NEWS AND INSIGHTS FROM THE WORLD OF ID SECURITY

NOVEMBER 2022

The VAULT

ARTIFICIAL INTELLIGENCE

FEATURED ARTICLE

Can we trust Artificial Intelligence?

Wibu-Systems

ALSO IN THIS ISSUE

Infineon Technologies
How biomentrics could impact the future of payments

Mühlbauer Group
Ready for the revolution?

Infineon Technologies
Secured NFC tags enhance brand experience

Can we TRUST *Artificial INTELLIGENCE*?

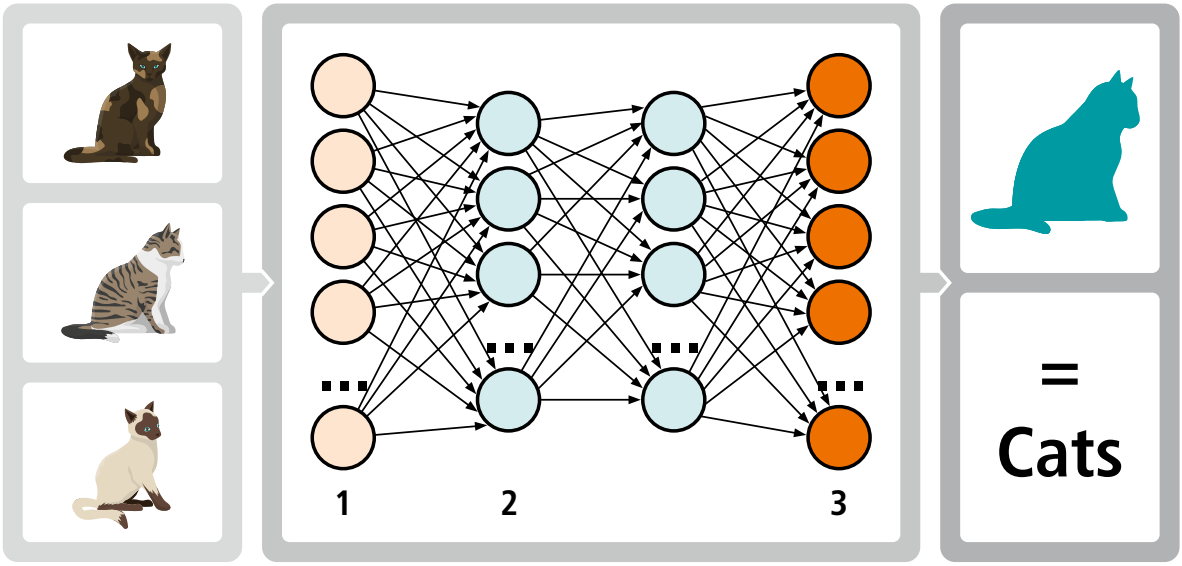
By Dr. Carmen Kempka, Wibu-Systems

The possibilities of artificial intelligence and machine learning seem endless. Neural networks and deep learning techniques are utilized nearly everywhere. Their actual or potential use cases range from speech recognition, malware detection, and quality testing to applications that could be critical for people's lives and limbs, like driver assistance systems or medical diagnostics.

In safety-critical environments like these, it is essential to use new and untested technologies in a responsible manner, especially those like AI that are not yet fully understood. An attack on an AI application in this context, or even a simple malfunction, could have life-threatening implications. An incorrect classification could lead to a wrong medical diagnosis and, by implication, incorrect treatment or, more directly, get a driver assistance system to cause the car to crash.

Moreover, especially in the medical sector, AIs are trained on sensitive patient data for which confidentiality and the patient's anonymity are paramount. This data could be a CT or MRT scan, or information about the patient's medical history. In addition, AI models are often trained with complex training parameters which, like the trained model itself, contain intellectual property.

All in all, protecting the machine learning lifecycle against tampering and unauthorized access to functions and data is a complex undertaking that requires sophisticated solutions. But before we look deeper into the attack surfaces and necessary protections for the machine learning lifecycle, we first need to investigate one important question:



The cat graphics created by macrovector / Freepik

Figure 1

How intelligent are AIs, really?

Neural networks and deep learning algorithms have been designed to imitate the way the human brain learns things. However, each AI is trained on a very limited selection of data – compared to a human being, at least. A neural network does not have the same experience as the human brain. It has no lifetime of adventures with all their ups and downs to process. It has no common sense to work with. It gets a very limited set of input data, tailored to a specific use case, like images of animals, traffic data, or CT scans, for which it learns to provide some classification.

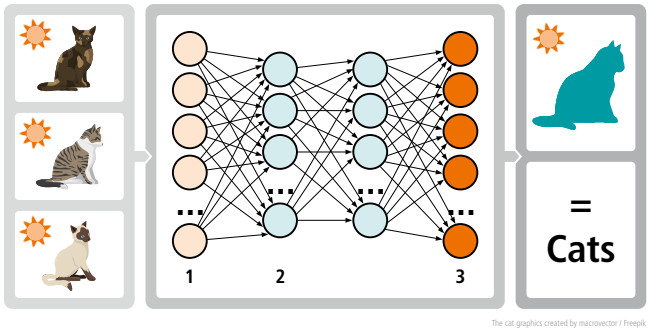
Most importantly, no AI actually "thinks" about its input data or the trained model in any way. There is no sanity check whether the input data or the inferred classification criteria make any sense at all.

Let's consider the following example: Imagine an AI that gets pictures as training data. Some of these show a cat and are labeled "cat" (Figure 1) while some show a dog and are labeled "dog". If the data to be classified after training is similar enough to the training data, the AI will distinguish cats from dogs correctly.

Now, imagine the cat images all have a sun in the picture (Figure 2) while the dog is always sitting in the rain. Now the AI will learn something like "cat-like animal and sun" means cat, and "dog-like animal and rain" means dog.

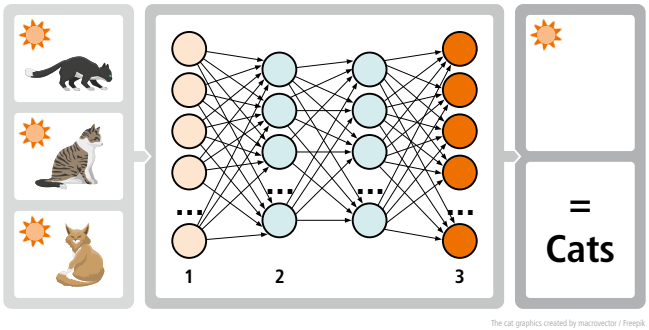
Even worse, if the cats and dogs are hard to distinguish or all cats/dogs look different, the AI will instead learn "sun means cat and rain means dog", not even considering the actual animals anymore in the classification process (Figure 3).

To make things worse still, instead of the sun and the rain, a potential attacker could color certain pixels in the training images to cause a certain classification behavior, even if these changes in the training data would not even be noticed by the human eye.



The cat graphics created by macrovector / Freepik

Figure 2



The cat graphics created by macrovector / Freepik

Figure 3

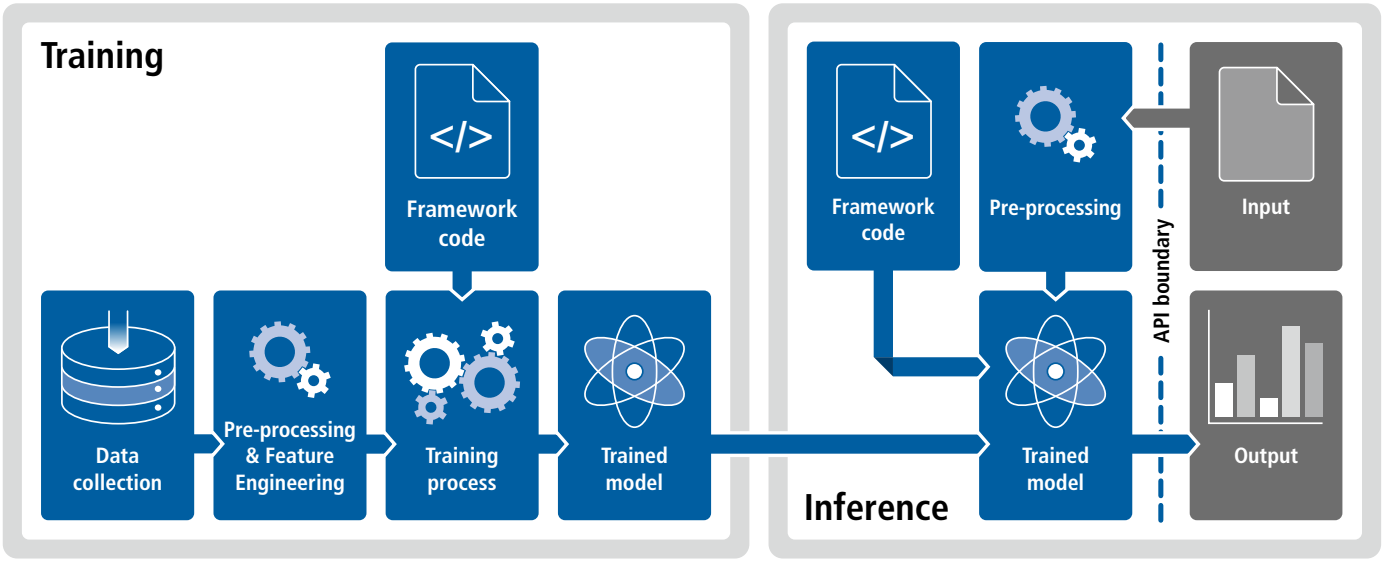


Figure 4

The ML lifecycle

In addition to neural networks and deep learning, there are several machine learning techniques based on math and statistics, such as separating data by a hyperplane or predicting data by putting a line through known points or by building decision trees. No matter which machine learning technique is used, there are common steps on the way from the raw training data to the trained, deployed, and used model. We call these steps the machine learning lifecycle, which can roughly be described as follows (Figure 4).

First, the raw training data needs to be preprocessed to provide the training algorithm with a homogeneous set of data. Preprocessing will, for example, scale all training images to the same size or delete unnecessary columns in tables. The actual training is then performed on the preprocessed data, resulting in a trained model which can be deployed and used for classification. In some cases, the model keeps training itself during use, utilizing the user's input as additional training data.

This can happen in the context of anomaly detection or clustering or the notion "people who looked at this also bought…", which means that this data – considered potential training data which could affect the quality of the model – must be protected and processed with similar care as the original training set.

Protecting the machine learning lifecycle

The machine learning lifecycle has a number of stakeholders who are interested in different protection targets: The data owner, who provides the training data, might want the data to stay confidential and anonymous. The machine learning engineer, who uses the training data to train the model, wants both the training data and the algorithms used for preprocessing and training to be of high quality and not tampered with, while the used training parameters, which often contain intellectual property, must stay confidential. The model owner, who deploys and provides the trained model, wants the intellectual property within the model to be protected and is interested in the correctness and integrity of the model, which requires the integrity of the whole machine learning lifecycle, including training data, training parameters, and algorithms. To realize business models or simply prevent model inference, the model owner might apply access controls and licensing techniques to the trained model. The customer who accesses the model to get a classification is interested in the correctness of the classification, which also requires the integrity of the whole machine learning lifecycle. The customer's query might contain data which requires confidentiality or which has the potential to be malicious and requires checking.

> *One peculiarity in the case of machine learning, especially neural networks, is that keeping the trained model confidential is not enough to prevent fraud.*

The role of software protection

The attack surfaces of the machine learning lifecycle are many. As mentioned above, any manipulation of any data or any algorithm used within the machine learning lifecycle can have fatal consequences. In addition, the confidentiality of sensitive data and intellectual property must be protected.

One peculiarity in the case of machine learning, especially neural networks, is that keeping the trained model confidential is not enough to prevent fraud. Unrestricted access to a trained model can be abused to train a second model using input/output pairs only, which can get very close to the original model in terms of classification behavior, or to evade the classification of the original model, for example, in the case of malware or deep fake detection. Therefore, limiting access to the trained model might be a reasonable or even necessary precaution.

Software protections safeguard applications from tampering and theft and enable the software provider to put in place business models like pay-per-use or a subscription. The protection suite developed by Wibu-Systems offers an all-round toolkit for the defense of both executables and data. While executables are protected from reverse engineering, we do not consider "security by obscurity" enough to protect an application. Executables or sensitive functions are encrypted using well-established cryptographic algorithms. In addition, cryptographic methods are utilized to protect the integrity of software and data. Functions and data are decrypted at runtime. Sensitive parts of the code can even be decrypted and executed and key material or certificates be securely transferred and stored in secure hardware. This does not only keep the key material secret, but it also prevents the manipulation of keys and certificates.

AxProtector Python

Due to the availability of open-source frameworks, as well as the popularity of the language, AI applications are often written in Python. AxProtector Python can protect Python applications from manipulation, reverse engineering, and unauthorized use. In addition to executables, AxProtector Python can also protect files like training data, confidential training parameters, and trained models. Data and code are decrypted and checked at runtime. With the ability of AxProtector Python to protect both the framework code used for training and the data used in the machine learning lifecycle, including training data, training parameters, and the trained model, AxProtector Python can protect the whole machine learning lifecycle from manipulation, theft of intellectual property, and unauthorized use.

This way, it can keep patients' data private or keep cars from speeding into pedestrians because of manipulated classifications, while protecting the complex training parameters of a neural network from being copied. In addition, the ability to license trained models allows for new business models for AI applications, such as pay-per-use access to a classification, a thirty-day trial period, or a monthly subscription.

Protecting the machine learning lifecycle is an essential step towards using artificial intelligence in a responsible way. It's not only software you protect, it's also protecting people. ⊠

WIBU SYSTEMS

Cybersecurity: Addressing the Global Challenge

A beacon for IT security innovations

A business enabler for all software and device makers

A guardian for all digital assets

Wibu-Systems and the House of IT Security support the local and international community for a more trustworthy, sustainable, and effective digital future.

Are you looking for thought leaders and inspiring partners to engage with?

Become a member of the IT Security Club

Are you looking for a rewarding career in IT security?

Send your application to Wibu-Systems

+49 721 931720
sales@wibu.com
www.wibu.com

SECURITY
LICENSING
PERFECTION IN PROTECTION