Trustable Machine Learning Systems

Proposal for BCTCS 2021 Presentations

Jay Morgan <j.p.morgan@swansea.ac.uk>

January 2020

Abstract

Machine Learning (ML) has had a remarkable impact on society. Everything from the phones in our pockets, to the cars that we drive, are being increasingly outfitted with this progressively sophisticated suite of algorithms. But while many of the most basic and fundamental algorithms from ML can be formally verified and tested for safety without much trouble, the same may not be said for Deep Learning (DL) – a prominent forerunner in the state-of-the-art for ML research. These DL models, while performing simple matrix-to-matrix operations at a micro-level, have evolved in scale far past what is tractable for current formal verification methods – all in the pursuit of improving accuracy and performance. This issue of tractability is unsettling considering that the existence of *adversarial examples* is well known in the ML community. These adversarial examples occur when very small changes to the input space result in a large change in the output space and cause a miss-classification made by the DL model. In the context of self-driving vehicles, small defects and visual artifacts in the sensor input of the DL model, could lead the vehicle to wrongly conclude a stop sign indicates to continue driving where it should have stopped. While the manufacturers will need to put safe-guards in place to prevent this from happening, we should formally prove the (non)-existence of these adversarial examples in the DL model itself. In this presentation, I present the foundational knowledge for understanding adversarial examples, how we can use the input space to dictate the search space for the existence of these examples, and demonstrate their presence with the use of SAT-solving. This work, as a free and open-source project, provides a framework for ML practitioners to verify their own architectures.

Bio

I am a 3rd year PhD candidate at Swansea University. Throughout my candidature period, I have been researching methods to make ML more ‘trustable’ for its usage in present day society, in which much of our lives are increasingly automated by this technology. Throughout this research I have been working in interdisciplinary settings with colleagues from different domains such as Quantum Chemistry, Corpus Linguistics, to Astrophysics. In these ventures, we have outlined key principles on how one may integrate prior expert knowledge into DL models to improve its performance, but verify the output matches what is expected by the expert. While my work covers many facets of ML, one method that may appeal to the Theoretical community is the verification of existence of so-called *adversarial examples*, very small, perhaps imperceptible, changes to the input of DL models that result in large changes to the output space and cause unexpected miss-classifications.