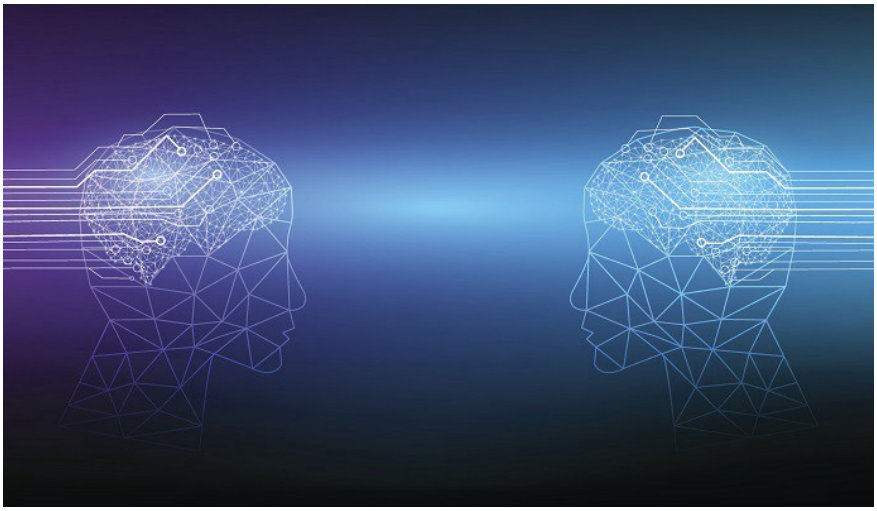
Artificial Intelligence in 2019

Eugenio Culurciello Follow
Dec 21 · 11 min read

Or how machine learning is evolving into AI



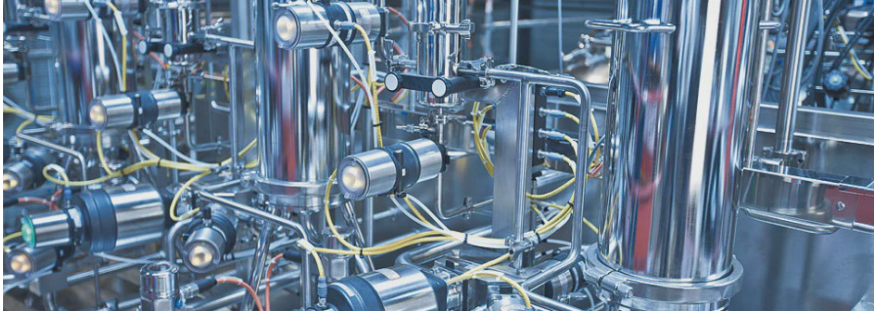
These are my opinions on where deep neural network and machine learning is headed in the larger field of artificial intelligence, and how we can get more and more sophisticated machines that can help us in our daily routines.

Please note that these are not predictions of forecasts, but more a detailed analysis of the trajectory of the fields, the trends and the technical needs we have to achieve useful artificial intelligence.

We will also examine low-hanging fruits, such as applications we can develop and promote today!

Goals

The goal of the field is to produce machines with beyond-human abilities. Autonomous vehicles, smart homes, artificial assistants, security cameras are a first target. Home cooking and cleaning robots are a second target, together with surveillance drones and robots. Another one is assistants on mobile devices or always-on assistants. Another is full-time companion assistants that can hear and see what we experience in our life. One ultimate goal is a fully autonomous synthetic entity that can behave at or beyond human level performance in everyday tasks.



Software

Software is defined here as neural networks architectures trained with an optimization algorithm to solve a specific task.

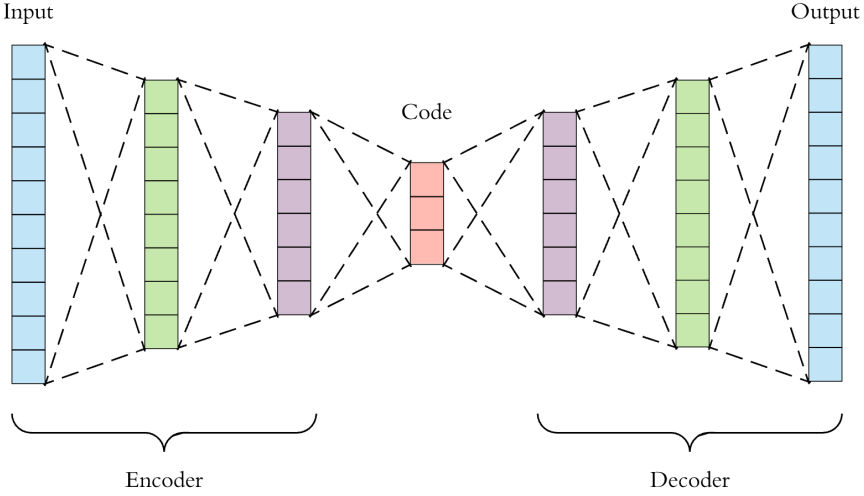
Today neural networks are the de-facto tool for learning to solve tasks that involve learning supervised to categorize from a large dataset.

But this is not artificial intelligence, which requires acting in the real world often learning without supervision and from experiences never seen before, often combining previous knowledge in disparate circumstances to solve the current challenge.

> *How do we get from the current neural networks to AI?*

1- Neural network architectures—when the field boomed, a few years back, we often said it had the advantage to learn the parameters of an algorithms automatically from data, and as such was superior to hand-crafted features.But we conveniently forgot to mention one little detail… the neural network architecture that is at the foundation of training to solve a specific task is not learned from data! In fact it is still designed by hand. Hand-crafted from experience, and it is currently one of the major limitations of the field. Neural network architectures are the fundamental core of learning algorithms. Even if our learning algorithms are capable of mastering a new task, if the neural network is not correct, they will not be able to. But there is much activity in this area, and we reviewed it here. The problem on learning neural network architecture from data is that it currently takes too long to experiment with multiple architectures on a large dataset. One has to try training multiple architectures from scratch and see which one works best. Well this is exactly the time-consuming trial-and-error procedure we are using today! We ought to overcome this limitation and put more brain-power on this very important issue.

2- Limitations of current neural networks—We have talked about before on the limitation of neural networks as they are today. Cannot predict, reason on content, and have temporal instabilities—we need a *new kind of neural networks* that you can about read here.



Encoder-Decoder neural network

Connecting to the topic of the previous section, neural networks are evolving into encoder-decoders, where the encoder is a network that compresses data into a short code (representation) and the decoder is expanding that representation to generate another larger representation (think of these as generated images, mental simulations, highlights on an image as bounding boxes and segmentation masks). We have talked about how to use such networks to localize and detect key-points in images and video extensively here; see also this analysis. This is also the main ingredient in predictive neural networks (more below).

3- Unsupervised learning—we cannot always be there for our neural networks, guiding them at every stop of their lives and every experience. We cannot afford to correct them at every instance, and provide feedback on their performance. We have our lives to live! But that is exactly what we do today with supervised neural networks: we offer help at every instance to make them perform correctly. Instead humans learn from just a handful of examples, and can self-correct and learn more complex data in a continuous fashion. We have talked about unsupervised learning extensively here.

4- Predictive neural networks—A major limitation of current neural networks is that they do not possess one of the most important features of human brains: their predictive power. One major theory about how the human brain work is by constantly making predictions: predictive coding. If you think about it, we experience it every day. As you lift an object that you thought was light but turned out heavy. It surprises you, because as you approached to pick it up, you have predicted how it was going to affect you and your body, or your environment in overall.

Prediction allows not only to understand the world, but also to know when we do not, and when we should learn. In fact we save information about things we do not know and surprise us, so next time they will not! And cognitive abilities are clearly linked to our attention mechanism in the brain: our innate ability to forego of 99.9% of our sensory inputs, only to focus on the very important data for our survival—where is the threat and where do we run to to avoid it. Or, in

the modern world, where is my cell-phone as we walk out the door in a rush.

Building predictive neural networks is at the core of interacting with the real world, and acting in a complex environment. As such this is the core network for any work in reinforcement learning. See more below.

We have talked extensively about the topic of predictive neural networks, and were one of the pioneering groups to study them and create them. For more details on predictive neural networks, see here, and here, and here.

5- Continuous learning—this is important because neural networks need to continue to learn new data-points continuously for their life. Current neural networks are not able to learn new data without being re-trained from scratch at every instance. Neural networks need to be able to self-assess the need of new training and the fact that they do know something. This is also needed to perform in real-life and for reinforcement learning tasks, where we want to teach machines to do new tasks without forgetting older ones. Continuous learning also ties in with *transfer learning,* or how do we have these algorithms learn on their own by watching videos, just like we do when we want to learn how to cook something new? That is an ability that requires all the components we listed above, and also is important for reinforcement learning.

For more detail, see our summary of recent results.

6- Reinforcement learning—this is the holy grail of deep neural network research: teach machines how to learn to act in an environment, the real world! This requires self-learning, continuous learning, predictive power, and a lot more we do not know. There is much work in the field of reinforcement learning and we already talked about this here and more recently here.

Reinforcement learning is often referred as the “cherry on the cake”, meaning that it is just minor training on top of a plastic synthetic brain. But how can we get a “generic” brain that then solve all problems easily? It is a chicken-in-the-egg problem! Today to solve reinforcement learning problems, one by one, we use standard neural networks:

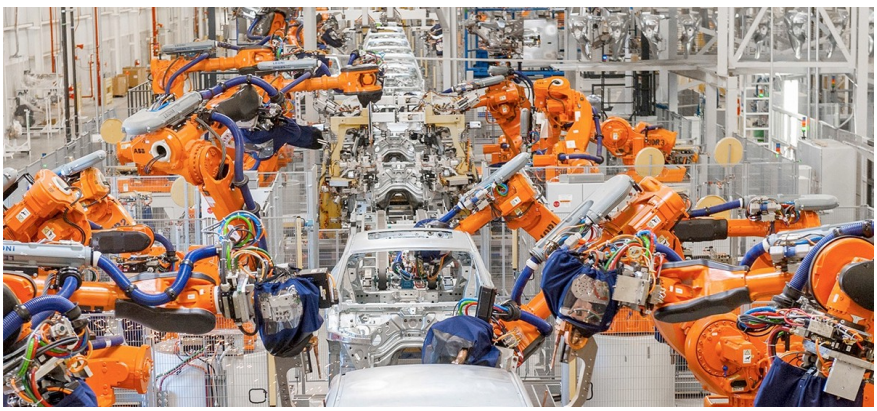
- a deep neural network that takes large data inputs, like video or audio and compress it into representations
- a sequence-learning neural network, such as RNN, to learn tasks

Both these components are what everyone uses because they are some of the available building blocks. Still, results are unimpressive: yes we can learn to play video-games from scratch, and master fully-observable games like chess and go—this year even trained overnight! —but I do not need to tell you that is nothing compared to solving problems in a complex world and machines that can operate like us.

We believe that predictive neural networks are indispensable for reinforcement learning. Curiosity, as it is called today in the field! Stay tuned for more!

8- No more recurrent neural networks—recurrent neural network (RNN) are falling out of vogue. RNN are particularly bad at parallelizing for training and also slow even on special custom machines, due to their very high memory bandwidth usage—as such they are memory-bandwidth-bound, rather than computation-bound, see here for more details. Attention based and especially convolutional neural networks are more efficient and faster to train and deploy, and they suffer much less from scalability in training and deployment.

We have already seen that convolutional and attention based neural network are going to slowly supplant speech recognition based on RNN, and also find their ways in reinforcement learning architecture and AI in general.



Hardware

Hardware for deep learning is at the core of progress. Let us now forget that the rapid expansion of deep learning in 2008–2012 and in the recent years is mainly due to hardware:

- cheap image sensors in every phone allowed to collect huge datasets—yes helped by social media, but only to a second extent
- GPUs allowed to accelerate the training of deep neural networks

And we have talked about hardware extensively before. But we need to give you a recent update! In the last 2 years saw a boom in the are of machine learning hardware, and in particular on the one targeting deep neural networks. We have significant experience here, having designed 5 generations of deep neural network accelerators (see the most recent at FWDNXT).

There are several companies working in this space: NVIDIA (obviously), Intel, Nervana, Movidius, Bitmain, Cambricon, Cerebras, DeePhi, Google, Graphcore, Groq, Huawei, ARM, Wave Computing, etc. All are developing custom high-performance micro-chips that will be able to train and run deep neural networks.

The key is to provide the lowest power and the highest measured performance while computing recent useful neural networks operations, not raw theoretical operations per seconds—as many claim to do.

But few people in the field understand how hardware can really change machine learning, neural networks and AI in general. And few understand what is important in micro-chips and how to develop them. A couple of ideas:

- Architectures: many see computer architectures as just an array of multipliers and adders. But not all architectures are the same. Some better than others can minimize memory bandwidth and keep all units occupied at all times.
- Compilers: many think the hardware is not important, that a neural network compiler is the key. But when you design your own hardware, the compiler is just interpreting the neural network computational graph in optimized machine code. Open-source compilers (many of which came out last year!) can only help so much, given that the most difficult step really depends on the secret architecture. While open-source compilers can be used as front-ends, there is still a lot of secret sauce at the intersection between hardware architecture and neural network graphs.
- Microchips: once an algorithm is important, the best way for optimizing performance per power is to make a custom microchip or ASIC or SoC. They can provide faster clocks and a smaller circuit area than FPGAs. FPGAs are including deep neural network accelerators now, expect them in 2019–2020, but microchips will always be better performers.
- Advances: there are several advances that will allow silicon deep neural network accelerators to gain 10–20x more performance per power easily, even if microchip scaling is not used. Look for advances in using lower numbers of bits, systems-on-a-package, advanced memories, to name a few.

About neuromorphic neural networks hardware, please see here. A commentary on imitating real neural networks is here.

Applications

We talked briefly about applications in the Goals section above, but we really need to go into details here. How is AI and neural network going to get into our daily life?

Here is our list:

- **categorizing images and videos**—already here in many cloud services. The next steps are doing the same in smart camera feeds—also here today from many providers. Neural nets hardware will allow to remove the cloud and process more and more data locally: a winner for privacy and saving Internet bandwidth.
- **speech-based assistants**—they are becoming a part of our lives, as they play music and control basic devices in our "smart" homes. But dialogue is such a basic human activity, we often give it for granted. Small devices you can talk to are a *revolution* that is happening right now. Speech-based assistants are getting better and better at serving us. But they are still tied to the power grid. The real assistant we want is moving with us. How about our cell-phone? Well again hardware wins here, because it will make that possible. Alexa and Cortana and Siri will be always on and always with you. Your phone will be your smart home—very soon. That is again another victory of the smart phone. But we also want it in our car and as we move around town. We need local processing of voice, and less and less cloud. More privacy and less bandwidth costs. Again hardware will give us all that in 1–2 years.
- **the real artificial assistants**—voice is great, but what we really want is something that can also see what we see. Analyze our environment as we move around. See an example here and ultimately here. This is the real AI assistant we can fall in love with. And neural network hardware will again grant your wish, as analyzing video feed is very computationally expensive, and currently at the theoretical limits on current silicon hardware. In other words a lot harder to do than speech-based assistants.But it is not impossible, and many smart startups like AiPoly already have all the software for it, but lack powerful hardware for

running it on phones. Notice also that replacing the phone screen with a **wearable glasses-like device** will really make our assistant part of us!

- **the cooking robot**—the next biggest appliances will be a cooking and **cleaning robot**. Here we may soon have the hardware, but we are clearly lacking the software. We need transfer learning, continuous learning and reinforcement learning. All working like a charm. Because you see: every recipe is different, every cooking ingredient looks different. We cannot hard-code all these options. We really need a **synthetic entity** that can learn and generalize well to do this. We are far from it, but not as far. Just a handful of years away at the current pace of progress. I sure will work on this, as I have done in the last few years~

> *This blog post will evolve, like our algorithms and our machines. Please check it again soon.*

About the author

I have almost 20 years of experience in neural networks in both hardware and software (a rare combination). See about me here: Medium, webpage, Scholar, LinkedIn, and more…

Donations

If you found this article useful, please consider a donation to support more tutorials and blogs. Any contribution can make a difference!