Researchers create privacy technique that protects sensitive data while maintaining performance

July 14 2023, by Adam Zewe

Credit: Pixabay/CC0 Public Domain

Imagine that a team of scientists has developed a machine-learning

model that can predict whether a patient has cancer from lung scan images. They want to share this model with hospitals around the world so clinicians can start using it in diagnoses.

But there's a problem. To teach their model how to predict cancer, they showed it millions of real lung scan images, a process called training. Those sensitive data, which are now encoded into the inner workings of the model, could potentially be extracted by a malicious agent. The scientists can prevent this by adding noise, or more generic randomness, to the model that makes it harder for an adversary to guess the original data. However, perturbation reduces a model's accuracy, so the less noise one can add, the better.

MIT researchers have now developed a technique that enables the user to potentially add the smallest amount of noise possible, while still ensuring that sensitive data are protected.

The researchers created a new privacy metric, which they call Probably Approximately Correct (PAC) Privacy, and built a framework based on this metric that can automatically determine the minimal amount of noise that needs to be added. Moreover, this framework does not need knowledge of the inner workings of a model or its training process, which makes it easier to use for different types of models and applications.

In several cases, the researchers show that the amount of noise required to protect sensitive data from adversaries is far less with PAC Privacy than with other approaches. This could help engineers create machine-learning models that provably hide training data, while maintaining accuracy in real-world settings.

"PAC Privacy exploits the uncertainty or entropy of the sensitive data in a meaningful way, and this allows us to add, in many cases, an order of

magnitude less noise. This framework allows us to understand the characteristics of arbitrary data processing and privatize it automatically without artificial modifications. While we are in the early days and we are doing simple examples, we are excited about the promise of this technique," says Srini Devadas, the Edwin Sibley Webster Professor of Electrical Engineering and co-author of a new paper on PAC Privacy.

Devadas wrote the paper with lead author Hanshen Xiao, an electrical engineering and computer science graduate student. The research will be presented on August 24 at the International Cryptology Conference (Crypto 2023).

Defining privacy

A fundamental question in data privacy is: How much sensitive data could an adversary recover from a machine-learning model with noise added to it?

Differential Privacy, one popular privacy definition, says privacy is achieved if an adversary who observes the released model cannot infer whether an arbitrary individual's data is used for the training processing. But provably preventing an adversary from distinguishing data usage often requires large amounts of noise to obscure it. This noise reduces the model's accuracy.

PAC Privacy looks at the problem a bit differently. It characterizes how hard it would be for an adversary to reconstruct any part of randomly sampled or generated sensitive data after noise has been added, rather than only focusing on the distinguishability problem.

For instance, if the sensitive data are images of human faces, differential privacy would focus on whether the adversary can tell if someone's face was in the dataset. PAC Privacy, on the other hand, could look at

whether an adversary could extract a silhouette—an approximation—that someone could recognize as a particular individual's face.

Once they established the definition of PAC Privacy, the researchers created an algorithm that automatically tells the user how much noise to add to a model to prevent an adversary from confidently reconstructing a close approximation of the sensitive data. This algorithm guarantees privacy even if the adversary has infinite computing power, Xiao says.

To find the optimal amount of noise, the PAC Privacy algorithm relies on the uncertainty, or entropy, in the original data from the viewpoint of the adversary.

This automatic technique takes samples randomly from a data distribution or a large data pool and runs the user's machine-learning training algorithm on that subsampled data to produce an output learned model. It does this many times on different subsamplings and compares the variance across all outputs. This variance determines how much noise one must add—a smaller variance means less noise is needed.

Algorithm advantages

Different from other privacy approaches, the PAC Privacy algorithm does not need knowledge of the inner workings of a model, or the training process.

When implementing PAC Privacy, a user can specify their desired level of confidence at the outset. For instance, perhaps the user wants a guarantee that an adversary will not be more than 1% confident that they have successfully reconstructed the sensitive data to within 5% of its actual value. The PAC Privacy algorithm automatically tells the user the optimal amount of noise that needs to be added to the output model

before it is shared publicly, in order to achieve those goals.

"The noise is optimal, in the sense that if you add less than we tell you, all bets could be off. But the effect of adding noise to neural network parameters is complicated, and we are making no promises on the utility drop the model may experience with the added noise," Xiao says.

This points to one limitation of PAC Privacy—the technique does not tell the user how much accuracy the model will lose once the noise is added. PAC Privacy also involves repeatedly training a machine-learning model on many subsamplings of data, so it can be computationally expensive.

To improve PAC Privacy, one approach is to modify a user's machine-learning training process so it is more stable, meaning that the output model it produces does not change very much when the input data is subsampled from a data pool. This stability would create smaller variances between subsample outputs, so not only would the PAC Privacy algorithm need to be run fewer times to identify the optimal amount of noise, but it would also need to add less noise.

An added benefit of stabler models is that they often have less generalization error, which means they can make more accurate predictions on previously unseen data, a win-win situation between machine learning and privacy, Devadas adds.

"In the next few years, we would love to look a little deeper into this relationship between stability and privacy, and the relationship between privacy and generalization error. We are knocking on a door here, but it is not clear yet where the door leads," he says.

More information: Hanshen Xiao et al, PAC Privacy: Automatic Privacy Measurement and Control of Data Processing, *arXiv* (2022).

DOI: 10.48550/arxiv.2210.03458