The Dimpled Manifold Revisited

Lukas Karner

Supervisors: Moritz Grosse-Wentrup, Sebastian Tschiatschek

The Dimpled Manifold Model

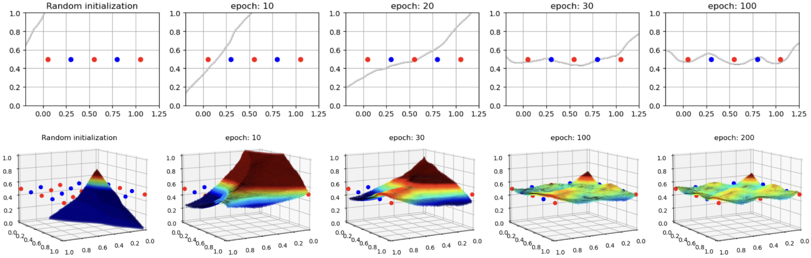

Figure 1: Exemplary behaviour of decision boundaries hypothesised in [3]

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. $\rightarrow$ Isometric Autoencoder [2].

Timeline

The plan was to finish phase 1 by end of November, it was actually finished today.

Further milestones:
Implement Isometric Autoencoder by end of December.
Conduct Experiments with it until February.
Start writing project paper 2-3 weeks before deadline.

What do we need?

For every dataset a classifier and an autoencoder.

A function to compute adversarial attacks.

A function to compute the projection onto the image manifold.

Classifiers

All classifiers could be trained successfully and achieved the performance claimed by the paper.

Problems encountered: No instructions for preprocessing, pre-trained weights for Cifar10 did not work.

Autoencoders

In the end all autoencoders could be trained successfully, and achieved good performance.

However there were several problems with this:

No instructions for preprocessing.

The reported performance on Cifar10 and ImageNet seems wrong.

Cifar10 architecture much too large $\rightarrow$ use much smaller architecture.

Had to add sigmoid output activation to the ImageNet autoencoder.

reported number of latent dimensions is wrong.

Attack Function

The paper used the *advertorch* [1] package, but this only supports very old versions of pytorch.

Hence I implemented the attack function myself, which worked without problems.

Projection on Manifold

To compute the projection on the manifold, at first, I computed the derivate of the output of an autoencoder with respect to the latent variables and then orthogonalised this set of vectors with the qr decomposition.

Problem: For ImageNet this matrix is 14GB large.

In order to not run out of RAM, I had to implement an in-place qr-decomposition.

Results

Additional Literature

Last week a new preprint article [5] was published which claims to have statistically significant results against the dimpled manifold model.

Further I found a second article [4] which also introduces a kind of Isometric Autoencoder.

Thank You for Your Attention!

References I

[1] Gavin Weiguang Ding, Luyu Wang, and Xiaomeng Jin. "AdverTorch v0.1: An Adversarial Robustness Toolbox based on PyTorch". In: *arXiv preprint arXiv:1902.07623* (2019).

[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: ().