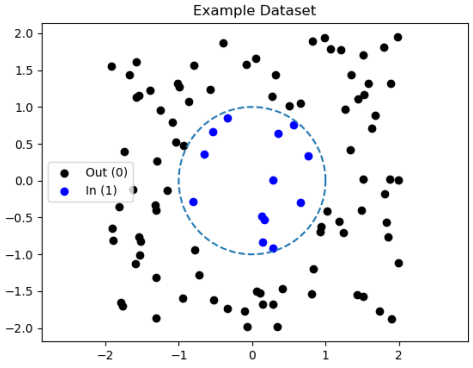
Assignment 1: PGD

Feb 10, 2025

Part 1: Theoretical example

You are given a dataset consisting of points within a square centered at the origin with a side length of 4. Points inside a circle of radius 1 are labeled 1, and points outside this circle are labeled 0.



You want to train a model to learn to determine whether a given point lies within the circle boundary. You decide to train a fully connected neural network with the following structure: an input layer with 2 nodes (for the X and Y coordinate of a point), two hidden layers with 100 nodes each, and an output layer with 2 nodes.

Suppose an attacker wants to use a Projected Gradient Descent (PGD) attack to trick the model. For given points, the attacker generates adversarial points by perturbing their position with respect to the constraint parameter ϵ, such that the model would classify these points to be the opposite side of the decision boundary, contrary to their true label. The L_2 norm is used for this exercise.

PGD: Projected Gradient Descent

- Use **constrained optimization** to find adversarial examples
  - Recall: $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^d$, such that $\mathcal{A}(x) \to \mathbf{x}'$, $\mathbf{f}(\mathbf{x}) = t, \delta(x', x) \leq \epsilon$
  - Let's use L_∞
- Projected Gradient Descent (PGD)
  - Initialize $\mathbf{x}^1 \leftarrow \mathbf{x}$ and $\eta \leftarrow \frac{\epsilon}{S}$
  - For $i = 1, \ldots, S$
    - Compute $\mathbf{x}^{i+1} \leftarrow \mathbf{\Pi}_C(\mathbf{x}^i - \eta \cdot \text{sign}(\nabla \mathcal{L}(\mathbf{f}, \mathbf{x}, \mathbf{c})))$, $\mathbf{\Pi}_C(\cdot)$ is a projection operator
  - Return x^S

Question 1: (1 pt)

- For the point (0.2, 0.8), what is the minimum epsilon value required to cause a misclassification? What about the point (-0.6, 0.6)?

Question 2: (1 pt)

- What is the minimum epsilon value required for every point within the circle to be misclassified?
- Use the minimum epsilon value computed just now on the point (0.2, 0.8) to generate an adversarial example. What is the maximum possible distance between such an adversarial example and the circle boundary?

Part 2: Coding assignment

In this assignment, we will launch the Projected Gradient Descent (PGD) attack on a realistic dataset.

Introduction

By perturbing images, we can create adversarial images that will cause the image classification models to underperform. The perturbations would be negligible to human perception but enough to confuse a neural network. This phenomenon highlights one of the many limitations of machine-learning models in comparison to the human brain.

We will conduct a PGD attack on the Resnet-50 model [HZRS15] pre-trained on an image classification task on the ImageNet-1k dataset. This dataset spans 1000 object classes and the training set contains approximately 1000 images for each class. All input RGB images are resized to 232 x 232, and the center 224 x 224 is cropped out, and then normalized for the mean and standard deviation of the ImageNet dataset. Hence the Resnet model takes 3 x 224 x 224 tensors as input and outputs a 1000-dimension vector. After applying the softmax function on the output vector, we can interpret the resulting vector as a probability vector over the 1000 possible classes. One may take the arg max as the model's predicted class or look at the top k classes for a more lenient evaluation metric.

$$f\colon [0,1)^{3\times 224\times 224} \to \mathbb{R}^{1000}$$

In particular, we will be using L_∞ norm for this exercise.

Environment Setup

Environment Setup

Make sure you are in the environment folder, then enter these commands in the terminal.

1. `conda create -n mls_assignment1 python=3.11.0`
2. `conda activate mls_assignment1`
3. `pip install -r requirements.txt`

Once your environment is set up, run the following command to verify that you have installed all required Python modules and files:

```
python launch_resnet_attack.py --test --batch_num 1 \
    --batch_size 10 --results 'test_launch'
```

For more references on conda environments, refer to Conda Managing Environments or the Conda Cheat Sheet.

Data Loading Instructions

The ImageNet dataset is available on Huggingface, but it is a gated dataset and you need a Huggingface account to access it (click here to create an account)[1]. Since it is an extremely large dataset, you may utilize streaming mode to avoid downloading large files to your computer, and the preprocessing procedure can be applied on the fly during the attack loop. This may lead to longer run times but significantly reduces the amount of storage required.

Once you have an account, head to the ILSVRC/imagenet-1k dataset (click here[2]) and select "accept conditions" to access its files and content. Then go to your account settings to create a user authentication token. Make sure to add ILSVRC/imagenet-1k to the list of readable datasets. In your command line interface, enter `huggingface-cli login` and paste your user token. Now you can freely access this dataset when executing the Python scripts from the command line.

If the Hugging Face CLI does not work, you can authenticate programmatically in your Python script using:

```
from huggingface_hub import login
login(token="your_hf_token")
```

This method allows authentication when running your script but should only be used if the CLI login is unavailable.

Please also make sure you have reproducible code for verification purposes.

[1] huggingface.co/join
[2] huggingface.co/datasets/ILSVRC/imagenet-1k

Question 3: Generate Adversarial Images (8pts)

In this task, you are asked to complete the relevant parts of `resnet_attack_todo.py` marked with `TODO` and generate adversarial images.

You are given an entry script `launch_resnet_attack.py` that provides a pre-trained model and a data loader, where the input data is loaded and pre-processed. Run the script with the line argument `--results 'adv_images'` (this name is an example, feel free to use your own informative filename for your assignment) to specify the filename to which the resulting adversarial images and relevant data are saved. You may also specify the batch size and the number of batches considering the amount of computer storage you have access to.

When executed, the script calls `ResnetPGDAttack.pgd_batch_attack()` specified in the file `resnet_attack_todo.py`, where the PGD attack is partially implemented. As is, the script will return the original images and save them under the `results` directory. Your task is to complete the implementation of `pgd_attack` to perform the gradient update step and the L_∞ norm projection step in each iteration of the PGD attack, so that the script will return adversarial images as desired. The unimplemented code sections are marked by `TODO` comments.

After executing the entry script, your directory structure should look something like this:

```
pgdattack
├── resnet_attack_todo.py
├── launch_resnet_attack.py
└── results
    └── adv_images
```

Evaluation

Make a plot of various epsilon values (at least three values between 0 and 0.1) vs accuracy on adversarial images. Clearly label the hyper-parameters used, such as the batch number, batch size, alpha value. See Figure 1 for an example[3]. **What trend do you observe?** You are graded on the accuracy of both your plot and your code completion for the `resnet_attack_todo.py` file.

Deliverables

For this question, submit the completed `resnet_attack_todo.py` file, as well as the files generated by running `launch_resnet_attack.py` as described above,

[3]The trend in the example plot is randomly generated and not representative of the expected result.

for three different epsilon values. The files will be saved within the `results` folder.

The plots you generate should be in the same PDF file used for the previous two theory questions. Along with the plots, explicitly specify the attacker configurations you used (3 epsilon values, alpha and number of steps value).

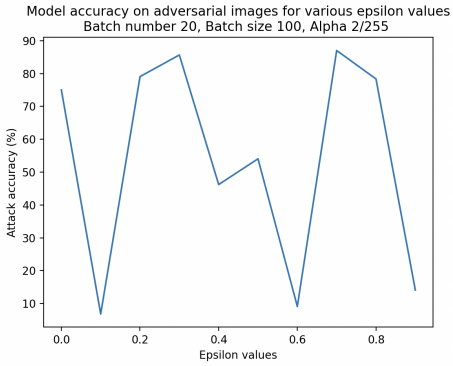


Figure 1: Example plot for Question 3

Question 4: Adversarial Training (5pts)

For this task, you are asked to finetune the Resnet model by feeding the adversarial images you just created.

By feeding adversarial images as extra inputs, the model learns to resist the PGD attack and predict correctly even when perturbed images are presented as inputs. Therefore, adversarial images can be used to improve model robustness[4]. Don't forget to clearly document your code!

Evaluation

Your model will be assessed against 2000 hidden adversarial images (i.e. test cases). You pass the test if your model achieves an accuracy above 50% on the hidden adversarial images. You excel in the test if it achieves an accuracy above 80%. You are graded on both the correctness of your code and your model's test performance.

Deliverables

For this question, submit the Python file with your code to fine-tune the Resnet model, and the fine-tuned model. To save your fine-tuned model, please save only its `state_dict()`[5].

Include the following command at the end of your fine-tuning, to save the model:

```
    torch.save({
    'model_state_dict': model.state_dict(),
}, "../fine_tuned_resnet")
```

Submission instructions

Upload the following:

1. A PDF file with answers to questions 1 and 2, and plots for question 3.
2. A ZIP file containing the following files
   (a) (Q3) The completed `resnet_attack_todo.py`
   (b) (Q3) `results` folder containing files generated by running `launch_resnet_attack.py` (One file for each attacker configuration used).
   (c) (Q4) Python file for fine-tuning Resnet.
   (d) (Q4) The fine-tuned Resnet model saved as `fine_tuned_resnet`

Please follow the instructions carefully to ensure the grading script functions correctly.

[4] The standard adversarial training procedure is actually more complex than this; it is a max-min optimization problem. If you are interested, see the original paper for more details [MMS+19].

[5] https://pytorch.org/tutorials/beginner/saving_loading_models.html

References

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[MMS+19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.