

Reliable Distillation Under Distribution Shift

Imane Farhat
Applied ML Lab
imane@aml.example

Abstract

We present a compact conference-paper scaffold for machine learning manuscripts with emphasis on reproducibility reporting, ablation structure, and concise experimental claims.

1 Introduction

Recent deployment settings reveal strong train-test mismatch. This paper studies distillation methods that remain stable when data drift and noisy labels co-occur.

2 Method

Given teacher logits z_t and student logits z_s , we optimize

$$\mathcal{L} = \alpha \mathcal{L}_{\text{CE}} + (1 - \alpha) T^2 \mathcal{L}_{\text{KL}}(\sigma(z_t/T), \sigma(z_s/T)).$$

We further add confidence gating for uncertain pseudo-labels.

3 Experiments

3.1 Setup

Datasets include two synthetic shifts and one real-world benchmark. Metrics are top-1 accuracy and expected calibration error.

3.2 Results

Model	Accuracy	ECE
Teacher baseline	81.7	0.082
Vanilla distillation	83.1	0.071
Proposed method	85.4	0.049

Table 1: Illustrative performance summary.

4 Discussion

Observed gains come primarily from lower-confidence regions, suggesting that gating mitigates harmful teacher overconfidence.

5 Conclusion

The scaffold supports clear conference submissions and can be adapted to different review templates while preserving technical structure.

References

1. Example reference A.
2. Example reference B.