Bhusan Chettri explains AI Adoption in real world application: Trust and Fairness are key ingredients

Bhusan Chettri summarises why AI systems showing impressive results may be deceiving & often lead to disappointments when deployed in real world.

LONDON, LONDON, UNITED KINGDOM, October 6, 2022 /EINPresswire.com/ -- Bhusan Chettri summarises why AI systems showing impressive results may be deceiving and often lead to disappointment (poor results) when deployed in real world conditions. He argues that trust, fairness and robustness are some of the important criterions that must also be considered when it comes to AI deployment in various safety-critical businesses such as medicine, finance and security.

Machine learning and Deep learning based AI models are heavily data-driven. This means that they learn to perform task exploiting patterns within the training data. Thus, quality of data is one important factor that needs to be considered before using them in building AI models. “With the availability of massive amount of training data, high-end computational resources and advancement in optimization algorithms for training AI systems, various research domains have witnessed tremendous success in past few years”, Bhusan Chettri says. He further adds that such success is often measured using some performance metric, for example mean accuracy. And, such metrics usually is a scalar number which explains nothing about how AI made a particular decision or obtained good results on a particular task. It might happen that AI model may have exploited some biases or features in the training data as a backdoor just to provide good accuracy. In simple words, the AI may have shown good results for wrong reasons i.e using cues not relevant to the problem. Dr. Chettri strongly argues that adoption of such AI systems in real-world application requires trust and they must perform fairly without any kind of biases. Trust, fairness, reliability and robustness are some of the important aspects any AI system must possess for its wide adoption. All these terminologies fall under the umbrella of interpretability a.k.a interpretable machine learning (sometimes referred to as

explainable AI).

Safety-critical applications cannot simply use AI systems (showing impressive results on some metrics) without understanding how it forms decisions under the hood. For example, domains such as finance, medicine/healthcare and security (to name a few) require sufficient explanations to ensure fairness, reliability
and trustworthiness of decisions they make. These AI models are black boxes. All that is known about them is that it takes certain input data and somehow produces the output. Bhusan Chettri further adds “Yes, in many applications, explanations may not be that important and all that matters would be just numbers - good results or say accuracy. Such businesses don’t care how their model arrived at such a decision because the result is pretty good and most importantly their client is happy. So all sorted.” However, this is not true with every domain as discussed earlier. There is a danger in using such black boxes without understanding their working mechanism. Also, research has shown that deep learning AI models can be fooled very easily by just making a tiny perturbation to its input data (which to humans is undetectable), yet the AI system produces completely different results. This is very dangerous. This means that an attacker can easily manipulate the input data passed to a neural network to produce the output they desire. Dr. Bhusan Chettri gives an example: imagine a scenario where a Text-to-Speech system takes the input text “I will call you tonight” but the system gets manipulated by an attacker to produce the speech that sounds “I will kill you tonight”. This field, not within the scope of today’s discussion however, is called Adversarial Machine Learning, a very hot topic in the field of AI that is actively being studied by researchers around the globe.

Interpretable machine learning is a field of machine learning aimed at understanding how an AI system makes a particular prediction. It aims at assisting its users towards finding answers/explanations to some of the key questions such as: is the AI system learning to solve problem using the actual cues within training data that it is supposed to exploit? which parts of the input components are contributing more towards high prediction/score? Check these blogs for more on this subject.

Narender Sharma
Blue Particle Solutions
email us here