Protecting the Intellectual Property of AI with Watermarking

July 20 2018, by Marc Ph. Stoecklin

(a) car

Prediction	Probability
automobile	0.99996
cat	0.0003
truck	0.0001
dog	0
ship	0

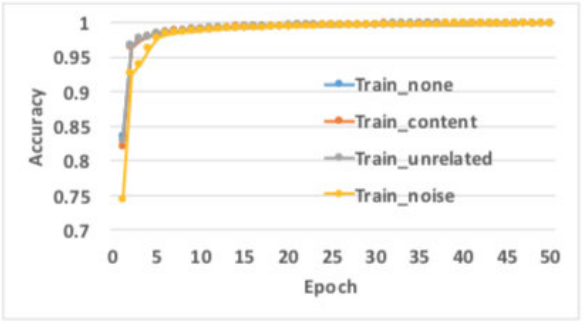
(b) Prediction results (top 5)

(c) car with $WM_{content}$

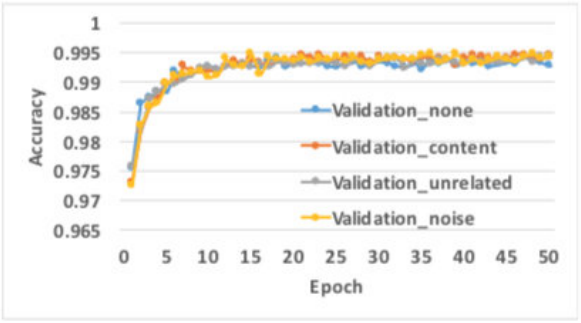
Prediction	Probability
airplane	1
bird	0
automobile	0
ship	0
truck	0

(d) Prediction results(top 5)

Figure 5: A case study of watermark verification

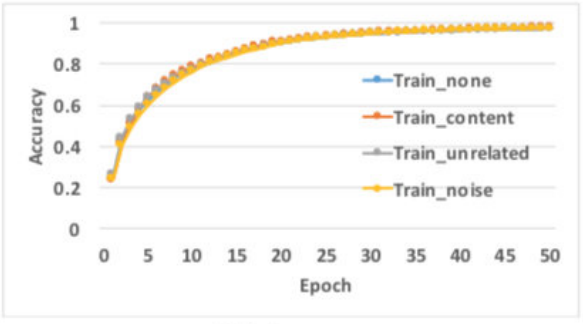


(a) Train accuracy

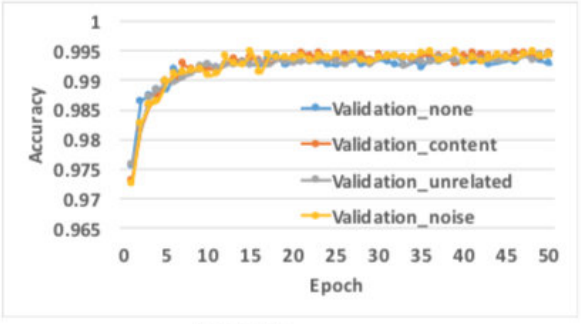


(b) Validation accuracy

Figure 6: Model accuracy over training procedure (MNIST)



(a) Train accuracy



(b) Validation accuracy

Model accuracy over training procedure. Credit: CIFAR10

If we can protect videos, audio and photos with digital watermarking, why not AI models?

This is the question my colleagues and I asked ourselves as we looked to develop a technique to assure developers that their hard work in building AI, such as deep learning models, can be protected. You may be thinking, "Protected from what?" Well, for example, what if your AI model is stolen or misused for nefarious purposes, such as offering a plagiarized service built on stolen model? This is an concern, particularly for AI leaders such as IBM.

Earlier this month we presented our research at the AsiaCCS '18 conference in Incheon, Republic of Korea, and we are proud to say that our comprehensive evaluation technique to address this challenge was demonstrated to be highly effective and robust. Our key innovation is that our concept can remotely verify the ownership of deep neural network (DNN) services using simple API queries.

As deep learning models are more widely deployed and become more valuable, they are increasingly targeted by adversaries. Our idea, which is patent-pending, takes inspiration from the popular watermarking techniques used for multimedia content, such as videos and photos.

When watermarking a photo there are two stages: embedding and detection. In the embedding stage, owners can overlay the word "COPYRIGHT" on the photo (or watermarks invisible to human perception) and if it's stolen and used by others we confirm this in the detection stage, whereby owners can extract the watermarks as legal evidence to prove ownership. The same idea can be applied to DNN.

By embedding watermarks to DNN models, if they are stolen, we can

verify the ownership by extracting watermarks from the models. However, different from digital watermarking, which embeds watermarks into multimedia content, we needed to design a new method to embed watermarks into DNN models.

In our paper, we describe an approach to infuse watermarks into DNN models, and design a remote verification mechanism to determine the ownership of DNN models by using API calls.

We developed three watermark generation algorithms to generate different types of watermarks for DNN models:

1. embedding meaningful content together with the original training data as watermarks into the protected DNNs,
2. embedding irrelevant data samples as watermarks into the protected DNNs, and
3. embedding noise as watermarks into the protected DNNs.

To test our watermarking framework, we used two public datasets: MNIST, a handwritten digit recognition dataset that has 60,000 training images and 10,000 testing images and CIFAR10, an object classification dataset with 50,000 training images and 10,000 testing images.

Running the experiment is rather straightforward: we simply provide the DNN with a specifically crafted picture, which triggers an unexpected but controlled response if the model has been watermarked. This isn't the first time watermarking has been considered, but previous concepts were limited by requiring accessing model parameters. However, in the real world, the stolen models are usually deployed remotely, and the plagiarized service would not publicize the parameters of the stolen models. In addition, the embedded watermarks in DNN models are robust and resilient to different counter-watermark mechanisms, such as fine-tuning, parameter pruning, and model inversion attacks.

Alas, our framework does have some limitations. If the leaked model is not deployed as an on-line service but used as an internal service, then we cannot detect any theft, but then of course the plagiarizer cannot directly monetize the stolen models.

In addition, our current watermarking framework cannot protect the DNN models from being stolen through prediction APIs, whereby attackers can exploit the tension between query access and confidentiality in the results to learn the parameters of machine learning models. However, such attacks have only been demonstrated to work well in practice for conventional machine learning algorithms with fewer model parameters such as decision trees and logistic regressions.

We are currently looking to deploy this within IBM and explore how the technology can be delivered as a service for clients.

More information: Jialong Zhang et al. Protecting Intellectual Property of Deep Neural Networks with Watermarking, *Proceedings of the 2018 on Asia Conference on Computer and Communications Security - ASIACCS '18* (2018). DOI: 10.1145/3196494.3196550

This story is republished courtesy of IBM Research. Read the original story here.

Provided by IBM