On The Relationship Between Data Manifolds and Adversarial Examples

Michael Geyer [1 2] **Brian Bell** [1 3] **David Glickenstein** [3] **Amanda Fernandez** [2] **Juston Moore** [1]

Abstract

In this work we study adversarial examples in deep neural networks through the lens of a predefined data manifold. By forcing certain geometric properties of this manifold, we are able to analyze the behavior of the learned decision boundaries. It has been shown previously that training to be robust against adversarial attacks produces models with gradients aligned to a small set of principal variations in the data. We demonstrate the converse of this statement; aligning model gradients with a select set of principal variations improves robustness against gradient based adversarial attacks. Our analysis shows that this also makes data more orthogonal to decision boundaries. We conclude that robust training methods make the problem better posed by focusing the model on more important dimensions of variation.

1. Introduction

The concept that robust models have the property of gradients aligned with human perception has been an area of recent research interest in the community (Ganz et al., 2022; Kaur et al., 2019; Shah et al., 2021). We hypothesize that the gradients of a perceptually aligned model are following a continuous manifold of valid images. In this work we are primarily interested in whether this property of *manifold alignment* implies *adversarial robustness*. The question of the converse, whether being robust implies manifold alignment, has been studied previously (Kaur et al., 2019; Ilyas et al., 2019). It has been demonstrated that models which are considered robust share the property that their input gradients are aligned with human perception, and thus the valid image manifold. If there exists a method of training which optimizes for manifold alignment and also provides robustness, it may lead to more explainable or more efficient ways of training robust models.

[*]Equal contribution [1]Los Alamos National Lab [2]University of Texas San Antonio [3]University of Arizona. Correspondence to: Michael Geyer <mgeyer@lanl.gov>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

We study this question by projecting a training set onto a well-known low dimensional manifold. This simplifies the problem of understanding the relationship between data manifolds and a models decision boundary. By forcing this structure, we are able to provide an empirical measure of how aligned a model's gradients are to this manifold.

For our purposes, Manifold Aligned Gradients (MAG) will refer to the property that the gradients of a model with respect to model inputs follow a given data manifold $\mathcal{M}$. This manifold could be known based on the dataset or could be approximated using the tangent space of a generative model. Other works have defined a similar relationship (Shamir et al., 2021), but we choose to take the simplest case by using Principal Component Analysis (PCA). We study this problem on MNIST, as it provides a simple test case while still being non-trivial. This allows us to define a low dimensional structure which is linear.

In this paper we present the following contributions:

1. We apply existing metrics for alignment to a well known data manifold, allowing for empirical measurement.
2. We demonstrate that adversarial training inherently serves to improve our defined metric for manifold alignment.
3. We show that directly optimizing this metric improves robustness against linear attacks.
4. We conclude that while adversarial robustness implies perceptually aligned gradients, the converse is not true for non-linear adversaries.

This is a preliminary work in which we show proof of concept on a simple dataset and technique. It is important to note that these results do not preclude the existence of an optimization method for manifold alignment which provides robustness against arbitrary attacks.

2. Related Work

There is a large body of work attempting to understand the phenomenon of adversarial examples (Akhtar & Mian,

2018). Modeling the relationship between adversarial robustness and perceptual alignment is a step towards improving this understanding. This section discusses the prior work that motivates our research direction.

The sensitivity of convolutional neural networks to imperceptible changes in input has thrown into question the true generalization of these models. Jo & Bengio study the generalization performance of CNNs by transforming natural image statistics (Jo & Bengio, 2017). Similarly to our approach, they create a new dataset with well-known properties to allow the testing of their hypothesis. They show that CNNs focus on high level image statistics rather than human perceptible features. This problem is made worse by the fact that many saliency methods fail basic sanity checks (Adebayo et al., 2018; Kindermans et al., 2019). Until recently, it was unclear whether robustness and manifold alignment were directly linked, as the only method to achieve manifold alignment was adversarial training. Along with the discovery that smoothed classifiers are perceptually aligned, comes the hypothesis that robust models in general share this property (Kaur et al., 2019). This discovery raises the question of whether this relationship is bidirectional.

Khoury & Hadfield-Menell study the geometry of natural images, and create a lower bound for the number of data points required to cover the manifold Khoury & Hadfield-Menell (2018). Unfortunately, they demonstrate that this lower bound is so large as to be intractable. Shamir et al propose using the tangent space of a generative model as an estimation of this manifold (Shamir et al., 2021).

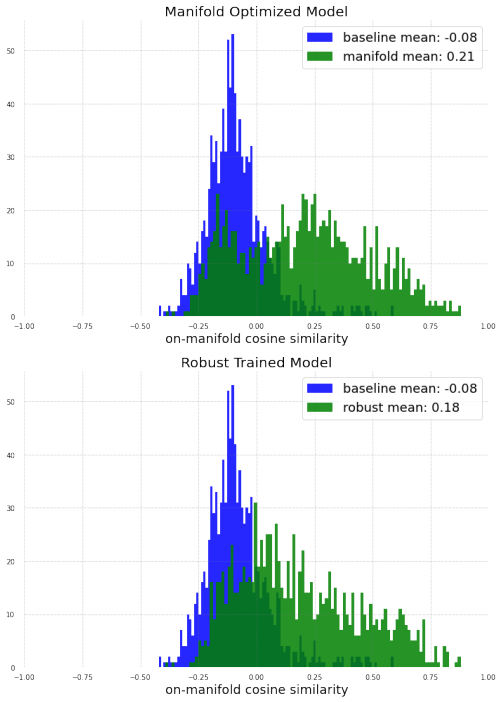

Figure 1. Comparison of on-manifold components between baseline network, robust trained models, and manifold optimized models. Large values indicate higher similarity to the manifold. Both robust and manifold optimized models are more 'on-manifold' than the baseline, with adversarial training being slightly less so.

3. Method

In order to provide an empirical measure of alignment, we first require a well defined image manifold. The task of discovering the true structure of k-dimensional manifolds in $\mathbb{R}^d$ given a set of points sampled on the manifold has been studied previously (Khoury & Hadfield-Menell, 2018). Many algorithms produce solutions which are provably accurate under data density constraints. Unfortunately, these algorithms have difficulty extending to domains with large d due to the curse of dimensionality. Our solution to this fundamental problem is to sidestep it entirely by redefining our dataset. We begin by projecting our data onto a well known low dimensional manifold, which we can then measure with certainty.

3.1. Data

To define our data, we first fit a PCA model on all training data, using k components for each class, where $k << d$. Given the original dataset X, we create a new dataset $X_{\mathcal{M}} := \{x \times \mathbf{W}^T \times \mathbf{W} : x \in X\}$. We will refer to this set of component vectors as $\mathbf{W}$. Because the rank of the linear transformation matrix, k, is defined lower than the dimension of the input space, d, this creates a dataset which lies on a linear subspace of $\mathbb{R}^d$. This subspace is defined by the span of $X \times \mathbf{W}^T$ and any vector in $\mathbb{R}^d$ can be projected onto it. Any data point drawn from $\{z \times \mathbf{W}^T : z \in \mathbb{R}^k\}$ is considered a valid datapoint. This gives us a continuous linear subspace which can be used as a data manifold.

Given that it our goal to study the simplest possible case, we chose MNIST as the dataset to be projected and selected $k = 28$ components. We refer to this new dataset as Projected MNIST (PMNIST). The true rank of PMNIST is lower than that of the original MNIST data, meaning there was information lost in this projection. The remaining information we found is sufficient to achieve 92% accuracy using a baseline Multylayer Perceptron (MLP), and the resulting images retain their semantic properties as shown in Figure 4.

3.2. Measuring the On-Manifold Component

Component vectors extracted from the original dataset are used to project gradient examples onto our pre-defined image manifold. Given a gradient example $\nabla_x = \frac{\partial f_\theta(x,y)}{\partial x}$ where f_θ represents a neural network parameterized by

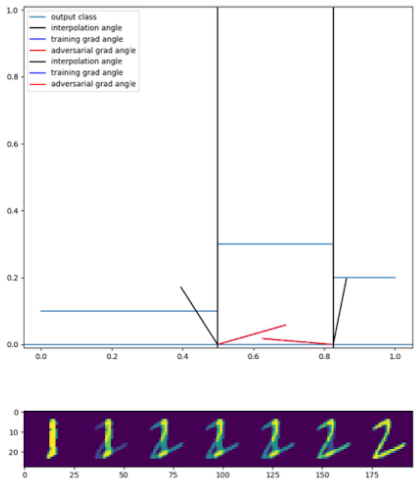

Figure 2. Plot demonstrating the angles at which a linear path interpolation crosses a decision boundary. The output class term is argmaxed, which results in a step functions. Black lines indicate the angle between the interpolation vector and the plane defined by the decision boundary. Crossing askew is a weak support for the dimpled manifold hypothesis presented by (Shamir et al., 2021).

weights θ. ∇_x is transformed using the coefficient vectors $\mathbf{W}$.

$$\rho_x = \nabla_x \times \mathbf{W}^T \times \mathbf{W} \tag{1}$$

The projection of the original vector onto this new transformed vector we will refer to as $P_{\mathcal{M}}$. The norm of this projection gives a metric of manifold alignment.

$$\frac{||\nabla_x||}{||P_{\mathcal{M}}(\nabla_x)||} \tag{2}$$

This gives us a way of measuring the ratio between on-manifold and off-manifold components of the gradient. Additionally, both cosine similarity and the vector rejection were also tested but the norm ratio we found to be the most stable in training. We use this measure as both a metric and a loss, allowing us to optimize the following objective.

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[L(\theta, x, y) + \alpha\frac{||\nabla_x||}{||P_{\mathcal{M}}(\nabla_x)||}\right] \tag{3}$$

Where $L(\theta, x, y)$ represents our classification loss term and α is a hyper parameter determining the weight of the manifold loss term.

4. Experiments

All models were two layer MLPs with 1568 nodes in each hidden layer. The hidden layer size was chosen as twice the input size. This arrangement was chosen to maintain the simplest possible case.

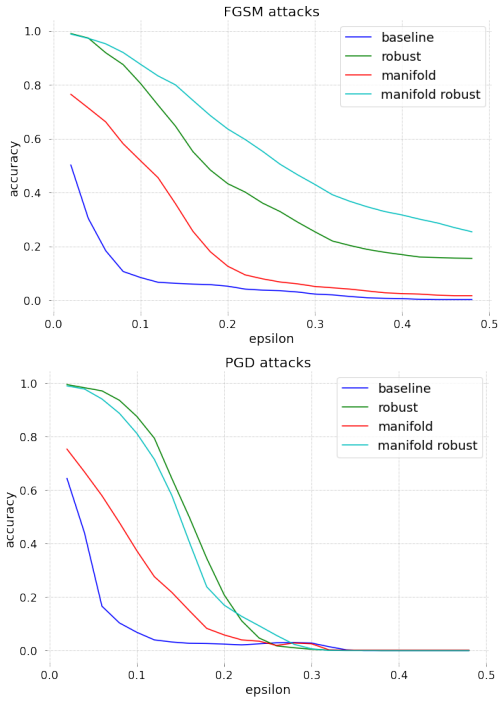

Figure 3. Comparison of adversarial robustness for PMNIST models under various training conditions. For both FGSM and PGD, we see a slight increase in robustness from using manifold optimization. Adversarial training still improves performance significantly more than manifold optimization. Another observation to note is that when both the manifold, and adversarial objective were optimized, increased robustness against FGSM attacks was observed. All robust models were trained using the l_∞ norm at epsilon = 0.1.

Two types of attacks were leveraged in this study: fast gradient sign method (FGSM) (Goodfellow et al., 2014) and projected gradient descent (PGD) (Madry et al., 2017). A total of four models were trained and evaluated on these attacks: Baseline, Robust, Manifold and Manifold Robust. All models, including the baseline, were trained on PMNIST. "Robust" in our case refers to adversarial training. All robust models were trained using the l_∞ norm at $\epsilon = 0.1$. Manifold Robust refers to both optimizing our manifold objective and robust training simultaneously.

Figure 1 shows the cosine similarity on the testing set of PMNIST for both the Manifold model and Robust model. Higher values indicate the model is more aligned with the manifold. Both models here are shown to be more on manifold than the Baseline. This demonstrates that our metric for alignment is being optimized as a consequence of adversarial training.

Figure 3 shows the adversarial robustness of each model. In both cases, aligning the model to the manifold shows an increase in robustness over the baseline. However, we do not consider the performance boost against PGD to be

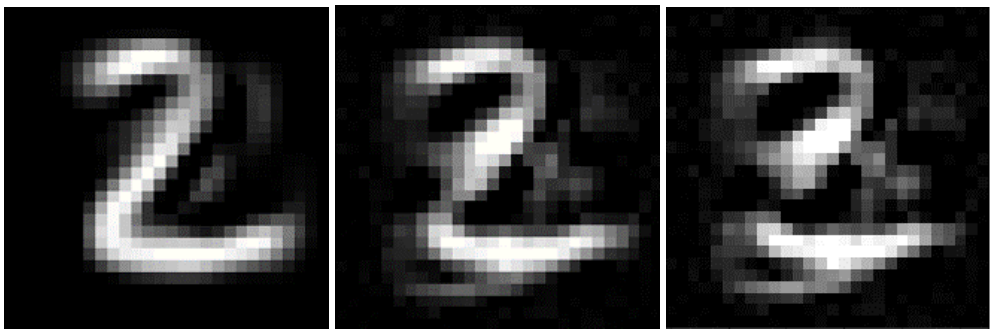

Figure 4. Visual example of manifold optimized model transforming 2 into 3. Original PMNIST image on left, center image is center point between original and attacked, on right is the attacked image. Transformation performed using PGD using the l_∞ norm. Visual evidence of manifold alignment is often subjective and difficult to quantify. This example is provided as a baseline to substantiate our claim that our empirical measurements of alignment are valid.

significant enough to call these models robust against PGD attacks. Another point of interest that while using both our manifold alignment metric and adversarial training, we see an even greater improvement against FGSM attacks. The fact that this performance increase is not shared by PGD training may indicate a relationship between these methods. Our current hypothesis is that a linear representation of the image manifold is sufficient to defend against linear attacks such as FGSM, but cannot defend against a non-linear adversary.

5. Conclusions

Here we present the simplest possible case of our hypothesis that manifold alignment implies adversarial robustness. Extending this to show results on more complex models and datasets is left to future work. In this early work, we only test against a linear manifold and show that it provides robustness against FGSM. We conclude that training a model to be aligned with a low dimensional manifold on which your data lies is related to robust training. While this model shows some properties of adversarial robustness, it is still vulnerable to PGD attacks. Additionally, a model trained to be robust using adversarial training shows manifold alignment under our definition.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Akhtar, N. and Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

Ganz, R., Kawar, B., and Elad, M. Do perceptually aligned gradients imply adversarial robustness? *arXiv preprint arXiv:2207.11378*, 2022.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Jo, J. and Bengio, Y. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

Kaur, S., Cohen, J., and Lipton, Z. C. Are perceptually-aligned gradients a general property of robust classifiers? *arXiv preprint arXiv:1910.08640*, 2019.

Khoury, M. and Hadfield-Menell, D. On the geometry of adversarial examples. *arXiv preprint arXiv:1811.00525*, 2018.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, 2019.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Shah, H., Jain, P., and Netrapalli, P. Do input gradients highlight discriminative features? *Advances in Neural Information Processing Systems*, 34:2046–2059, 2021.

Shamir, A., Melamed, O., and BenShmuel, O. The dimpled manifold model of adversarial examples in machine learning. *arXiv preprint arXiv:2106.10151*, 2021.