# Lofexidine Case Study Report

Lucas Bulczynski, Colin Song, Emi Pollard, & Alex Katopodis

March 22, 2024

## 1 Introduction

Deep learning architectures have rapidly become ubiquitous in modern software systems. The underlying neural networks are highly nonlinear approximations of complex functions that are great at handling large amounts of unstructured data. Their applications range from creative language/image synthesis, all the way to well-defined object detection. In certain cases, it can be crucial that a model has high specificity.

Neural networks are known to consistently misclassify *adversarial examples*, or inputs created by applying small but intentionally worst case perturbations to normal inputs so that they output incorrect answers with high confidence [1].

Constructing adversarial examples presents a challenging task. There are a practically infinite amount of ways to slightly perturb an image, of which only a tiny subset are adversarial. All algorithms for generating adversarial examples we will discuss are gradient based. If we can compute the gradient of the network output $\nabla_x F(x)$, we can perturb in the image in the direction that maximizes the chances of it being classified as $y_{\text{adv}}$ [1].

We will explore how adversarial examples can be efficiently constructed and how they are harmful in the context of computer vision. We'll peer into the math of gradient estimation and implement a black-box attack on a real model from scratch.

## 2 Discussion

This is the discussion.

## References

[1] Ian Goodfellow, Jonathan Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

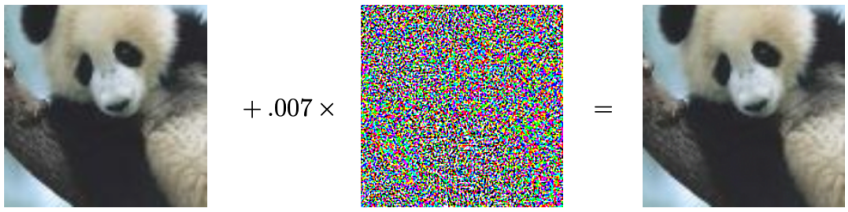


Figure 1: A demonstration of an adversarial example from Goodfellow et al.[1]. The resulting image is not discernible from the original, but receives a starkly different classification.