The path to AGI is unsupervised

Konstantinos Kogkalidis [6230067]

October 17, 2018

1 Introduction

Artificial General Intelligence (AGI), also referred to as Strong AI, is a theoretical artificial construct that is characterized by holistic, broad intelligence, 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 Even though many of these original criticisms are now dismissed as outdated, the general feeling of negativity and disillusionment towards AGI has persisted.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 neural architectures of computation, and a comparison between the paradigms of supervised and unsupervised learning. Unsupervised techniques will be promoted as a stronger, more promising alternative to their supervised counterparts. Motivating reasons for this position 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 , 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 $f : A \rightarrow B$. 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), 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. We begin by briefly reviewing the most important of those.

Emergence Performing computation via the interaction of many small, interconnected units falls within the standard paradigm of emergence; 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. Several kinds of neural networks satisfy the conditions set by cognitivist hypotheses of emergence for human-level intelligence [1, 2].

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 [3]. 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 [4]. 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, allowing them to generalize over to previously unseen situations (i.e. data).

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 [5] . 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

[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).

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 value to a network that made use of all the capacitors on Earth's surface to simulate a baby (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 rotation 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 [6].

Internal Feedback Handling temporal dependencies within sequential data may be accomplished by recurrent networks. 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 [7], but also a very natural and intuitive addition to any interactive system that hopes to capture intelligence [8].

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 [9].

The big breakthrough, however, came by the very recent proposal of attention-based mechanism [4] . 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 associative memory. Associative memory (both auto or hetero-associative) is considered a fundamental property of human intelligence [10, 2]; 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. 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.

Many impressive implementations of attentional mechanisms exists. some of which unique in their ability to make use of a memory matrix, external but tied to the computational framework [11, 12]. In doing so, these architectures gain 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 from many standpoints; for one, it shows progress on the path of automated code synthesis, one of the high barriers of AI for a long while. At the same time, it may be used to bring more complete, actually implementable models of computational cognition back to the forefront of the discussion.

3 The Bad

However, supervised learning is not all roses; let us focus on where it goes wrong when it comes to AGI.

Data As stated earlier, supervised learning heavily relies on the existence of a plethora of annotated data pairs. These may have indeed become more numerous lately; the world, however, is for the most part unlabeled. Trees do not usually come with a sign that reads "*Tree*" stapled on them, and neither do animals or any other sorts of object or situation. 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 really can be represented as a data sample, and the corresponding reaction as a mapping from that sample, a key issue arises. Fist of all, who decides on the encoding scheme? 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 a world construed ad-hoc. Granted, its ability for generalization could be extraordinary; but it would still be incompetent of further learning, a trait that unambiguously renders it unintelligent.

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.

4 The Alternative

Given the few but extremely severe limitations presented above, the focus turns on how to make use of just the nice things offered by machine learning while escaping the pitfalls of supervised learning on our quest to AGI. The answer is hidden in supervised learning's underappreciated cousin, namely unsupervised learning.

Unsupervised learning architectures utilize neural computation techniques to organize their input in a coherent way by only inspecting their structural form. Classical applications of unsupervised learning include K-means clustering, encoder/decoder systems and feature compression algorithms [13]. A more modern definition would claim that unsupervised architectures revolve around learning to represent data in such a way as to allow its reconstruction given only part thereof [14].

4.1 Bypassing Limitations

Unsupervised learning suffers from none of the aforementioned issues.

Data By definition, unsupervised learning requires no example output data, thus avoiding the need for a labeled dataset. Unlabeled data are not just abundant; as the digitalized world of the 21st century tends to occupy more and more of humanity's time, unlabeled data are being generated at an unconceivably exploding pace. On top of the purely digital data, sensors could at any time be glued to a machine, allowing it to absorb the real physical world surrounding it. The continuous presence of a human oracle is not needed to instill knowledge into an unsupervised system.

Loss Function Unsupervised learning can do away with the necessity of a predefined loss function in two ways. To begin with, constructing efficient representation schemes is task-independent and can be tied with a multitude of downstream tasks; in fact, so-called unsupervised pretraining used to be common practice in the last decade, and many specialized systems still rely on unsupervised representations [15]. This already completely resolves the issue mentioned earlier; even if we are driven by many, time-varying 'loss functions', they can all still be accomodated by a unified, lower-level representation.

At the same time, generative networks, a variation of unsupervised systems that rely on sampling and approximate inference to perform variational reconstruction of some given input, are actually capable of learning their own objective functions [16, 17]. Such architectures are the exemplar of the benefits to be gained by adopting unsupervised techniques. They are the digital equivalent of a brain that learns by analogy how to construct singular mental representations for

objects belonging in the same category, and to even imagine new such objects by knowing which characteristics are its defining constants and which are the features that may vary. What is also astonishing is that a long-term change in the form of the input being provided would result in a gradual shift in the loss function being learned; is this not the kind of adaptability we would expect to see in an actually inteligent system?

4.2 Breaking Boundaries

Unsupervised learning brings more to the table than just making up for supervised learning's shortcomings.

Learning Examples vs. Learning from Examples A common point of doubt against machine learning has been the nature of learning performed. I would like to argue that unsupervised learning nullifies this doubt, at least to an extend. First off, it has been pointed out that unsupervised learning requires no labeled data; yet it may still benefit from it, in a way that is completely congruent with human learning. A toddler does not need to inspect a thousand pictures of a cat, each time being reminded "*this is a cat!*", before inferring what constitutes a cat. Similarly, an unsupervised system is able to distinguish between categories by simply being shown many unlabeled instances of them (serving as the basis for structural separation) and a few labeled ones (allowing the already distinct classes to be identified with a name) [18]. In that way, part of the weight of learning has been moved from the specific examples themselves to the learner and its environment.

Imagination and Creativity Human intelligence is not bound to knowledge transfer; we may inspect our environment and learn from examples, but we also generate new knowledge by performing leaps of intuition. These leaps require, if at least momentarily, breaking constraints set by past knowledge, sometimes using current information; they are a product of our creative ability and our imagination. There is a lack of consensus on where creativity originates from; but approximate bayesian inference, as used in unsupervised systems, is the most suitable applicable theory for modeling it. Let's use our ability of imagination to consider that, just like bayesian networks, a probability distribution function describes the generative process of our thoughts and ideas, with most of the probability mass at each junction point being directed towards conventional world theories, guided by our past observations and knowledge. Such a system would warrant that our thought patterns usually do not stray too far into uncharted territory, as this would equate following a chain of improbable junctions; yet, every once in a while, following unlikely paths leads to new equilibria or optimal points, forcing a redistribution of the probability mass and reshaping our world views.

5 Conclusion

Human intelligence may be thought of as the sum of a few core properties; perception, planning, memory, learning, common sense and self-awareness. The first four have already successfully been integrated within AI systems, either in the form of supervised or reinforcement learning architectures. Common sense is the next big thing to tackle; and unsupervised learning provides us the necessary framework to do just that. The world is full of spatio-temporal regularities, the manifestions of which we simply assume for the majority of the time. Unsupervised learning's ability to reconstruct the world (and hopefully, soon enough, predict its future trajectories [14]) gives us reason to hope for new, exciting advancements in the upcoming years. Even though

it does not yet address the crucial issue of meta-cognition, it at least gives a realistic, feasible direction to follow today.

References

[1] M. P. Penna, P. K. Hitchcott, M. C. Fastame, and E. Pessa, *Emergence in Neural Network Models of Cognitive Processing*, pp. 117–126. Cham: Springer International Publishing, 2016.

[2] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities.," *Proc Natl Acad Sci U S A*, vol. 79, pp. 2554–2558, Apr 1982. 6953413[pmid].

[3] H. L. Dreyfus, *What Computers Can't Do*. Harper & Row, 1972.

[4] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," *CoRR*, vol. abs/1406.6247, 2014.

[5] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, pp. 303–314, Dec 1989.

[6] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence," *Sci Rep*, vol. 6, p. 27755, Jun 2016. 27282108[pmid].

[7] H. Siegelmann and E. Sontag, "On the computational power of neural nets," *Journal of Computer and System Sciences*, vol. 50, no. 1, pp. 132 – 150, 1995.

[8] J. Hawkins. Times Books, 2004.

[9] G. Hinton, "Lecture notes in advanced machine learning, lecture 10," 2013.

[10] M. S. Fanselow and A. M. Poulos, "The neuroscience of mammalian associative learning," *Annual Review of Psychology*, vol. 56, no. 1, pp. 207–234, 2005. PMID: 15709934.

[11] S. Sukhbaatar, a. szlam, J. Weston, and R. Fergus, "End-to-end memory networks," in *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds.), pp. 2440–2448, Curran Associates, Inc., 2015.

[12] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwinska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, A. P. Badia, K. M. Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu, and D. Hassabis, "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, pp. 471 EP –, Oct 2016. Article.

[13] Z. Ghahramani, *Unsupervised Learning*, pp. 72–112. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[14] Y. LeCun, "Predictive learning, invited talk at thirtieth conference on neural information processing systems.".

[15] D. Erhan, Y. Bengio, A. C. Courville, P. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[18] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 594–611, Apr. 2006.