Adaptive Neighbourhoods for the Discovery of Adversarial Examples

Jay Morgan, University of Toulon

13th October 2022

A thank you to my collaborators

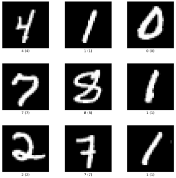
Adeline Paiement
University of Toulon

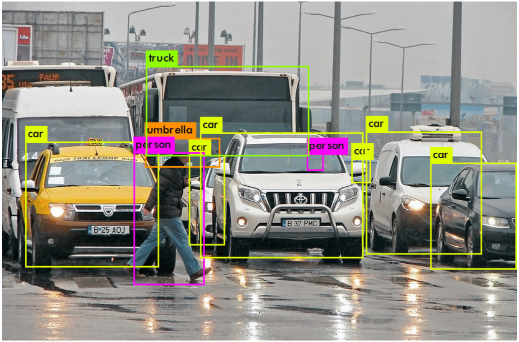
Arno Pauly
Swansea University

Monika Seisenberger
Swansea University

Deep Learning models

The abilities of Deep Learning models have only continued to improve, and the range of tasks they can perform is growing: from simple digit recognition, to simultaneous detection of multiple objects in a scene.

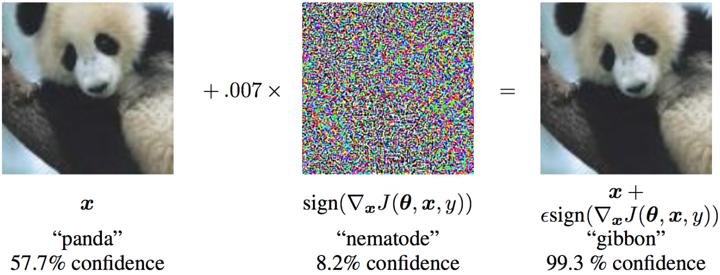




(Potdar, Kedar and Pai, Chinmay and Akolkar, Sukrut, 2018)

Adversarial Examples

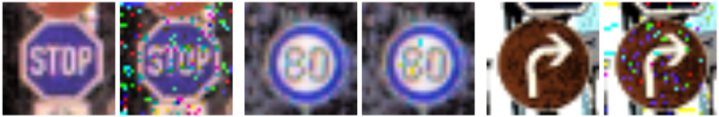
Adversarial examples are created by changing pixel values in the input image, resulting in an output image that looks almost identical but the Deep Learning model predicts and entirely different class for this output image.



(Goodfellow, Ian J and Shlens, Jonathon and Szegedy, Christian, 2014)

Motivating Principles

For safety critical systems, miss-classifications are more catastrophic.



"stop"
to "30m speed limit"

"80m speed limit"
to "30m speed limit"

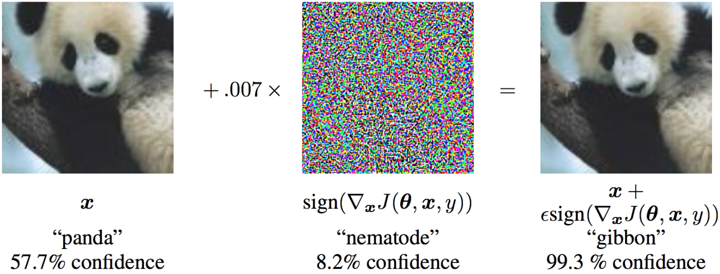
"go right"
to "go straight"

(Huang, Xiaowei and Kwiatkowska, Marta and Wang, Sen and Wu, Min, 2017)

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

Fast Gradient Sign Method (FGSM)



(Goodfellow, Ian J and Shlens, Jonathon and Szegedy, Christian, 2014)

Projected Gradient Descent (PGD)

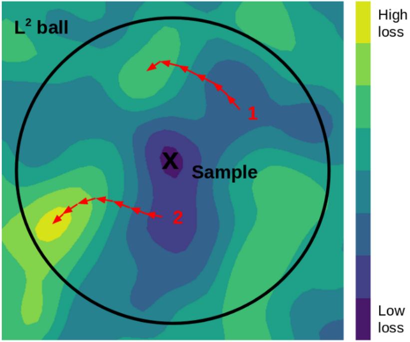
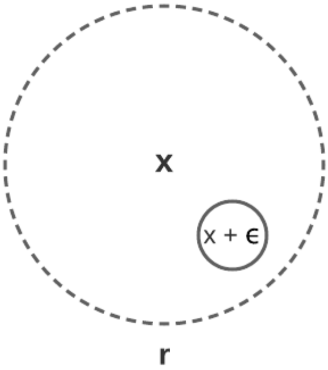


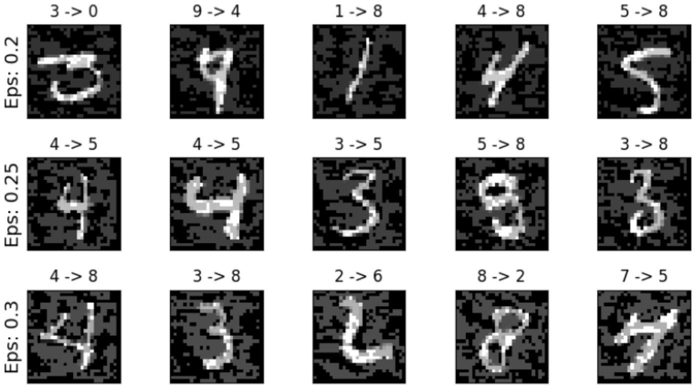
Figure:
https://towardsdatascience.com/know-your-enemy-7f7c5038bdf3

(Madry, Aleksander and Makelov, Aleksandar and Schmidt, Ludwig and Tsipras, Dimitris and Vladu, Adrian, 2017)

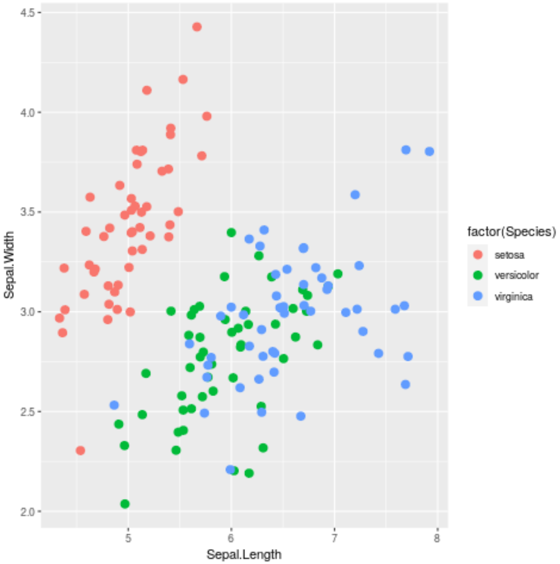
What do we learn from these methods?



Amount of change is important

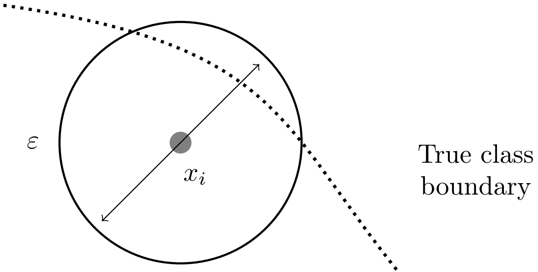


How to decide maximum perturbation for non-image representations



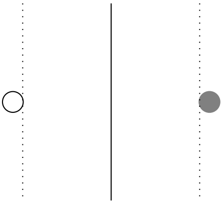
Our method – Adaptive Neighbourhoods

Perturbations shouldn't pass class boundaries

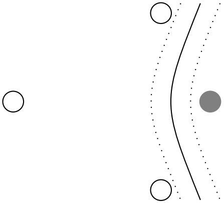


Example where a data point x_i lies close to the class decision boundary. In these situations, too large ε values, may push the synthetically generated point over true class boundaries.

Estimated boundaries can be deceiving



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



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

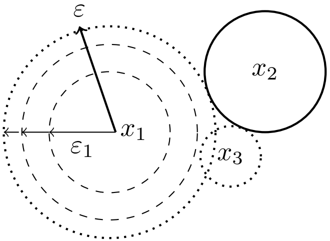
Estimating Sparsity/Density

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

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).

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

Iterative expansion to create 'adapted neighbourhoods'



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

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

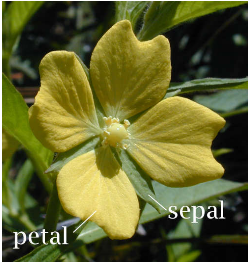
Results

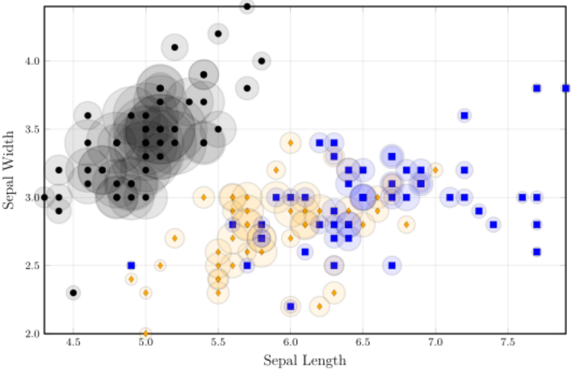
Aim of Experimentation

We'd like to answer the following:

1. Does using adaptive neighbourhood provide any benefit? Why use it at all?
2. Can existing methods work for non-image based datasets, or do we need to design new methods entirely?

Classification of Iris flowers – problem statement





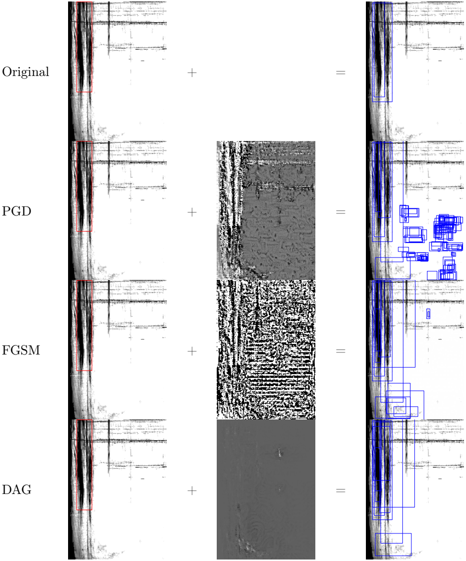
(Jay Morgan and Adeline Paiement and Arno Pauly and Monika Seisenberger, 2021)

Attack and defence results for the Iris dataset classification task

Defence	None	Attack FGSM	PGD	FGSM+AN	PGD+AN
None	0.9745 (0.0413)	0.9278 (0.0618)	0.8572 (0.1036)	0.7764 (0.0813)	0.8461 (0.0968)
FGSM	0.9811 (0.0396)	0.9408 (0.0757)	0.8468 (0.1080)	0.7873 (0.0785)	0.8448 (0.0698)
PGD	0.9867 (0.0400)	0.9462 (0.0740)	0.8680 (0.0740)	0.8508 (0.0746)	0.8759 (0.0823)
Random+AN	0.9936 (0.0193)	0.9272 (0.0620)	0.8274 (0.0918)	0.7935 (0.0822)	0.8454 (0.0864)
FGSM+AN	0.9936 (0.0193)	0.9406 (0.0745)	0.8420 (0.0987)	0.8140 (0.1085)	0.8588 (0.1157)
PGD+AN	0.9936 (0.0193)	0.9472 (0.0642)	0.9472 (0.0642)	0.8679 (0.0899)	0.8753 (0.0864)

What we learn here then is that our adaptive neighbourhoods is able to strengthen the form of adversarial attack and defence.

Adversarial examples in a Solar Burst Detection task – problem statement



(Jay Morgan, 2022)

Attack and defence results for the solar bursts task

		Attack					
Defence	None	FGSM	FGSM+AN	PGD	PGD+AN	DAG	DAG+AN
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
FGSM+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

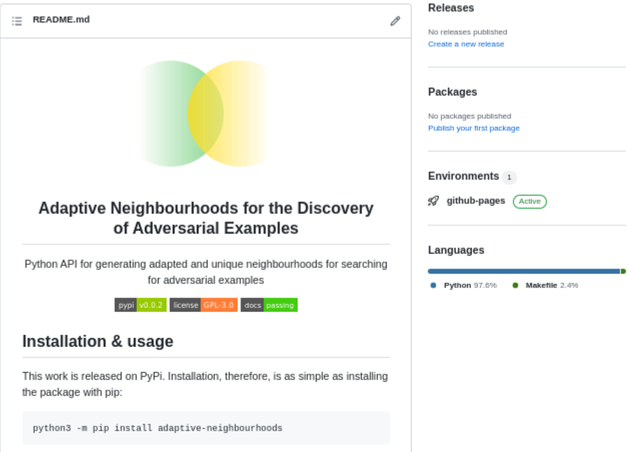
Like our previous task, we see that, through the combination with adaptive neighbourhoods, the attack is more successful. And likewise the defence is more powerful.

Summary of Results

We'd like to answer the following:

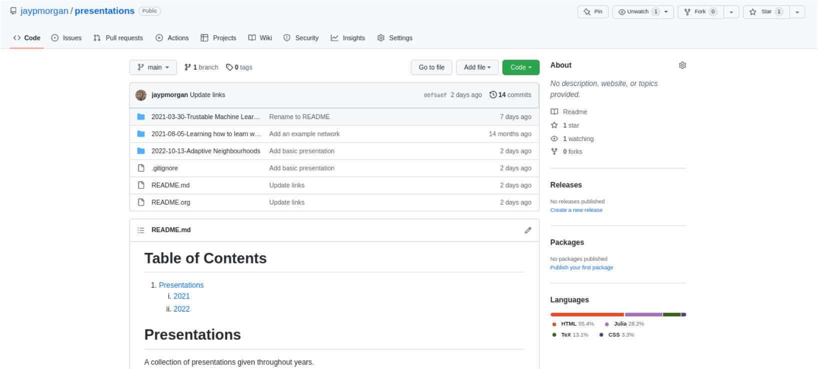
1. Does using adaptive neighbourhood provide any benefit? Why use it at all? - Adaptive neighbourhoods is an effective method that compliments existing adversarial generation methods such as FGSM & PGD.
2. Can existing methods work for non-image based datasets, or do we need to design new methods entirely? - Through the use of adaptive neighbourhoods, one can meaningfully define searchable regions for datasets other than image-based data where adversarial examples can be visually inspected.

Source code



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

Link to the Slides



https://github.com/jaypmorgan/presentations

Thank you!

References

Goodfellow, Ian J and Shlens, Jonathon and Szegedy, Christian (2014). *Explaining and harnessing adversarial examples*, arXiv preprint arXiv:1412.6572.

Huang, Xiaowei and Kwiatkowska, Marta and Wang, Sen and Wu, Min (2017). *Safety verification of deep neural networks.*

Jay Morgan (2022). *Strategies to use Prior Knowledge to Improve the Performance of Deep Learning.*

Jay Morgan and Adeline Paiement and Arno Pauly and Monika Seisenberger (2021). *Adaptive Neighbourhoods for the Discovery of Adversarial Examples*, CoRR.

Madry, Aleksander and Makelov, Aleksandar and Schmidt, Ludwig and Tsipras, Dimitris and Vladu, Adrian (2017). *Towards deep learning models resistant to adversarial attacks*, arXiv preprint arXiv:1706.06083.

Potdar, Kedar and Pai, Chinmay and Akolkar, Sukrut (2018). *A Convolutional Neural Network based Live Object Recognition System as Blind Aid.*