Data Poisoning - the poisoned apple for AI

Learning processes and machine learning models have vulnerabilities that attackers can exploit. The goal of the attacks is to steer the statements of an AI application in a certain direction. This results in targeted false statements caused, for example, by the infiltration of manipulated data. This method is referred to as "data poisoning". It involves a number of techniques to influence the behavior of AI.

Adversarial attack: when neural networks come to false conclusions

Adversarial attacks on image recognition using neural networks are particularly impressive. Here, manipulations of image data lead to false-looking results in the recognition of image objects by artificial neural networks. An example: The neural network indicates that the image of a turtle represents a rifle. This erroneous classification is achieved by manipulating pixel values in the image in a way that is imperceptible to the human eye, overlaying the image with a noise pattern. While humans can easily recognize a turtle on the "noisy" image, the neural network gets into trouble. Human perception is fundamentally different from the neural network's decision making based on mathematical rules. Humans identify a turtle by visually familiar pattern groups such as head or feet. The neural network, on the other hand, recognizes objects for a classification of an image via the mathematical comparison of individual pixels, their learned neighborhood with other pixels and the color values for red, green and blue (RGB).

The "noise" corresponds to a significant change in input values (RGB) of individual pixels. Even if these represent minimal mathematical deviations, they can lead to a wrong decision by the individual neuron in the neural network. The attacker's goal is to create noise that causes the individual neurons in the staggered decision process to tip into a wrong decision with a predominantly high probability. The result is a misclassification of the subject of the image.

Other well-known examples lead to misinterpretation in traffic sign recognition in autonomous driving systems. Adversarial attacks are also characterized by great creativity on the part of the attackers. Recent examples encode noise from an image information into 3D printed models. The result is an object whose three-dimensional shape contains noise that leads the neural network to make incorrect decisions during image recognition.

What is Data Poisoning?

The quality of the information provided by machine learning models is significantly influenced by the data with which they are trained or queried. If these are not systematically checked for correctness, attackers can deliberately inject manipulated data to compromise the model's statements. Data poisoning can thus be applied to data to be analyzed by the model or to data used to train AI models. Potentially at risk are almost all known AI methods, from deep learning in neural networks, to supervised learning in statistical regression-based methods. When attacking training datasets, attackers try, for example, to specifically change awards("labels") or manipulate values in datasets. Attackers can disguise these manipulations by not falsifying all training data, but by interspersing modified data sets in a statistical distribution in training data. Depending on the number of training data and the distribution of the manipulation, there is the possibility to steer the expressiveness of the model in a direction desired by the attacker. The attack can take place over the entire data supply chain. This often has a large attack surface in practice: manipulation of data at the data source, man-in-the-middle attack during data transfer, or API attacks compromise in the cloud data store or data versioning system. Skilled attackers modify data records over a long period of time. The delta of these changes is kept minimal in each case. This makes the attack difficult to detect via monitoring systems and filters for statistical deviations. Attackers run the risk of discovering far too late that there is a problem with the reliability of the data for the AI model and that data has been manipulated.

These are the dangers of data poisoning

There is an active research community working on data poisoning worldwide. The demonstrated attacks are mostly related to proof-of-concepts in the context of scientific studies. These demonstrated attacks are very well documented in their methodological description and are usually accompanied by approaches to minimize risk and defend against data poisoning. Scientific work with data poisoning is an important component for the further development and improvement of AI methods.

In 2016, a public AI experiment (https://www.welt.de/kultur/article153688321/Wie-der-Microsoft-Bot-uns-den-Spiegel-vorhaelt.html) by Microsoft failed due to data poisoning. The development team of the chat bot Tay planned to improve the system's ability by actively communicating in dialog with Twitter followers, thus using Unsupervised Learning to expand the system's capabilities of a natural linguistic conversation. Tay learned his communication skills from the comments and messages of his followers on Twitter. Shortly after the system launched on Twitter, a group of users realized that Tay's behavior could be influenced by what he said in comments. The clincher was a post on the Internet discussion board 4Chan. Users suggested that Tay could be overwhelmed with racist and insulting comments, thus steering the training data and Tay's statements in a negative direction. The data poisoning quickly took effect. 16 hours after Tay appeared on Twitter, the chatbot had exchanged over 95,000 messages with its data poisoing mob. Each of those messages was used to train the system. In retrospect, the experiment sharpened the focus on data poisoning. The problem lay in the setting of Unsupervised Learning via an open Twitter community. The bot acted as an open gateway and thus for unfiltered learning of the chatbot via a public social media platform. Negative examples like Tay lead to more careful planning of building training systems with public data interfaces. Machine learning is protected against data poisoning by an organized Internet mob by means of filters and monitoring.

Protection against data poisoning

Blind trust in data is the gateway for data poisoning. In addition, each AI model can serve as a "parent model" for new ones. This means that an unnoticed attack on learning data is passed on in the process. If the learning model is transferred, the "poisoned" data will also be included. Therefore, it is important to protect data for these learning models. There are numerous working approaches around the world to learn from experiences with ML security attacks and develop effective methods to defend against them. One of these is the Adversarial ML Threat Matrix collaboration, which has published an Adversarial Threat Landscape (https://github.com/mitre/advMLthreatmatrix) for Artificial-Intelligence Systems. It builds on the established MITRE Att&CK Framework (https://attack.mitre.org/), the globally accessible knowledge base on tactics and techniques of such attacks. However, there are also systemic limitations for attackers: to inject poisoned data, it is necessary for systems to be re-trained on a regular basis. Only if the training data comes from sources to which the attacker has access can the training be poisoned and the attacker influence the AI model.

Conclusion

It has proven very difficult in the past to detect and reliably defend against data poisoning attacks. Attackers can even effectively bypass multiple defenses applied in parallel. One of the most promising defenses against adversarial attacks is training with AI to prevent the manipulation. During the training phase, examples of adversarial attacks are integrated to increase the robustness of the system. However, if these are very large and complex, it delays the training time of the model. If only weak attacks are integrated as examples for performance reasons, the system remains more vulnerable to strong, effective attacks. The danger of such defensive techniques is primarily that they give a false sense of security

At present, neural networks still have to be examined in depth and samples analyzed in the event of anomalies. Human expert knowledge is one of the essential criteria for the safe defense against manipulations on AI training data.

In addition, efforts are being made to develop standards for test procedures in Germany. A standardization roadmap AI (https://www.din.de/de/forschung-und-innovation/themen/kuenstliche-intelligenz/fahrplan-festlegen) has already been presented for this purpose. In the future, it will be more important than ever to be able to define generally applicable criteria and instruments to make AI systems sufficiently verifiable and secure.

Captions

Figure 1
Schematic of an attack surface of an ML system.
Source: asvin GmbH

Figure 2
Microsoft's chatbot Tay on Twitter.
Source: Twitter International Company

Figure 3
Proper statistical selection of training data sets protects against data poisoning. In this example, almost 25% of the data sets need to be manipulated to significantly increase the error rate of the trained model.
Source: asvin GmbH

About asvin

asvin provides solutions to ensure the security and provenance of software throughout its lifecycle. This includes services and analytics that monitor data and software supply chains, support secure roll-out of over-the-air software updates, and generate Software Bill of Materials (SBOM).

Learn more at www.asvin.io, on Twitter and LinkedIn.

About the Author:

Mirko Ross, born in 1972, is an internationally recognized activist, expert, speaker, publicist and researcher in the field of cybersecurity and the Internet of Things. He began investigating security vulnerabilities in IT systems at the age of 14. Rather than a career in hacking, he chose the "good side" of the cybersecurity industry. Ross was a member of the European cybersecurity authority ENISA's Internet of Things Security Expert Group until 2020 and is an expert advisor to the EU Commission. He is also active in international committees and research projects in cybersecurity and blockchain technologies. Mirko is still closely connected to the positive hacker and maker movement and promotes non-profit projects in the field of Open Data and IT education. In 2018, he founded asvin.io, a company with the goal of increasing cybersecurity in the Internet of Things and providing software solutions for this purpose. asvin was awarded Best Cybersecurity Startup in German-speaking countries at it-sa in 2020. Mirko lives in the countryside and works in Stuttgart.

Press Contact

BCW
asvin@bcw-global.com

Mirko Ross, asvin CEO
m.ross@asvin.io