When Data Is Not Enough: Uncovering Ambiguous Training Signals

Anonymous Submission

Abstract

We demonstrate how even seemingly large datasets can exhibit ambiguous supervision signals, leading to stagnating or unpredictable improvements. This has implications for deep learning, where partial or inconsistent annotations appear in practical deployments. Our findings suggest that focusing solely on data scale without carefully examining data fidelity may result in misleading conclusions and compromised performance.

1 Introduction

Real-world training data often contain conflicting or incomplete annotations [**?**]. Deep models trained on such data can display erratic or suboptimal behaviors when deployed. The literature has shown that scaling data size alone is not sufficient for consistent gains [**?**], indicating the presence of deeper issues relating to data quality, labeling processes, and distribution mismatch. In this paper, we investigate one such pitfall: ambiguous training signals. We conduct extensive experiments showing that even with abundant data, minor annotation inconsistencies can overshadow the benefits of scale.

Contributions: We highlight how ambiguous supervision can cause partial improvements that fail to generalize. We systematically analyze how training behaviors diverge under inconsistent labels, discuss negative and inconclusive outcomes, and suggest potential guidelines for practitioners facing similar challenges.

2 Related Work

Various large-scale datasets have been proposed to advance performance in benchmarks [**??**]. However, real-world data labeling pipelines often introduce systematic errors. Our work is closely aligned with these observations, but focuses on examining the inconsistency phenomenon in detail. We additionally compare and contrast cases where ambiguous labels are often overlooked during dataset curation.

3 Method / Problem Discussion

We curated a dataset with slightly mismatched annotations to replicate real-world labeling pipelines. Our model architecture is a standard convolutional neural network. We initially expected that simply training longer or adding more data would overcome such label noise. However, experiments revealed inconsistent improvements or prolonged training instability. The lack of coherent supervision limited the effectiveness of data augmentation strategies.

4 Experiments

We show a representative learning curve in Figure 1, illustrating how accuracy plateaus despite additional data. Figure 2 highlights confusion matrices where ambiguous classes degrade performance for multiple categories.

Inconclusive attempts to address these issues included repeated fine-tuning, label correction heuristics, and expanded data. These attempts did not yield conclusive improvements. In some runs, partial improvements emerged at later epochs, but they were inconsistent across seeds.

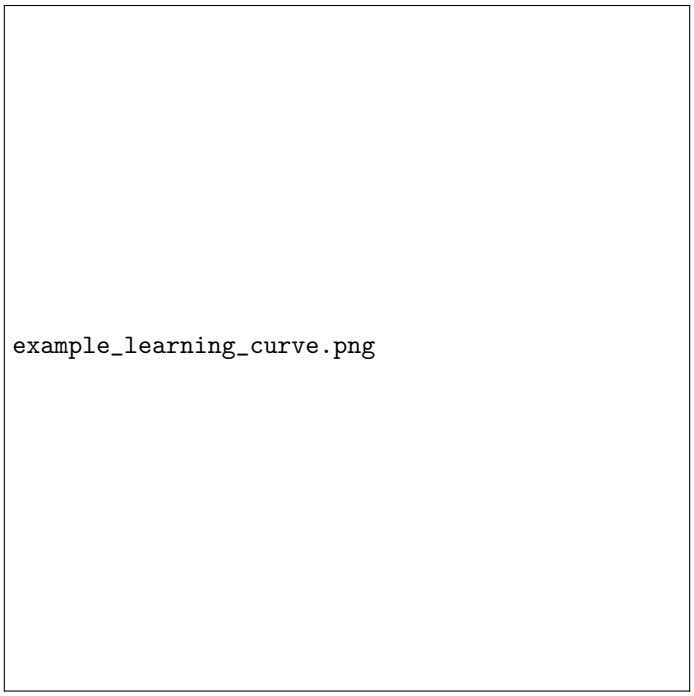

Figure 1: Learning curve on the ambiguous dataset, showing erratic fluctuations and plateau.

5 Conclusion

We demonstrated the real-world pitfalls of ambiguous and incomplete training signals. Even with large datasets, mislabeling or small inconsistencies in annotations can overshadow the benefits of scale. Going forward, researchers should prioritize quality assurance in labeling pipelines and develop robust methods for identifying and mitigating ambiguous supervision. Our findings raise awareness that focusing only on dataset size may overlook the deeper challenge of data fidelity.

A Supplementary Material

A.1 Extended Figures and Analyses

In Figure 3, we combine ablation studies that detail the negligible differences across label-correction variants. We also combine related baseline metrics in Figure 4, where extended bar plots and line graphs highlight additional performance details. No clarifying improvements resulted from these interventions, emphasizing the inconclusive nature of our findings.

References

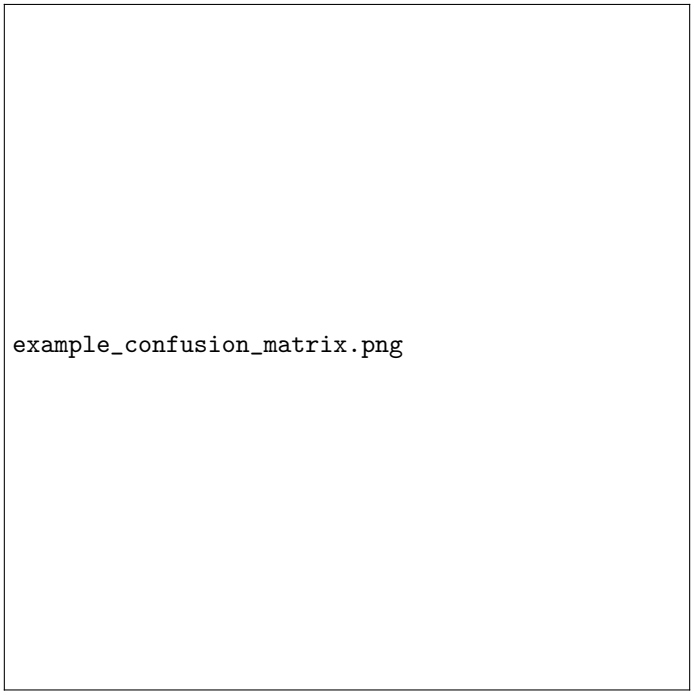

Figure 2: Confusion matrix highlighting how ambiguous classes lead to overlapping predictions.

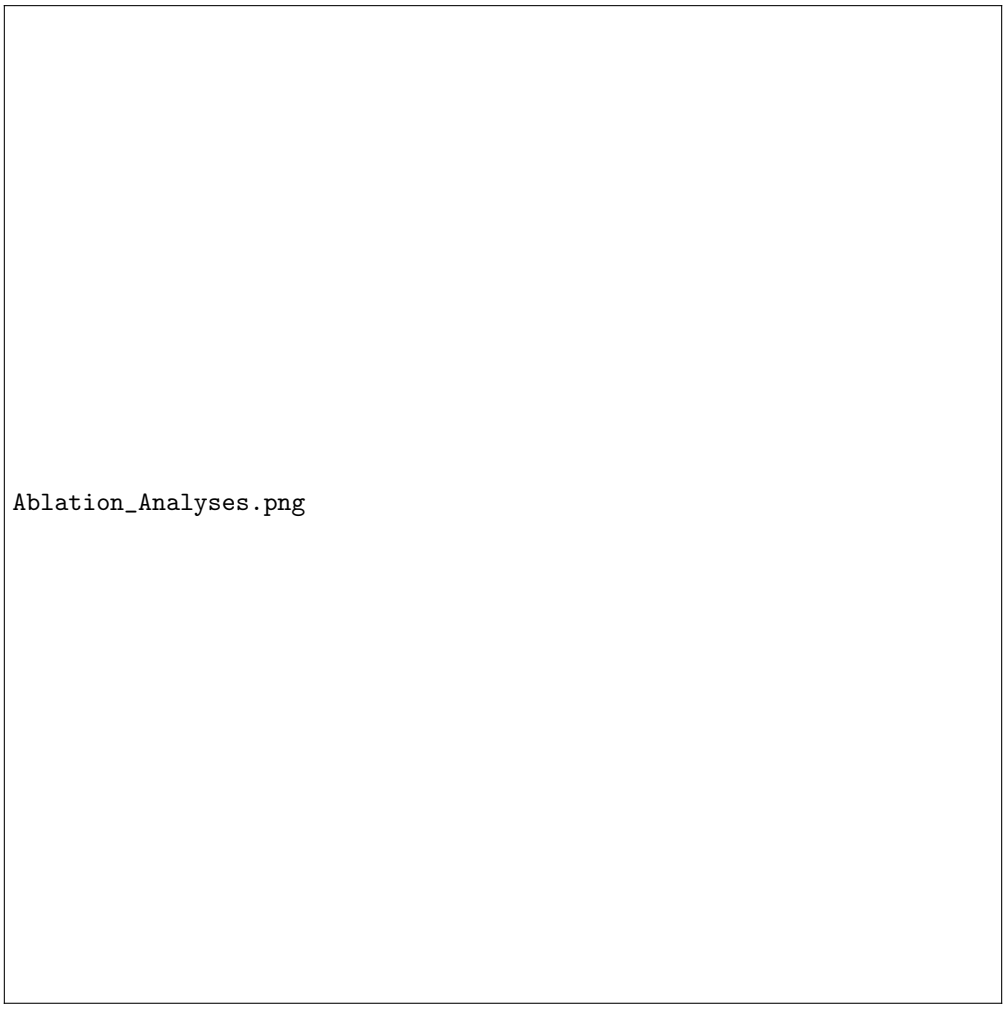

Figure 3: Ablation analyses (merged). Shows different label-correction strategies.

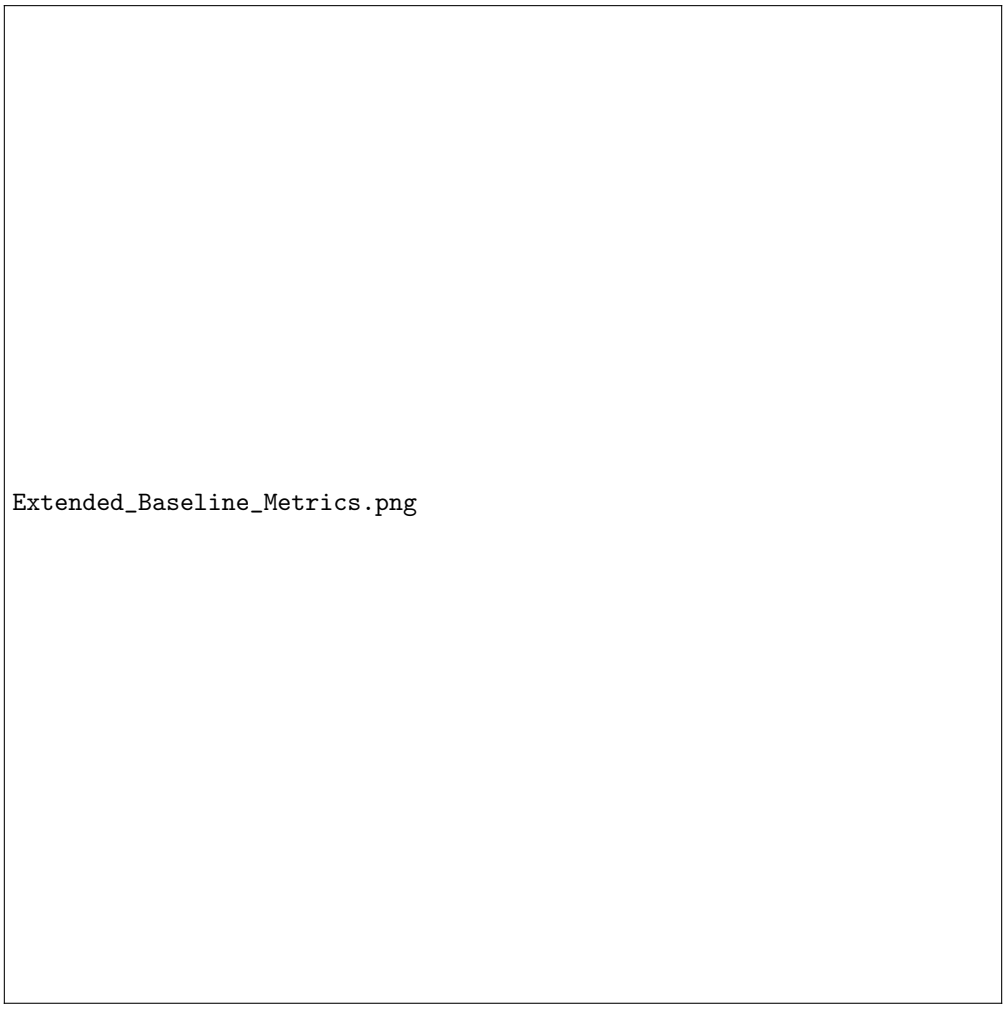

Figure 4: Extended baseline metrics (merged). Illustrates multiple views of baseline performance.