Experimenting with Adversarial Robustness: Guiding Neural Nets to Learn Human-Centric Features Through Creating a Minimally Identifiable Dataset

Anonymous Author(s)
Affiliation
Address
email

Abstract

We seek to construct a "minimally identifiable dataset" (a dataset which a human is only barely able to classify) to explore whether embedding "human-centric" robust features as part of the data creation process can aid in adversarial defense.

1 Introduction

The advance of machine learning models has significantly improved computer vision tasks such as image classification. However, it was quickly discovered that simple adversarial attacks by perturbing an image can "fool" a neural network into predicting an incorrect label [7]. To a human, such perturbations can be indiscernible, or just simply regarded as noise. For several years, many defenses have been proposed but all beaten down by various attacks in quick order. As an example, a team published a paper recently in which they defeated many proposed defences published in various conferences [8].

In 2019, Andrew Ilyas et. al. published a paper in which they divide the features of a data set into "*robust features*" and "*non-robust features*" [4]. The authors define "non-robust features" as features that are "highly predictive, yet brittle and (thus) incomprehensible to humans" [4]. Examples of non-robust features include noisy patterns, minor perturbations and single pixel modifications [6]. The authors then claim that classifiers *do* rely on these non-robust features and as a result, this dependence will increase the adversarial vulnerability of the model [4].

To verify and strengthen Ilyas et. al's idea, we propose to further explore how to teach a machine what "human-centric" robust features are. We note that one could likely take a modified image from some commonly used datasets (such as MNIST), significantly reduce its dimensions, and still be able to correctly classify it. Since machine learning models are agnostic to human preferences unless it is trained to recognize them [4], we hypothesize that any level of information beyond what is minimally necessary for a human to correctly classify the image can result in an attack vector for the adversary, since non-robust features can be exploited to make "buggy" predictions [4].

In this research, we seek to modify an existing dataset so that the common Networks trained using the modified dataset are more robust to adversarial attacks.

2 Related Works

Ilyas et. al. note that robust features are correlated with the label in spite of adversarial attacks [4]. For example, a robust feature of the hand-written digit "1" is a straight stroke downwards. Non-robust features are highly predictive, yet imperceptible to humans [4]. But what humans regard as legitimate features *and* what humans regard as noise can both help the classifier improve accuracy during

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

training. The authors then conduct an experiment that use only "useful" and "robust" features from the penultimate layer of their deep neural network and construct a new training dataset on which they trained a classifier, and achieve improvements in defending adversarial attacks. The authors showed that they could improve the robustness of a dataset by removing the non-robust features.

In 2016, S. Dodge and L. Karam published a paper "Understanding How Image Quality Affects Deep Neural Networks" [2], in which they provided an evaluation of four state-of-the-art deep neural network models for image classification under five quality distortions (blur, noise, contrast, JPEG, and JPEG2000 compression) and found that the classification accuracy of the networks *can* be significantly affected by these distortions, especially by blur and noise. The authors showed that the reduced performance under low quality images is common over existing classifier models [2]. Based on the fact that humans, in most of cases, can identify low quality images correctly, we believe that classifiers are trained to partly rely on imperceptible features.

3 Methodology

3.1 Hypothesis

We hypothesize that an image classifier with a given architecture is more robust to adversarial attack when trained using a dataset that is minimally human-identifiable. That is, the removal of detailed features that do not affect a human's ability to recognize said images improve robustness of the trained model.

3.2 Data Preparation Methodology

We used the MNIST dataset and ran various combinations of downscaling and upscaling, blurs, and contrast changes. After visually inspecting the results, we selected a sequence of image distortions that remove as much unnecessary detail as possible without severely compromising the ability of a human to label the resulting data.

The final distortions applied were the following in sequence: 1) a downscaling of the MNIST dataset with bilinear interpolation of the image from 28×28 to 10×10; 2) an upscaling of the resulting 10×10 images back to 28×28 but with nearest interpolation of the pixels to preserve the blocky nature of the 10×10 images; 3) a four fold increase in the contrast of the image.

5,000 images of the MNIST training dataset were manually relabelled by looking at only the distorted images. Of the 5,000 relabelled images, 4,736 were correctly labelled (compared with the ground-truth label of the original MNIST image). These 4,736 images were used as a training dataset (called "modified MNIST"). There was no change made to the MNIST test set for any of the experiments.

After the distortion, some images retained recognizable features such that they were easily distinguishable from others (See Figure 1), while for other images, a few crucial features were lost such that the image's correct category became ambiguous for the labeller (See Figure 2). But even among those images whose correct cateogry became ambiguous, for the most part, it was clear that there were only a few viable choices. By training with the modified images that were correctly labelled (where correct is defined vis-à-vis the original labels), images that did not contain human-recognizable features were implicitly discarded.

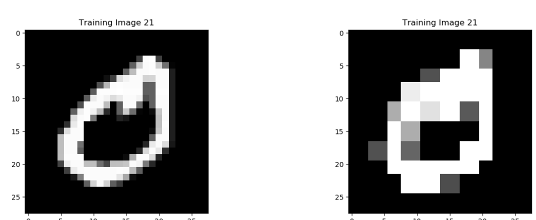

Figure 1: Recognizable features preserved: Original (Left) and Modified Image (Right)

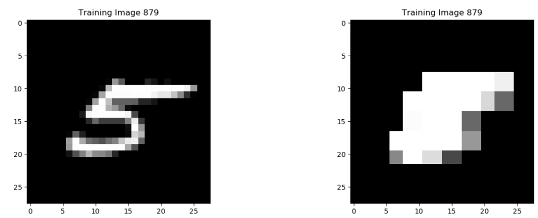

Figure 2: Most recognizable features removed: Original (Left) and Modified Image (Right)

4 Experiments

We trained a CNN image classifier and a LeNet image classifier each using 1) the raw MNIST dataset and 2) the modified MNIST dataset. The architecture of the CNN used is shown below and the LeNet follows the design of LeNet-5 [5].

Table 1: CNN Architechure

Layer	Detail
conv1	Conv2d(1, 32, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
conv2	Conv2d(32, 64, kerne_size=(5, 5), stride=(1, 1), padding=(2, 2))
fc1	Linear(in_features=3136, out_features=1024, bias=True)
fc2	Linear(in_features=1024, out_features=10, bias=True)

In addition, starting with the classifiers trained on the full MNIST dataset, we trained a third model using the modified MNIST dataset by freezing all but the last layers. This is termed "MNIST + Modified MNIST" in the table below. On the trained models, two adversarial attacks were implemented and tested against the models: the Carlini Wagner attack [1] and the Fast Gradient Sign Method attack (FGSM) [3]. The table below lists the architecture, training data used and attack method along with the test accuracy under the attack method.

Table 2: Experimental Results

Architecture	Training Data	Attack Method	Test Accuracy
CNN	MNIST	None	99.36%
CNN	MNIST	FGSM	6.49%
CNN	MNIST	CW	52.00%
CNN	Modified MNIST	None	94.30%
CNN	Modified MNIST	FGSM	19.40%
CNN	Modified MNIST	CW	68.00%
CNN	MNIST + Modified MNIST	None	98.80%
CNN	MNIST + Modified MNIST	FGSM	19.40%
CNN	MNIST + Modified MNIST	CW	73.00%
LeNet	MNIST	None	99.13%
LeNet	MNIST	FGSM	27.20%
LeNet	MNIST	CW	30.00%
LeNet	Modified MNIST	None	90.93%
LeNet	Modified MNIST	FGSM	31.70%
LeNet	Modified MNIST	CW	43.00%
LeNet	MNIST + Modified MNIST	None	98.37%
LeNet	MNIST + Modified MNIST	FGSM	30.40%
LeNet	MNIST + Modified MNIST	CW	47.00%

5 Discussion and Analysis

The experimental results above indicate that test accuracy is higher for both architectures if either the Modified MNIST dataset was used in training directly, or if it was used to retrain the final layer. We can also see from Figure 3 that in general, the models trained using the Modified MNIST dataset are more robust to tampering. A higher number of pixels need to be distorted in order to reduce the accuracy of the models trained using the Modified MNIST dataset.

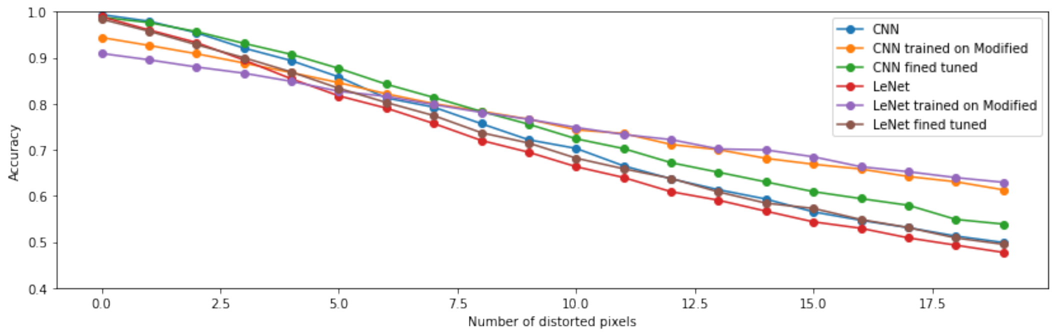

Figure 3: Robustness of Model: Number of Distorted Pixels vs. Test Accuracy

However, from the image outputs of successful attacks, the modifications to images that are able to fool the image classifiers do not visibly change based on training data. Taking an example from the Modified MNIST CNN model under FGSM attack, we see that a slight modification to pixels that are not related to the number 9 (see Figure below) causes the model to misclassify the image as a 4. We can see that the adversarial vulnerability persists.

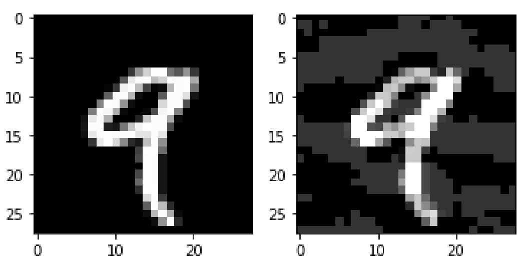

Figure 4: Successful attack (Right) on CNN trained using Modified-MNIST dataset causing model to misclassify 9 as 4

The higher test accuracy using the modified MNIST dataset provides some evidence of Ilyas et. al.'s claim that robustness is not only a property of the architecture, but also a property of the data itself [4]. It also implies that given a target test accuracy, the perturbations of the images used to attack models trained with the modified MNIST dataset must be larger. In practice, this improves the ability to detect malicious attacks on the model.

6 Conclusion

Our experiments conclude that for simple images such as those found in MNIST, there is some empirical evidence to support that the models trained using "minimally identifiable datasets" are more robust to FGSM and CW attacks. However, the degree of the improved robustness is moderate, and further experiments with larger datasets and stronger distortions are required to test the strength of the effect. Furthermore, removing "non-robust features" is a simple task with the MNIST dataset due to its solid colour background; it is more difficult in practice to carry this experiment out for more complex images.

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017.

[2] Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks, 2016.

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[4] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features, 2019.

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[6] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, Oct 2019.

[7] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

[8] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses, 2020.