Contextual Fusion For Adversarial Robustness

Aiswarya Akumalla
Department of Medicine
University of California, San Diego
La Jolla, CA 92093
aakumall@eng.ucsd.edu

Seth Haney
Department of Medicine
University of California, San Diego
La Jolla, CA 92093
sethdhaney@gmail.com

Maxim Bazhenov
Department of Medicine
University of California, San Diego
La Jolla, CA 92093
mbazhenov@ucsd.edu

Abstract

Mammalian brains handle complex reasoning tasks in a gestalt manner by integrating information from regions of the brain that are specialised to individual sensory modalities. This allows for improved robustness and better generalisation ability. In contrast, deep neural networks are usually designed to process one particular information stream and susceptible to various types of adversarial perturbations. While many methods exist for detecting and defending against adversarial attacks, they do not generalise across a range of attacks and negatively affect performance on clean, unperturbed data. We developed a fusion model using a combination of background and foreground features extracted in parallel from Places-CNN and Imagenet-CNN. We tested the benefits of the fusion approach on preserving adversarial robustness for human perceivable (e.g., Gaussian blur) and network perceivable (e.g., gradient-based) attacks for CIFAR-10 and MS COCO data sets. For gradient based attacks, our results show that fusion allows for significant improvements in classification without decreasing performance on unperturbed data and without need to perform adversarial retraining. Our fused model revealed improvements for Gaussian blur type perturbations as well. The increase in performance from fusion approach depended on the variability of the image contexts; larger increases were seen for classes of images with larger differences in their contexts. We also demonstrate the effect of regularization to bias the classifier decision in the presence of a known adversary. We propose that this biologically inspired approach to integrate information across multiple modalities provides a new way to improve adversarial robustness that can be complementary to current state of the art approaches.

Preprint. Under review.