

# The Dimpled Manifold Revisited

Lukas Karner

Supervisors: Moritz Grosse-Wentrup, Sebastian Tschitschek

# Adversarial Examples

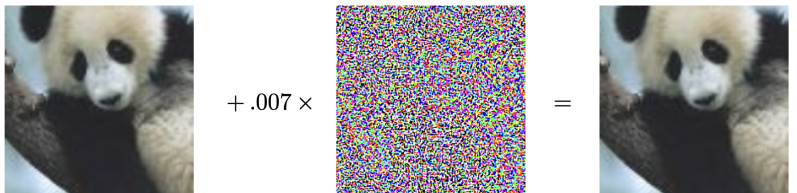

$$\begin{array}{ccc} \text{panda image} & + .007 \times \text{noise image} & = \text{adversarial image} \\ x & \text{sign}(\nabla_x J(\theta, x, y)) & x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \\ \text{"panda"} & \text{"nematode"} & \text{"gibbon"} \\ 57.7\% \text{ confidence} & 8.2\% \text{ confidence} & 99.3\% \text{ confidence} \end{array}$$

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

# The Dimpled Manifold Model

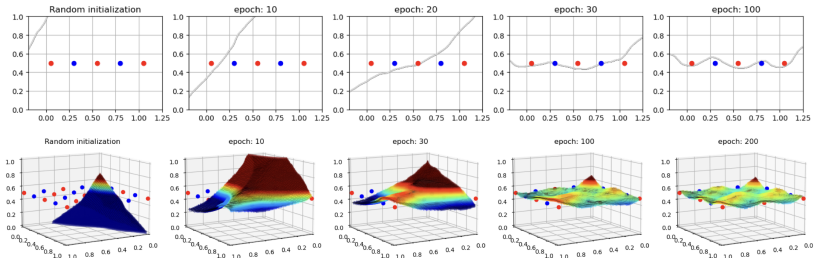


Figure 2: Exemplary behaviour of decision boundaries hypothesised in [3]

# Project Objective

The project consists of 2 phases:

1. To reproduce results from the paper [3] introducing the Dimpled Manifold Model.
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. → Isometric Autoencoder [2].

# Milestones

- ▶ Phase 1: Reproduce results from original paper [3] of the Dimpled Manifold Model.
- ▶ Phase 2A: Find a way to leverage the Dimpled Manifold Model against adversarial attacks.
- ▶ Phase 2B: Implement the hypothesised defence.

# Phase 1

Goal: Reproduce original paper's results on MNIST, CIFAR10 and ImageNet.

Necessary steps:

- ▶ train/load classifiers for each dataset
- ▶ train autoencoders for (subset of) each dataset
- ▶ implement adversarial attack
- ▶ implement projection on manifold extracted from autoencoder

# Phase 1 - Evaluation

Several difficulties encountered with the paper:

- ▶ Wrong hyperparameters reported?
- ▶ Wrong claims about model architecture?
- ▶ Unclear accounts of model performance.
- ▶ Could not successfully load pre-trained CIFAR10 model.

Solution: Figure everything out yourself.

Result: All tasks completed, gained a lot of practical experience, took much more time than expected.

# Phase 2A

Goal: Find a novel defence against attacks using the Dimpled manifold Hypothesis.

Literature: Many existing defences use denoising methods, but they are not completely satisfying (vulnerable to white box attacks).

Idea: Use hypothesis that adversarial noise is orthogonal to data manifold. → Use Isometric Autoencoder [2].



## Phase 2A - Evaluation

The goal was successfully completed, with the Isometric Autoencoder I found exactly what I was looking for!

So far I am not aware of similar existing approaches.

Possible alternative to I-AE is Isometric Representation Learning [4] (in contrast to I-AE this is peer-reviewed and published).

Also here I learned a lot of new things / read many papers.

## Phase 2B (in progress)

Goals:

- ▶ Implement Isometric Autoencoder and/or Isometric Representation Learning. (in progress)
- ▶ Evaluate their mathematical properties, i.e. check if they behave as expected on artificial data. (in progress)
- ▶ Apply them to adversarial examples. (tbd)

## Phase 2B - Evaluation (so far)

The loss functions for the I-AE have been implemented.

First attempts to train an I-AE on MNIST were not successful (reconstruction error decreases, but not the I-AE losses).

First attempt to train an I-AE on artificial data was not successful.

# Results and Code Demo

# Discussion

A new paper [5] claims to refute the Dimpled Manifold Model - no response by Shamir et. al. yet.

The DMM paper is not very rigorously formulated, in high dimension everything is "almost orthogonal".

Latent space dimensions in the used Autoencoders seem to be much too large.

# Future Work

Finish my remaining tasks for this project.

Use coefficients of QR decomposition to determine manifold dimension.

Compare I-AE for projection on manifold to projection using the QR decomposition of the tangential matrix.

Thank You for Your Attention!

# References I

- [1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (2014).
- [2] Amos Gropp, Matan Atzmon, and Yaron Lipman. “Isometric autoencoders”. In: *arXiv preprint arXiv:2006.09289* (2020).
- [3] Adi Shamir, Odelia Melamed, and Oriel BenShmuel. “The dimpled manifold model of adversarial examples in machine learning”. In: *arXiv preprint arXiv:2106.10151* (2021).
- [4] LEE Yonghyeon et al. “Regularized Autoencoders for Isometric Representation Learning”. In: *International Conference on Learning Representations*. 2021.



## References II

- [5] William Zhao and Subha Nawer Pushpita. “Extensions on The Dimpled Manifold Hypothesis”. In: ().