The Dimpled Manifold Revisited

Lukas Karner

Supervisors: Moritz Grosse-Wentrup, Sebastian Tschiatschek

Adversarial Examples

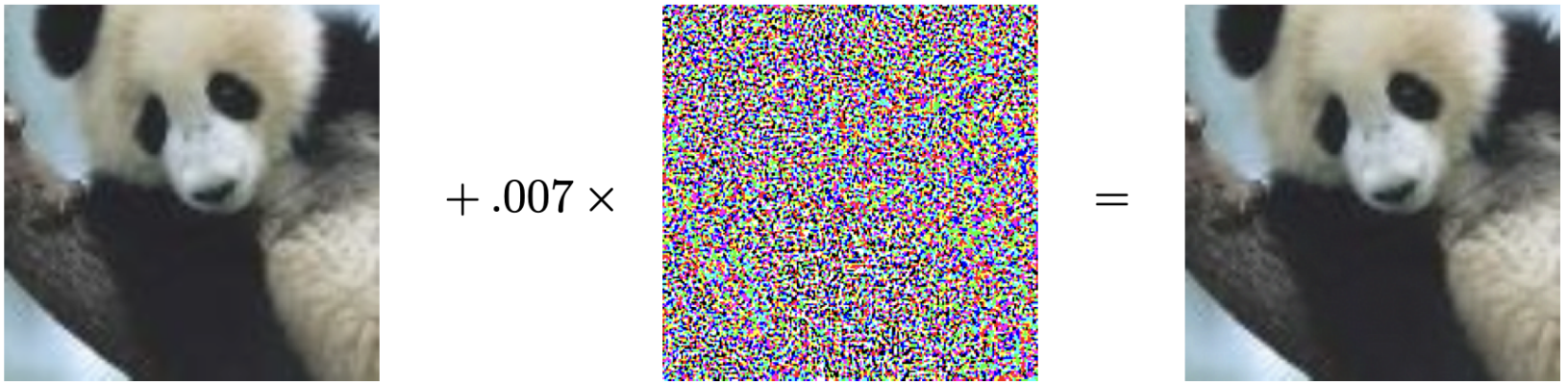

Figure 1: A classical adversarial example from [2].

Proposed Explanations and Defences

Some of the more recently proposed explanations are:

The Non-Robust Features Hypothesis [5].

The Boundary Tilting Hypothesis [10].

The Dimpled Manifold Model [9].

There is a plethora of proposed defences, a survey can be found in [8], but none are perfect so far.

The Dimpled Manifold Model

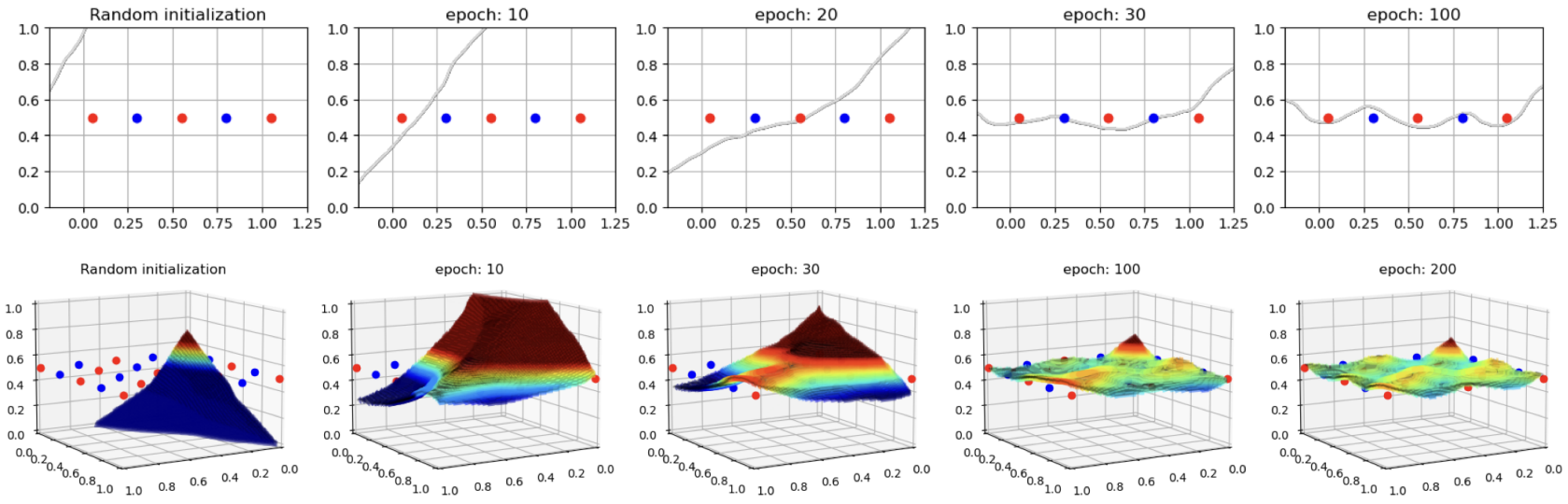

Figure 2: Exemplaric behaviour of decision boundaries hypothesised in [9]

Project Objective

The project consists of 2 phases:

1. To reproduce results from the paper [9] introducing the Dimpled Manifold Model. (in progress)

2. To explore possibilities to change the objectives and architectures of neural networks in the light of the Dimpled Manifold Model in order to make them more robust to adversarial examples.

Denoising as Defence

Many techniques have been proposed, most notably:

Denoising Autoencoders [4].

JPEG compression [1].

Feature Squeezing [11].

High-Level Representation Guided Denoiser [7] (HGD, winner of the 2017 NIPS contest [6]).

However, all of them have pros and cons and are still vulnerable to adaptive adversaries. (for an overview see [8])

Denoising Orthogonally?

Can we use the geometric hypothesis of the Dimpled Manifold Model to design a better denoiser?

Yes! If the image manifold is a linear subspace, then PCA would be a perfect denoiser beacuse a classifier trained on the latent space would have a decision boudary that is orthogonal to the manifold instead of clinging to it - making attacks essentially impossible!

But what about non-linear manifolds? Kernel PCA is expensive. Could we use some kind of autoencoder?

Isometric Autoencoder

The Isometric Autoencoder (I-AE) was introduced in [3] and is a novel regularisation method that enforces an autoencoder to behave simillar to PCA locally around the manifold.

In particular in promotes the decoder to be an isometry and the encoder to be its pseudo-inverse by enforcing conditions that are necessary for this to be fullfilled.

In this way it could be possible to achieve orthogonality of the decision boundary to the manifold similarly to the example from before.

Comparison to Other Methods

I-AE does not rely on heuristics or sampling (such as HGD), but it works provably (locally around the manifold).

It is a regularisation method that can be applied to any autoencoder architecture.

In contrast to HGD one can work not only in the original space, but also in the latent space.

If the PCA-like properties hold also off the manifold, it can be used to make classifiers robust to any attack.

Open Questions

What is the geometry of the encoder away from the manifold?

Does it still project (nearly) orthogonally on the manifold?

How quickly do these properties degrade when one moves away from the manifold?

If the encoder does not behave as desired, how can one improve it?

Can the I-AE indeed mitigate adversarial attacks as expected?

Steps Toward Answers

At first examine the behaviour of the I-AE on suitable artificial data in a medium to high-dimensional setting. This involves studying the mapping's properties directly, but also launching adversarial attacks on models trained with the proposed method. In particular benchmark the I-AE against PCA.

Further examinations would then be done on real data, in particular one can compare the performance of the I-AE against HGD (for which code is only available for ImageNet).

Thank You for Your Attention!

References I

[1] Nilaksh Das et al. "Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression". In: *arXiv preprint arXiv:1705.02900* (2017).

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

[3] Amos Gropp, Matan Atzmon, and Yaron Lipman. "Isometric autoencoders". In: *arXiv preprint arXiv:2006.09289* (2020).

[4] Shixiang Gu and Luca Rigazio. "Towards deep neural network architectures robust to adversarial examples". In: *arXiv preprint arXiv:1412.5068* (2014).

References II

[5] Andrew Ilyas et al. "Adversarial examples are not bugs, they are features". In: *Advances in neural information processing systems* 32 (2019).

[6] Alexey Kurakin et al. "Adversarial attacks and defences competition". In: *The NIPS'17 Competition: Building Intelligent Systems*. Springer, 2018, pp. 195–231.

[7] Fangzhou Liao et al. "Defense against adversarial attacks using high-level representation guided denoiser". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1778–1787.

References III

[8] Gabriel Resende Machado, Eugênio Silva, and Ronaldo Ribeiro Goldschmidt. "Adversarial machine learning in image classification: A survey toward the defender's perspective". In: *ACM Computing Surveys (CSUR)* 55.1 (2021), pp. 1–38.

[9] Adi Shamir, Odelia Melamed, and Oriel BenShmuel. "The dimpled manifold model of adversarial examples in machine learning". In: *arXiv preprint arXiv:2106.10151* (2021).

[10] Thomas Tanay and Lewis Griffin. "A boundary tilting persepective on the phenomenon of adversarial examples". In: *arXiv preprint arXiv:1608.07690* (2016).

References IV

[11] Weilin Xu, David Evans, and Yanjun Qi. "Feature squeezing: Detecting adversarial examples in deep neural networks". In: *arXiv preprint arXiv:1704.01155* (2017).