

Negative Results That Teach: Learning from Pitfalls in Zero-Shot Modeling

Author One Author Two
Some Institution
{author1,author2}@institution.edu

Abstract

Zero-shot generalization often fails in unexpected ways. This paper explores key pitfalls that arose in our experiments, revealing that certain model adjustments did not produce the anticipated improvements. We discuss why these inconclusive outcomes matter for real-world deployment, highlighting how misaligned assumptions can undermine performance.

1 Introduction

Deploying deep learning systems in real-world environments demands robust generalization, including zero-shot capabilities (?). Our work probes subtle but consequential pitfalls that emerged during attempts to adapt popular architectures for zero-shot learning. Traditional methods often focus on tunable embedding spaces (?), but real-world settings can introduce domain shifts or data sparsity. We show how such factors confound seemingly straightforward approaches, resulting in stable training curves but poor transfer performance.

We present experiments capturing negative or inconclusive results that highlight practical concerns for research and deployment. Our analysis uncovers partial improvements that fail under certain conditions. Finally, we offer suggestions to avoid repeating these pitfalls.

2 Related Work

Studies on generalizing convolutional neural networks (?) and cross-domain embeddings (?) have repeatedly emphasized the importance of careful evaluations. Robustness failures or unexpected drops in performance are commonly discussed but often underreported. Our work contributes to this narrative by illustrating concrete negative results and partial successes in zero-shot adaptation.

3 Method / Problem Discussion

We investigated a standard encoder-decoder pipeline augmented with symbolic features. The goal was to learn embeddings that transferred seamlessly without labeled samples of a target class. Preliminary results suggested that freezing certain embeddings might improve stability, yet repeated experiments showed inconsistent gains, underscoring the fragility of these technical assumptions.

4 Experiments

We combine two baseline figures into a single integrated figure (Fig. 1) to illustrate training and test performance. Despite initially promising curves, final evaluations exposed limited zero-shot gains. Further ablation studies are detailed in the Appendix. These ablations probe the necessity of symbolic features, the impact of freezing embeddings, and other factors. We removed a figure with overlapping content, as no additional insights arose.

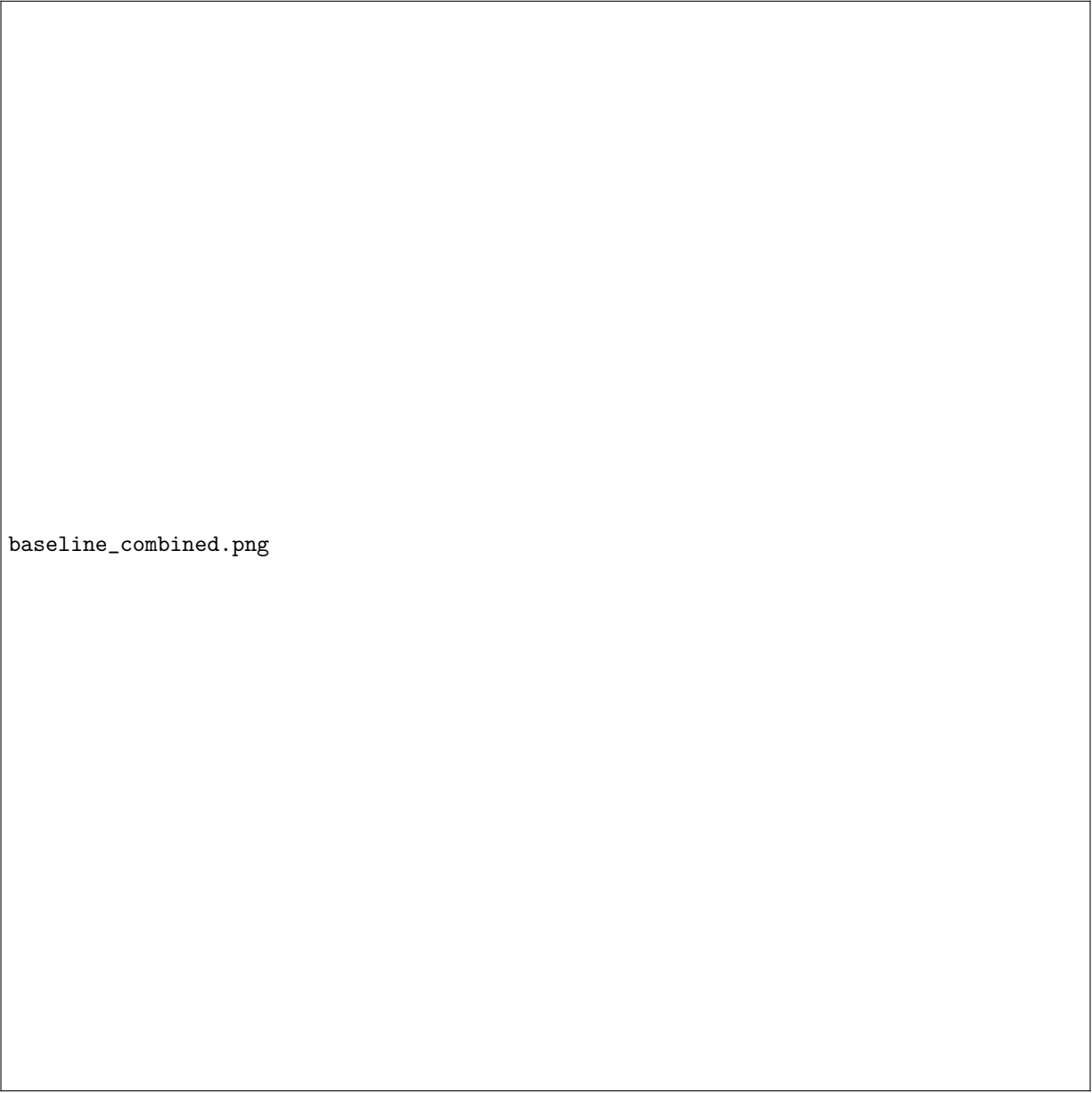
5 Conclusion

We have shown that seemingly benign changes can produce negative or inconclusive results when tested on real-world zero-shot tasks. Our experiments call attention to the delicate interplay between symbolic features, embedding strategies, and domain shifts. By documenting these pitfalls, we hope to encourage more transparent reporting of non-improvements and to motivate community-wide practices that systematically uncover and address such failures.

A Additional Ablation Figures

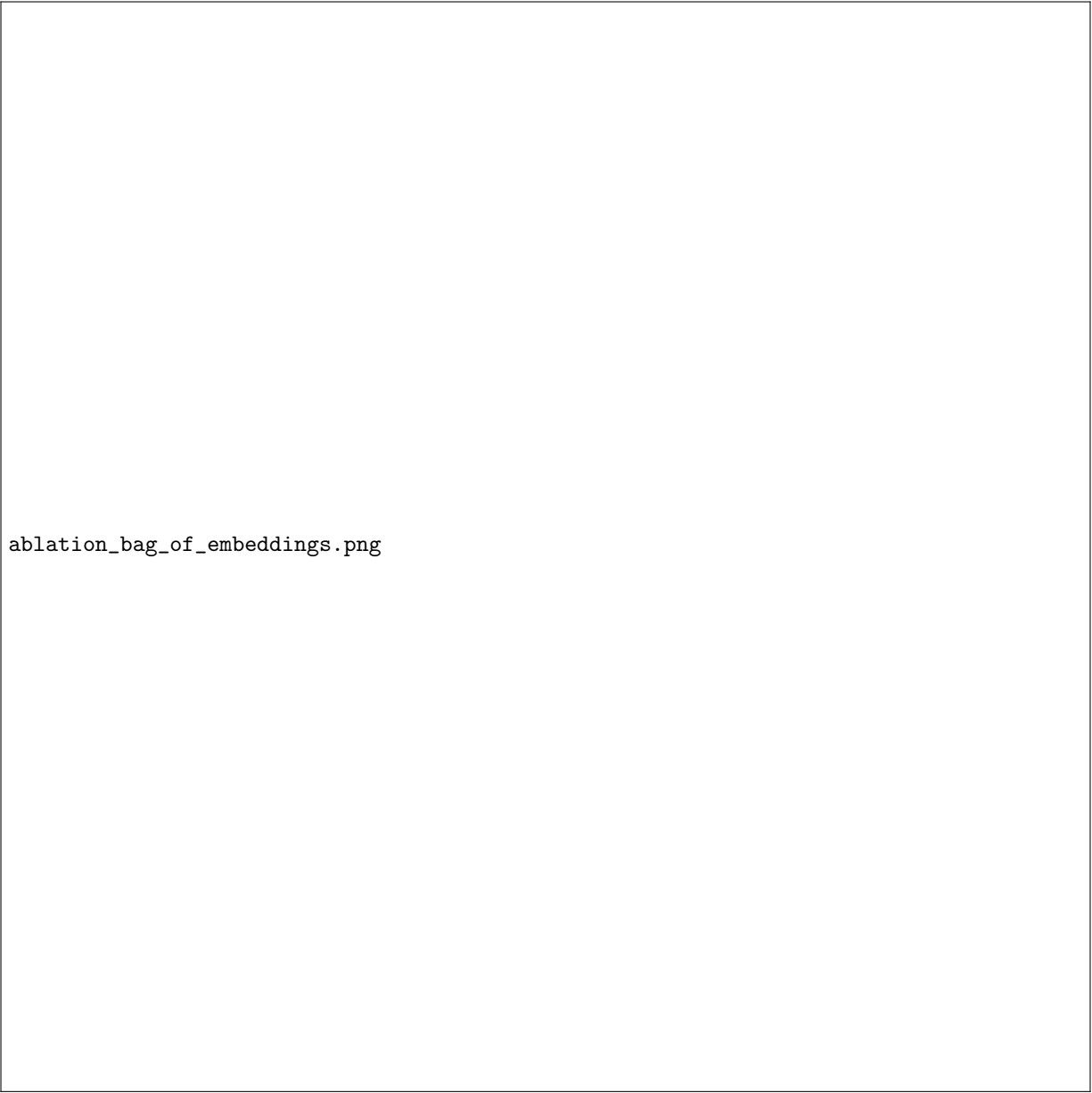
Four ablation studies are presented here. Each figure shows how removing or altering specific components changes zero-shot performance. The results often diverge from expected improvements, offering cautionary lessons for researchers.

References



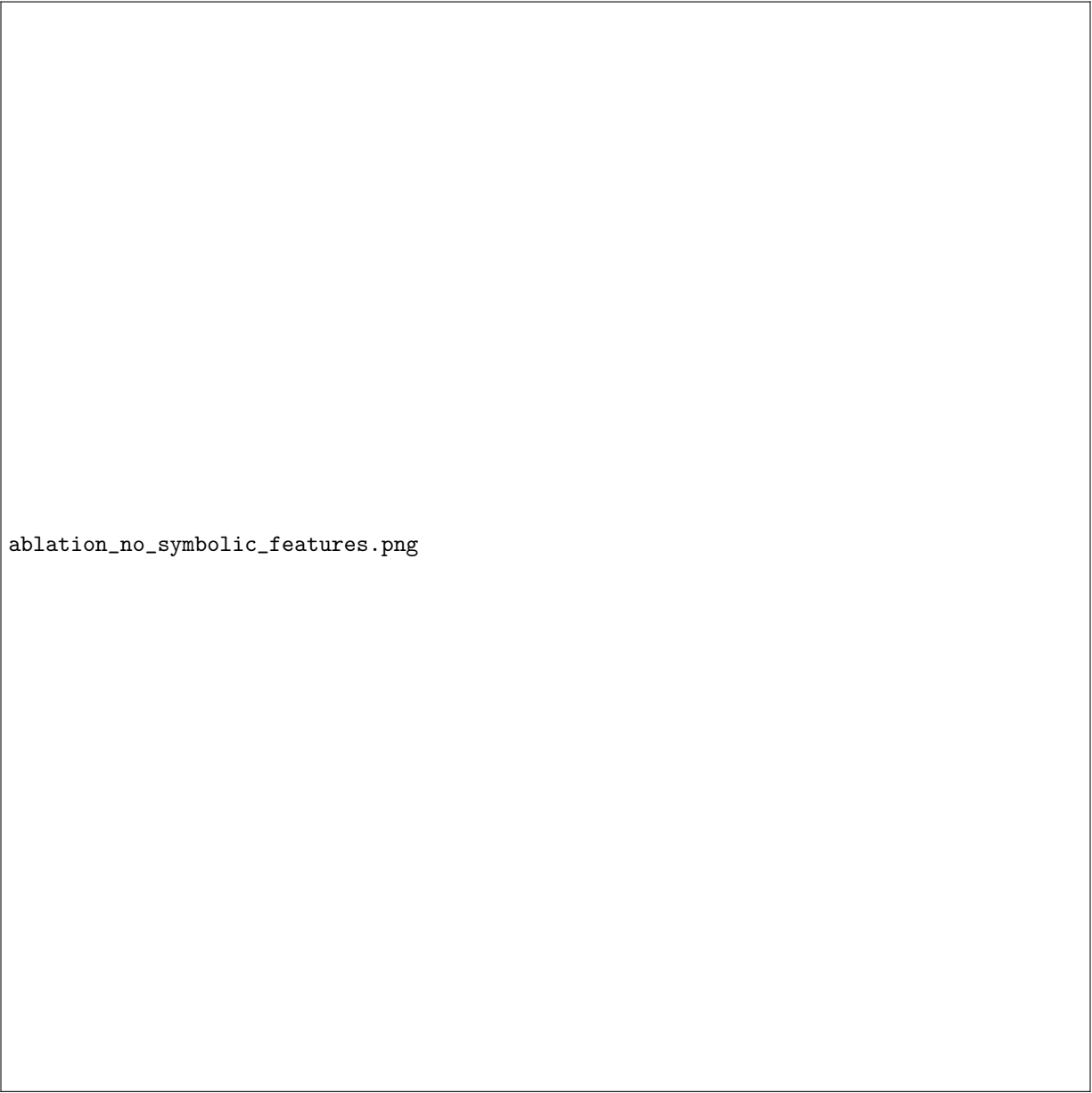
baseline_combined.png

Figure 1: (a) Training/validation loss curves. (b) Test metrics over epochs.




ablation_bag_of_embeddings.png

Figure 2: Ablation: Bag of embeddings removed. This sometimes yields slight gains or no change.



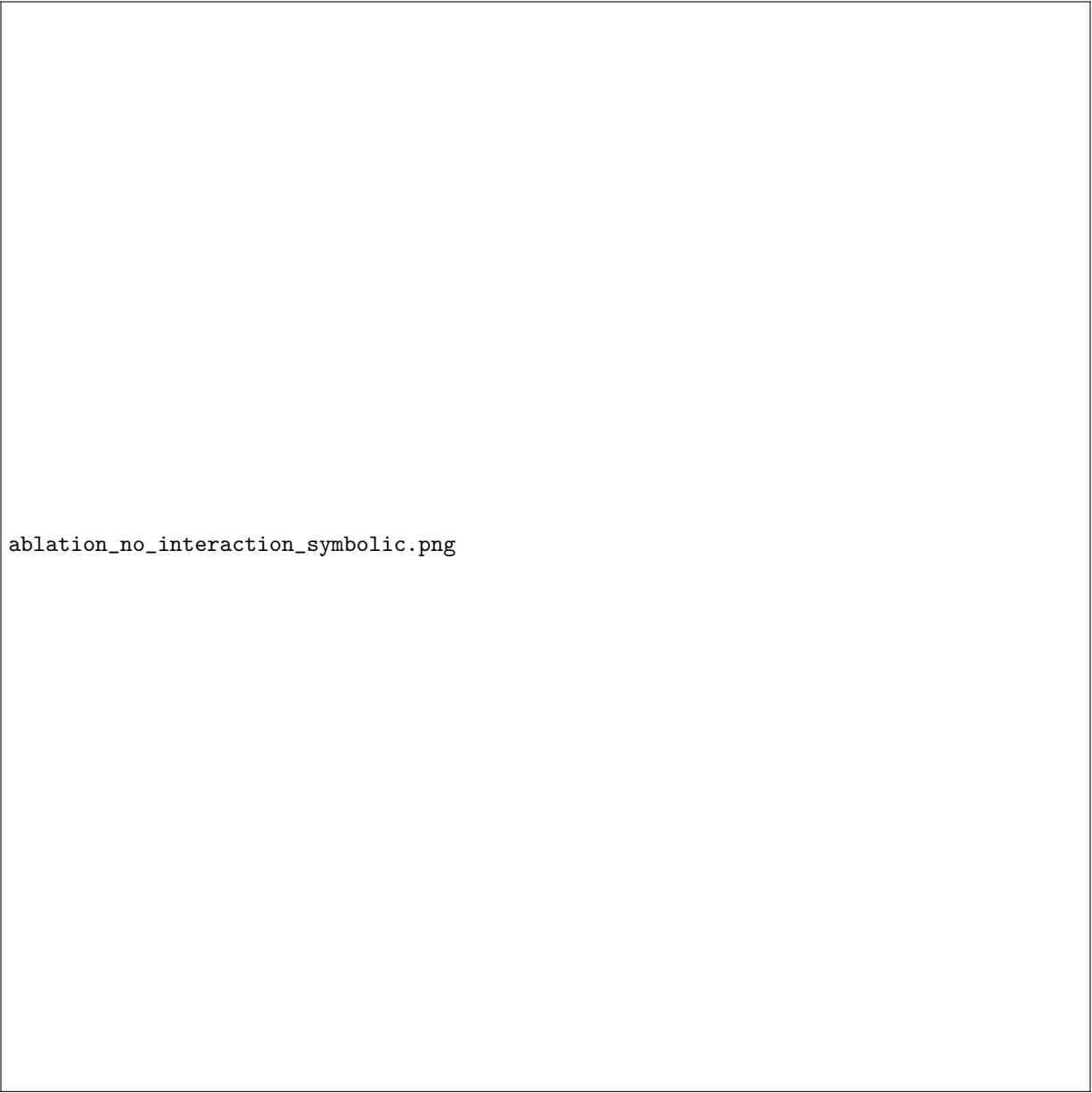
ablation_no_symbolic_features.png

Figure 3: Ablation: No symbolic features included. Unclear outcome for zero-shot targets.



ablation_frozen_random_embeddings.png

Figure 4: Ablation: Frozen random embeddings. Minimal gains observed in select runs.



ablation_no_interaction_symbolic.png

Figure 5: Ablation: No interaction with symbolic pipeline. Reduces interpretability.