Evasion Attacks

1 min

While AI has continually improved on tasks related to classification and identification, there exist attacks that are designed to break those essential functions. Attacks targeting this fall under the classification of “evasion” attacks. The exact form of such evasion attacks varies significantly from model to model. Generally, however, the results of these attacks are the same.

For example, Tesla’s autopilot feature. There has been a large amount of research focused on the feature. To ensure safe vehicle operations, the Tesla autopilot feature uses advanced AI combined with an extensive array of external detection systems to identify potential road threats, as well as vital signage to ensure the vehicle operates lawfully. Recently, a group of researchers found it was possible to leverage a projector to create two-dimensional content that would trick Tesla’s autopilot feature. While these projections were fake to a human observer, the vehicle was fooled into thinking the images were real. By projecting fake signs and even human silhouettes directly onto the road, the AI behind the autopilot feature would change how the car was driving. In some situations, this leads to the vehicle driving erratically and unlawfully.

While the model is trained on good data, ruling out a potential poisoning attack, purposefully malicious inputs, such as those in the described evasion attacks, pose a significant threat to many AI systems.