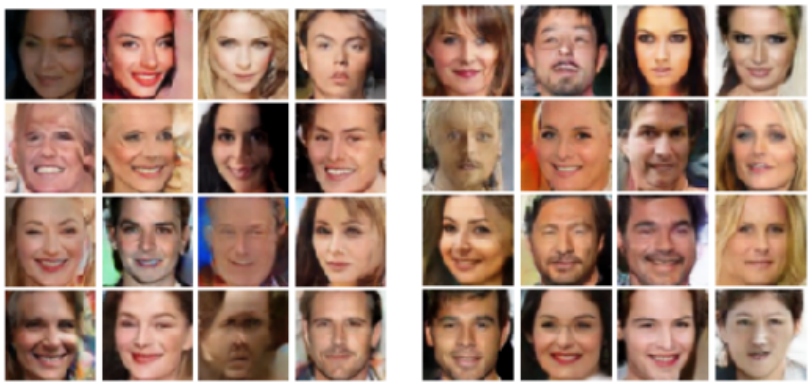
Assignment 3 - CS747

Hoang-Dung Bui
NetID: G01301478

October 31, 2021

1 Part 1

The result GAN models with two loss functions: GAN loss and Least square GAN loss.



(a) Output of GAN (b) Output of LSGAN

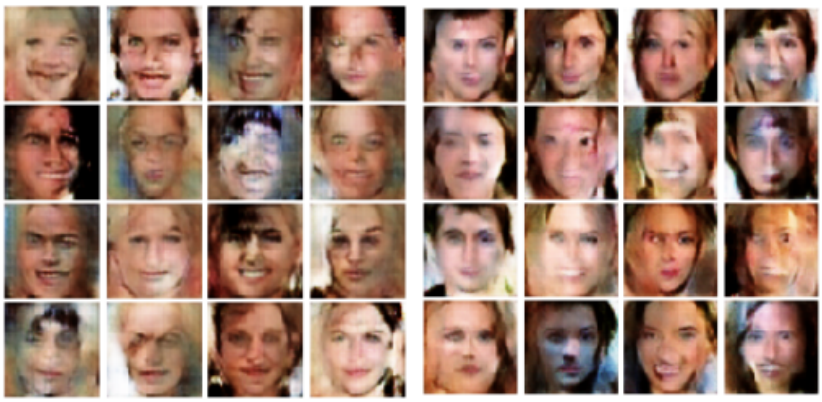
Figure 1: The outputs of GAN and LSGAN

As comparing the two results with the same epoch (5) and iteration (49200):

- The GAN with the least square GAN loss function provided a better result. It converged faster than regular GAN, and the finial outputted pictures are clear and similar to human faces.
- The loss values for Discriminator and Generator networks of LS GAN were: 0.207 and 0.3321. Meanwhile, the values for the GAN were only 0.6409 and 0.7069
- During training, there were many times the instances of model collapse in the GAN training. In the cases, we can recognize a human face in the outputted pictures. For GAN, at iteration 5550, it is difficult to recognize the human face in the image in second row, third column. For LSGAN, on the first several iterations, we could not also recognize the faces. After 10000 iterations, the collapse phenomenon reduced significantly, especially for LSGAN model.
- The spectral normalization has a significant influence to improve the image quality in the GAN network. It normalized the weight in each layers in the GAN,

Extra credits: I built new models which can handle the original images with dimensions of 128x128. The idea was I still keep the same input dimension for the generator, and adding one more convolutional transpose layer to its model. It will help to extend the image back to size of 128x128. The kernel size is 4, and the stride is 2, and padding is 1. In similar way, we also added a convolutional layer to the original model with the kernel size is 4, stride = 2 and padding = 1. Thus, the convolutional layers still outputted the image's dimension of 4x4. The results of the two layers were followed:

As we can see, the quality of the LSGAN with original images were similar, but for the GAN, the images quality was slightly poorer to the GAN with image size of 64x64.



(a) Collapsed output of GAN (b) Collapsed output of LSGAN

Figure 2: The collapsed outputs of GAN and LSGAN

(a) The output of GAN (b) The output of LSGAN

Figure 3: The outputs of GAN and LSGAN with image dimensions of 128x128

2 Part 2 - Adversarial Attack

In this section, we used the pretrained Alexnet model to work with the *imagenette*. Because the pretrained Alexnet model worked with the *imagenet* dataset with 1000 classes, so we changed the output layer's dimension from 1000 to 10, and retrained the network with several epochs (number of epochs = 4). The used loss function was from [1], which can handle the output tensor of the models and the label vector. After that, we tested the trained model with the valuated dataset, and it reached 93,77 % accuracy.

We designed two attacked mechanism: *fast gradient sign method - fgsm* and *iterative gradient method - igm*. Some images and their adversarial ones were shown in the Fig. 4. The difference can not recognized by human, however, it reduced the accuracy of the networks to 44.83%.

To defense the neural network, we used the adversarial training, it means we trained both the original images with the perturbed images. Before training with the adversarial defense, the model reached 44.83% with adversarial images. After doing the adversarial training, the accuracy was improved to 70.01% on the adversavery data.

(a) The original image (b) The output of IGM

Figure 4: The images after performing FGSM and IGM

References

[1] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? *CoRR*, abs/1906.02629, 2019.