Towards Explainable and Robust Deep Learning Architectures

Eduardo Almeida Soares
Lancaster University

Introduction

Deep neural networks have made a tremendous success as they could achieve high accuracy on different complex applications.
However:

- Several vulnerabilities to adversarial attacks.
- Deep learning tends to make wrongly overconfident predictions on modified.
- Black-box nature makes extremely difficult to audit their decisions.

Security aspects of machine learning are extremely important specially on high stake applications.
Therefore, we propose a prototype-based method that is able to detect changes in the data patterns and detect imperceptible adversarial attacks on real time.

Imperceptible Adversarial Attacks

Image adversarial attacks are focused on perturbations that cause misclassification by a deep classifier while are imperceptible to humans.

PerC algorithm which creates adversarial examples by perturbing images through its perceptual color distance

Figure 1.
Imperceptible adversarial attack.

Results

RADNN

The proposed RADNN is equipped with a mechanism that allows real-time concept drift detection (data pattern detection) due its density-based nature and its design based on prototypes. The RADNN approach is described as a feedforward neural network. The training architecture that is composed by the following layers:

- Features layer;
- Density layer;
- Conditional probability layer;
- Prototype identification layer.

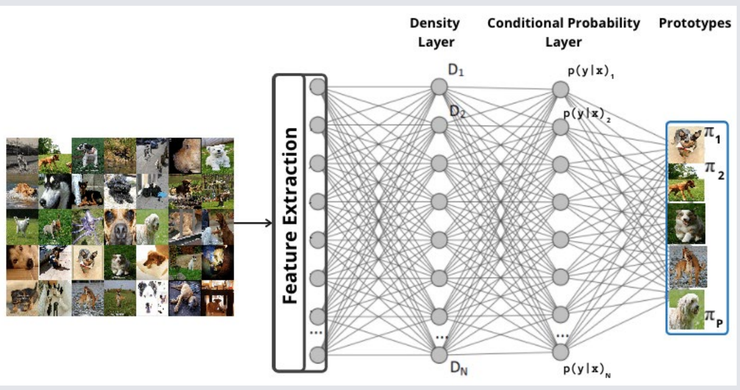

Figure 2.
RADNN Training architecture.

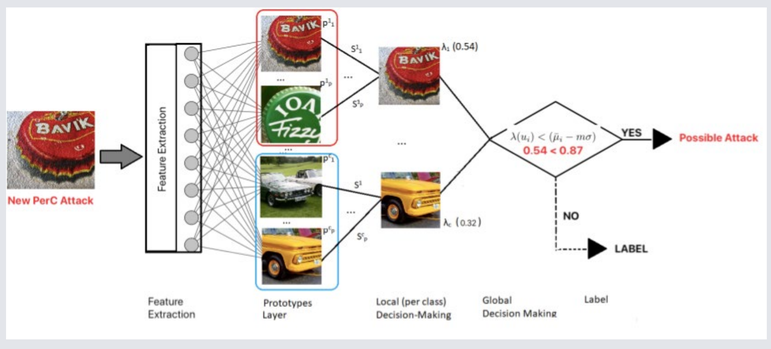

Figure 2.
RADNN decision structure.

RADNN calculates the similarity of the new unlabeled data sample to the existing prototypes. If the similarity score is lower than the 3-sigma rule it means that the new data belongs to a new data distribution type, which may characterize an attack. The similarity is calculated as:

$$S(x_i, \pi_j) = \frac{1}{1 + \frac{||x_i - \pi_j||^2}{||\sigma_j||^2}}$$

Where:
S(x, π) is the Similarity
π is the prototype
σ is the variance

To evaluate the robustness of RADNN to imperceptible attacks, we considered 1000 images from the Imagenet dataset attacked by the PerC algorithm. Deep Learning approaches as VGG--16 and ResNet were also used during this experiment.

Method	Detection rate (%)
RADNN	**97.2%**
DenseNet-201 [18]	57.48%
ResNet-152 [11]	46.23%
VGG–16 [14]	38.31%
AlexNet [10]	37.72%

Table 1.
Results for robustness testing.

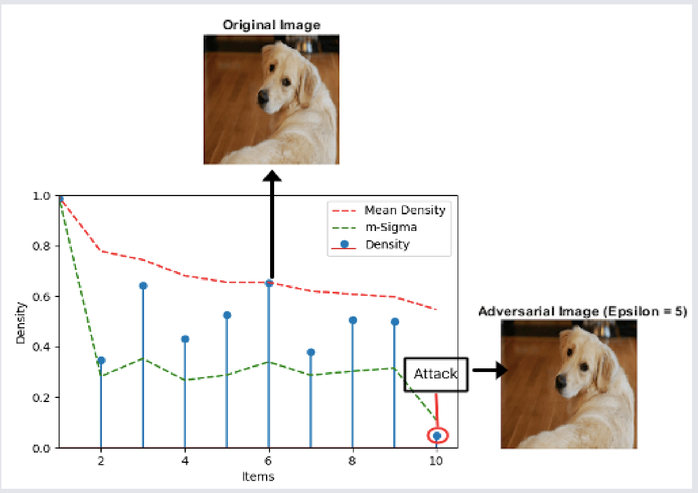

Figure 3.
RADNN attack detection based on the similarity drop.

Discussion

We introduced RADNN algorithm. The algorithm has a robust design that is able to detect imperceptible to human attacks due its density-prototype-based nature. The experiment have shown that differently from traditional approaches that need to be trained on the attacks to obtain high performance in terms of detection, the RADNN is able to detect attacks without *prior* training on it due its confidence system based on similarities.

References

1. Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). Neural Networks, 130, 185-194.
2. Angelov, P., & Soares, E. (2021). Detecting and learning from unknown by extremely weak supervision: exploratory classifier (xClass). Neural Computing and Applications, 1-13.
3. Zhao, Z., Liu, Z., & Larson, M. (2020). Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1039-1048).

Acknowledgments

This work is supported, in part, by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]