Neuro-Bayesian Methods for Realtime Scene Perception

by

Austin J. Garrett

Department of Electrical Engineering and Computer Science

Proposal for Thesis Research in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

December 2019

Author .....
Department of Electrical Engineering and Computer Science
December 11, 2019

Certified by .....
Vikash Mansinghka
Research Scientist
Thesis Supervisor

Accepted by .....
Placeholder
Chairman, Department Committee on Graduate Theses

Neuro-Bayesian Methods for Realtime Scene Perception

by

Austin J. Garrett

Submitted to the Department of Electrical Engineering and Computer Science
on December 11, 2019, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

In this thesis, I consider Bayesian approaches to doing inference over structured scenes. This includes the design and implementation of a probabilistic language for representing semantic relations between multiple objects. I further consider a strategy for integrating neural techniques into MCMC-based techniques for inverse graphics, where a generative model is constructed to predict the behavior of neural detectors, allowing for error correction and semantic structuring of the output. I finally discuss the development of a framework for realtime particle filter based inference based on the Robot Operating System and Gen probabilistic programming language, as well as its integration into the broader Cora project for probabilistic modeling of intuitive physics.

Thesis Supervisor: Vikash Mansinghka
Title: Research Scientist

Chapter 1

Thesis Proposal

1.1 Introduction

The past decade has seen a flourishing of highly powerful computer vision techniques, due to the success of deep learning techniques accelerated on massively parallel hardware.

Yet the shortcomings of neural techniques have only grown increasingly urgent in proportion to their dominance. While these approaches have resulted in the amortization of traditionally very difficult and high-dimensional computer vision problems, the lack of semantic structuring in their output has led to great challenges in their deployment in practical domains where interpretability and composability are desperately needed to integrate these components into larger frameworks.

Meanwhile probabilistic techniques offer a more principled approach to modeling structured data, and have seen a rapid growth in the collective ecosystem. Such environments are desirable for their expressibility, corresponding to the fact that they often represent intuitive models of rationality inspired from cognitive and neuroscience which increasingly rely on such modeling to explain the human reasoning.

DATASET??

This part is a highly relevent research goal, due to the undervaluing of semantic structuring of information by the neural network community. In some sense we are trying to provide a solution to a problem that is not fully visible to the community, or

at the very least is thought to be "too hard" for current techniques. Hard grounding metrics, or at least a fuller range of explicitly specified common-sense tasks will go a long way in changing the over-skeptical perspective toward Bayesian techniques trying to answer this question. Thus, an important part of the thesis work will be identifying the gap that exists between types of problems that we'd like to solve (common sense reasoning), and how neural networks currently fail to address these problems, as well as providing as much of a solution as we can in the time available, which we believe will be compelling enough to demonstrate the utility of Bayesian methods in modern CV pipelines.

1.2 Related Works

1.2.1 Neural Approaches

TODO: replace me

Deep Object Pose Estimation

TODO: replace me[?]

PoseCNN

TODO: replace me

1.2.2 Inverse Graphics

TODO: Picture, Analysis by Synthesis, Informed Sampler

1.2.3 Bayesian Inference over Structured Scenes

Existing perspectives on Bayesian methods in vision systems operate under a fully-generative analysis-by-synthesis approach to modeling vision. This includes a full rendering pipeline and likelihood defined over

Contact Graph Representation

TODO: FIG: Contact graph representation (maybe from Ben's poster?)

Existential Doubt

TODO: FIG: NeurIPS figure

LAFI

TODO: LAFI figure

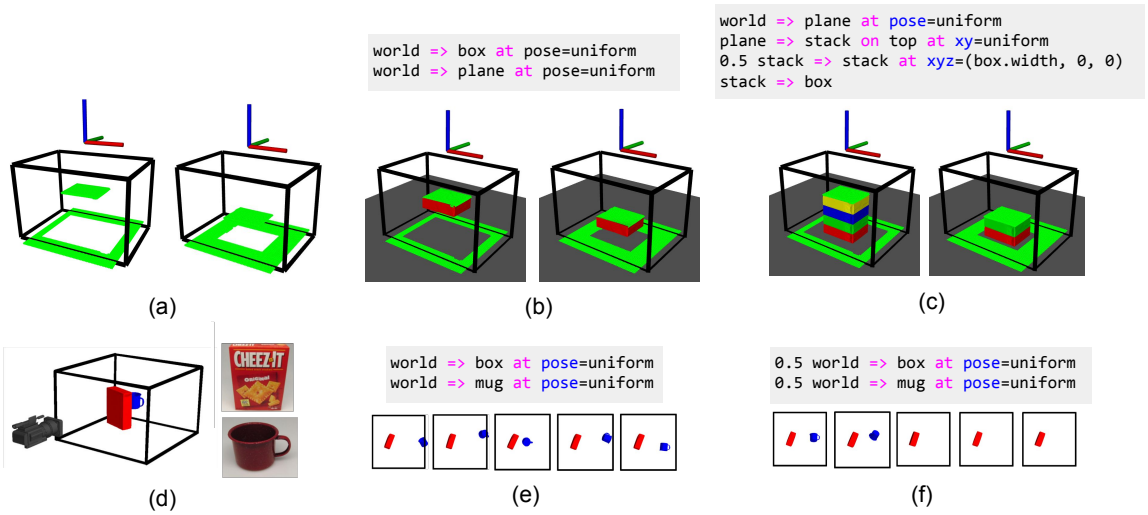

Figure 1-1: Two scenarios with prior knowledge specified as programs in our probabilistic scene description language. (a) shows two depth measurements made by a depth camera. (b) shows a program in which boxes have random poses in 3D space, and the resulting inferred pose of the box. (c) shows a program that assumes boxes are in stacks that rest on the floor, and the inferred number of boxes that explain the data for each observation. (d) shows another scenario, where a depth camera observes a box that is occluding a mug. (e) shows a program that asserts that both objects exist (but at unknown poses). The resulting inferences show the mug must exist somewhere behind the box to be consistent with this knowledge and the observation. (f) shows a program that allows for either object to not exist, and the resulting joint inferences about the mug's existence and pose.

1.3 Proposed Work

Task	Expected Completion
Development of realtime ROS-based particle filter framework with visualization	Jan 15th
Two-level neuro-predictive generative model and associated inference procedure	Jan 30th
Some project involving replacing the neural component with more distributed detectors that can be modeled probabilistically (hierarchical generalization of the research from January). This needs to be broken up into more digestible tasks, and we need consideration of exactly what form this takes.	April 30th

TODO: FIG: Bayes' net diagram SUBFIG A: static structure, dynamic parameters SUBFIG B: dynamic structure, dynamic parameters

1.3.1 Neuro-predictive Generative Modeling

Modeling a rendering pipeline in a fully Bayesian way suffers from certain computational challenges. While guaranteed to converge to a correct posterior with unlimited compute time, the non-asymptotic properties of MCMC are poorly-understood. Preliminary work in inverse graphics techniques applied to structured scenes supports the hypothesis that naive analysis-by-synthesis approaches that leverage a full rendering pipeline often fail to explore all the modes of the posterior in reasonable time bounds.

Concretely, for scene graph $\mathcal{S}$ and continuous parameterization $\vec{\nu}_{\mathcal{S}}$, A rendering-based likelihood relies on modeling the image data using a full rendering pipeline $R(\mathcal{S}, \vec{\nu}_{\mathcal{S}}) = X$, where X is a rendered image. For robustness, the likelihood is modeled not directly on X, but on a noisy function of the pixel data. The noise is modeled is a mixture of a uniform distribution on the range of possible values and a normal distribution with mean centered at the true rendered pixel value $R(\mathcal{S}, \vec{\nu}_{\mathcal{S}})$ with fixed variance σ^2, leaving the full likelihood on noisy image data Y as

$$p(Y|\mathcal{S}, \vec{\nu}_{\mathcal{S}}) = \prod_{r=1}^{R} \prod_{c=1}^{C} \left(0.1 \cdot \frac{1}{D} + 0.9 \cdot \mathcal{N}(Y_{r,c}; R(\mathcal{S}, \vec{\nu}_{\mathcal{S}})_{r,c}, \sigma) \right) \tag{1.1}$$

1.1 is very high-dimensional, making it prone to capture by local minima for very long periods of time. Neural networks empirically demonstrate strong performance at locating strong *maximum a posteriori* estimates in bounded compute time. Thus, we may consider combining neural techniques with MCMC to improve performance of inference.

One perspective may consider neural networks as amortizations of the inverse generative procedure. Consider a neural detector $\phi_{\mathcal{S}}$ that outputs poses under a given scene graph structure $\mathcal{S}$ (often a full unconstrained 6D pose), such that $\phi_{\mathcal{S}}(R(\mathcal{S}, \vec{\nu}_{\mathcal{S}})) = \vec{\nu}_{\mathcal{S}}$. Given image data Y, we can use the corresponding estimate $\phi_{\mathcal{S}}(Y)$ as an initialization for some MCMC technique. However, this relies on $\phi_{\mathcal{S}}$ robustly predicting the mode of the distribution. This assumption is not guaranteed, and in fact when neural networks do fail, they often fail catastrophically, estimating wildly incorrect poses that suffer from the same convergence issues as an unlearned proposal distribution. Part of the issue lies in the fact that there this approach contains no way to reason about common failure modes of the neural detector. All initializations are treated as equally valid,

Alternatively, we might observe that the failure modes of these neural techniques are often quite predictable. When objects are too close or too far away, at strange angles, or heavily occluded, the neural detector is much more likely to fail. In this case,

Alternatively,

$$p(\phi(Y)|\mathcal{S}, \vec{\nu}_{\mathcal{S}}) =$$

Such a model would have a likelihood defined in terms of a conditional distribution over scene parameters given a static scene structure.

Prior on scene structures

1.3.2 Fixed Scene Structure, Dynamic Parameters

The first task involved with modeling neural

- An end-to-end system that uses DOPE+Gen+ROS to track a fixed set of ob-

jects, integrating prior knowledge about temporal consistency of object trajectories via state space modeling. For this, we can use a simple particle filter over object trajectories. That will help us nail down the end-to-end integration aspect, and also serves as the minimal example of 'filtering the output of a neural network via prior knowledge'

- An end-to-end system that that uses DOPE+Gen+ROS to infer scene structure, integrating prior knowledge about scene graphs (e.g. contact relationships) as well as temporal consistency. This version will assume a static scene graph, but can be adapted into an online algorithm by running the algorithm on the past e.g. 10 time steps of data in sliding windows.

- An end-to-end system .. that includes prior knowledge about temporal consistency of scene graphs (e.g. scene graphs change relatively infrequently).

1.3.3 Dynamic Scene Structure, Dynamic Parameters

TODO: replace me

Reversible Jump MCMC

TODO: replace me

1.3.4 Engineering Infrastructure

TODO: replace me

Realtime Particle Filtering

TODO: replace me

Visualization

TODO: replace me

Cora Project

TODO: replace me

1.3.5 Extensions

TODO: replace me

1.4 Conclusion

TODO: replace me

1.4.1 Placeholder

TODO: replace me