# Building Machines That Learn and Think Like People (2016)

Tom Rochette <tom.rochette@coreteks.org>

December 21, 2025 — 77e1b28a

### 0.1 Context

### 0.2 Learned in this study

### 0.3 Things to explore

- Decomposition of concepts (extracting the components/concepts in a grid of pixels and reason based on these and not the pixels)
- How can decomposition be done? Where to start and where to stop? How can components be identified? Is it necessary that the components be labeled by a human and the machine trained to recognized instead of the machine generating components and the human giving them labels?

## 1 Overview

### 1.1 1 Introduction

- Two different computational approaches to intelligence
  - Statistical pattern recognition
  - Model-building

### 1.2 1.1 What this article is not

- Neural networks have been applied in compelling ways to many type of machine learning problems, demonstrating the power of gradient-based learning and deep hierarchies of latent variables
- We believe that reverse engineering human intelligence can usefully inform AI and machine learning
- Some exciting progress
  - Probabilistic machine learning
  - Automated statistical reasoning techniques
  - Automated techniques for model building and selection
  - Probabilistic programming learning

### 1.3 1.2 Overview of the key ideas

- Ingredient: Two pieces of developmental start-up software
  - Intuitive physics
  - Intuitive psychology
- Ingredient: Learning is a form of model building
  - Explaining observed data through the construction of causal models of the world
- Compared to state-of-the-art algorithms in machine learning, human learning is distinguished by its richness and its effeciency

- We suggest that compositionality and learning-to-learn are ingredients that make this type of rapid model learning possible
- Ingredient: How the rich models our minds build are put into action, in real time

## 1.4 3 Challenges for building more human-like machines

- Two challenges for machine learning and AI:
  - Learning simple visual concepts
  - Learning to play the Atari game Frostbite

## 1.5 3.1 The Characters Challenge

- While humans and neural networks may perform equally well on the MNIST digit recognition task and other large-scale image classification tasks, it does not mean that they learn and think in the same way. There are at least two important differences
  - People learn from fewer examples
  - People learn richer representations
- People learn more than how to do pattern recognition: they learn a concept
- In addition to recognizing new examples, people can also
  - generate new examples
  - parse a character into its most important parts and relations
  - generate new characters given a small set of related characters
- The Characters Challenge: learn more from a lot less and capture human-level learning abilities in machines

## 1.6 3.2 The Frostbite Challenge

- Requires temporally extended planning strategies
- The neural networks are trained anew for each game, meaning the visual system and the policy are highly specialized for the games it was trained on (so no inter-game reusability of the trained/learned policy)
- The DQN was trained on 200 million frames from each of the games, which equates to approximately 924 hours of game time (~38 days), or almost 500 times as much experience as the human received (about 2h)
- The differences between the human and machine learning curves suggest that they (the machines) may be learning different kinds of knowledge, using different learning mechanisms, or both
- We speculate that people do this by inferring a general schema to describe the goals of the game and the object types and their interactions, using the kinds of intuitive theories, model-building abilities and model-based planning mechanisms we describe (I think that machine learning and human learning is simply very different because they start from scratch and we already have an immense amount of knowledge to build on top of)
- The DQN (deep Q neural network) is also rather inflexible to changes in its inputs and goals: changing the color or appearance of objects or changing the goals of the network would have devastating consequences on performance if the network is not retrained

## 1.7 4 Core ingredients of human intelligence

## 1.8 4.1 Developmental start-up software

## 1.9 4.1.1 Intuitive physics

- At the age of 2 months and possibly earlier, human infants expect inanimate objects to follow principles of persistence, continuity, cohesion and solidity
- Young infants believe objects should move along smooth paths, not wink in and out of existence, not inter-penetrate and not act at a distance

- At around 6 months, infants have already developed different expectations for rigid bodies, soft bodies and liquids
- By their first birthday, infants have gone through several transitions of comprehending basic physical concepts such as intertia, support, containment and collisions

## 1.10 4.1.2 Intuitive psychology

- Pre-verbal infants distinguish animate agents from inanimate objects
- The presence of eyes, motion initiated from rest and biological motion are cues that are often sufficient but not necessary for the detection of agency
- Infants also expect agents to act contingently and reciprocally, to have goals, and to take efficient actinos toward those goals subject to constraints
- At around three months of age, infants begin to discriminate anti-social agents that hurt or hinder others from neutral agents and they later distinguish between anti-social, neutral, and pro-social agents
- It seems to us that any full formal account of intuitive psychological reasoning needs to include representations of agency, goals, efficiency, and reciprocal relations
- Intuitive psychology provides a basis for efficient learning from others, especially in teaching settings with the goal of communicating knowledge efficiently
- Whether or not there is an explicit goal to teach, intuitive psychology let us infer the beliefs, desires, and intentions of the teacher

## 1.11 4.2 Learning as rapid model building

- Many of the most impressive examples of human learning are better characterized as rapid model building than gradual improvements in pattern recognition
- It is important to mention that there are many classes of concepts that people learn more slowly
- Concepts that are learned in school are usually far more challenging and more difficult to acquire, including mathematical functions, logarithms, derivatives, integrals, atoms, electrons, gravity, DNA, evolution, etc.
- There are domains that machines are more apt learners than people, such as combing through financial or weather data
- But for the vast majority of cognitively natural concepts - the types of things that children learn as the meanings of words - people are still far better learners than machines
- The richness and flexibility of model building suggest it is a better metaphor than learning as pattern recognition
- Furthermore, the human capacity for one-shot learning suggests that these models are built upon rich domain knowledge rather than starting from a blank slate

## 1.12 4.2.1 Compositionality

- Compositionality is the classic idea that new representations can be constructed through the combination of primitive elements

## 1.13 4.2.2 Causality

- Causality is about using knowledge of how real world processes produce perceptual observations
- "Analysis-by-synthesis" theories of perception maintain that sensory data can be more richly represented by modeling the process that generated it
- Not all causal processes present tractable learning problems, and in most cases it is crucial to find the right level of causal description

## 1.14 4.2.3 Learning-to-learn

- Transfer learning, multi-task learning or representation learning refer to ways that learning a new task can be accelerated through previous or parallel learning of other related tasks

- Bayesian Program Learning (BPL) transfers readily to new concepts because it learns about object parts, sub-parts, and relations, capturing learning about what each concept is like and what concepts are like in general
- It is crucial that learning-to-learn occurs at multiple levels of the hierarchical generative process
- Further transfer occurs by learning about the typical levels of variability within a typical generative model
- Human players can transfer what they have learned in playing other video games because they immediately parse the game environment into objects, types of objects, and causal relations between them
- To produce machines that learn like humans and as fast as humans do, we might also have to build machines that learn what humans learn
- We believe that adopting a more compositional, causal forms of knowledge representation helps both humans and machines get the most from learning-to-learn
- We want to emphasize more generally that we believe all of the core ingredients for learning rich models articulated in this section - compositionality, causality, and learning-to-learn - can be incorporated into deep learning systems, and that these ideas will only benefit from being integrated together

### 1.15 4.3 Thinking Fast

### 1.16 4.3.1 Approximate inference in structured models

- Computing a probability distribution over an entire space of programs is usually intractable, and often even finding a single high-probability program poses an intractable search problem
- It has been proposed that humans can approximate Bayesian inference using Monte Carlo methods, which stochastically sample the space of possible hypotheses and evaluate these samples according to their consistency with the data and prior knowledge
- For domains where program or theory learning happens quickly, it is possible that people employ inductive biases not only to evaluate hypotheses, but also to guide hypothesis selection

### 1.17 4.3.2 Model-based and model-free reinforcement learning

- There is substantial evidence that the brain uses model-free learning algorithms in simple associative learning or discrimination learn tasks
- Considerable evidence suggests that the brain also has a model-based learning system, responsible for building a “cognitive map” of the environment and using it to plan action sequences for more complex tasks
- Model-based planning is an essential ingredient of human intelligence, enabling flexible adaptation to new tasks and goals; it is where all of the rich model-building abilities earn their value as guides to action
- We conjecture that a competent player can easily shift behavior (toward a new goal) appropriately, with little or no additional learning, and it is hard to imagine a way of doing that other than having a model-based planning approach in which the environment model can be modularly combined with arbitrary new reward functions and then deployed immediately for planning

## 2 5 Responses to common questions

- Comparing the learning speeds of humans and neural networks on specific tasks is not meaningful, because humans have extensive prior experience
  - Successful learning-to-learn - or at least, human-level transfer learning - is enabled by having models with the right representational structure, including the other building blocks (discussed in this paper)
  - There are a set of core cognitive ingredients for human-like learning and thought. Deep learning models could incorporate these ingredients through some combination of additional structure and perhaps additional learning mechanisms, but for the most part have yet to do so

- Biological plausibility suggests theories of intelligence should start with neural networks
- Language is essential for human intelligence. Why is it not more prominent here?

# 3 6 Looking forward

- We believe that deep learning and other learning paradigms can move closer to human-like learning and thought if they incorporate psychological ingredients including those outlined in this paper

# 4 See also

# 5 References

- arXiv:1604.00289v1 [cs.AI]