The path to AGI is unsupervised

Konstantinos Kogkalidis [6230067]

October 14, 2018

1 Introduction

Artificial General Intelligence (AGI), also referred to as Strong AI, is the notion of an artificial system of computation that is characterized by holistic, broad intelligence. Its counterpart is Weak AI, which might achieve human-level performance on narrow tasks but lacks the capacity of expanding its abilities beyond those. AGI has been a core concept within the field of AI from its very inception (perhaps even more so then than in recent years). Key researchers in the 1950s were overly eager to make bold claims about the imminence of strong AI before the end of the century, prompting a variety of responses from philosophers. Philosophical considerations related to strong AI were in fact raised before the establishment of both AI and Computer Science as fields of research, and this does not come as a surprise, given the intriguing nature of the subject and its strong ties with yet deeper philosophical questions, such as the nature of human intelligence, the source of consciousness, the descriptive ability of formal logics and mathematics, or the relations between computation and intelligence amongst others.

As the initial forecasts started proving utterly ungrounded, concerns were voiced about the suitability of the particular systems being used. The result was a vast divide between defenders of symbolism and connectionism. As both subfields underwent prolonged periods of dismissal, infamously known as the AI winters, scientific and engineering progress largely stagnated. A side-effect of this was the development of a challenging attitude towards the plausibility of strong AI as a whole from the parallely evolving field of AI philosophy. Even though there were numerous arguments and discussions both for and against strong AI, the viewpoints remained consistently narrow and focused on nowadays outdated techniques and specifics.

The purpose of this paper is to pose a fresh look at the recurring issue of strong AI, bringing recent technical advancements at the forefront of the discussion. It will be argued that predictive learning, an emerging subfield of deep unsupervised learning, is one of the most promising approaches (if not perhaps the only) towards constructing general artificial intelligence. A brief overview of predictive learning, in comparison to standard deep and unsupervised learning, will be presented in Section 2, for the sake of establishing a common ground and to facilitate the discussion to follow. Motivating reasons, as well as potential pitfalls, will be examined in the Section ?.

2 Background

Machine Learning, and more recently Deep Learning, have been by far the most striking breakthrough in AI for the last couple of decades. Three major subfields of Machine Learning are Supervised Learning, Reinforcement Learning and Unsupervised Learning. The distinction between the three will be briefly summarized below.

2.1 Supervised Learning

Supervised Learning is inarguably the most accomplished subfield of Machine Learning, with its architectures achieving astounding results in a broad range of tasks, spanning from sensory processing to statistical inference. The standard problem formulation for machine learning systems revolves around performing a search over a parameter space for a parametrically fixed function in the presence of a set of examples, where each example consists of an input-output value pair. This search is guided by an evaluation metric that assigns an error value (or cost) depending on the distance between the currently predicted output and its ground truth value.

2.2 Reinforcement Learning

Reinforcement Learning revolves around optimizing an interactive system's action planning, given some information about its environment, on the basis of maximizing a pre-defined or dynamically approximated reward function. Reinforcement Learning bears similarity to Supervised Learning, in the sense that the learning is facilitated by some external guidance (in this case the simulation environment, rather than input-output examples).

2.3 Unsupervised Learning

Contrary to the above cases, unsupervised architectures simply attempt to organize the input data in a coherent way, by just utilizing their underlying structural form. Classical applications of unsupervised learning include clustering and feature compression. A more modern definition would claim that unsupervised architectures attempt to reconstruct their input, given any observed part of it. **to-do**

Predictive Learning Predictive Learning is an as of yet not concretely defined subclass of Unsupervised Learning. The term has gained traction in the last couple of years, due to its use by LeCunn in a number of talks. Intuitively, the core concept is the application of unsupervised techniques with the goal of inferring a model of the system's environment, complete with a means of describing its temporal evolution and its relation to the system's actions. The idea has remained at a rather abstract level, and no particular implementations of it exist at this point in time. However, its significance and potential for paving the road to AGI will be showcased in the following sections.

3 The Story So Far

Deep Learning hails from connectionism and thus inherits the sum of its benefits for modeling actual intelligence. At the same time, the most popular implementations suffer from a task-specific narrowness that hinders their suitability as models for AGI. A brief summary of both the pros and cons of deep learning, as practiced today, is exposed below.

3.1 The Good

Emergence Performing computation via the interaction of many small, interconnected units falls within the standard paradigm of emergence **to-do**; i.e. the notion that complex behavior can burst out of a system of individually simplistic building blocks that are rather uninteresting when inspected in isolation. The complexity and capacity of systems, according to this world

view, far exceeds that of the sum of their parts; it rather lies on their connectedness, their unique means of information transfer and interaction.

Biological Plausibility The building blocks, in this case, are artificial neurons. These are heavily inspired by their carbon-based counterparts, bringing forth an argument towards their biological plausibility as means of constructing artificial intelligence. However, dissimilarities due to engineering concessions, in combination with a general lack of understanding of the exact functionality that describes biological neurons, weakens this argument. Still, it makes for an interesting point to keep in mind; it is certainly not unreasonable to hope that, with the field of neuroscience undergoing rapid evolution, more convincing matchings between artificial and biological neural systems will become possible in the coming decade.

Hierarchical Representations A chained application of neural transformations (or layers) is the core theme of deep neural networks. These construct increasingly higher-level representations of the input data with each subsequent transformation (or layer). From an intuitive standpoint, this hierarchical representation scheme matches the flow of information as it is seemingly performed by human beings; we are able to go from perceptive-level input to more complex, abstract concepts, without having to consciously process signals from the lower levels when operating at a higher cognitive level.

Relevance and Generalization A key issue within symbolic AI has been establishing relevance in the presence of an abundance of information **to-do**. Recognizing which subset of the given input is in fact important for distinguishing between different scenarios and guiding decision making would traditionally require a full processing of the full range of the input; something that seems computationally unfeasible and does not match the way humans argue about the world. The aforementioned hierarchical nature of deep architectures essentially allows for different levels of computation, with each level automatically picking out relevant properties of the previous level's outputs. Additionally, the weighted nature of deep connections could be translated as an efficient means of context-dependant selection, even more so in the case of recurrent or attention-driven systems **to-do**. Just as humans learn by performing a continuous (i.e. non-sharply separable) distinguishing between classes of situations, without focusing on the minutia of each subcase, neural networks assign weight only on the features deemed important for each particular task. This enables them to generalize over to previously unseen data (i.e. situations), which brings them closer to human intelligence.

Adaptability Human intelligence does not collapse in the partial absence or distortion of information. In that same way, neural systems are characterized by extreme noise tolerance, and their property of graceful degradation; as their input becomes incomplete or less reliable, their output quality deteroriates but in a smooth, continuous way. Their ability to operate under noisy data grants them an adapatability that posits yet another argument for their ability to simulate human intelligence.

Universal Approximation Neural networks have been proven to be able to approximate within any error margin, however small, any function that is computable **to-do**. Assuming that there is nothing hyper-computational in the way the human brain works, this entails that a neural network would in theory be able to simulate the function that defines human cognition,

regardless of how complex that might be[1].

This is indeed a big assumption to make, but it is mandatory if one abides by any of the material-centric monistic philosophies of mind. Granted that hyper-computation is isomorphic to computation over the field of reals, of which no physical analogue exists (at least none that we are aware of), given the quantized nature of the universe as we know it, it could be argued that hyper-computation is in fact not physically plausible. Additionally, arguing against the materialistic nature of cognition is an even weaker and wholly non-evidential assumption to make that essentially cancels out the entire discussion; if there is something beyond the physical plane (of which we have no immediate access over), there would not be a way for us to construct artificial intelligence to begin with.

On the more practical side, the universal approximation theorem only asserts that a neural approximator exists for any function, but makes no claim on what the required size of such a network would be. There would probably be little use to a network that made use of all the capacitors on Earth's surface to simulate a young child (assuming we ever even knew how to train such a network, if we are to further ground this argument to reality).

In any case, having established that there is no theoretical limit to the computational capacity of neural networks is still a necessary (albeit insuficcient) argument for them being candidates as the implementional backbone of AGI.

Memory Memory has been a long-standing issue for neural networks. Despite their amazing performance at information processing, they have been for a long while considered incapable of information storage and retrieval. Two major breakthroughs attempted to tackle this problem.

Recurrent networks partially bypassed it by allowing memory to be implicitly encoded into the processing steps of computation **to-do**. Even though this does seem like an ad-hoc solution for memory-centric operations, it is at the same time of crucial importance as it allows for temporal connectedness. Considering the issue that internal feedback loops are obviously occurring within our thought patterns (e.g. an external stimulus provokes a thought, which in turn causes another thought, which in turn prompts an action), internal recurrency is a very natural addition to any interactive system that hopes to capture intelligence.

The big breakthrough, however, came by the very recent proposal of the Neural Turing Machines **to-do**, a neural architecture that is unique in its ability to learn how to correctly make use of a memory matrix, external but tied to the computational framework. Aside from being able to perform algorithmic inference (learning entire programs, if at least simple, rather than just functions), neural turing machines utilize auto-associative methods, i.e. methods of content-based addressing. Addressing by content is in full accord with our ability to recover low or high-level 'snapshots' from our long-term memory when being presented the appropriate sensory stimuli or triggers. In comparison, addressing by value, as performed by traditional software, requires explicitly stating the position of each particular memory (which would require that a dynamically changing look-up table is somehow implemented within the human brain, a rather far-fetched scenario).

3.2 The Bad

Data

Loss Functions

[1] Hyper-computation in this case refers to computability of non-Turing-computable functions, rather than hyper-efficient computation (i.e. in the sense of a quantum computer).

Deductive vs. Constructive

4 Unsupervised Learning

5 Conclusion