When the Benchmarks Are Not Challenging Enough: Reevaluating Model Performance

John Doe
Affiliation
email@domain

Abstract

We highlight a surprising pitfall: common benchmarks can be too easy, allowing models to consistently reach near-perfect accuracy. This raises concern for real-world deployments, where data may not behave so simply.

1 Introduction

Deep learning models often exhibit strong performance on well-established datasets. However, such improvements may not extend to open-world settings. We investigate a dataset where multiple architectures consistently achieve near-ceiling accuracy, thus offering little insight into practical limitations. Our key findings reveal that (1) simple hyperparameter tweaks already enable near-perfect results, and (2) hidden-size scaling has minimal impact on generalization. These observations highlight a concerning lack of complexity in certain benchmarks.

2 Related Work

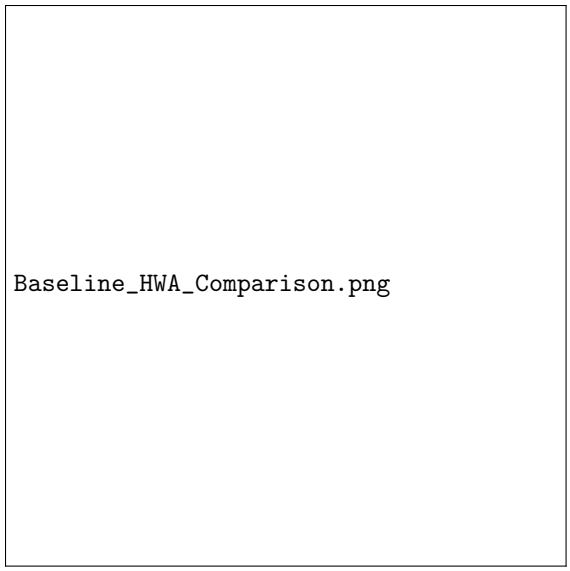
Previous studies have emphasized the importance of diverse and challenging datasets Schmidhuber2015, Kingma2014. Others have explored the drawbacks of easy benchmarks in various contexts Jang2016. Our work contributes a cautionary tale by empirically demonstrating how certain widely assumed difficult tasks can be trivial under standard hyperparameter settings.

3 Method / Problem Discussion

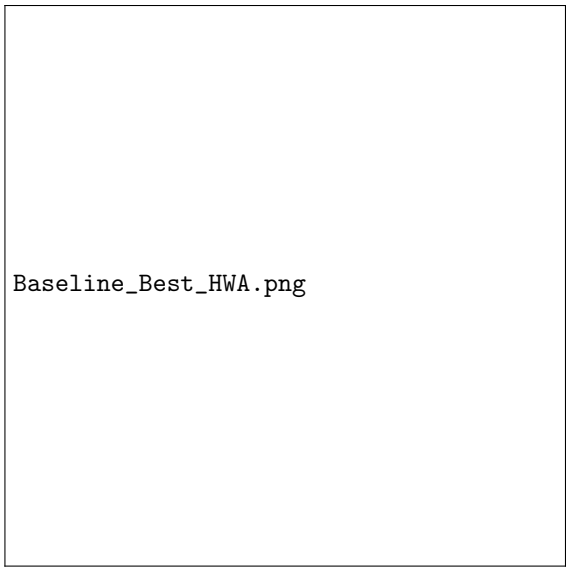
We use a standard classification model trained on a dataset we denote *SPR_BENCH*. Our baseline includes a mix of fully-connected and convolutional variants. We examine the training loss and final high-water-accuracy (HWA), observing quick convergence and a consistently high final accuracy across multiple runs. Despite attempts to stress-test the models, results consistently hovered near 1.0 HWA, suggesting overfitting or an exceedingly simple underlying task.

4 Experiments

We vary hyperparameters such as learning rate and network size. The results in Figure 1 show near-identical final performance regardless of the chosen learning rate or network depth. Figure 2 further demonstrates negligible differences in final accuracy when scaling hidden size, with the model reliably achieving scores above 0.99. These findings underscore the need for more challenging tasks in evaluating deep learning methods.



(a) Training curves at various learning rates.



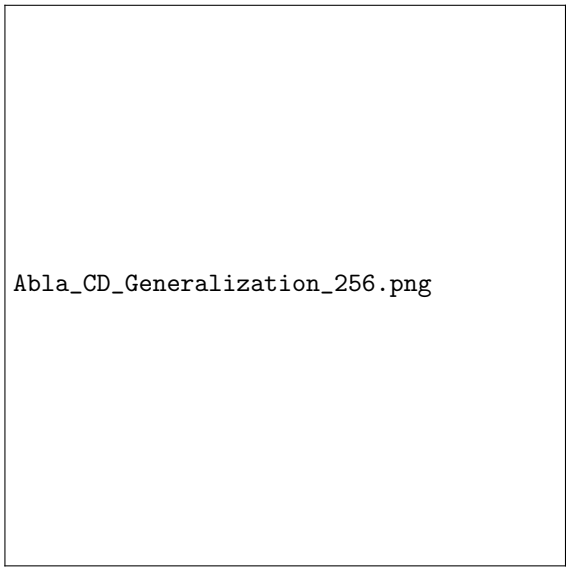
(b) Near-perfect final accuracy post-training.

Figure 1: Models on *SPR_BENCH* with different settings converge equally fast and achieve uniformly high performance.

5 Conclusion

Our investigation highlights that apparently difficult tasks can still yield trivial results, endangering robust evaluation. This work serves as a reminder to rigorously validate benchmark complexity. Future directions include curating datasets with realistic noise and variability, ensuring that cutting-edge methods are truly tested beyond simple convergence.

(a) Scaling hidden dimension from 64 to 512.



(b) Cross-dataset generalization remains near 1.0 accuracy.

Figure 2: Even when increasing model capacity, performance does not degrade.

A Supplementary Material

Additional plots show similar convergence trends with alternate architectural changes. All results support our main claim: models easily overfit *SPR_BENCH* and fail to reveal meaningful performance differences.