This is the codebase of the GM loss.

Environment settings

The codebase is tested under the following environment settings:

- cuda: 11.0
- python: 3.8.10
- pytorch: 1.7.1
- torchvision: 0.8.2
- scikit-learn: 1.0.1

MNIST

The code uses the network of 6 convolutional layers and a linear layer with 2-D / 100-D output.

2-D

To train the 2-D model with GM loss, please run the following code inside the ./mnist/ folder. The training procedure needs around 20 minutes on a Titan X gpu.

```
CUDA_VISIBLE_DEVICES=0 python mnist_release.py --loss gm --lr 0.01 --edim 2
```

In the default setting, α is set to be 1.0 and λ is set to be 0.1.

The best accuracy is around 99.29% for GM loss.

After running the training code, the feature distribution on the testing set will be shown in ./mnist/mnist_gm_test.png.

If you want to train with the softmax loss, please run the following code inside the ./mnist/ folder. You will get the best accuracy of around 99.02%.

```
CUDA_VISIBLE_DEVICES=0 python mnist_release.py --loss softmax --lr 0.01 --edim 2
```

100-D

To train the 100-D model with GM loss, please run the following code inside the ./mnist/ folder. The training procedure needs around 20 minutes on a Titan X gpu.

```
CUDA_VISIBLE_DEVICES=0 python mnist_release.py --loss gm --lr 0.02 --edim 100
```

In the default setting, α is set to be 1.0 and λ is set to be 0.1.

The best accuracy is around 99.70% for GM loss.

See Table 1 in the paper for reference.

CIFAR10

The code uses the ResNet20 network.

To train the model with GM loss, please run the following code inside the ./cifar/ folder. The training procedure needs around 70 minutes on a Titan X gpu.

```
CUDA_VISIBLE_DEVICES=0 python cifar10_release.py --loss gm
```

In the default setting, α is set to be 0.3 and λ is set to be 0.1.

The best accuracy is around 92.80% for GM loss.

See Table 2 in the paper for reference.

Kernel Density

The code implements the computation of the detection AUC-ROC score for semi white-box FGSM attack on a ResNet50 pretrained on the ImageNet dataset with Softmax+KD detection. The AUC-ROC score should be around 74%.

The code is tested on 4 Titan X gpus.

Please download the ImageNet training and validation sets and extract them to [ImageNet Dataset Folder].

To obtain the AUC-ROC score, simply run the following command in the ./KD/ folder.

```
CUDA_VISIBLE_DEVICES=0,1,2,3 python ROC.py --data [ImageNet Dataset Folder]
```