Adaptive Neighbourhoods for the Discovery of Adversarial Examples

Jay Morgan, University of Toulon

13th October 2022

A thank you to my collaborators

Adeline Paiement
University of Toulon

Arno Pauly
Swansea University

Monika Seisenberger
Swansea University

Deep Neural Networks

What is a deep neural network

Adversarial Examples

What is an adversarial example

Motivating Principles

Outline for this talk

1. Look at existing solutions
2. Our complimentary method
3. Some results on two tasks:
   - Iris Dataset
   - Solar Burst Detection
4. Some conclusions

Existing Methods for Creating Adversarial Examples

Fast Gradient Sign Method (FGSM)

Projected Gradient Descent (PGD)

Carlini & Wagner (C&W)

What do we learn from these methods?

Figure

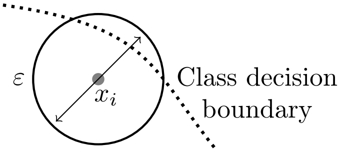


Figure: Example where a data point x_i lies close to the class decision boundary. In these situations, too large ε values, may push the synthetically generated point over true class boundaries.

Figure 2

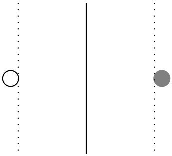


Figure: Sparse regions of the manifold may appear simple due to the lack of information.

Figure 3

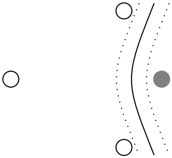


Figure: More data points enable more precise estimation of the class boundary.

Our method

How does our method work?

Estimating Sparsity/Density

$$\varphi(x;\overline{x}) = \frac{1}{\sqrt{1+(\varepsilon r)^2}}, \text{ where } r = \parallel \overline{x} - x \parallel \tag{1}$$

Providing the RBF's width parameter is suitably chosen, we achieve a good measure of the density through the sum of the RBFs centred on all data points X^c of class c (Eq.~2).

$$\rho_c(x) = \sum_{x_j \in X^c} \varphi(x; x_j) \tag{2}$$

Expansion

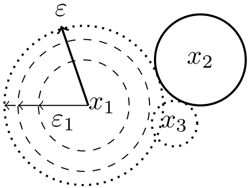


Figure: Iterative ε-expansion process in a binary class scenario. The two classes are distinguished by the dotted and solid circles.

$$\Delta\varepsilon_i^n = e^{-\rho_{c(i)}(x_i)\cdot n} \tag{3}$$

Iris Dataset

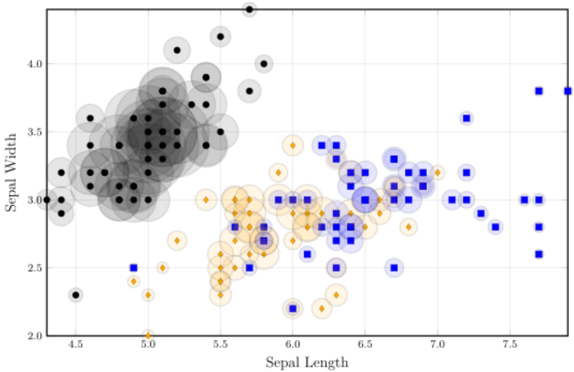


Figure: Proposed adaptive neighbourhoods for the Iris dataset. The three classes of flower are represented by different shaped markers. The size of the neighbourhood for each sample is indicated with a circle centred on the data point. Intersections between neighbourhoods of different classes are not real but are visualisation artefacts coming from the 2D projection of 4 dimensions.

Training

$$\mathcal{L}_{total} = (1 - \alpha)\mathcal{L}_{cls} + \alpha\mathcal{L}_{adv}$$

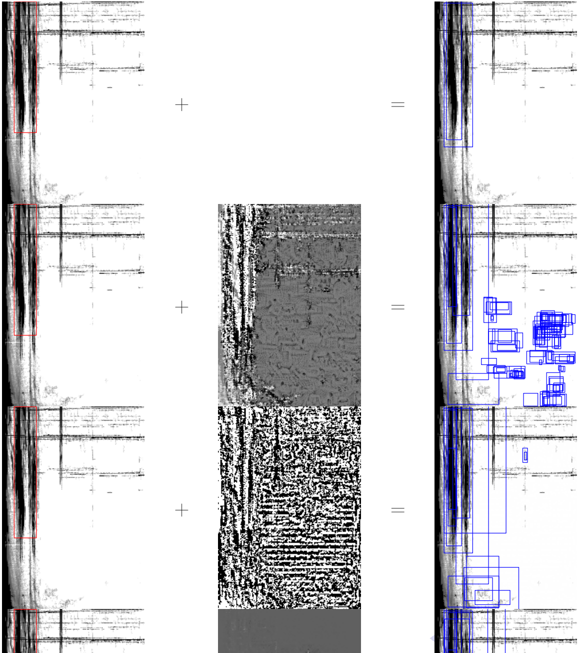
where $\mathcal{L}_{cls}$ and $\mathcal{L}_{adv}$ are the cross-entropy losses of the un-perturbed and perturbed data, respectively

Results

Table: F_1 score of DNN for the Iris dataset using various adversarial defence methods. Scores are in the format: mean (standard deviation) over 10 k-folds. Bold font face indicates the best form of attack for each type of defence method.

		Attack			
	None	FGSM	PGD	FGSM+AN	P
	0.9745 (0.0413)	0.9278 (0.0618)	0.8572 (0.1036)	**0.7764 (0.0813)**	0.84
	0.9811 (0.0396)	0.9408 (0.0757)	0.8468 (0.1080)	**0.7873 (0.0785)**	0.84
	0.9867 (0.0400)	0.9462 (0.0740)	0.8680 (0.0740)	**0.8508 (0.0746)**	0.87
AN	0.9936 (0.0193)	0.9272 (0.0620)	0.8274 (0.0918)	**0.7935 (0.0822)**	0.84
N	0.9936 (0.0193)	0.9406 (0.0745)	0.8420 (0.0987)	**0.8140 (0.1085)**	0.85
N	0.9936 (0.0193)	0.9472 (0.0642)	0.9472 (0.0642)	**0.8679 (0.0899)**	0.87

Adversarial Training for Solar Burst Detection



Results

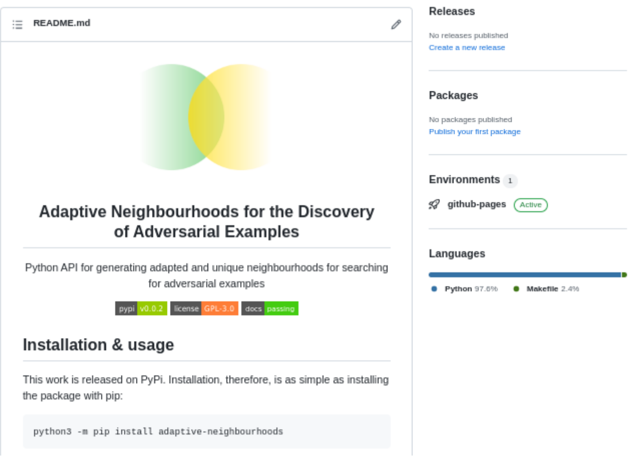
Table: F_1 score performance on the WAVES dataset using Faster R-CNN. Numbers highlighted in a bold font face indicate the best achieving adversarial attack for each form of defence.

		Attack					
Defence	None	FGSM	FGSM+AN	PGD	PGD+AN	DAG	DAG+A
None	0.568	0.539	0.486	0.198	**0.105**	0.399	0.251
FGSM	0.463	0.458	0.178	0.013	**0.012**	0.055	0.028
GSM+AN	0.480	0.465	0.462	**0.007**	**0.007**	0.043	0.023
PGD	0.421	0.425	0.379	0.391	0.359	0.378	**0.259**
PGD+AN	0.364	0.359	0.330	0.339	0.324	0.330	**0.212**

Summary of Results

This is the summary

Source code



https://github.com/jaypmorgan/adaptive-neighbourhoods
https://gitlab.com/jaymorgan/adaptive-neighbourhoods
https://git.sr.ht/~jaymorgan/adaptive-neighbourhoods

Thank you

References