

2 Related Work

Extensive literature has examined overfitting in vision tasks and discrepancies between validation accuracy and real-world transfer. Several prior investigations argued that small domain shifts can produce large drops in performance. While *data augmentation* can remediate some issues, the underlying mismatch often persists. These studies resonate with our findings, though our approach focuses on color-shape tasks and the misleading nature of validation success.

3 Method

We employ a two-channel network to represent visual shape and color features. The baseline setting uses a standard training regimen. A second, dual-channel variant separately encodes color and shape before merging them in a final layer. Our experiments isolate which factors contribute to overfitting.

4 Experiments

We evaluate training/validation performance on the baseline and dual-channel models. Figure 1 combines the baseline training loss curves (left) and validation metrics (right). Despite signs of steady improvement, the model fails to generalize to novel shape-color combinations. Further, Figure 2 presents an analogous comparison for the dual-channel variant. Validation metrics briefly approach near-perfect accuracy, yet consistent test failures show the learned representations rely on overly simplistic cues. Additional ablation studies are provided in the appendix.

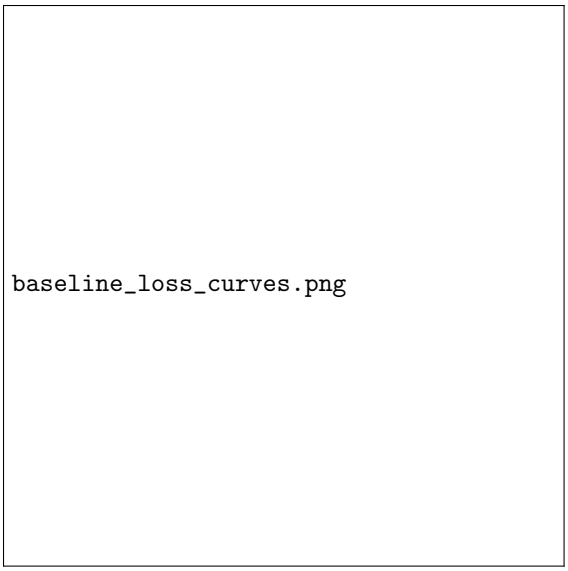


Figure 1: **Baseline results.** (Left) Baseline training and validation losses over time. (Right) Validation metrics under different hyperparameters. Despite apparently smooth convergence, more diverse test sets reveal pitfalls.

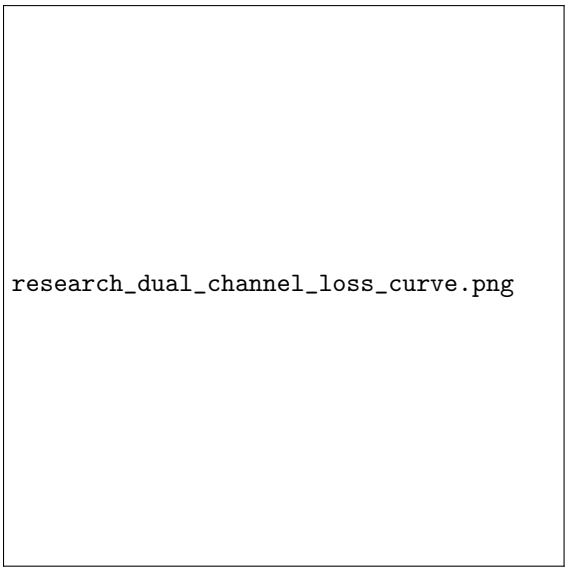
5 Conclusion

We highlight the discrepancy between validation metrics and genuine generalization in color-shape learning. Our analysis shows how strong validation results can be illusory, prompting practitioners to incorporate broader test scenarios in training pipelines. Future work could explore more robust data augmentation and refined network architectures to avoid these pitfalls.

A Additional Ablation Studies

Confusion matrices and ablation experiments detail how removing shape or color inputs impacts performance. See the supplementary figures for visualization of the shape-only model and further breakdowns of training stability. While some ablations partly alleviate overfitting, they still fail to resolve the underlying mismatch between validation and real-world performance.

References



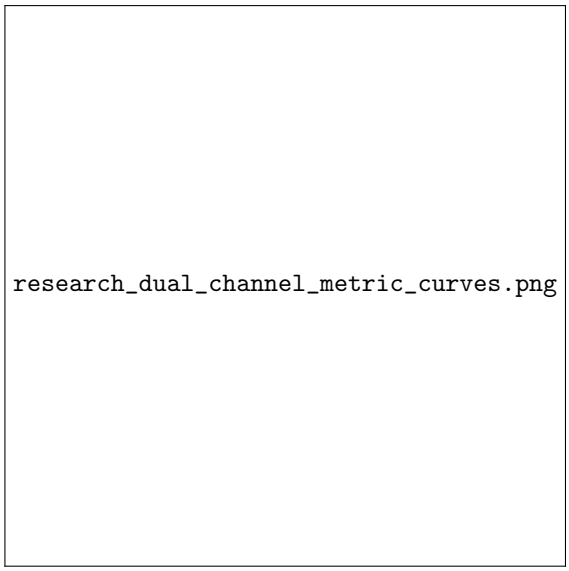


Figure 2: **Dual-channel results.** (Left) Loss curves signifying rapid overfitting. (Right) Validation accuracy peaks before dropping, suggesting a potential mismatch between validation distributions and real-world conditions.