Survey on adversarial attack against machine learning systems and defensive strategy

Section 1 History, definition and impact of adversarial attacks. Why is it important to machine learning / security communities ? Does adversarial attack relate to fundamental issues of machine learning theory ?

Section 2 Adversarial Attack Technique

2.1 Mathematical definition of general adversarial attack techniques

2.2 Categories of adversarial attack techniques, focus of different categories of attack methods, advantages/shortage. For example, we categorise them into “black-box” and “white-box” attack techniques, we can add more detailed categorisation.

Section 3 Defensive strategies against adversarial attack techniques

3.1 Purpose of defensive strategies: why and how. Does defensive strategies relate to any fundamental theory, for example, generalisation error of ML models, robust statistics, differential privacy ?

3.2 Defensive learning algorithms against white-box attacks

3.3 Defensive learning algorithms against black-box attacks

3.4 Detection of adversarial samples

3.5 Open issues

Section 4 Benchmark of state-of-the-art defensive learning methods against white- and black- box attack techniques

4.1 Evaluation criterions of effectiveness of defensive learning methods. For example, transferability ?

4.2 Experimental results: a big table, with rows as each defensive learning method and cols corresponding to each attack techniques

4.3 Analysis on experimental results: which defensive strategy is the most effective method against which attack method and why ?

Section 5 Discussion and Conclusion