# Artificial Intelligence: limits of Deep Learning approaches and ideas toward a human-like intelligence

**Michele Puppin** ID 1227474 michele.puppin@studenti.unipd.it

(Dated: July 1, 2021)

In this essay I will present the paper *Deep Learning: A Critical Appraisal* authored by Marcus [1] to summarize the challenges that Deep Learning approaches are facing nowadays. I will briefly introduce Artificial Neural Networks architectures functioning and I will highlight the limiting aspects of this technology. Then, I will present the paper *Building Machines That Learn and Think Like People* authored by Lake et al. [2] to briefly introduce some ideas to combine the strengths of recent Neural Network advances with more structured cognitive models to aim to a human-like intelligence.

## I. INTRODUCTION

Artificial Intelligence (AI) research has gained more and more attentions since its dawn in the fifties of the last century. Recent developments and the increasing computational power available has lead to an extensive use of AI for many applications. Investigating AI not only allows to automatize tasks in everyday world applications, but also to shed some light in understanding how human cognition works.

Among several approaches, Artificial Neural Networks (ANN), and in particular Deep Learning (DL), are the state-of-the-art algorithms to tackle tasks such as image recognition, speech recognition and game playing. In these tasks, these algorithms have reached super-human performances. Despite its success and its human-brain inspiration, DL is intrinsically different from human intelligence. In this scenario, new paradigms can be proposed with the aim to build truly human-like learning and thinking machines.

## II. DEEP NETWORKS ARCHITECTURE

ANN are biologically inspired computing systems composed by a set of processing units called *neurons*, organized in a network. An artificial neuron receives a signal, processes it and can signal neurons connected to it. The strength (weight) of the connection, analogously to the synaptic strength, encodes how much a neuron influences another neuron. By adapting network structure, i.e. finding the optimal set of weights, the network can learn to perform desired tasks and to develop internal representations of the environment.

ANN are mostly used as classification systems in a supervised learning framework. While for linearly separable problems shallow architecture are sufficient, linearly inseparable problems can be solved by networks with a deep structure, namely a number of hidden layers grater than one. DL architectures are able to ap-

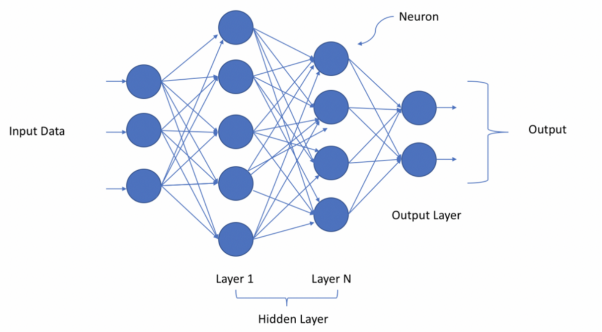


**FIG. 1** Feedforward ANN schematic representation. Neurons are organized in layers: an input layer, some hidden layers and an output layer. Connections are characterized by a weight.

proximate a function, allowing to find a mapping from an input to an output. Learning in this architectures is achieved mainly with backpropagation algorithms which adjust the weights to reduce the error between the output and the target. The hierarchical structure of the network plays a fundamental role, allowing to approximate very complex functions. In deeper layers the network is able to encode complex and abstract information. Some features implicitly embedded in input data emerge only in the deepest layers and can be used for machine learning tasks. Numerosity perception, for example, emerges spontaneously as a high-order feature in deep neural networks. [3]

In supervised learning applications of DL, a training set is used to teach the model to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. Ideally, with infinite data and computational resources, DL would be able to learn any mapping. Nevertheless, in real world applications the amount of data is limited; often the algorithm has to generalize beyond the specific set of data it has been trained on. The algorithm does not just work as a memory storage of the training set, indeed, but it learns regularities in data. For this sake, several supplementary techniques, called regularization, are implemented. For example, the number of hidden layers can be limited and weight decay rules can be implemented by introducing different penalty terms for the weights in the error function. Moreover, dropout can be used to prevent the network to strongly depend on specific neurons. Finally, early stopping can be exploited to avoid overfitting by stopping the training procedure as soon as the performances of the model on the validation data are not improving. [4]

## III. DEEP LEARNING CHALLENGES

Although in many fields DL has achieved super-human abilities, this technology presents many limitations.

Firstly, human beings can learn abstract relationships in a few trials. Such an abstraction can be acquired both by providing explicit definition and by inferring implicit information. Instead, DL algorithms require a large number of examples to learn simple rules and lack a mechanism for learning abstractions through explicit, verbal definitions.

Secondly, despite the adjective *deep*, the patterns extracted by DL have proved to be more superficial than expected. For example, when trained to play the brick-breaking Atari game *Breakout*, the algorithm learns that digging a tunnel through the wall is the most effective technique to beat the game. Indeed, it doesn't really understand the concept of tunnel or wall, but it just learns specific contingencies for particular scenarios.

Moreover, DL has no intrinsic capability to analyze hierarchical structures. For instance, DL language models represent sentences as sequence of words and do not understand language hierarchical structure. This because, DL learns correlations between sets of features that are themselves nonhierarchical such as the sequential position of words presented in a sequence.

Then, DL algorithms are far from drawing open-ended inferences from real-world knowledge with performances comparable to human beings. For example, machine reading systems can easily deal with tasks in which the answer to a question is explicitly contained in the text. Instead, they struggle with tasks that require to infer concepts that go beyond what is explicitly contained in the text.

DL algorithms are often referred to as a *black box* that can not be interpreted or understood. It is not accessible to describe in a meaningful way how an output is produced in response to a given input. Nevertheless, AI is expected to be characterized by transparency and responsibility. Being able to explain how an AI algorithm has achieved to solve a task is fundamental to build cognitive models. Also, AI algorithms should be able to justify their choices when applied to fields like medical diagnosis and financial trades. Many domains in which AI is being, or is going to be largely exploited, i.e. self-driving vehicles, can easily have legal and ethical implications.

Human beings make a large use of prior knowledge when learning new tasks. Albeit the use of pre-trained models and *recycling* techniques has proved to be effective [5], DL does not provide intrinsic integration with prior knowledge. In fact, both for technical difficulties and for explicit choice of developers, DL is designed as self-contained and isolated from other, potentially useful knowledge.

In addition, DL has shown to be unable to discriminate causation from correlation. Indeed, it can easily learn complex correlations emerging from data, but it can not build any representation of causality. Often correlation is far from being enough to interpret results and spurious correlation frequently emerges from data (here some examples of spurious correlations [6]).

One of the domain in which DL has by far outperformed human beings is game playing. In fact, DL algorithms can achieve optimal performances in highly stable environments, like games with well defined and immutable rules. In several variable and unreliable scenarios, such as politics and economy, performances can be drastically reduced.

Despite high-level achievements, often DL algorithms can not be fully trusted, indeed they can be easily fooled. This is the case of adversarial attacks. By injecting an appropriate small percentage of noise in the input we can produce changes that are not even distinguishable for human being but still can fool DL algorithms. For example,

self-driving vehicles can be tricked by placing stickers on road signs creating issues for transport security, cybersecurity and survaillance. [7]

Eventually, in human cognition context plays a key role. The context can be defined as any information (implicit, explicit or background) that can be used to characterize the situation of a person, a place or an object. A context-aware agent uses context to provide relevant information to the user, where relevancy depends on the user's task. Context-awareness is not well integrated in DL and would be useful in drawing conclusions based on the environment and adapt to its changes. [8]

## IV. BEYOND DEEP LEARNING FOR A HUMAN-LIKE AI

Looking at the aspects discussed above, we can state that we are far from creating a human-like AI, both for the sake of engineering application and cognition science. In the following paragraphs, some crucial aspects, inspired by cognitive capabilities present early in human development, are considered with the aim of designing more human-like learning and thinking machines.

First of all, in early stages of development, human beings show a foundational understanding of several core domains, such as number, space, physics and psychology.

Knowledge of intuitive physics has been found in very young children and it is used for solving many everyday physics-related tasks. In fact, they are able to expect object to show behaviours compatible with basic physics law and use these expectations to guide learning process. Several approaches has been inspected to integrate this kind of basic knowledge. Intuitive physical reasoning could be modeled as an inference process over a physics simulator. Current ANN are trained to make predictions without simulating physics, but could be trained to emulate a general-purpose physics simulator.

Intuitive psychology, as well, has important influence on human learning and thought. Infants can distinguish animate agents from inanimate objects and expect agents to act contingently and reciprocally, to have goals, and to take efficient actions towards those goals. Integrating intuitive psychological reasoning in AI means to include representations of agency, goals, efficiency, and reciprocal relations. This can be incorporated in ANN as built in or may arise as a simple inductive bias, integrating reasoning about more abstract concepts of agency.

Moreover, DL algorithms implement learning merely as a process of gradual adjustment of connection strengths to implement pattern recognition on a large training set. Human beings, instead, show richness and flexibility in learning tasks. This suggest that learning should be implemented as model building. To this aim, three ingredients should be considered. Firstly, AI algorithms should integrate compositionality, namely the idea that new representations can be constructed through the combination of primitive elements. Secondly, causal models should be implemented in order to distinguish the direction of causality from examples. Finally, with strong priors learning a new task can be accelerated through previous learning of other related tasks. Bayesian Program Learning (BPL) represents a solution to integrate these ideas. In these algorithms concepts are represented as probabilistic generative models expressed as structured procedures in an abstract description language. BPL has shown to be capable of learning a large class of visual concepts from just a single example and generalizing in ways that are mostly indistinguishable from people. [9]

Eventually, human beings show optimal timing performances in perceiving and acting. ANN, as well, are very fast to produce an output once trained. Rich model-building learning mechanism, instead, are often slow and could be accelerated integrating ANN that "learn to do inference". By learning to recognize patterns in these inferences, the outputs of inference can be predicted without having to go through costly intermediate steps. In this way, it could be explained how human minds can understand the world so well, so quickly.

## V. CONCLUSIONS

In current AI research, DL represents a popular and promising implementation with optimal performances in many domains. In a context of general excitements for DL encouraging results, the paper authored by Marcus [1] presents several limitations both in terms of reliability in engineering applications and of providing useful insights on human cognition. The weaknesses highlighted suggest that this approach in not sufficient for designing a human-like AI.

To this aim, alternative approaches should support and integrate DL systems. The paper authored by Lake et al. [2] presents some key ingredients of human cognition that should be considered to design systems where learning is based on the process of model-building. Indeed, cognition is about using these models to understand and explain the world and to plan actions.

Only by analyzing actual technology limitations and promising new approaches, modern research will be able to design a more human artificial intelligence.

## REFERENCES

[1] G. Marcus. Deep learning: A critical appraisal. *ArXiv*, abs/1801.00631, 2018.
[2] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people, 2016.
[3] Ivilin Stoianov and Marco Zorzi. Emergence of a "visual number sense" in hierarchical generative models. *Nature neuroscience*, 15:194–6, 02 2012.
[4] Jan Kukacka, Vladimir Golkov, and Daniel Cremers. Regularization for deep learning: A taxonomy. *CoRR*, abs/1710.10686, 2017.
[5] Alberto Testolin, Ivilin Stoianov, and Marco Zorzi. Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nature Human Behaviour*, 1(9):657–664, Sep 2017.
[6] Spurious correlations. https://www.tylervigen.com/spurious-correlations.
[7] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.
[8] Mario Pichler, Ulrich Bodenhofer, and W. Schwinger. Context-awareness and artificial intelligence. *ÖGAI Journal*, 23:4–, 04 2004.
[9] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.