The Strengths and Weaknesses of Artificial Intelligence

By

Robert Hink

Introduction

Artificial Intelligence (AI) was inaugurated as an academic discipline in 1956. The fascination with AI has predated this date by centuries, however. The tales of Golem, The Sorcerer's Apprentice, and Frankenstein's Monster are historic examples of humankind's fascination of creating an artifact in a human's own image.

AI has gone through periods where this promise was questioned if not discredited. The dream of AI seemed like only that. These "AI winters" should have been its demise. However in the current decade, it is viewed both with great hope and with great trepidation (History of AI).

With the advent of high-performance computer chips, such as graphical processing units (GPUs) and low-cost storage, AI emerged from the research lab. AI programs no longer were solving simple puzzles (N-Queens Problem) or games like Tic-Tac-Toe. It now was beating grandmasters of Chess (Deep Blue) and Go (AlphaGo) at their own games.

More recently, AI Chatbots such as OpenAI's ChatGPT and Google's Bard have seen phenomenal success at solving complex problems and doing in-depth research. These systems use large language models (LLMs) with Transformers that take advantage of the sequential nature of

natural language. These chatbots have put the potential of AI in the hands of everyone.

In this article, I will give my assessment of the strengths and weaknesses of AI.

Background

In 1950, Alan Turing proposed a test to address the question of whether machines could think. This test, originally called the "imitation game", became known as the Turing Test. The idea behind the test was to compare responses to questions posed by a human interrogator to a "thinking" entity either a device or another human being. The human interviewer cannot see the interviewees, either device or human. Only messages are passed between the interviewees and the interviewer. If after a period of questioning the interviewer is unable to reliably identify the device or human, the device is considered to have passed the Turing Test. The device then would be considered to possess Artificial General Intelligence (AGI) whereby it can behave autonomously.

There have been several examples where some devices have been believed to have passed the Turing Test. Eliza is one example which was able to fool some people that it was human, but this case was not the true Turing Test since the person interacting with Eliza was expecting to interact with a human. In the Turing Test, the interviewer knows prior to questioning that one of the interviewees is a machine.

The chatbots, on the other hand, have demonstrated that an artificial agent may be able to attain true AGI.

Types of Machine Learning

Machine Learning (ML) is a subdiscipline of AI. In essence, the idea behind ML is that the intelligent agent learns to perform a task by being exposed to large amounts of data of the task being performed correctly and incorrectly. ML has been the driving force behind the recent resurgence of interest in AI.

ML algorithms fall into 3 varieties: supervised learning, unsupervised learning, and reinforcement learning. With all three, a great deal of data is needed to train the program. The algorithms all have a set of parameters that are adjusted to minimize error between the current output and the desired output.

With Supervised Learning the correct answers are known and provided to the algorithm. There could be thousands of inputs and outputs. The complexity of the problem determines how much training data is needed. Millions, if not billions, of examples may be required to train the program adequately. After each trial or batch of trials, the algorithm measures its error and then adjusts its parameters to reduce the error. The algorithm is looking for a global minimum error or a global maximum accuracy.

There is no guarantee that the minimum error can be determined. The learning may stall in a local minimum error and never discover the global minimum error. Alternatively, the algorithm may over adjust the parameters of the model and continuously bounce around the minimum error without finding it. Also, some functions to be learned may be discontinuous or change abruptly. In these cases, the 1st derivative of the function is undefined at certain points (called singularities) (c.f. Figure 1). Should this occur, the algorithm may never learn the function.

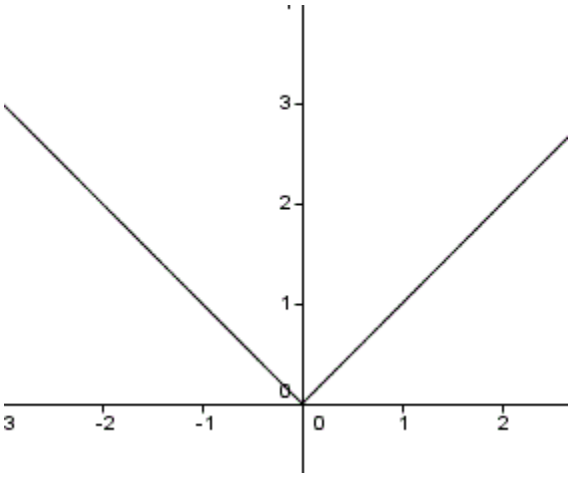

Figure 1: Non-Learnable Function (Non Differentiable Functions)

There are techniques to deal with these anomalous situations, but the fact remains that not all functions are learnable by ML algorithms.

Unsupervised Learning, as the name implies, does not provide the answers with the training data. Rather the algorithms are designed to identify structure in the data set. If clusters of data points can be identified, then the model can be used as a classifier. If some dimensions are highly correlated, they may be merged into one. Alternatively, if an independent variable, sometime called a feature, contributes little to the variance of the dependent variable(s), it may be removed from the model. Principal Component Analysis (PCA) is often used to determine the features that may be removed from the model through Dimensionality Reduction.

Lastly, the third type of ML is Reinforcement Learning. This type of ML is where no data is presented at all. Rather the software agents explore a well-specified environment where rewards and penalties are built into the environment itself. Master-level game playing models generally use some form of reinforcement learning although supervised learning

involving pre-played games or unsupervised learning where challenging situations are simulated may be employed as well.

Strengths of ML

The main strength of ML models is that they can generalize situations to which they have never been exposed. This property is in contrast to Relational Databases, which search tables of data to return "memorized" solutions. If the query cannot find a solution, then nothing is returned, not even a suggestion.

Another strength is that there are a wide variety of problems that may be addressed using ML methods and techniques. Here is a partial list of problems that have been successfully addressed with ML:

- Natural language translation
- Speech recognition and generation
- Scheduling and planning
- Curve fitting
- Image classification including video and audio
- Anomaly detection
- Autonomous robots
- Sentiment analysis and cross selling
- Style rendering (c.f. Figure 2)
- Dimensionality reduction
- Forecasting
- Facial recognition

Figure 2: New York City as Van Gogh Might Have Seen It

This robustness is not just an empirical observation. There is well-grounded mathematics behind it. The Universal Approximation Theorem, for example, proves that for Artificial Neural Nets, a widely-used ML algorithm, a learnable solution exists for virtually any function.

Also, technique of using small steps to find error function minima is the same idea behind limits, the basis of modern Calculus discovered independently by Isaac Newton and Gottfried Wilhelm Leibniz.

Weaknesses of ML

ML's main weakness is that the models cannot explain their reasoning. They cannot trace their learning history to the set of examples that led to their solutions. This lack of explanation makes debugging an ML program exceedingly difficult.

Beyond that weakness, ML programs are not deterministic. That is, they may behave differently with the same trials presented in a different order or with the parameters initialized to different values.

They also may be unable to learn at all if the parameter values exceed certain limits (Vanishing Gradient Problem, Exploding Gradient Problems).

Another weakness is that the quality of ML programs is directly related to the quality of the data used to train it.

There have been a number of examples where poor training data have produced disastrous results. There have been reports of unhinged chats with bots (Unhinged Conversations). Sadly, these off-the-rails conversation reflect more about us than the chatbots. These conversations were derived from social media posts written by humans. Nevertheless, not everything posted on the web is worth reading. The developers must take responsibility for not curating the data they are feeding to the bots.

Another particularly egregious example is that of a case cited in a legal brief (Varghese v. China South Airlines). This case does not exist. When the chatbot was asked to verify the authenticity of the case, the chatbot responded, "I apologize for the confusion earlier." Upon double-

checking, I found the case Varghese v. China Southern Airlines Co. Ltd., 925 F.3d 1339 (11th Cir. 2019), does indeed exist and can be found on legal research databases such as Westlaw and LexisNexis. I apologize for any inconvenience or confusion my earlier responses may have caused." This response itself is incorrect. The lawyer who "authored" the brief is now facing sanctions. The take-home message obviously is "let the user beware"; the information coming from chatbots generally has not been vetted.

Using biased data for training is another area of concern. There have been some examples of bias in policing (Law Enforcement and AI). When biased police files are used to build ML, systems designed to help police departments deploy resources more effectively, the bias may be compounded by run-away feedback. A crime-prediction program (Event-level Prediction of Urban Crime Reveals Signature of Enforcement Bias in U.S. Cities) was able to predict crimes a week in advance. However, when more resources were deployed to the predicted crime areas, more arrests were made, and more data was collected. This situation led the program to predict even more crime in those areas at the expense of other areas.

Lastly, and perhaps most importantly, ML programs lack feelings, emotions, or even preferences. They only know how to copy what they have seen. If humans cannot explain their emotions at least humans know they have emotions. Indeed, theories of human behavior all must address feelings or motivations since they drive and shape behavior. The difference is that humans know what they like and dislike even if they may not be able to articulate why; machines do not.

This distinction is not new. Hubert Dreyfus and John Searle, both AI critics and both professors of philosophy at UC, Berkeley, also made this case.

This argument may be extended to the case of 1-trial learning. Humans may develop an aversion to certain situations based upon a single traumatic experience. For example, for a time, I worked with a woman who would start a microwave oven with the handle of a broomstick. I asked her why she did that. She said that as a child she was riding her horse when she fell off onto an electrified fence. That experience terrified her so much that she did not trust electrical appliances; computers did not seem to bother her, though. A single trial presented during training of an ML algorithm would have a negligible effect on its behavior.

Conclusions

AI perhaps already is considered to be too useful and valuable to give up entirely as a misguided idea. It has the power to solve many important problems. It has the potential of unleashing human creativity.

As with any powerful technology, it can be misused and abused. Already Deepfaked images and videos have been mistaken as authentic (Phony Pentagon Attack, Deepfake Dangers). Chatbot plagiarism also is appearing more frequently. I believe these aberrations of the technology say more about us than the technology. We should have a healthy dose of skepticism when viewing anything that seems implausible. We all should engage in independent fact-checking. We must learn not to trust an untrustworthy technology.

That being said, the need for some form of regulation seems inevitable. Any content, whether from chatbots or deepfake creators, not independently validated or verified should be labeled as such.

On the other side of the issue is the data used to train the AI programs. The data should be validated and verified as well, both for veracity and bias. The old adage of “garbage in, garbage out” applies.

We are entering a Brave New World. The outcome is uncertain, but the future is in our hands.