HOW TO LEARN PERMANENTLY

ACHILLES HEEL OF ARTIFICIAL NEURAL NETWORK

Mykola Rabchevskiy
Mar 12 ♡ 1

Despite the lack of a generally accepted definition of AGI, the ability to permanently learn (including self-learning), explicitly or implicitly, is one of the requirements that are present in any of the many variants of the definition of AGI. Therefore, AGI developers need to understand the underlying consequences of such a requirement so that the implemented approach does not contradict this requirement. Otherwise, the used technique will be unacceptable after spending resources on something that obviously cannot lead to success.

Learning at a technological level means adding new knowledge to the existing knowledge set, removing a piece of knowledge that turned out to be inadequate, or correcting a corresponding knowledge portion. It is natural and obvious for systems acceptable in practice to require that the addition, modification, and removal of a piece of knowledge does not distort the rest of the accumulated knowledge.

In turn, to correct or delete a piece of knowledge, it is necessary to specify it somehow. In other words, the entire set of knowledge must be fragmented into parts with a unique designator for each fragment. In such designators, it is easy to see what distinguishes symbolic systems from systems with distributed memory - unique pointers to pieces of

Correcting the system's undesirable behavior in the symbolic system is reduced to modifying individual fragments of knowledge while the rest remain unchanged. Taken together, this constitutes what is called ***explainability*** and ***correctability*** (explainability has another meaning - the ability to explain the decisions made to the user of the system; this is a different but no less important aspect).

Correction of behavior by modifying the knowledge accumulated by the AGI system is a particular case of ***transferring knowledge*** from one knowledge carrier to another, in the case of correction - from a human to the AGI system. A similar mechanism can be used to implement exchanging personal knowledge, which is inevitably different for distinct instances of a permanently learning AGI system. This makes it possible to organize the effective collective accumulation of knowledge - precisely as it happens in human society.

Neural networks (in the sense that this term is being put into now) implement a distributed memory model in which any change in information potentially affects all numerical parameters, which determines the accumulated knowledge and skills. There is also no possibility (or at least extremely difficult) of finding the causes of undesirable system behavior, behavior correction, and the transfer of knowledge between neural networks. Given the above, it becomes clear that neural networks cannot be used as the AGI system's ***core***.

Of course, known ANN variants with parameters changed during the system's running (changing link weights and possibly other parameters). Formally, this means the possibility of permanent learning, but since the use of distributed memory does not allow changing knowledge ***selectively***, not only is the integrity of previously obtained knowledge not guaranteed but on the contrary, their permanent corruption in the learning process is assured. This, in most cases, makes such systems unacceptable for practical use.

The inappropriateness of neural networks as the basis of AGI at first glance may seem absurd since the human brain is obviously a neural network. But this is an imaginary contradiction: the models of neural networks used at present (namely, we are talking about them) are radically different from natural neural networks. The human brain, unlike the ANN, is capable of permanent learning, is capable of manipulating symbolic information, and demonstrates an evident ability to add new knowledge and modify existing knowledge without damaging the rest of the knowledge and skills. The learning process technology differs no less radically: in artificial neural networks, an algorithm external to the network calculates corrections to the network parameters. It is obviously

completely impossible in the case of a natural nervous system - a calculator external to the brain for corrections to the parameters of neurons and connections obviously does not exist. The list of differences is easy to continue. Still, the essence boils down to that, due to natural evolution, the ***brain acquired the structure and functionality typical of symbolic systems*** - with the separation of the functions of storing information in symbolic form and information processing functions.

The "analog" nature of neurons does not mean using a distributed memory combined with an information processing system (as in the ANN), just as the ***analog*** computer basic ***elements*** do not interfere with the use of a ***discrete representation*** of information in the form of binary values. Digital computers are also a product of the evolution of analog computers that used distributed memory combined with an information processing system. To build digital (that is, symbolic) systems from analog elements, it is sufficient to have basic nonlinear active components that can form flip-flop elements with ***two stable states***.

The above, of course, does not exclude the ANN's use as a component of the AGI system, which is not responsible for accumulating knowledge but implements a beneficial transformation of data according to fixed rules formed in training. Training, in this case, plays the role of a specific way of programming such components. Correcting the rules for processing information in the ANN requires re-training, which is usually very time-consuming and, accordingly, expensive. But due to the cheapness of replicating the trained component, this is not a significant obstacle.

An obvious option for using the ANN as part of an AGI system is the role of a "***smart sensor***" - an intermediary between actual sensors and the core of the AGI system, in particular, for processing the preliminary transformation of visual information into a symbolic form. Apparently, it is possible to use the ANN as a converter, transforming the actuators' command into a detailed chain of actions, such as an "***artificial innate reflex***."

Naturally, the question arises: how then to explain the intensive development of ANN, their improvement and application despite the unsuitability for the role of the basis of AGI? There is no need for AGI in many applications - it is enough (and possibly cost-effective) to use systems that are incapable of permanent learning. This applies to those massive Internet services where erroneous decisions do not lead to severe consequences, and the lack of explainability and correctability is acceptable. Examples are advertising, recommendations, and annotation systems based on object recognition in images and videos, recognition of user behavior patterns, computer games, and so on.

Second, the possibility of permanent learning includes ***self-learning***, and self-learning, in turn, may include ***experimentation*** when the AGI system decides to perform actions with a ***knowingly unknown*** result. In critical systems, this is unacceptable, so there are concerns about the possibility and loss of control over too intelligent and too independent AGI systems. This does not contribute to rapid progress in the development of AGI. The design of the AGI system should provide for ***two modes*** of activity: the ***learning*** (or self-learning) ***mode***, in which the system is allowed (and even encouraged) to experiment with expanding the volume of knowledge and skills (for example, driving a car on a training ground without passengers), and the ***operational mode*** when experimentation is prohibited; this complicates the system.

Many problems arise in terms of the exchange of knowledge between AGI systems - along with the possibility of effective ***collective accumulation of knowledge and skills***, we got the danger of ***misinformation***, ***malicious manipulation***, etc. In the end, we have a whole bunch of problems known with human society.

Nevertheless, there are areas where the use of AGI seems to be reasonably practical and effective, so developments in this direction will obviously not be stopped. The heterogeneity and imperfection of humanity guarantee the ***inevitability of the emergence of versions of the real AGI***, despite public doubts about the dangers of its use. The history of the development of technology suggests that any new technology creates opportunities for progress and the possibility of using technology to the detriment of society. Still, the net effect of innovations, in the end, turns out to be positive.

SUMMATION

- *The Achilles heel of ANN, which prevents their use as the basis of AGI, is distributed memory, fused with information processing.*
- *Modern ANNs can be useful as an AGI component.*
- *The need to use a symbolic approach to create AGI leads to the increased complexity of the AGI system, but it's worth it.*

Subscribe

← Previous Next →

Write a comment…

Ready for more?

Subscribe