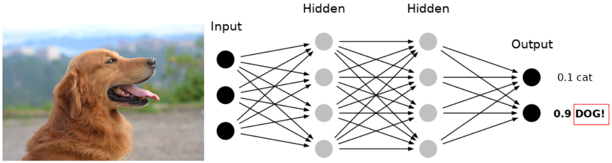
Adversarial Attacks on Neural Networks

Xiaolu Hou

Facutly of Informatics and Information Technologies
Slovak University of Technology in Bratislava

Neural networks

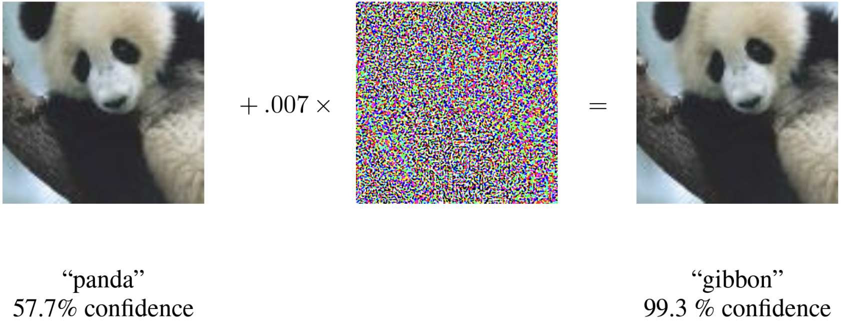
- AI algorithm for classification problems, e.g. image recognition
- A network of interconnected nodes or neurons where a signal is transmitted from input neurons toward output neurons



src: https://blog.finxter.com/the-magic-of-neural-networks-how-they-work/

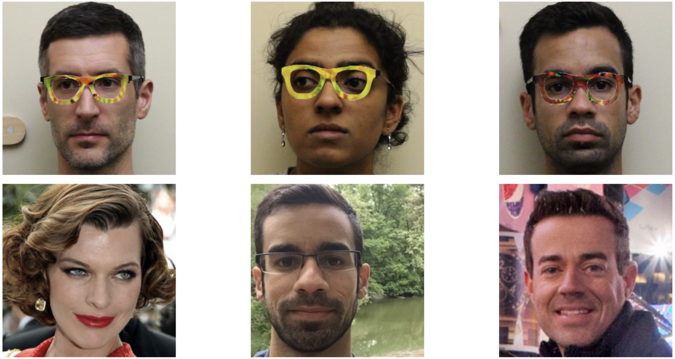
Adversarial examples

- Inputs formed by applying small but intentionally *worst-case* perturbations to examples from the dataset
- Aim: (targeted) misclassification with high confidence



Goodfellow, Ivan, et al. Explaining and Harnessing Adversarial Examples, 2014.

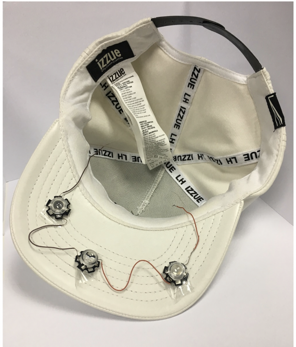
Impersonation with eyeglasses



- Optimization problem
  - Predict the attacker as the target
  - Printability with a commercial printer, smoothness, robustness
- Attacker gets the values of the pixels on the frame and prints the eyeglasses

Sharif, Mahmood, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, 2016.

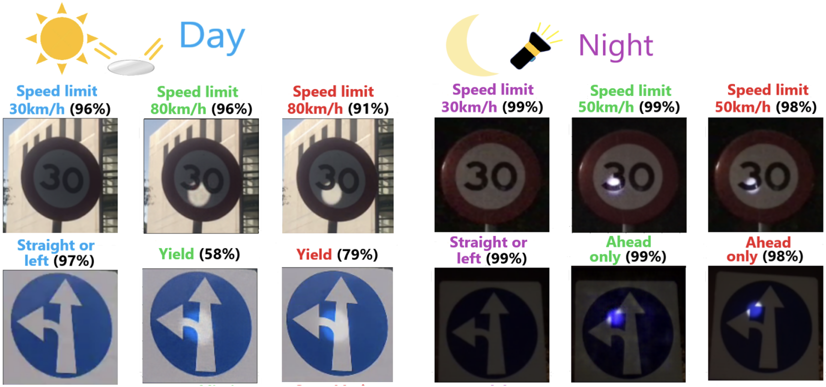
Invisibility with infrared light



- Face searching: video split into frames, extract face portion by identifying landmarks using a land marking model, located face is cropped out for later use
- Infrared light generated by an IR LED, cannot be observed by humans but can be captured by camera sensors
- With enough amount of infrared on the face, no landmarks can be found

Zhou, Zhe, et al. Invisible mask: Practical attacks on face recognition with infrared, 2018.

Road sign misclassification with lights



- Figures: original, simulation, attacked
- Day: mirror; Night: flashlight
- Public neural networks for GTSRB, LISA dataset

Hsiao, Teng-Fang, at al. Natural Light Can Also Be Dangerous: Traffic Sign Misinterpretation Under Adversarial Natural Light Attacks, 2024.

Thanks for your attention!

E-mail: `houxiaolu.email@gmail.com`