Your World First

C/M/S/
Law.Tax

Artificial Intelligence

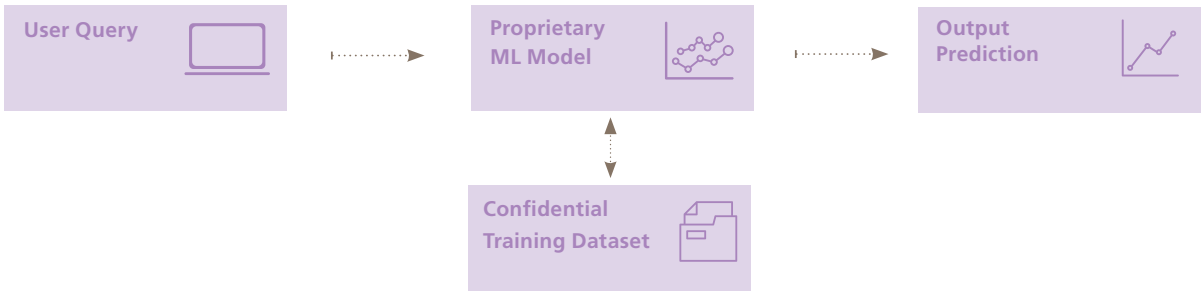
Artificial intelligence and the implications of reverse engineering

AI and machine learning are some of the great transformational technologies of our generation with societies becoming more reliant on them every day. It has been assumed that without knowing the algorithms powering the machine learning model, competitors cannot replicate the technology. But with the development of 'reverse engineering' are the secrets in these models safe from discovery? What can technology developers do to protect their models?

Once an area of computer science limited to research papers and doctoral dissertations, top tech companies are now racing to provide exciting new products and services powered by the latest machine learning models and techniques. Whether it's enhancing your entertainment services (e.g. **Netflix**), customizing your shopping experiences (e.g. **Amazon**), or powering your self-driving cars (e.g. **Tesla**), artificial intelligence/machine learning innovators are disrupting the industry like never before. Not only have machine learning architectures been adapted for specific consumer goods and services, but the models themselves have even been packaged up and commercialized, leading to machine-learning-as-a-service (MLaaS) and AI-as-a-service (AIaaS) products being launched for anyone to use (see **Microsoft's Azure ML service**, **Google's Cloud AI**, **BigML** etc.)

But just how secure how are these models?

Machine learning models are often presented as 'black box' services in the cloud, hidden behind user interfaces such as mobile apps, websites, or application programming interfaces (APIs). Given an input, a black box will return an output without exposing the decision-making process hidden within.



It is often assumed that machine learning models presented as black box services in the cloud are secure; that is, the machine learning model is not revealed to competitors. It is assumed that, without explicit knowledge of the type of algorithms powering the technology, its parameters, or the datasets it has been trained on, machine learning models are very difficult to decipher and reproduce. It is thought that by maintaining their back-end code and datasets under strict confidentiality, e.g. as trade secrets, a company may prevent competitors from using their proprietary algorithms, datasets, and parameters, and thus can keep a competitive edge.

However, new abilities to reverse engineer these machine learning algorithms are starting to emerge.

In a 2016 paper entitled "Stealing Machine Learning Models via Prediction APIs" by researchers from Cornell Tech, the EPFL, and the University of North Carolina, research showed that with nothing more than repeated queries to a machine learning model, it was possible to effectively reproduce a machine learning-trained AI via a 'model extraction attack'. The research team demonstrated it was possible to infer the hyperparameters of a deep neural network hosted on Big ML and Amazon ML on AWS services by observing responses to a sequence of queries sent to an API of a deep neural network service, when the responses included prediction or confidence values associated with the responses. These values allowed the researchers to infer and estimate to a high degree of accuracy the internal parameters of the neural network, effectively "reverse engineering" the neural network and making it potentially reproducible. As such, by using their inference method any proprietary information or trade secret protection in the original deep neural network is lost.

Researchers from the Max-Planck Institute for Informatics have also similarly demonstrated the ability to infer information from black-box models in their 2018 paper entitled "Towards Reverse-Engineering Black-Box Neural Networks". The research team showed that by building 'metamodels' – i.e. a model trained to predict model attributes – they were able to predict attributes relating not only to the model architecture, but also to training hyperparameters with high accuracy.

The latest research into reverse-engineering machine learning architectures shows that it is more important than ever before to ensure that any proprietary machine learning models or data is either protected or kept secure. The implications of research such as those described above are numerous – such as those of privacy and security - and not least is the legal implication into proprietary intellectual property. Products and services whose commercial importance relies on the strength of the machine learning algorithms powering them for example now cannot rely on simple confidentiality or trade secrets to protect them. Indeed, trade secrets are a weak form of intellectual property, as once the confidential information loses its confidential nature as a result of being made public through a mistake or by a malicious party making the information public, the trade secret is lost and cannot be restored.

Companies investing in machine learning and AI technologies must now seek other ways to protect their machine learning assets. Patents are one option – in return for disclosing the nature of their machine learning architectures, it is potentially possible to secure up to 20 years legal protection against the unlicensed reproduction or use of that architecture. Additionally, technical security must be tightened to ensure model extraction attacks such as those above cannot glean crucial information from public-facing machine learning models. For a commercial world that is increasingly reliant on ML and AI technologies, protecting our technological assets is crucial – especially when our black boxes are steadily becoming more transparent.

Contact

Rachel Free
Of Counsel, Patent Attorney
T +44 20 7067 3286
E rachel.free@cms-cmno.com

Leon Zhang
Trainee Patent Attorney
T +44 20 7367 2958
E leon.zhang@cms-cmno.com