Interference Attacks

2 min

In many situations, an AI model may be trained using sensitive or confidential information. In these cases, it is crucial not only that we protect our resultant AI model but also the training data used to create the model.

While an AI model may appear to be a one-way path, issues in the model's design could expose confidential training data via inference attacks. When a model is trained, there are a variety of connections, links, and weights assigned to internal logic, nodes, and pathways throughout the model. While these features may seem inaccessible to an average user, it may be possible to unravel and detect these connections through careful input manipulation.

Given a trained model, it may be possible for an attacker to send a large number of varied inputs, and by carefully observing the outputs, map out and extract information about how the model is handling information and even what data the model used for training.

For example, if we take a basic classification model designed to identify animals, we could identify information about the initial training model by passing in various inputs. Suppose a model was trained entirely with animals from a specific forest region. If we were to pass a desert-dwelling lizard to this model, there would be a significant chance that the model would fail to classify the lizard. Let's continually send new images of desert-dwelling animals to the model, and we'll be able to determine that the model was trained using specific regioned animals. Furthermore, depending on how much data we have, it could be possible to narrow down the data category. For example, if we know that the model was trained with forest animals in a specific region, we could send animals from various forest regions and continue to narrow down the area the model was trained on.

Now, this type of manipulation is broader than forest identification. Individuals can use this tactic to expose sensitive information in other models, i.e., financial trends, where its model is trained on individuals' economic patterns. Using strategic input manipulation, a person can reveal sensitive information about an individual whose data was used in the financial AI model training.