The path to AGI is unsupervised

Konstantinos Kogkalidis [6230067]

October 16, 2018

1 Introduction

Artificial General Intelligence (AGI), also referred to as Strong AI, is a theoretical artificial intelligence that is characterized by holistic, broad intelligence, that is comparable to that of humans. Its counterpart is Weak or Narrow AI, which is an artificial system that might achieve (or even surpass) human level performance on narrow tasks, but lacks the capacity of expanding its skills and abilities beyond those.

AGI has been a theme that has bothered philosophers since before the establishment of AI or Computer Science as fields of research. This does not come as a surprise, given the intriguing nature of the subject and its ties with yet deeper philosophical questions, relating but not limited to the nature of the human mind, the source of consciousness, the descriptive ability of formal logics and mathemtics, or the connections between computation and intelligence.

From the field's very early early days of inception, AI researchers would theorize about the possibility of Strong AI and in fact even make bold claims about its imminence before the end of the century. As these initial forecasts started proving utterly ungrounded, some first discussions were initiated on the suitability of the particular systems being used, with a divide forming beteween defenders of symbolism and connectionism. With both fields eventually undergoing prolonged periods of dismissal, infamously known as the AI winters, engineering progress largely stagnated. As a result, a challenging attitude towards the plausibility of Strong AI developed, with many philosophers launching staunch attacks against the idea of genuine intelligence being reproducible through software. Even though many of these original criticisms are now dismissed as outdated, the general feeling of negativity and disillusionment towards AGI has persisted.

This is perhaps unreasonable, given the extreme breakthroughs that we have had the chance to experience within the field during the last decade. The purpose of this paper is to adopt a fresh view at the recurring issue of AGI, bringing recent technical advancements at the forefront of the discussion. The focal point will be modern neural architectures of computation, and a comparison between the paradigms of supervised and unsupervised learning. Motivating reasons as well as potential pitfalls and shortcomings of these paradigms, under the scope of AGI, will be discussed in the sections to follow.

1.1 Background

Machine Learning has been by far the most striking breakthrough in AI for the last couple of decades. New benchmarks, some well above human-level performance **to-ref**, are set on various kinds of tasks ranging from language understanding to game playing on a monthly basis. Although these have garnered quite a lot of attention, most of the focus has remained on the tasks and applications themselves, with few ongoing discussions on whether this marks the beginning of a new (potential) path towards AGI. This, of course, begs the question of why this is the case,

especially considering the extreme enthusiasm that used to emanate from the field during its first few decades, despite the disproportionally faster pace of recent progress. To put it plainly, why are people not excited?

The most obvious reason would be that, nowadays, people are not as easily convinced as they were, while scientists are far less eager to start spouting grandiose promises. History has shown that you can not really predict the future (at least within AI research), and there seems to be a repeating pattern of hype followed by stagnation. In that sense, might it be the case that we have finally learned our lesson on how to remain calm and laid back in the face of astounding progress? I doubt this is the case; what is happening, rather, is that the kind of progress we are currently experiencing is on the a side of AI that is more in line with practical applications and results rather than actual intelligence.

When people hear of machine learning, the thing that instantly comes to mind is supervised learning. Supervised learning, in a sentence, uses a set of labeled data pairs $S : \{(a_i, b_i) \in A \otimes B\}$ to approximate an underlying generator function $f : A \to B$. While it's true that supervised learning has been spearheading machine learning, with most of the popular success stories coming that way (with a few triumphant exceptions **to-ref**), supervised learning is almost wholly engineering. It gives no answer to any question (and in fact does not even attempt to); when presenting a problem to a supervised system, it will yield back a solution, but not the actual process of solving it; it is truly one of the most efficient ways of crunching data, but at the end of the day it's just that. It offers little in terms of scalability; while we might improve upon our machinery, or come up with a more stable training algorithm, or enhance the representational efficiency using neat tricks, the paradigm will always be the same: you find some labeled data, you throw them to your network and then it works (or doesn't). But that paradigm is extremely narrow and corresponds to only a very small subset of the problems out there in the world. And while it's true that you can sometimes manage to convert your problem to a nail when your only tool is a hammer, real intelligence requires adapting the solution to the problem and not the other way around.

This by no means is to imply that supervised learning as a whole is a failed effort; on the contrary it is a truly remarkable advancement, and the next section is invested in admiring why.

2 The Good

2.1 Machine Learning

Obviously, machine learning is a descendant of connectionism. As such, it does inherit all of its benefits, regardless of whether we are talking about supervised or unsupervised techniques.

Emergence Performing computation via the interaction of many small, interconnected units falls within the standard paradigm of emergence **to-ref**; 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.

Emergence is commonly occuring in nature, with the most pronounced example being biological systems. Individual organic life is enabled by the functionalities of many organs, which- however complex- are only capable of very minor, specialized tasks. At an even lower level, these organs are composed of yet smaller, further specialized subparts, whose role becomes blurry when removed from their system. The same holds for the human brain; functional neuroscience suggests that intelligence is permitted by the pathways formed between neurons in the brain.

At a higher level, complex societal structure can arise out of clusters of individuals who are unremarkable on their own, with the most striking cases being insect colonies.

Biological Plausibility The building blocks, in the case of machine learning systems, 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 abstract concepts, without having to consciously process signals from the lower levels when operating at a higher cognitive level. Functionally, it is exactly as if cognition is also organized in a layered manner, even if we cannot make claims about any similarity between the structural form of a neural network and the brain.

Relevance and Generalization A key issue within symbolic AI has been establishing relevance in the presence of an abundance of information **to-ref**. 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-ref**. 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-ref**. 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.

2.2 Supervised Learning

Aside from the pros of machine learning as a whole, supervised learning has made tremendous leaps towards tackling long-standing issues of intelligence modelling, some of them recent but others already quite old.

Sensory Processing Convolutional neural networks have been long established as the de-facto model of image processing. This framework allows for shift (and, more recently, rotation **to-ref**) invariant processing of minimally processed images, for purposes such as recognition or segmentation. It advanced image processing by a large margin, while remaining to a large degree faithful to its biological equivalent, the visual cortex. It is a prime example of how inspiration from biological systems is quite often the most efficient way of progressing artificial intelligence, and one of the big successes of supervised learning.

Internal Feedback As a means for handling temporal dependencies within sequential data, recurrent networks were conceived **to-ref**. Recurrent networks account for past data by feeding the output of their transformations into their future applications, in a step-wise fashion. In that sense, they accomplish a sort of statefulness and internal feedback, which might at first seem ad-hoc, but after closer inspection reveals itself as crucially important for general intelligence. Internal feedback loops are obviously occuring within human thought patterns; it is not uncommon for an external stimulus to provoke a thought, which in turn acts as the trigger for another thought, which finally prompts a reaction. Considering this, recurrency is more than a boost in computational efficiency **to-ref**, but also a very natural and intuitive addition to any interactive system that hopes to capture intelligence.

[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).

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. Recurrent architectures partially bypassed this by implicitly merging computation with memory through their distributed statefulness **to-ref**.

The big breakthrough, however, came by the very recent proposal of attention-based mechanisms **to-ref**. Skipping technical details, attention-based mechanisms can be described as allowing pattern matching and context-sensitive selection to happen on the basis of value similarity. On the side of sensory processing, this accounts for focusing on particular spatial or temporal regions of some high-dimensional sensory input depending on the contexts, in full accord with how human attention works. More importantly, however, it provides a tangible means of content-based addressing within neural architectures, consequently enabling auto-associative memory. Auto-associative memory is considered a fundamental property of human intelligence **to-ref**; we are admirably able of recovering 'snapshots' (either high or low level) from our long-term memory, when presented with the appropriate sensory stimuli or internal triggers (consider for instance the method of loci **to-ref**, or how a particular sound or smell can bring forth memories of people, places or events). In full opposition to this, computation would traditionally rely on value-based addressing, which requires explicitly stating the position of each particular memory (i.e. tying a variable to a memory address), which would require a dynamically changing look-up table to be somehow implemented within the human-brain, a scenario that seems at best far-fetched.

An example of one of the impressive proposals implementing attentional mechanisms are Neural Turing Machines **to-ref**, an architecture that is unique in its ability to correctly make use of a memory matrix, external but tied to the computational framework. In doing so, it gains the capacity of performing algorithmic inference (learning entire programs, if at least simple, including read/write operations, rather than just functions). This is a truly remarkable achievement that can bring computational cognition back to the forefront of the discussion.

3 The Bad

However, not all is great and well within machine learning. Leaving aside overarching issues that have been plaguing the field's motivations, ideals and scientific standards for a while (see e.g. **to-ref** for an overview), let us focus on where supervised learning in particular goes wrong as the candidate framework for AGI.

Data As stated earlier, supervised learning heavily relies on the existence of a plethora of annotated data pairs. These may have indeed become fairly abundant in the digitalized 21st century world. We still, however, operate (for the most part at least) in the real physical world, which is not labeled. We could not, therefore, throw an infant robot, even if fully equipped with the most efficient supervised algorithm, out there in the open and hope for it to learn how to think or act.

More practically, expecting a system to generalize beyond a narrow domain would require us to be able to constantly provide it with new data. In the best case, that is costly and counter-efficient. In the worst case, it is simply unfeasible; assuming that every situation we encounter be represented as a data sample, and the corresponding reaction be a mapping from that sample, who decides on the particular encoding? Even if we invested a few decades for constructing an unambiguous, universal and properly sound system of labeling everything conceivable, and then trained a machine on the generated data, wouldn't it simply fall victim to the symbol grounding problem? It might mimic human intelligence, and even encapsulate the sum of human knowledge, but its knowledge would be superficial in the sense that it was built out of a static image of the

world. Granted, its ability for generalization could be extraordinery but it would still lack the ability to learn further.

Loss Functions Another related problem, intrinsic to the way supervised learning works, is the presupposition of a loss function. But is there a loss function that universally guides intelligence? One could argue that this is in fact plausible; that we have some internal metric of success, just like a supervised algorithm, that we seek to maximize throughout the course of our lives via our thoughts and actions. Depending on which field we side with, we could find a multitude of potential candidate functions. An evolutionary biologist would claim that such a function operates on a species-wide level and concerns preservation. Zooming in to the individual, a hedonist would argue that it is the sought of joy and pleasure that guides life, to which an ethical idealist would counter argue that it is morality and personal ideals that we are based on, instead. Enumerating the various schools of thought would of course be a fruitless effort; but it is this exhaustive nature of the list that actually hints at something here. There is a plasticity in what we set as our drives; and even though particular trends might be discernible within specific societies, this plasticity is a strong indicator of the dynamic nature of the loss function of intelligence (if there is one such). And since supervised learning algorithms can only really work if given a loss function that is preset and static, i.e. set in stone before and throughout the algorithms life-cycle, we have another good reason to doubt their feasibility as truly intelligent beings.

Deductive vs. Constructive

4 The Alternative

5 Conclusion

subsectionSupervised 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.

5.1 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).

5.2 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-ref**

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.