An Intelligent Approach Inspired by Cognitive Process for General Purpose Learning Machines

Stone Fang (Student ID: 19049045)

Contents

1 Introduction

background

research question

organization of this article

2 Part A: Critical Review

This part is a review of recent studies on computational infant learning models and theories. The review is focused on visual physics intelligence such as object tracking and physical reasoning because it is the fundamental of human intelligence [citation_need]. First the main ideas of the studies are briefly described, and then a critical review will be conducted from cognitive science perspective.

2.1 Review of Recent Work

2.1.1 Visual Physics Reasoning

The paper of (Riochet et al., 2018) mainly has two parts. In the first part, the authors proposed IntPhys is a benchmark to diagnose AI systems on visual intuitive physics reasoning tasks, inspired by studies of intuitive physics on infants. In the second part, the authors proposed two unsupervised deep neural networks "infant" learning models on intuitive physics.

2.1.1.1 Diagnostic test benchmark The first part aims at answering a basic problem: evaluation. That is, how can we be sure that a system has certain level of understanding of physics? The authors argue that end-to-end visual tasks such as 3D structure recovery, object tracking, or visual question answering (VQA) are not suitable for such evaluation because of 1) dataset bias and 2) noise measure. For example, a VQA system does not perform well could be because of, not poor understanding of physics, but a bad language model. In order to solve this problem, the IntPhys benchmark is proposed, containing a set of videos as "unit tests" which are independent of any end-to-end task. The tests are organised into four categories: 1) Object permanence, 2) Shape constancy, 3) Spatio-temporal continuity, and 4) Energy/Momentum. It is inspired by the of "violation of expectation" revealed by psychologists, that is, an infant or animal will be "surprised" by visual scenes which is physically impossible. Following this idea, the "physical plausibility" score is introduced to measure the level of surprise. This score is expected to be lower on video clips violating physical laws than that comply.

2.1.1.2 Two "infant" learning models In the second part, two "infant" learning models are proposed. They are both unsupervised/self-supervised neural network models to learn intuitive physics merely from visual presentations in first-person viewpoint. For the sake of simplification, the models are not allowed to interact with the settings, though this is not the real situation of infants or animals. The first model is a Convolution Neural Network (CNN) encoder-decoder model of resnet-18[citation] pre-trained on ImageNet[citation], and the other is a conditional

Generative Adversarial Network (GAN)[citation]. The models are trained on the objective of next frame prediction, so there is no requirement for data labeling. The two models are tested on both short-term (5 frames) and long-term (35 frames) tasks. Then, the plausibility score for a frame is computed on the basis of comparison between the prediction $\hat{f}_t$ and the ground truth f_t.

2.1.2 Unsupervised Intuitive Physics

An remarkable nature of infant learning is the fact that physical representations and laws are learned without explicit supervision. To mimic this functionality on machines, a recent study of (Ehrhardt, Monszpart, Mitra, & Vedaldi, 2018) proposed an approach to predict meaningful physical parameters, such as object position and velocity, from raw visual representation of real data without any unsupervised or simulator. Instead of mere prediction of future frames, the goal of this study is to construct an intelligent agent that can learn physical states such as positions and velocities of objects, as well as physical laws enabling the prediction of changes of such states (not changes of appearance) over time.

The solution consists of two steps. The first step is to build a tracker that learns to discover objects and extract the positions. The object detection problem is formalised as to learn a function $\Phi(\mathbf{x}_t) = u_t \in \mathbb{R}^2$ where $\mathbf{x}_t \in \mathbb{R}^{H \times W \times 3}$ is an RGB video frame containing a single object and u_t is the extracted 2D position of the object. This function is learned without any label or a-prior knowledge about that object. More concretely, the function $\Phi(\mathbf{x}_t)$ is modeled as follows:

1. calculates a scalar score $f_v \in \mathbb{R}$ at each pixel v with a shallow Convolutional Neural Network (CNN), generating a heat map.
2. f_v is normalised to a probability distribution s_v by softmax function.
3. the location u is extracted as the expectation over s_v, that is, $u = \sum_v v s_v$.

Then, the algorithm to learn Φ consists of the following parts:

1. causality principle: in a video of real physical process, the trajectory of objects are causal and smooth, that is, physically plausible; on the other hand, if the frames in a video are randomly shuffled, the trajectory should be not. Therefore, a discriminator network $D(\Phi)$ is incorporated that can classify between true physical ordering of frames and random shuffled ones, which results in the classification loss of $\mathcal{L}_{disc}$.
2. equivariance principle:
3. penalty on distribution:

This tracker is scalable to large datasets and robust on a variety of objects in different kinds and shapes without further specifications.

Then, the second step is to build a predictor that can predict the positions through time by extrapolation of the positions extracted in the first step.

dataset of Roll4Real containing videos of balls rolling on different surfaces.

The research team also conducted a research to do similar unsupervised learning from past experience.

2.1.3 Make Meanings from Sensory Input

In (Evans, Hernandez-Orallo, Welbl, Kohli, & Sergot, 2019) the authors give an answer to a fundamental question concerning human and machine learning: what is the meaning of "make sense" of raw sensory input. In their study, such approach is modeled as unsupervised program synthesis.

2.1.4 Curiosity Driven & Self-aware Agents

Infants are experts at generating structured behavior from unstructured settings by playing without any supervision or even explicit external rewards. Inspired by this, a study (Haber, Mrowca, Wang, Fei-Fei, & Yamins, 2018) was conducted to model such ability with an agent which is intrinsically motivated by "implementation" of curiosity and self-awareness.

2.1.5 AlphaGo

Deep Reinforcement Learning

2.2 critical Discussion: Cognitive Perspective

This part criticize the intelligence of algorithms mentioned above from a cognitive perspective. There is no unique definition of intelligence. In this paper, an AI algorithm is not evaluated by how well it performed on certain tasks; instead, it is criticized from the cognitive viewpoint. To be precise, it will be evaluated on whether it provides insights into the mechanism of cognitive process.

// follow the approach in (Yeap, 2011) mapping be examined by three characteristics of cognitive mapping: fragmented, incomplete and imprecise.

2.2.1 Criteria from Cognitive Perspective

> *A striking property of natural intelligences is their ability to perform accurate and rapid predictions of physical phenomena using only noisy sensory inputs. Even more remarkable is the fact that such predictors are learned without explicit supervision; rather, natural intelligences induce their internal representation of physics automatically from experience. (Ehrhardt et al., 2018)*

2.2.2 Visual Physics Reasoning

In work of (Riochet et al., 2018) as introduced in section 2.1.1, two unsupervised "infant" learning models are proposed. The advantage of these models is the absence of supervision or external reward, which is more likely as the situation of real infant learning. On the other hand, these proposals are still task-oriented models, because they are trained to learn frame predication on raw pixel inputs. as a result of the "black-box" natural of neural network, it is hard to tell what is inside such model, and more essentially whether the model learns the concept of "object" and

the law of "object permanence" as humans do. On some extent, learning predictions on visual sequences is a similar task to learning a language model on raw texts as word sequences, which still have difficulties to capture the complex structures and reasoning in humans' language.

3 Part B: An Intelligent Process

Object segmentation and tracking.

3.1 Model of Cognitive Process

principles:

- similar pixels form a object
- pixels moving and changing together
- probabilistic
- knowledge
- memory
- error correction

3.2 Discussion

placeholder

Reference

Ehrhardt, S., Monszpart, A., Mitra, N. J., & Vedaldi, A. (2018). Unsupervised intuitive physics from visual observations. *CoRR*, *abs/1805.05086*. Retrieved from http://arxiv.org/abs/1805.05086

Evans, R., Hernandez-Orallo, J., Welbl, J., Kohli, P., & Sergot, M. (2019). *Making sense of sensory input*. Retrieved from http://arxiv.org/abs/1910.02227

Haber, N., Mrowca, D., Wang, S., Fei-Fei, L., & Yamins, D. L. K. (2018). Learning to play with intrinsically-motivated, self-aware agents. *Proceedings of the 32Nd international conference on neural information processing systems*, 8398–8409. Retrieved from http://dl.acm.org/citation.cfm?id=3327757.3327931

Riochet, R., Castro, M. Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., & Dupoux, E. (2018). IntPhys: A framework and benchmark for visual intuitive physics reasoning. *CoRR*, *abs/1803.07616*. Retrieved from http://arxiv.org/abs/1803.07616

Yeap, W.-K. (2011). A computational theory of human perceptual mapping. *CogSci*.