Task 7.1

Taxonomy of Attacks, Defences, and Consequences in Adversarial Machine Learning

SIT719

Sanket Thakur

The taxonomy is structured in a conceptual manner, based on and incorporating previous AML survey works. A hierarchy containing main forms of attacks, defences, and consequences. Components in machine learning can be objects of adversary attacks using different methods and systems awareness. For the purpose of this report we are going to look at understanding few of the important

Attack types in Adversarial Machine Learning

Black box attacks

Scenarios in which the attacker does not have full access to the policy network are defined in Black-box adversarial attacks. The taxonomy classifies black-box attacks into two major classes

White Box Attacks

Scenarios in which the attacker has access to the underlying training policy network of the target model define the white-box adversarial attacks. The study showed that even the incorporation of tiny perturbations in the training policy may have a dramatic effect on model efficiency.

Grey Box Attacks

This type of adversarial attack is also a limited knowledge attack. Grey-box means that any knowledge about the device, its design or whatever is understood by an individual. Most common solutions use publicly open architectures such as Google.

Poisoning

The data or model is changed implicitly or specifically by poisoning, commonly known as causative attacks. In Indirect Poisoning, before pre-processing, adversaries without access to pre-processed data used by the target model would instead poison the data. The evidence is altered by data injection or data alteration in direct poisoning, or the paradigm is directly altered by logic corruption.

Evasion Attack

The opponent addresses a constrained optimisation issue in Evasion Attacks to locate a minor input disruption that induces a significant change in the loss function and results in misclassification of output. Gradient-based search algorithms such as

- Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS),
- Fast Gradient Sign Method (FGSM),
- or Jacobian-based Saliency Map Attack (JSMA) usually provide this.

L-BFGS was the first algorithm used to produce misclassifications using feedback disturbances that were imperceptible to human observers by a machine vision device model.

Defences against the attacks

Data Sanitization and Robust Statistics provide Protections against Poisoning Attacks. Adversarial examples are defined in Data Sanitization by measuring the effects of examples on classification results. In a method known as Reject on Negative Effect, examples that cause high error rates in classification are then excluded from the training collection. Robust statistics use limitations and regularisation methods to reduce possible disruptions in the learning model caused by poisoned data, rather than seeking to find poisoned data.

Different model robustness enhancements, including adversarial preparation, gradient masking, defensive distillation, ensemble approaches, function pressing, and reformers/autoencoders have protections against Evasion . While used in the Testing (Inference) phase as defences against attacks, these defences are deployed by the defender in the preparation phase that precedes testing (evasion)

Conclusion

In this article we have learned about different ways of adversarial attacks, and also looked at promising ways to protect against these threats.
When we introduce machine learning algorithms, this is undoubtedly something to keep in mind. We ought to protect against these adversarial threats instead of implicitly depending the models to deliver the right outcomes and still consider twice before we accept the decisions taken by these models.

Bibliography

Tabassi, E. et al., 2019. A Taxonomy and Terminology of Adversarial Machine Learning. *NIST.*