Research on Machine Learning and Cognitive Science

Yu-Zhe Shi

April 16, 2021

Abstract

Currently I'm interested in the reciprocative research of Machine Learning and Cognitive Science to find a way towards Artifical General Intelligence. My projects generally exploit neural-symbolic models intergrating neural perception with logical reasoning. More concretely, the former serves as a denoiser to handle raw data and the latter helps constraint the hypothesis space of the model, with both of which being optimized jointly. Also, the learning paradigm partially mimics human learning that extracts highly abstracted explanable knowledge from only a few noisy raw examples, with the help of commonsense and background knoweldge. In summary, this kind of hybrid model forms a computational cognitive model that learns from little noisy raw data and has the capacity to solve human-level tasks like abduction, induction, abstraction, explanation and planning.

1 Introduction

I have been focused on developing the neural-symbolic learning paradigm to integrate neural perception with logical reasoning. My models are consistented of two components: 1) the neural network model to map the raw data space into a symbolic space; 2) the logic reasoning model exploiting human background knowledge. The former handles raw data like images, videos, or natural language, which are extremely noisy and are hopelessly tackled by pure logic models. The latter executes logic reasoning in the symbolic space, more concretely, it takes the output of perception module as **Observation Fact** O and predefined **Background Knowledge** BK as input, tries to make an **Interpretation** Δ which makes

$$BK \cup \Delta \models O \tag{1}$$

where $\models$ denotes logic entailment. We introduce the background knowledge in two ways: 1) the **Ingredient Knowledge**, a.k.a. atomic predicate or grounding, describes the property of an instance, such as `door(X)` meaning *X is a door*; 2) the **Causal Knowledge**, a.k.a. first-order logic, describes the relationship between instances, such as `open(A, X)` showing that *A opens the door X*. Though we represent the knowledge as logic programs, we can also execute logic reasoning with statistical models, because O, BK, Δ can be Likelihood, Prior Distribution and Posterior Distribution in the bayesian formulation respectively. Knowledge module adds constraint to the perception module by forcing it to learn a distribution that not far from the initial distribution of background knowledge, and the learned Δ is consistent to BK. What's more important is that the knowledge learned by our hybrid model is highly comprehensible by human since it is represented in logic programs or Bayesian models instead of tons of parameters in a deep neural network. Note that this learning paradigm is sort of close to the way of human learning, which extracts highly abstracted explanable knowledge from only a few noisy raw samples, with the help of commonsense and background knoweldge. In summary, the neural-symbolic paradigm of learning forms a computational cognitive model that learns from little noisy raw data and has the capacity to solve human-level tasks like abduction, abstraction, explanation and planning.

Inspired by some previous work (e.g. from groups lead by Zhi-Hua Zhou, Stephen Muggleton, Josh Tenenbaum, Song-Chun Zhu, and Brenden M. Lake) on this under-researched topic, I launched a few research projects on my own and luckily had discussions with several researchers in the community.

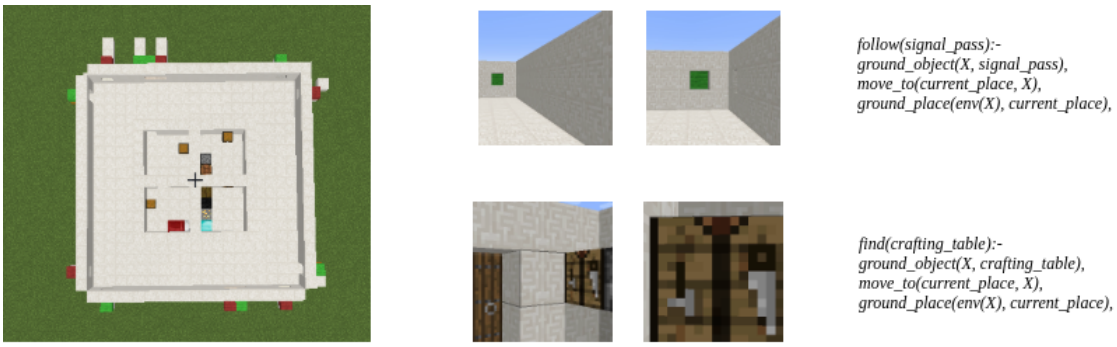


Figure 1: **Demonstrations of our Minecraft Experimental Environment and simple task samples.**

2 Do It Yourself: Abductive Visual-Policy Grounding

2.1 Motivation

- **Motivation:** *Thanks to the learning-by-explanation nature of human cognition, we can learn to explore a physical world like Minecraft well even with very little guidance.*
- Given first-person video demonstrations and human-level background knowledge, we create a learner that jointly learns a perceptive module that maps raw data into symbolic space and logic-program-represented policy that inferred from the observations of the agent, as shown in Figure 1.

2.2 Challenges

- Local perceptive observations (compared to 2D-grid and god's perspective tasks, where the model obtains global observation of the world) which requires additionary modeling of the environment;
- Passive information aggregation (comparing to god's perspective tasks, the model only gets the information simultaneous to the agent in the demonstration);
- Changing background (compared to our third-person view counterparts, where the background is always still);
- Unpredictable variations of visual objects (e.g. scale, aspect-ratio, while the objects in third-person videos perform regular and predictable motions).

2.3 Background Knowledge Examples of the Task

- Ingredient Knowledge:
  - The usage of the items:
    usage(bed, sleep).
    usage(crafting_table, [X|_], production).
    door(livingroom, corridor).
  - The object-environment relationships:
    in(gold_ore, wareroom).
- Causal Knowledge:
  - Visual stimuli in the style of Skinner Box:
    causes(pass_signal, move_to).
    implies(entrance_signal, env(door)).

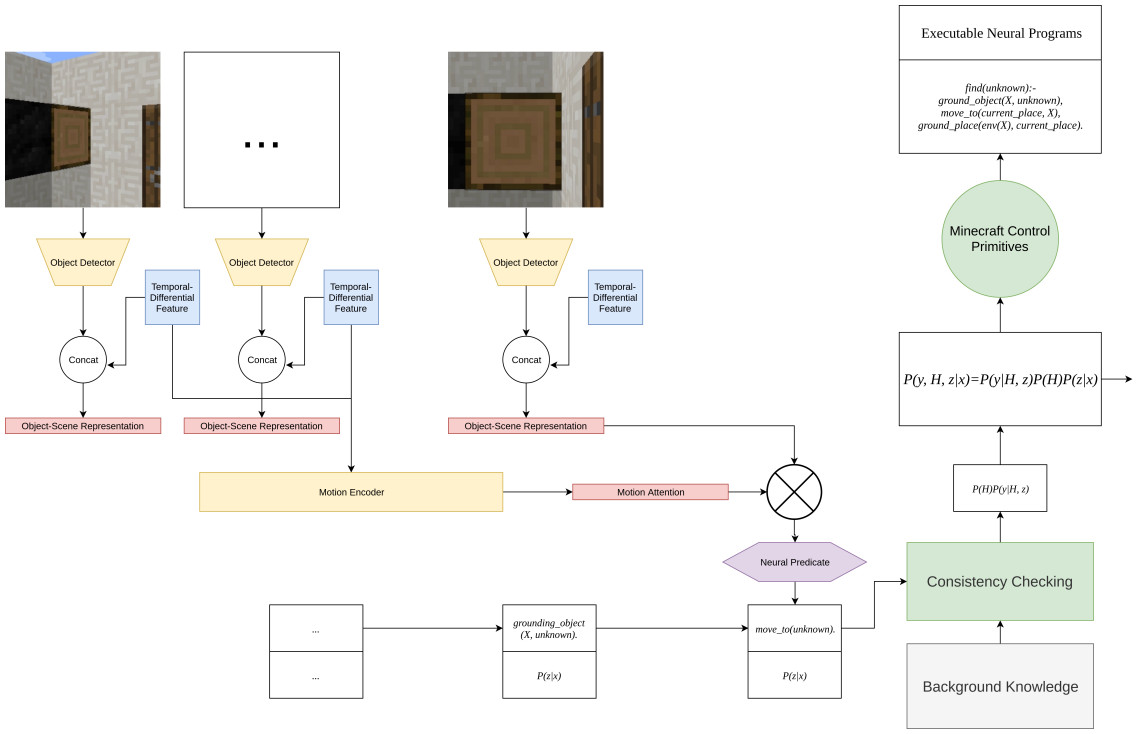


Figure 2: **Architecture of our Abductive Neural-Symbolic Model.**

- Program Primitives:
  - Ungrounded logical predicates that indicates the operations of the agent: *move_to(current_place, X).*

2.4 Contribution

- We introduced a novel abductive task: Exploring a Minecraft world by making Explanations from First-Person Visual perception, Background Knowledge and Human Commonsense;
- We proposed a Neuro-Symbolic Framework that leverages background knowledge and visual perception to incrementally learn neural predicates, as shown in Figure 2;
- We introduced two novel downstream tasks: Novel Visual Concept Grounding and Novel Human Policy Grounding;
- We developed a plug-in experimental environment based on Minecraft and Microsoft Malmo.

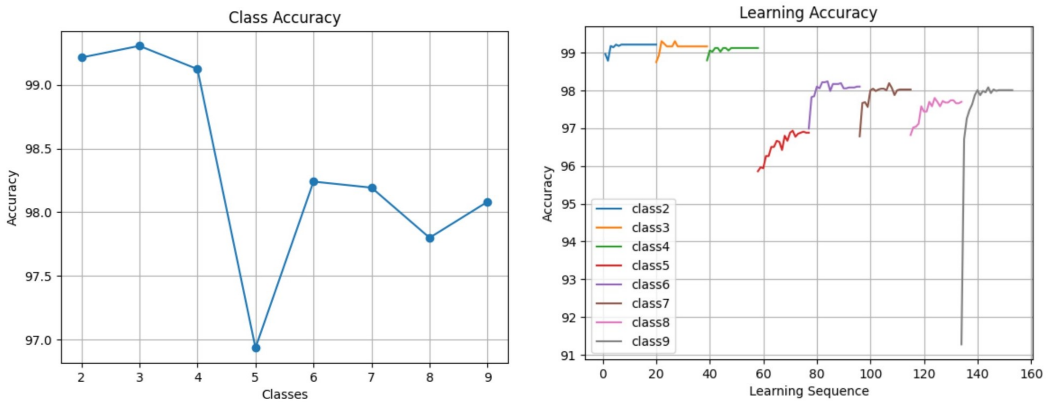
3 Abductive Novel Object Invention for Incremental Learning

3.1 Motivation

Can a machine learner detect instances belonging to classes that never been seen nor known before from raw data and label them autonomously with the help of background knowledge?

3.2 Task

The learner starts with a CNN classifier that recognizes hand-written symbols `0, 1, 2, +, -, =` and background knowledge about successor relation. The learner aims to recognize image sequences like `0++-=1`, where `+` and `-` represents successor and predecessor relation respectively. However, there exists numbers that



(a) Class Accuracy (b) Accuracy over Iterations

Figure 3: **Results of Novel Handwritten Digit Recognition Task.**

the learner has never seen in the training phase and it is expected to label the novel instances with rational logic concepts derived from known knowledge.

3.3 Method

I feed the learner with image sequences like `0+++=3` and it tries to estimate the distribution of novel objects by both the classification score and logical consistency of rules defined by a Probabilistic Context-Free Grammar. The former serves as the likelihood of the digit being consistent to prediction while the latter is a prior distribution of the positions of the symbols in the sequence. We detect novel instances by Bayesian Inference and Expectation Maximization. Then the learner abducts an atomic predicate representing the relationship between the novel class and known classes, e.g. `new_digit(X):-succ(2,X)`. Ultimately, it labels the novel instances with the predicates and update the classifier under the supervision of revised labels.

3.4 Experiments

Figure 3 shows results of Novel Handwritten Digit Recognition Task with (a) accuracy over novel classes and (b) accuracy over iterations.

4 Human-Level Abuctive Learning and Planning

4.1 Motivation

A novice Minecraft player can explore the world well from only visual observation given very little guidance thanks to the rich background knowledge (or commonsense). Can a machine learner do the same?

4.2 Task

I prepare several first-person video sequences recording a rational human player playing Minecraft and define some logic rules representing the human commonsense, e.g. Perspective Relationship and Relative Motion. The learner aims to learn a perception model that predicts the motion of the agent in the video, and inference the intentions of human players. Aftermore, the learner is asked to control an agent to solve tasks that both the environments and the goals are different from the training ones.

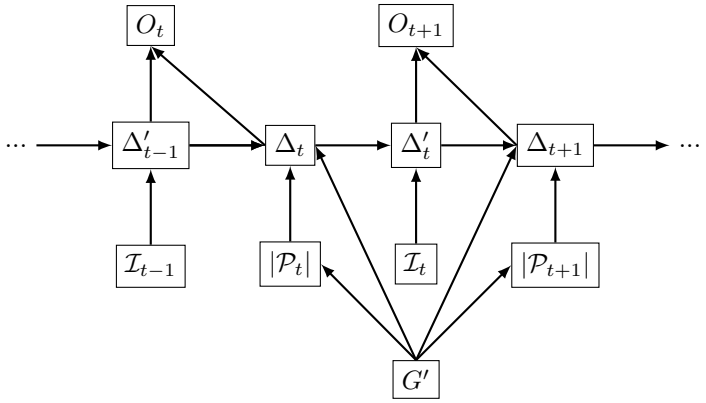


Figure 4: **Dependency Model for Abductive Planning.** This graph illustrates a part of the dependency model for solving subgoal G' (corresponding to one of the critical points) in a task, which is the prerequisite of the starting states and the selection of the operations. The motion state at Δ_t relies on observation O_t and the predecessor state Δ'_{t-1}. Note that the state transition $P(\Delta_t|\Delta'_{t-1}), P(\Delta'_t|\Delta'_t), P(\Delta_{t+1}|\Delta'_t)$ are all defined by background knowledge.

4.3 Method

The learner models the agent operations following the idea of Qualitative Simulation, which describes the motion not step-to-step but in a high-level state-to-state way, yielding greater representative power. The operations of our agent is characterized by a set of operation functions $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ and a set of abstract motion states Δ. The agent starts with an empty pushdown automata for storing motion, once operates a function by pushing it into the stack, it runs into a **balanced dynamic state** until it observes stimuli to end the current state and transit to another state. To realize the transition, the agent must execute pop operation to the automata and select push another function into the stack. The we exploit a sequence-to-sequence model (Siamese CNN, RNN, and Transformer are experimented) to translate the video sequence to a sequence of motion state transition signals. Then the learner tries to inference subgoals of the player by abducting an interpretation from motion observations and background knowledge. We solve the inference as Inverse Planning that try to maximize the likelihood of subgoals given action trajactories, environment and the goal. section 4 shows the probabilistic dependency model of the variables. Finally, the learner generates logic programs representing the strategies learned from human players.

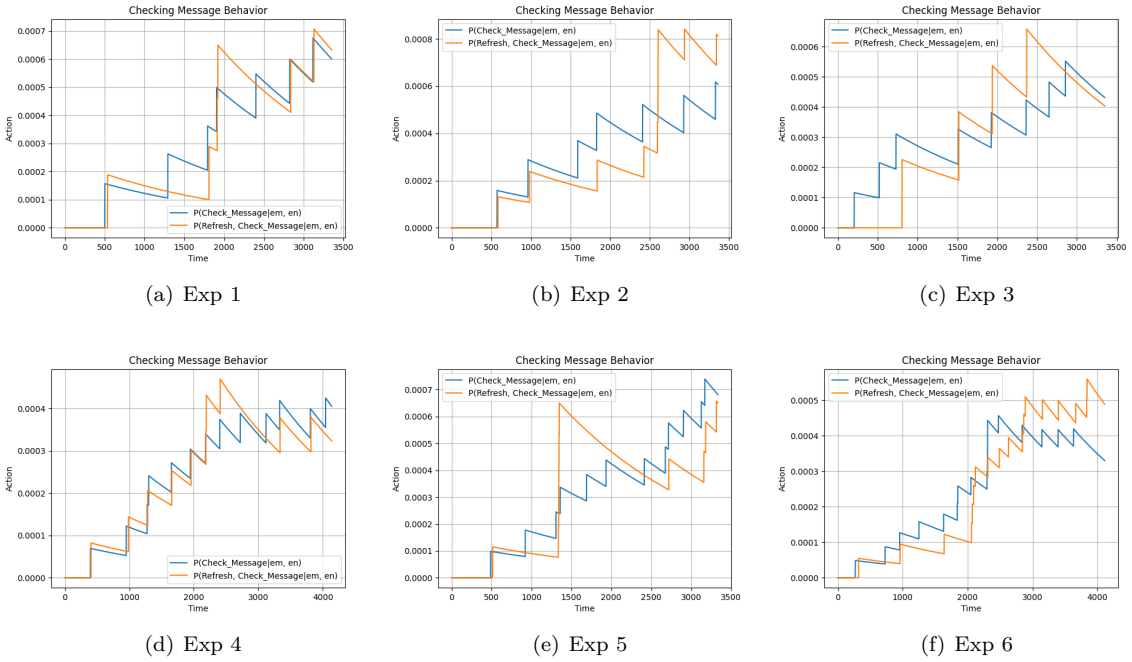
5 Waiting for the Bus: A Human-Centered Computing Perspective

5.1 Motivation

When people waiting for something that is not determined to happen or not before a specified deadline, there are two perspective: 1) the long-term perspective: will it happen before the deadline? 2) the short-term perspective: will it happen in the next time interval? Obviously the former is a static version of subjective hope while the latter is a dynamic, marginalized view of subjective probability. How do these two kinds of perspectives affect human subjective probability when waiting for a bus? Can we design an intelligent bot that helps users feel better when waiting for some event?

5.2 Case Study: Waiting for the Bus

We design a 2×6 experiment making participants waiting for a schoolbus in a virtual environment simulating the campus, within 6 scenes and 2 ways of prompts (in short-term and long-term respectively). We record the videos and interactions with the computer of 28 participants.



(a) Exp 1 (b) Exp 2 (c) Exp 3

(d) Exp 4 (e) Exp 5 (f) Exp 6

Figure 5: **Exponential Moving Average smoothed observations of Participant # 4.**

We apply micro-expression recognition models to the videos as groundtruth of emotion states of the participants and develop a bayesian model inferencing emotion states from human interactions with the computer via inverse planning. We aim to develop a bot that selects appropriate prompting strategy adaptively for different users according to their behavior on the interface, thus providing users with a better experience. Figure 5 shows the Exponential Moving Average smoothed observations of user behavior.

6 Interpretative Neural Feature Primitives for Image Classification

6.1 Motivation

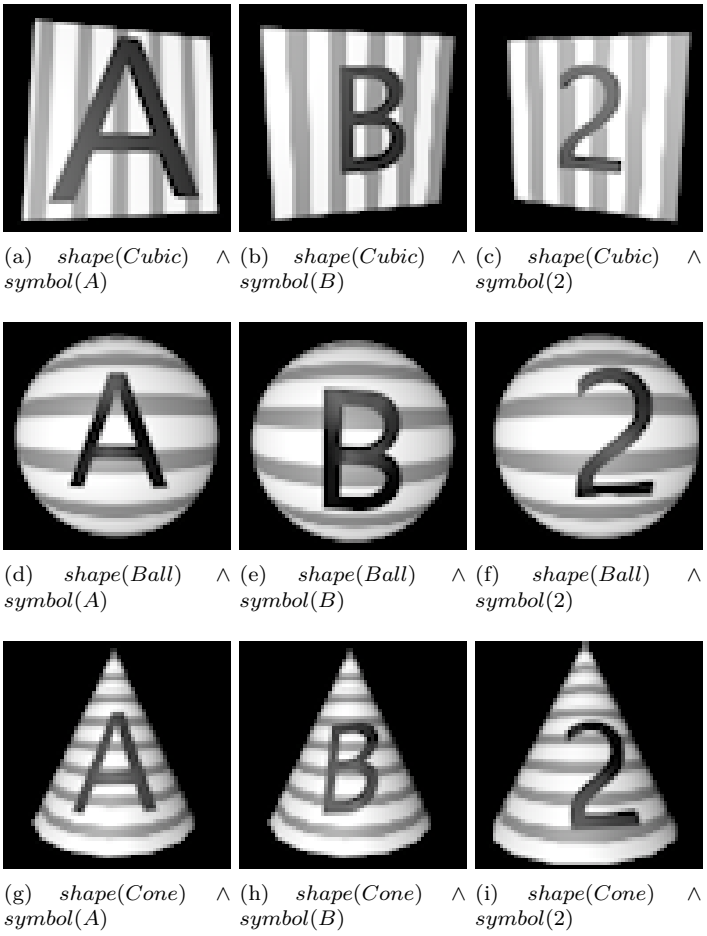
Human can recognize instances by executing association between different visual concepts using simple features such as shape, texture, color or symbol. I call these simple explanable features primitives. Can we train a neural network that detects such feature primmitives, and has compositional generalization ability over feature primitives?

6.2 Task

We construct a toy image dataset with objects in diverse shapes, textures, colors, and with different symbols on them, using C4D. This leads to a conceptual experiment over the hypothesis that the 4 features can serve as feature primitives.

6.3 Method

We train 4 CNN classifiers, one in each simple feature family. For a particular sample, we order the other samples by calculating their similarities to it. Experiments show that our method extremely reduces the open-world risk for novel instances and we can link them with known classes by applying association over some features. Referenced to Figure 6, given only the 5 samples from the first row and the first column, our model has the capability to scale to other unseen classes (the remainders in the matrix).



(a) $shape(Cubic) \wedge symbol(A)$ (b) $shape(Cubic) \wedge symbol(B)$ (c) $shape(Cubic) \wedge symbol(2)$

(d) $shape(Ball) \wedge symbol(A)$ (e) $shape(Ball) \wedge symbol(B)$ (f) $shape(Ball) \wedge symbol(2)$

(g) $shape(Cone) \wedge symbol(A)$ (h) $shape(Cone) \wedge symbol(B)$ (i) $shape(Cone) \wedge symbol(2)$

Figure 6: **An illustration of our toy dataset.** There are two primitive features exploited in the example: the row space is the shape feature space, with *Cubic, Ball, Cone* from the top to the bottom; the column space is the symbol feature space, with *Letter A, Letter B, Digit 2* from the left to the right.