IDEAS Initial Project Proposal

Shay Manor

Guardrail Distillation of a Large Vision Teacher to an Edge-Deployed Student

Idea

Train a small real-time student model from a large accurate teacher with guardrails to flag when the student is wrong compared to the teacher in cases where the teacher is correct. This will find a measure to find the benefit of running the teacher model over the student model (most cases this is low but in edge cases, this would be higher). The guardrail outputs an image risk score and per-pixel failure heatmap to trigger fallback.

Problem

Distillation optimizes for average cases, but safety-critical failures are often edge cases (night, motion blur, glare). Standard confidence heuristics often fail under a distribution shift (i.e., change in brightness, location, camera, etc.) so an edge model may be confidently wrong.

Goal

Add a cheap test-time trust signal to reduce unsafe silent failures without running the teacher model for inference.

Prior Research

1. **KD:** compress teacher into students with soft targets (Hinton et al., 2015)
2. **Structured KD for segmentation:** distill structured relations for dense predictions (Liu et al., CVPR 2019)
3. **Confidence/Uncertainty:** MSP, Calibration, MC dropout, etc. are often wrong under distribution shift

Base Paper: Structured Knowledge Distillation for Semantic Segmentation (Liu et al., CVPR 2019) to train a compact student.
New Contribution: add the grounded guardrail (trained offline) to predict cases where student is wrong (compared to the teacher) on-device. At runtime, this allows for controllable risk-coverage trade-offs and gives a failure heatmap without the teacher.

Hypothesis:

1. Better safety (better AURC) for the same compute compared to MSP/entropy/temperature scaling. Similar performance to MC Dropout but significantly cheaper in compute.
2. More robust under distribution shift (e.g., night, fog, rain)

Novelty:

The teacher-grounded supervision is the novelty as predicting when the student is more wrong than the teacher is expected to outperform simply predicting when the student is wrong or other uncertainty strategies.

Implementation Strategy

1. Task: front-camera drivable-area segmentation and obstacle detection (Cityscapes dataset in-domain, ACDC and Dark Zurich for domain shift)
2. Baseline: student with no KD, student with KD, student with structured KD, MSP, temperature scaling, MC Dropout, etc.
3. Teacher: high-capacity segmentation model. Students: 3–5 small architectures of varying size.
4. Train a guardrail network to predict the teacher–student risk gap and output a per-pixel failure heatmap, using only student inputs and image features.
5. At runtime, if the predicted gap exceeds a threshold, flag the frame and trigger a fallback policy.

Related Work

- ShiftKD: https://arxiv.org/abs/2312.16242
- Failure Prediction by Learning Confidence: https://arxiv.org/abs/1910.04851
- Structured Knowledge Distillation for Semantic Segmentation: https://openaccess.thecvf.com/content_CVPR_2019/papers/Liu_Structured_Knowledge_Distillation_for_Semantic_Segmentation_CVPR_2019_paper.pdf