Project Report
Lower bound of perturbation for FGSM attack

Artur Pak

November 2023

Abstract

One of parameters for generation of adversarial examples is ϵ, which defines the upper bound for infinite norm of perturbations. This project establishes lower bound for ϵ for a successful evasion attack in a particular setting. This bound is a linear function of margins of logits in the output of the classifier. See source code in appendix.

Keywords: Machine Learning, Adversarial Attacks, FGSM

1 Introduction

Today machine learning algorithms are used daily in a wide range of human activities. As the time goes by, we rely even more on them - they become our tools in both routine and highly important tasks. However, these tools are vulnerable to **adversarial attacks** - methods which aim to fool the algorithms or extract sensitive information from them. **Adversarial ML** - a branch of machine learning - studies these attacks and devises means to increase robustness and security of our tools.

Generally adversarial attacks can be divided into three types: poisoning, extraction and evasion attacks. The latter one receives a lot of attention due to relevance and success rate of these attacks - they aim to alter input data in a way, that will fool the algorithms; perturbed datapoints are called **adversarial examples**. Methods of producing these perturbations can be further divided into white- and black-box methods. White-box methods require access to the classification model, which is a very unlikely condition in real-life situations. This is why most of the recent practical advances are connected to black-box methods and ways to protect against them [3]. White-box methods are easier to operate and understand, so this project is dealing with one of them.

In context of image classification **adversarial examples** - are specifically perturbed input data for classification problems that make the algorithm do wrong predictions, while being barely noticeable to human eye. One intuitive method to create adversarial examples is **Fast Gradient Sign Method** (FGSM) [1]. As an input, it takes 4 parameters:

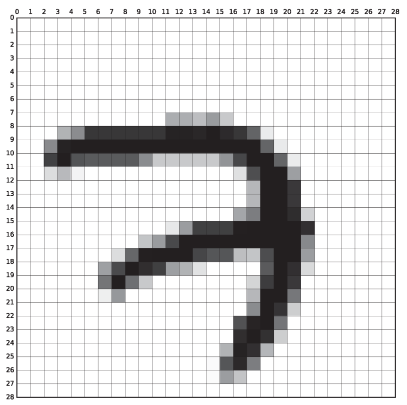
- classification model - a model which we intend to fool
- image - input data for the classification model which the algorithm will perturb
- ϵ - an upper bound for the perturbation on each parameter
- target class (optional) - a class which we want the image to be classified as after perturbations

2 Motivation

Evasion attacks want to create perturbations as low as possible to not be detected, yet large enough to ensure success of the attack. Balancing these two factors may pose a challenge when choosing ϵ. This is why **predicting a lowest possible ϵ for a successful attack could be useful**. This project shows that margins from the output layer of the classifier may be used to find the lower bound for the optimal ϵ.

3 Experiment description

Dataset


(a) MNIST sample belonging to the digit '7'.

(b) 100 samples from the MNIST training set.

Figure 1: Samples from MNIST dataset

For these experiments I chose MNIST dataset. The dataset contains gray-scale images of handwritten digits in 28×28 resolution. Training set contains 60 000 samples, and test set contains 10 000 samples.

3.1 Classifier

Classifier in this experiment is a simple 3-layer ReLU neural network. It has an input layer, two 64-node layers, and an output layers with 10 nodes (one for each class). The network was trained with Adam optimizer, over 3 epochs. Test accuracy is 96.8%.

3.2 Adversarial Attack

Targeted FGSM attack was implemented by the following algorithm:

Algorithm 1 Fast Gradient Sign Method

```
    Input: model, img, target_idx, $\epsilon$, steps
    Output: perturbed_img
 1: $\Delta \leftarrow$ random noise
 2: cur_idx $\leftarrow$ model.predict(img)
 3: perturbed_img $\leftarrow$ img + $\Delta$
 4: for $i = 1$ to $steps$ do
 5:     logits $\leftarrow$ model.forward(img)
 6:     loss $\leftarrow$ nll_loss(logits, target_idx) − nll_loss(logits, cur_idx)
 7:     optimize $\Delta$ w.r.t. loss
 8:     $\Delta \leftarrow max(\Delta,\ \epsilon)$
 9:     perturbed_img $\leftarrow$ img + $\Delta$
10: end for
```

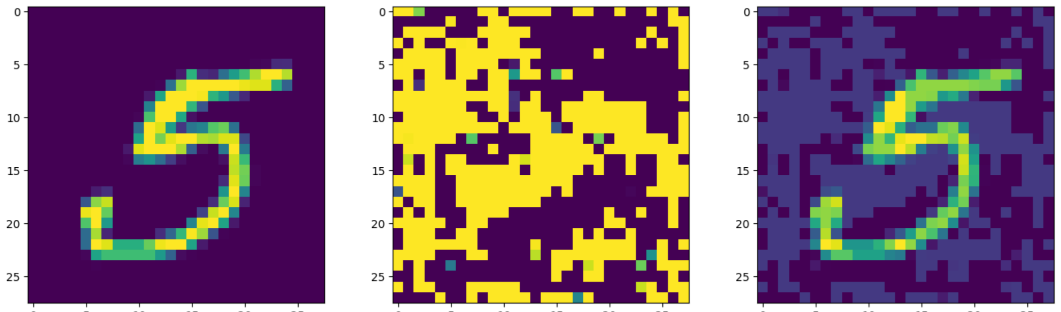
3.3 Experiment

In the experiment, $m = 10$ images from each class are selected randomly. For each image the following operations are performed:

1. feed the image to the classifier and get logits
2. calculate margins (δ) as difference between the highest logit and the rest of the logits
3. for each class use binary search on the interval $[0, 1]$ to find least ϵ that results in successful attack
4. record corresponding (δ, ϵ) pairs

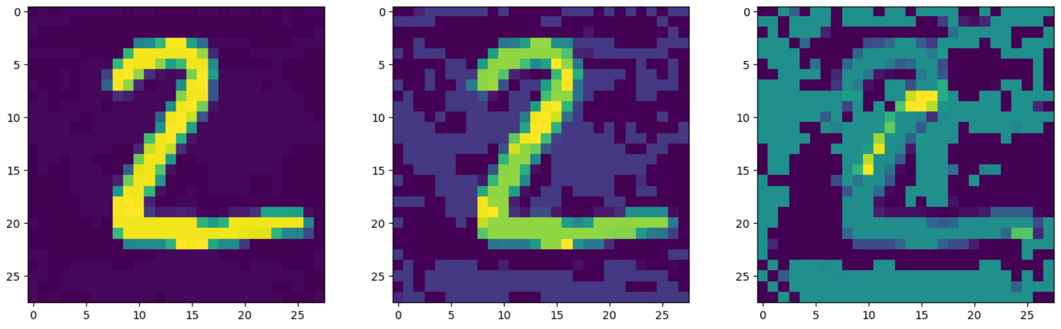
4 Results

The following figures display adversarial examples:



Base | Pred: 5 $\quad$ Δ | Target: 8 $\quad$ Perturbed | Pred: 8

Figure 2: Illustration of base image, noise, and perturbed image after FGSM attack with $\epsilon = 0.1$



$\epsilon = 0.01$ | Pred: 2 $\quad$ $\epsilon = 0.1$ | Pred: 3 $\quad$ $\epsilon = 0.5$ | Pred: 7

Figure 3: Images of 2 with different extents of perturbation

As can be seen form the figures, implemented FGSM can create perturbations to make the neural network make bad prediction. However, in several cases it may require a higher extent of perturbation to be classified as an intended class. Moreover, in some instances $\epsilon = 1$ is not enough to make the attack successful. In other cases binary search gives estimation of ϵ with precision up to 2^{-10}.

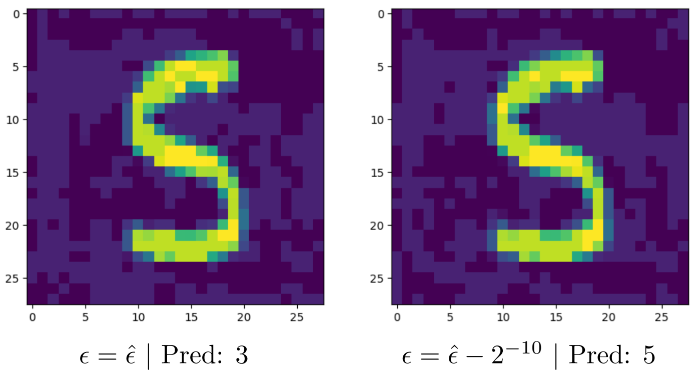


Figure 4: Image of 5 with perturbations. $\hat{\epsilon}$ is found by binary search.

It is reasonable, that if margin is 0, the optimal ϵ is 0 too. So, if we assume that the bound f is indeed a linear function of margin δ, it will look like $f(\delta) = C\delta$. Hence we may approximate C by finding (δ, ϵ) pair with the lowest ratio between the coordinates. In this experiment it is 0.00204801105925972. As can be seen on the figure, other points follow this bound as well.

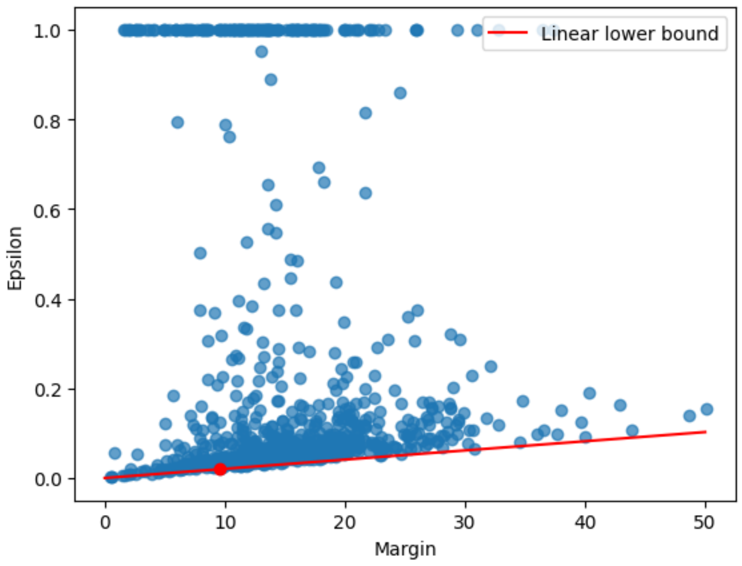


Figure 5: (ϵ, δ) pairs. Red line indicates the lower bound for optimal ϵ.

5 Conclusion

As observed from the experimental results, lower bound of optimal value for the ϵ can be inferred from the margins δ of the output of the classifier. In the setting examined in this project, the lower bound is a linear function $\epsilon \geq f(\delta) = C\delta$.

6 Future work

This study can be improved and extended in the following ways:

1. Explore other setups, with other datasets and classifiers. This may lead to other relations between margins and epsilons, or confirm linear realtion.
2. Extend concept of ”extent of perturbation” to other adversarial attacks, likely black-box methods. Explore possibility of inference of optimal perturbations from margins there.
3. Find theoretical derivations for the lower bound, that will explain empirical results obtained in this project.
4. Find ways to infer a tighter bound for the optimal ϵ by other means. One possible way could be inference of ϵ via datamodeling framework.

   As Ilyas et. al have already shown [2], there is a possibility for a model which can predict an outcome of a model based on a training set. If we define the outcome as ”the least sufficient ϵ for a successful FGSM attack”, results of Ilyas et. al suggest that the model in question may exist.

References

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples.* 2015. arXiv: `1412.6572 [stat.ML]`.

[2] Andrew Ilyas et al. *Datamodels: Predicting Predictions from Training Data.* 2022. arXiv: `2202.00622 [stat.ML]`.

[3] Oliver Smith and Anderson Brown. “Comprehensive Review on Advanced Adversarial Attack and Defense Strategies in Deep Neural Network”. In: *International Journal of Research in Intelligent and Advanced Systems* 8.4 (2023). Received: 06 March 2023; Accepted: 16 March 2023; Published: 5 May 2023. DOI: `10.51584/IJRIAS.2023.8418`.

Appendix

Part a: link to the github with the code

https://github.com/usaginoki/pytorch_mnist_fgsm_pet_project