Wybrane metody

- Poisoning attack Umieszczenie na obrazach w jednym konkretnym miejscu znaczniki, dla każdej klasy inny znacznik. Sieć neuronowa nauczy się tego znacznika. Jak tylko taki znacznik w nieodpowiednim miejscu sieć zauważa i źle klasyfikuje.
- Model stealing Metoda kradnięcia modelu na podstawie API modelu
- Fast gradient sign method Dodajemy szum do zdjęcia z jakąś wagą. Szum jest stworzony na podstawie funkcji straty i używając back propagation. Wrzucamy zdjęcia funkcja straty mówi nam jak bardzo jesteśmy odlegli żeby zmienić klasyfikacje, liczymy back propagation jaki szum dodać. Człowiek nie zauważy szumu.
- Wizualizacja cech Możemy wizualizować cechy. Możemy patrzeć jak dany pixel jest postrzegany przez sieć i jak bardzo decyduje on o klasyfikacji. Zamieniane są kluczowe punkty, decydujące o klasyfikacji. Tych punktów jest tylko parę.
- One-pixel attack Zamieniany jest tylko jeden pixel i całkowicie oszukujemy sieć Jest atakiem black box. White box znamy sieć neuronową, możemy liczyc gradienty szumy. Black nie mamy dostepu do sieci, algorytm genetyczny modyfikujący doprowadzając do zmienienia jednego pixelu najbardziej kluczowego.
- Algorytmy genetyczne są ogółem wykorzystywane do oszukiwania sieci. Np generowanie obrazów, które w ogóle nie przypominają rzeczy a jest klasyfikowane.
- Adversarial patch Generowanie specjalnej nalepki nakładamy na obraz przed kamerę i sieć wariuje.
- Naturalne ataki Sieci neuronowe potrafią się uczyć niechcianych korelacji. Baza danych ma ileś zdjęć reprezentujących jakieś cechy np 1400 zdjęc bobasa w pieluszce. Jeżeli damy zdjecie bobasa na którym tej pieluchy nie ma sieć i tak wskaże że obraz należy do klasy z pieluchą

Jak się bronić?

Nauczyć się ataków. Sami genrujemy atak i uczymy sieć, że to jest atak.