# A Computational Cognitive Model for General Purpose Autonomous Learning Machines

Stone Fang (Student ID: 19049045)

## Contents

**1 Part A: Critical Review** **2**
1.1 Review of Recent Work . . . 2
1.1.1 Visual Physics Reasoning . . . 2
1.1.2 Unsupervised Intuitive Physics . . . 3
1.1.3 Make Meanings from Sensory Input . . . 4
1.1.4 Curiosity Driven & Self-aware Agents . . . 4
1.2 Critical Discussion: Cognitive Perspective . . . 5
1.2.1 Evidences from Cognitive Science . . . 5
1.2.2 Visual Physics Reasoning . . . 5
1.2.3 Unsupervised Intuitive Physics . . . 6
1.2.4 Make Meanings from Sensory Input . . . 6
1.2.5 Curiosity Driven & Self-aware Agents . . . 6

**2 Part B: An Intelligent Process** **7**
2.1 A Cognitive Learning Algorithm . . . 7
2.1.1 Principles and Hypothesis . . . 7
2.1.2 Algorithm . . . 7
2.2 Potential Applications . . . 8
2.2.1 Object Detection . . . 8
2.2.2 Object Recognition . . . 8
2.2.3 Image Understanding and Visual Reasoning . . . 8
2.2.4 Knowledge Accumulation . . . 8
2.3 Discussion . . . 8

**Reference** **9**

# 1 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.

## 1.1 Review of Recent Work

### 1.1.1 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.

**1.1.1.1 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.

**1.1.1.2 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 $\widehat{f}_t$ and the ground truth $f_t$.

### 1.1.2 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 $\Phi$ 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: this principle claims that if a frame $\mathbf{x}_t$ is applied by a transformation $g$, then the detector's output should be equivalent to be applied by the same transformation, that is, $\Phi(g\mathbf{x}_t) = g\Phi\mathbf{x}_t$. This is called a Siamese branch and the training loss is noted as $\mathcal{L}_{siam}$.
3. penalty on distribution: Because $s_t$ is modeling the distribution of the object's position, it is reasonable to expect it have a peaky shape. To encourage the sharpness of the distribution, the entropy is taken as the penalty adding to the overall loss: $\mathcal{L}_{ent} = -\sum_{v \in \Omega} s_v \log(s_v)$.

Finally the overall loss is a weighted sum of the three items above: $\mathcal{L} = \lambda_d \mathcal{L}_{disc} + \lambda_e \mathcal{L}_{ent} + \lambda_s \mathcal{L}_{siam}$.

For multiple objects, the detector works in a one-by-one way: first, it detects a single object, and then masks it from the frame, and then repeats the process. However, the proposal assumes the number of objects is given.

This tracker is claimed to be well scalable to large datasets, and robust on a variety of objects in different kinds and shapes without prior knowledge..

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.

This work also proposed a 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(Ehrhardt, Monszpart, Mitra, & Vedaldi, 2019).

### 1.1.3 Make Meanings from Sensory Input

A significant characteristic of human learning is emerging symbolic represented concept from raw sensory data. In (Evans, Hernandez-Orallo, Welbl, Kohli, & Sergot, 2019) the authors give an answer to this fundamental question concerning human and machine learning: what is the meaning of "make sense" of raw sensory input. In their study, such intuitive approach is formalised by unsupervised program synthesis.

The authors believe that "making sense" of sensory data means interpreting the data by conceptualised and symbolic representation such as objects, which is more than the prediction of future values. The authors also criticise the neural network to be perceptive but not **apperceive** because, while it is trained well to classify images, it can not integrate such classification with other knowledge and do further reasoning. It is also proposed that, inspired by Kant's theory of the "synthetic unity of apperception", a theory that can make sense of data must satisfy four unity conditions, namely spatial, conceptual, static, and temporal.

It also proposed a model called "Apperception Engine", which satisfies such constraints and is able to learn from small amount of data thanks to its strong inductive bias. This is achieved by introducing a casual programming language, Datalog, and synthesizing a Datalog program from sensory input that explains the input and also complies with the four unity conditions. Therefore, the inductive bias comes from two aspects: specification of Datalog language and Kantian unity conditions. Datalog is based on Inductive Logic Programming and the Kantian conditions are expressed by formal logical expressions.

This model is designed as a *general purpose* apperception system, and has been tested on a range of domains. It is claimed to significantly out-perform neural network baselines, and particularly, to reach human-level in the sequence induction IQ tests, which would be more notable considering the system is not specifically designed to solve such tasks.

### 1.1.4 Curiosity Driven & Self-aware Agents

Infants are experts at generating orderly action 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. In this study, an agent are designed to learn how to play like an infant via deep reinforcement learning approach.

The agent is constructed by two interacting models: a world-model and a self-model. The world-model is a predictor of physical events based on its observational inputs, while the self-model is an estimator of the loss of world-model based on both the input and potential actions. The action of the agent is always chosen to maximise the loss of the world-model, that is, antagonize the world-model. To aim this, the self-model is introduced to estimate such loss, which is claimed to be a representation of self-awareness on its internal state, and a mathematical formalisation

of "interestingness" and "curiosity". The two components are both modeled by convolutional neural networks.

In the execution of this algorithms, the agent always pays attention to more challenging objects once it has learned some knowledge. In other words, it gets "bored" on the things that it has already "understood" well and always tries to challenge itself, resulting in a curiosity-driven behavior.

In the future work, the authors are planning to build computational models more than a robust learning agent, but precisely quantitatively comparable to human children's development.

## 1.2 Critical Discussion: Cognitive Perspective

This part criticises the learning algorithms mentioned above from a cognitive perspective. In this paper, an AI algorithm is not evaluated by how well it performed on certain tasks as most AI researchers do; instead, it is criticised from the cognitive science viewpoint. To be precise, it will be evaluated on whether it provides insights into the mechanism of cognitive process. As demonstrated in the work of (Yeap, 2011), the mapping algorithm is examined by comparing to three characteristics of cognitive mapping: fragmented, incomplete and imprecise. Similar idea is also expressed in (Dupoux, 2016), in which a reverse engineering approach is proposed in order to, from AI studies, gain scientific insights about underlying mechanisms of humans' psychological process, especially language acquisition.

### 1.2.1 Evidences from Cognitive Science

Studies in cognitive science uncovered some characteristics of humans' learning. Comparing to AI, the natural intelligence differs in the following aspects significantly:

- structured representation: it is proven that infants have the concept of objects, understand the physical law such as object permanence, and can organise raw information into categories. (Goswami, 2019; Stangor & Walinga, 2014)
- absence of supervision and reward: infants learn the physical concepts and rules without explicit supervision and reward; instead, natural intelligences build internal representation of physics inductively and automatically from raw perceptional data and experience. (Ehrhardt et al., 2018; Stangor & Walinga, 2014)

### 1.2.2 Visual Physics Reasoning

In work of (Riochet et al., 2018) as reviewed in section 1.1.1, 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. However, 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. Therefore, we can gain little inspiration from this "black-box" system on the cognitive process of how human infants learns. 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.

### 1.2.3 Unsupervised Intuitive Physics

The study reviewed in section 1.1.2 is a novel approach in intuitive physics learning. It assumes the form of learning function $\Phi$, which is a form of knowledge on the concept of object. That is, for this algorithm, the concept of object is modeled as a-prior knowledge instead of learnable variable from data.

Furthermore, the proposal assumes the number of objects is given on multiple objects positioning, which is a drawback because this is a specific instead of general assumption, and makes the algorithm less practical in real world applications. From cognitive perspective, human learns the concept of object and the ability of detection first, and the the number of objects second. Therefore, when a infant learns the detection of object, it is impossible to know the number of objects in advance.

### 1.2.4 Make Meanings from Sensory Input

The model reviewed in section 1.1.3 is unsupervised and independent of domain knowledge, which are significant advantages. The core of it is the Datalog language which provides strong inductive bias. Datalog is not embedded by any prior concepts; instead, all it preset rules is its logical form and four unity constraints, and the concepts are learned from input data. It was tested in a variety of domains and performed well, which in some extent demonstrates its nature of "general intelligence".

However, the Datalog language is a strong bias, it equivalently presets strong rules on the model, which have not been proven by cognitive studies. Furthermore, though the model was tested on different domains, they are all tasks more relying on logic and reasoning which is suitable for a logic based system. By contrast, humans can do both logical and intuitive tasks with one single system – the brain. The model may need more tests on intuitive tasks such as face recognition and image segmentation.

### 1.2.5 Curiosity Driven & Self-aware Agents

The study reviewed in section 1.1.4 is different from the others above because it focuses on the "motivation" aspect of a learning system. In a addition to a "world-model" modeling the outside world, the AI agent employs a "self-model" to model the inner state as a simulation of humans' curiosity and self-awareness. With this approach, the agent can learn without any extrinsic supervision or reward.

From the perspective of cognitive science, however, the biggest concern on this approach is the machine equivalent of humans' curiosity and self-awareness. There has not been sufficient evidence to prove that the algorithm works the same way as humans' cognitive process. The "self-model" estimates the loss of "world-model" and thus provides the knowledge of inner state, but curiosity and self-awareness are much more than estimation. To prove the self-awareness, at least a mirror test (Wikipedia contributors, 2019c) should be performed on this agent.

# 2 Part B: An Intelligent Process

In this part, an artificial will be discussed aiming at inspiring the understanding on human infants' learning process. First, a cognitive intelligent algorithm for object segmentation and tracking will be introduced and analysed following the approach in (Yeap, 2011). Second, a variety of questions will be investigated by this approach.

## 2.1 A Cognitive Learning Algorithm

As summarised before, human infants learning in early stage is entirely unsupervised, without explicit supervision or extrinsic reward. As a result, the proposed approach should also have these characteristics. To minimise any prior knowledge implied in the model, the input view of an infant is the only starting point, and the bias is preferred to more general hypothesis rather than specific ones.

### 2.1.1 Principles and Hypothesis

The following principles form the base of the algorithm:

- Association: or correlation/relativeness. It is a fundamental cognitive process initially identified by Pavlov's famous experiment of dogs on classical conditioning (Stangor & Walinga, 2014). Though it is mostly used to describe structured behaviour in psychology, in this paper this principle is extended to a universal extent. That is, if some things often occur simultaneously, they might also do so in the future. The more frequent their occurrence happens in the past, the more confident we assert it in the future.
- Attention: or selectiveness. Attention is a necessary condition of visual recognition (Goswami, 2019). With this cognitive mechanism, human can concentrate on a selected discrete aspect of information while ignore others (Wikipedia contributors, 2019b). This is essential because it prevents the human brain from disorientation in the unmeasurable amount of sensory input.
- Abstraction: or conceptualisation. Abstraction is one of the key characteristics of human behaviour (Wikipedia contributors, 2019a). It is the way of hiding information and thus managing complexity (Abelson & Sussman, 1996). Abstraction can also be seen as the process of compression, which is an implementation of minimum description length (MDL) principle (Henderson & Muggleton, 2014).

The principles introduced above is quite fundamental and general in human learning behaviour. As a result, the model following these principles is expected to have minimal inductive bias.

### 2.1.2 Algorithm

The input of the model is time sequences of raw pixel images $\mathbf{x} = \{x_t\}$.

1. Attention: Extract changing area by simply calculate the difference between two consecutive frames $\Delta x_t = x_t - x_{t-1}$. Obviously, in common cases, the changing area is caused by moving objects. However, no prior concept of "object" or "motion" is introduced into

this model; instead, only “change” is a presumption and the concepts of “object” and “motion” are learned.

2. Associative and structured Representation: Construct an representation for extracted area
3. Abstraction: Refactor the representations to emerge new abstractions

## 2.2 Potential Applications

### 2.2.1 Object Detection

### 2.2.2 Object Recognition

### 2.2.3 Image Understanding and Visual Reasoning

### 2.2.4 Knowledge Accumulation

## 2.3 Discussion

placeholder

# Reference


Abelson, H., & Sussman, G. J. (1996). *Structure and interpretation of computer programs* (2nd ed.). Cambridge, MA, USA: MIT Press.

Dupoux, E. (2016). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *CoRR*, *abs/1607.08723*. Retrieved from http://arxiv.org/abs/1607.08723

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

Ehrhardt, S., Monszpart, A., Mitra, N. J., & Vedaldi, A. (2019). Unsupervised intuitive physics from past experiences. *CoRR*, *abs/1905.10793*. Retrieved from http://arxiv.org/abs/1905.10793

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

Goswami, U. (2019). *Cognitive development and cognitive neuroscience : The learning brain.* Retrieved from http://ezproxy.aut.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cat05020a&AN=aut.b27297470&site=eds-live

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

Henderson, R., & Muggleton, S. (2014). Automatic invention of functional abstractions. In *Latest advances in inductive logic programming* (pp. 217–224). https://doi.org/10.1142/9781783265091_0023

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

Stangor, C., & Walinga, J. (2014). *Introduction to psychology – 1st canadian edition*. Retrieved from https://opentextbc.ca/introductiontopsychology

Wikipedia contributors. (2019a). *Abstraction — Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Abstraction&oldid=922555014

Wikipedia contributors. (2019b). *Attention — Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Attention&oldid=921297171

Wikipedia contributors. (2019c). *Mirror test — Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Mirror_test&oldid=921100578

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