Abstract:
Computer vision has had tremendous success recently in perception tasks, especially in cloud applications. An embodied system brings the additional objectives of memory, compute and energy optimization in addition to system accuracy. As opposed to processing in the cloud, an embodied agent has to perform in the wild without access to additional infrastructure. In this thesis, we address four major limitations of visual perception in embodied systems.

From Images to Videos: When dealing with image streams, there is often redundancy in the temporal neighborhood since objects are often repeated. Thus, a frame-level analysis leans towards the objects present in a larger number of frames rather than a fair evaluation for all objects in a video. To mitigate this, we propose an evaluation metric that clubs together different appearances of each object from multiple frames. In our work, we show how this results in metrics that are sensitive to bias against any objects in the videos.

Resource Constrained Object Detection: When object detectors run at different speeds while processing an image stream, a number of frames have to be skipped to maintain real-time operation. Thus, a faster but less accurate detector may be able to process more frames than a slower detector which may be more accurate at the frame level. It is unclear, however, which of these object detectors would be able to detect more objects. This shows the dependence of object detection performance on the speed at which the algorithms are run. In this work, we propose a resource-aware evaluation metric that jointly reasons about speed and accuracy of object detectors for a true evaluation of object detection performance in an embodied system.

Generalization to Unseen Environments: Owing to the large number of parameters in deep neural networks, models often over-fit to the data distribution on which the models are trained. Even as perception improves on a certain dataset, the test environment changes continuously for an embodied agent. There is often limited visibility into the test data distribution in a deployed system. Thus, one has to ensure that the perception module in an embodied agent is aware and adaptive of this shift. We propose a method to use sparse data from an alternate modality (LiDAR) in order to adapt a pre-trained depth prediction network to the current context and provide accurate depth even in unseen environments.

Energy Efficiency with Multimodal Sensing: Visual perception is an energy intensive task. Even human beings rely on their model of the environment or alternate senses for perception, e.g., using sound as a cue for visual engagement while resting. In this work, we observe a similar phenomenon in a multimodal obstacle detection system. We show how a low power ultrasonic sensor is able to provide a low energy sensing mechanism to trigger visual perception. We take a detailed look at the available runtime modes for an obstacle detection system and perform an accuracy energy tradeoff to demonstrate different possibilities.