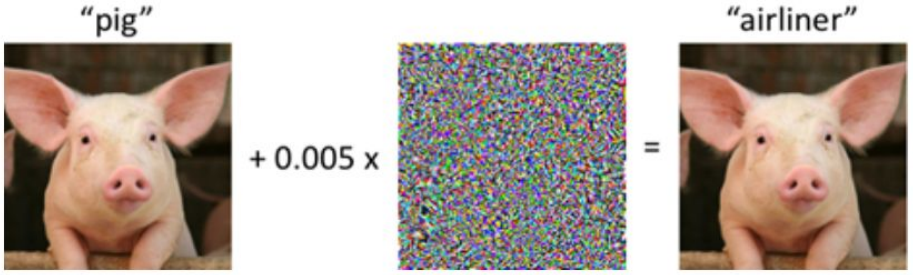
Adversarial Robustness through Saliency-based Noise Injection

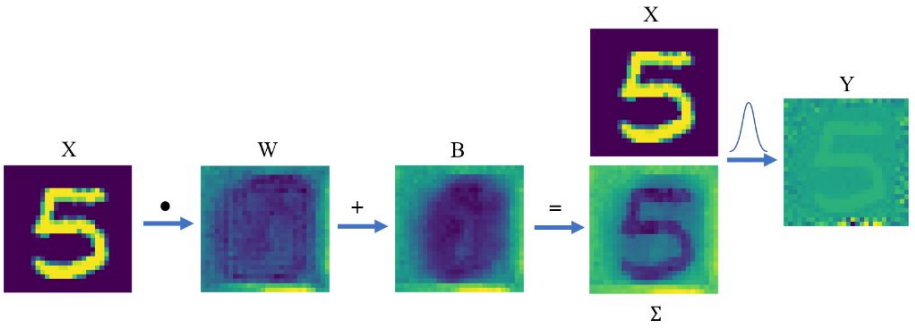
Introduction: Deep learning models have become a common feature in industry as well as everyday life. They are used to make decisions in autonomous driving, healthcare, security, and more. However, it has been shown that these models are vulnerable to small adversarial perturbations to their input, which can completely change their prediction. This limits the deployment of deep learning models in industrial applications. This research topic aims at training more robust neural networks using human intuitions about saliency.



@https://www.microsoft.com/en-us/research/blog/adversarial-robustness-as-a-prior-for-better-transfer-learning/

Possible Objectives:

1. Learnable injection of noise into the inputs (images) of a neural network to weaken adversarial attacks (Different Injection methods)
    a. Linear
    b. Non-Linear
    c. With human prior (saliency)



Requirements: Strong knowledge of deep learning; experience in programming with Python

Contact: leo.schwinn@fau.de, dario.zanca@fau.de