The Dimpled Manifold Revisited

Lukas Karner

Supervisors: Prof. Moritz Grosse-Wentrup, Prof. Sebastian Tschiatschek

About The Project

- Adversarial examples are small perturbations that cause models to misclassify images.
- Several explanations for them have been proposed, but none are satisfying so far.
- This project seeked to reproduce results from the recently proposed *Dimpled Manifold Model*.
- A further goal of the project was to design a new defence exploiting the insights provided by the DMM.
- The project consisted of the following phases:
  - **Phase 1: Reproducing original results.**
  - **Phase 2A: Leveraging geometrc insights.**
  - **Phase 2B: Implementing the new defence.**

What are Adversarial Examples?

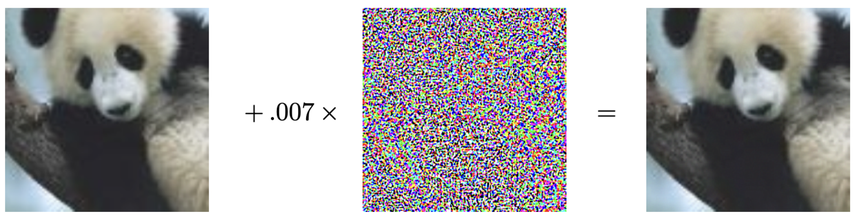

Fig. 1: A classical adversarial example from [1]

Proposed Explanations for Adversarial Examples

Some of the more recently proposed explanations are:

- The Non-Robust Features Hypothesis [3].
- The Boundary Tilting Hypothesis [5].
- The Dimpled Manifold Model [4].

The Dimpled Manifold Model

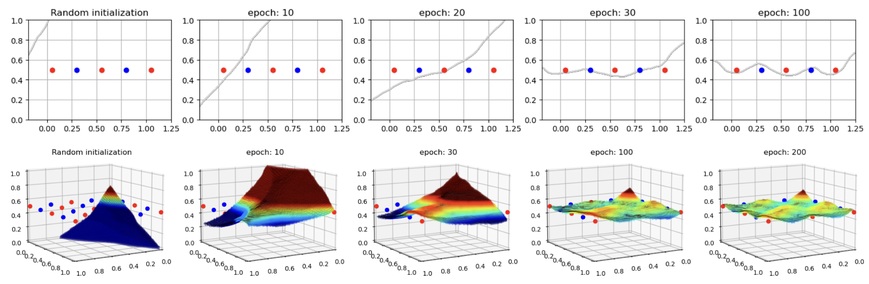

Fig. 2: Exemplaric behaviour of decision boundaries hypothesised in [4]

Phase 1 - Reproducing Original Results

Here we need the following things:

- **Classifiers and Autoencoders:** One of each for every dataset. After some problems everything worked.
- **Adversarial Attack Function:** Implemented on my own. Very valuable learning experience!
- **Manifold Projection Function:** Implemented using QR-decomposition. Encountered RAM problems.

Conclusion and Future Work

- The main results of the the paper [4] could be reproduced successfully.
- This is not a confirmation of the DMM!
- The Isometric Autoencoder could be implemented, but not trained.
- Use as defence remains an open question.

Adversarial Examples

Standard ℓ_2 Attacks

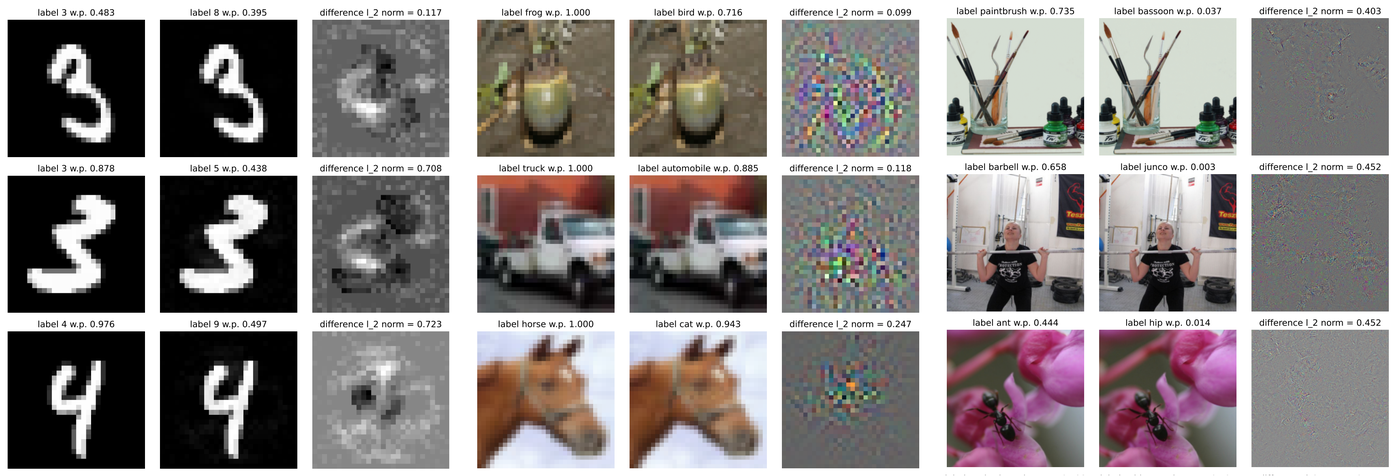

Fig. 3: Adversarial attacks on MNIST test data

Fig. 4: Adversarial attacks on CIFAR test data

Fig. 5: Adversarial attacks on ImageNet test data

Attacks with on/off Manifold Projection

Attacks on Training Data

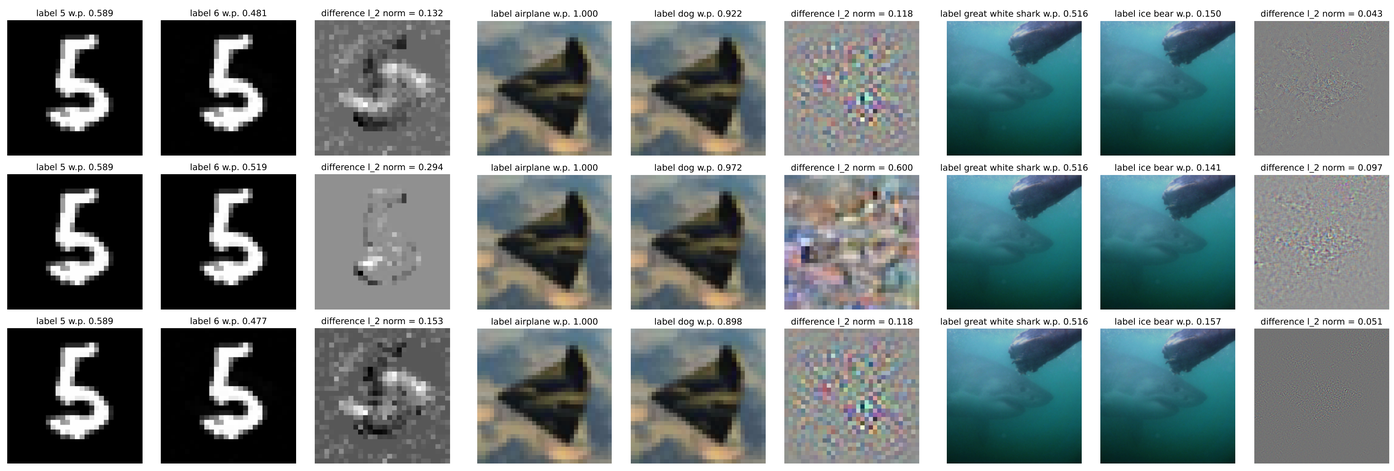

Fig. 6: Adversarial on/off manifold attacks on MNIST training data

Fig. 7: Adversarial on/off manifold attacks on CIFAR training data

Fig. 8: Adversarial on/off manifold attacks on ImageNet training data

Attacks on Test Data

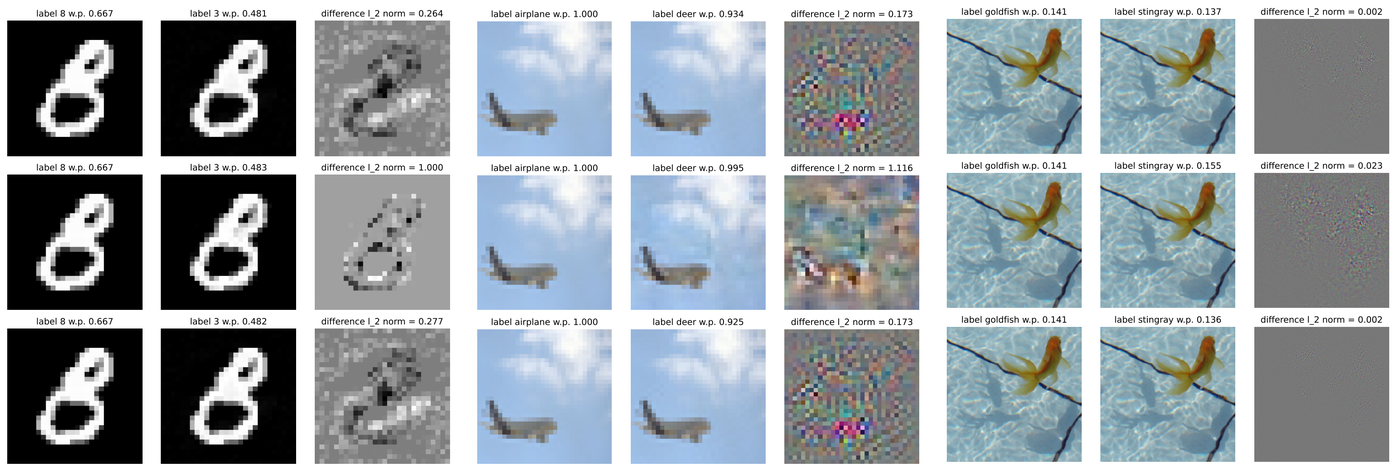

Fig. 9: Adversarial on/off manifold attacks on MNIST test data

Fig. 10: Adversarial on/off manifold attacks on CIFAR test data

Fig. 11: Adversarial on/off manifold attacks on ImageNet test data

Phase 2

Part A - Leveraging Geometric Insights

Existing Defences

Image denoising methods as defences each have their advantages and disadvantages and are still vulnerable to adaptive adversaries.

Denoising Orthogonally?

The DMM says that adversarial examples are orthogonal to the image manifold. $\Rightarrow$ Use orthogonal projection on manifold as defnece!

The Isometric Autoencoder

The Isometric Autoencoder [2] is a regularisation method that enforces an autoencoder to act as orthogonal projection onto the manifold.

Part B - Implementing the New Defence

- Only partially completed in the given time of the project.
- The regularisation of the Isometric Autoencoder could be implemented, but not trained successfully.

References

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

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

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

LaTeX TikZposter