1 Idea

Increase robustness of deep convolutional image classifier against adversarial examples [3] by augmenting the training set using interpolated images created by a Generative Adversarial Network (GAN) [2].

2 Procedure

First we randomly sample two images of a training dataset and find the two latent vectors for which GAN's generator creates images reasonably close to the two original images. We can then average those two latent vectors using different weights (such as 0.1 and 0.9) and input that to the generator to generate a new interpolated image. The label will be automatically generated using the weights. After training the classifier using this new data, we check if it is now more robust towards adversarial attacks. This is to some extent inspired by mixup [8][9].

3 Methods

Finding corresponding latent vectors can be done using gradient descent with the image itself as target [6] or feature vectors extracted by a (pretrained) model [4]. Latent Vectors could be cached for reuse. Robustness can be tested using DeepFool [7] or a similar tool.

4 Datasets

We will try our method on MNIST, fashion-MNIST and CIFAR-10. This should let us test the method without running into hardware constraints.

5 Baselines

- To get a baseline, we will first create a Convolutional Neural Network (CNN) and test how robust it is against adversarial attacks. We will then train a gan to augment our dataset to then train the before mentioned CNN using the augmented dataset and compare the two CNNs.
- We will test if using a GAN augmented dataset improves robustness over the simple technique of blurring the two images into one.
- We will test if using a VAE instead of a GAN improves results.
- We will compare our method to some other adversarial training methods [1][5].
- Besides comparing robustness, we will also test if this method improves the accuracy of the classifier.

References

[1] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness, 2021.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

[4] Arxiv Insights. Editing faces using artificial intelligence.

[5] Hyeungill Lee, Sungyeob Han, and Jungwoo Lee. Generative adversarial trainer: Defense to adversarial perturbations with gan, 2017.

[6] Zachary C. Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks, 2017.

[7] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states, 2018.

[9] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2017.