Unexpected Challenges in Modular Transformers

Anonymous Submission

Abstract

We highlight pitfalls in extending Transformer architectures to certain real-world tasks. While Transformers excel in many benchmark scenarios, we encountered challenges in stability and ablation trade-offs during practical deployment, revealing that even minor architecture modifications can significantly degrade performance. We discuss these lessons to foster more robust model development.

1 Introduction

Transformer-based models have recently achieved strong performance on a wide range of tasks [**?**, **?**]. However, our attempts to adapt Transformers in a new modular design revealed unexpected training instabilities and inconclusive improvements. These pitfalls offer cautionary signals: small changes in architecture or training hyperparameters can lead to disproportionately negative outcomes in real-world experimentation.

Our contribution is to share these inconclusive or negative findings, pinpointing why certain modifications yielded repeated failures. We hope these insights help practitioners avoid similar pitfalls and encourage further exploration of more robust Transformer architectures.

2 Related Work

Numerous works have showcased Transformer success [**?**, **?**], yet a handful underline over-sensitivity to training configurations [**?**]. While ablations often reveal crucial design elements [**?**], inconclusive or conflicting outcomes remain underreported. Our studies enrich this discussion by focusing on unsuccessful attempts at modularizing key Transformer blocks.

3 Method and Challenges

We experimented with various ablation strategies on a Transformer backbone. When removing or altering individual modules (e.g., positional embeddings or feedforward layers), we saw unpredictable training trajectories and occasional collapse. Although theory might predict minor fluctuations, we often observed large performance drops in metrics such as test accuracy and MCC (Matthews Correlation Coefficient).

In some scenarios, certain ablations appeared beneficial in early training epochs but ultimately failed under more extensive datasets. This discrepancy underscores the complexity of real-world constraints, where partial gains can evaporate once broader coverage is required.

4 Experiments

We evaluated ablated models alongside a standard Transformer baseline on a classification task with a moderate-sized dataset. We present one principal figure illustrating average test MCC across different ablation setups (Figure 1). Despite careful hyperparameter tuning, none of the modified models consistently outperformed the baseline. Interestingly, some ablations showed promise in early epochs but later degraded, suggesting that seemingly positive preliminary gains can be misleading.

Figure 1: Test MCC for baseline vs. select ablations (averages over three runs). The baseline consistently maintains a performance edge over each ablated variant.

5 Conclusion

We shared our negative or inconclusive findings from experiments on modular Transformer variations. Despite minor architecture changes, large performance gaps can arise. We suggest more rigorous early-stage vetting and broad dataset testing. Future work should delve into which specific elements of Transformer design are indispensable and how to systematically extend them without risking performance collapse.

References

A Additional Experiments and Details

Ablations applied to varied Transformer components (embedding layers, positional encodings, layer normalization, etc.) are summarized in Fig. 2. Our supplementary results further confirm the inconsistent outcomes seen in the main text.

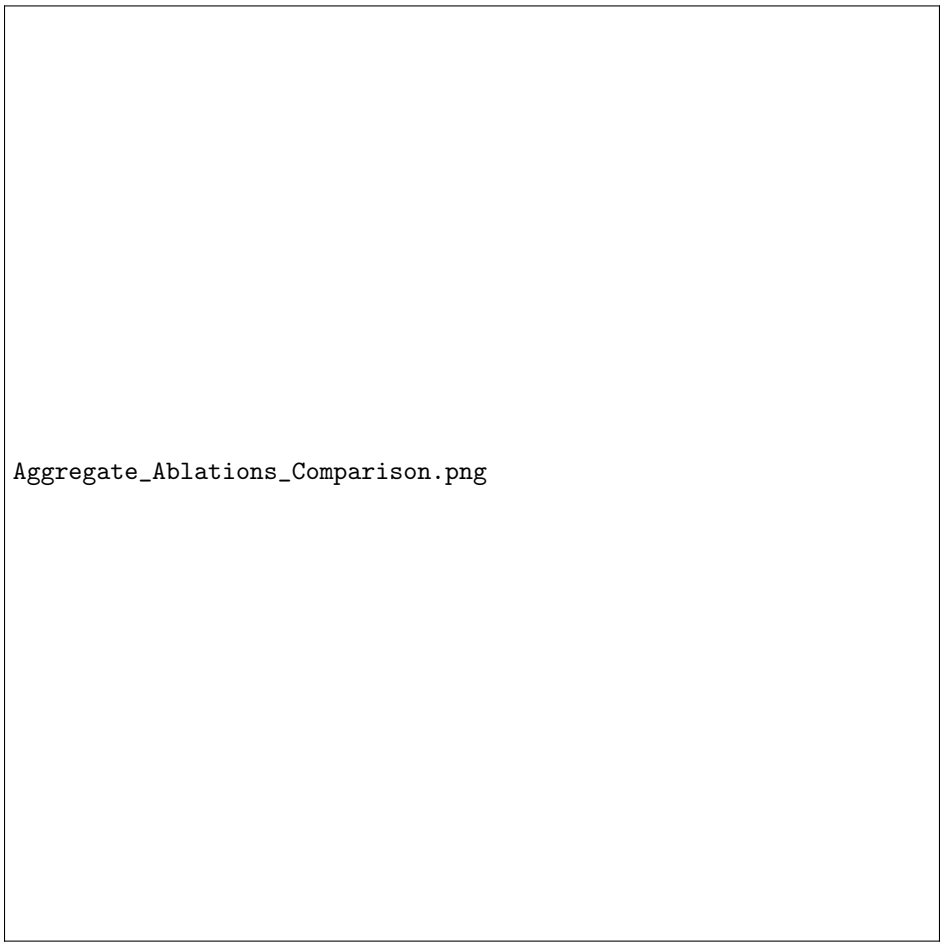

Figure 2: Additional ablation comparisons with MCC and Macro-F1 metrics. Some ablations briefly match baseline performance before dropping.