Deep Fake Detection

Zehao Hui, Zhaoyuan Fu, Haoxiang Sun
{hzh98, fuzy, shx95}@bu.edu

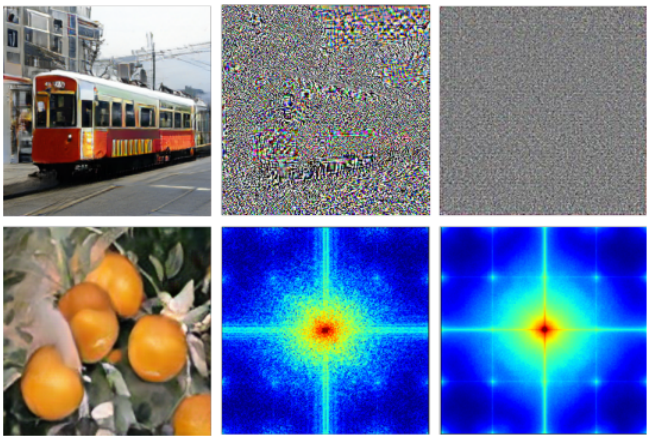


Figure 1. Examples of GAN synthetic images together with their not visible artifacts. From top to bottom: artificial fingerprint and its averaged version, Fourier spectrum and its averaged version.

1. Task

In recent years, people have proposed and implemented a large number of methods for artificially synthesizing pictures and media files based on deep learning. Generative Adversarial Networks (GANs) in particular have brought huge quality improvements. Using GANs it is even possible to regenerate images as well as modify existing ones. Based on these functions, some practical software or programs are gradually developed, such as improving the clarity of pictures, or intelligently retouching pictures. However, the technology can also be used for malicious purposes, such as generating fake profiles on social networks or generating fake news. Users are easily confused by GAN-generated images because they may differ from real images by a small amount. Therefore, there is an urgent need for automated tools that can reliably distinguish between authentic and manipulated content. This is also the purpose of our project.

2. Related Work

The content involved in [3] is mainly an overview, and some common deep fake detection methods are mentioned. A “contrastive learning based approach” has been raised in [1], which aims to make the method more generalized and robust in practical scenarios. The performance of this method was tested against other methods to illustrate the merits of the method. In [2], Seven different GAN detectors are mentioned by the authors, and these detectors are divided into three categories according to the detection method. The authors tested the performance of these detectors on the same dataset and visualized the results with images.

3. Approach

To start with, we mainly focus on two detection methods as shown in Figure 1: Learning spatial domain features from GAN’s artificial fingerprint, and Learning frequency domain features from GAN’s Fourier spectrum.

We aim to re-implement multiple different detectors that belong to above two methods, as the original authors didn’t make their codes available. These detectors include but are not limited to: Xception and DenseNet in the first method, and SRNet, Spec, Co-Net in the second method.

After that, we will pick the best performing detector of each type, and then deploy Ensemble Learning to combine the two best performing algorithms.

In the end, we draw conclusions by comparing the combined results with the results from each detector by itself.

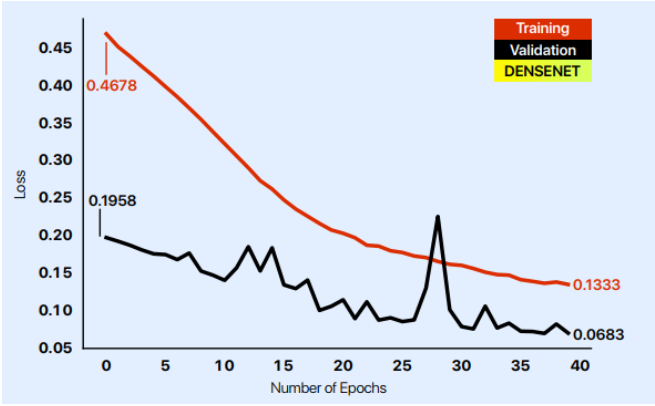
4. Dataset and Metric

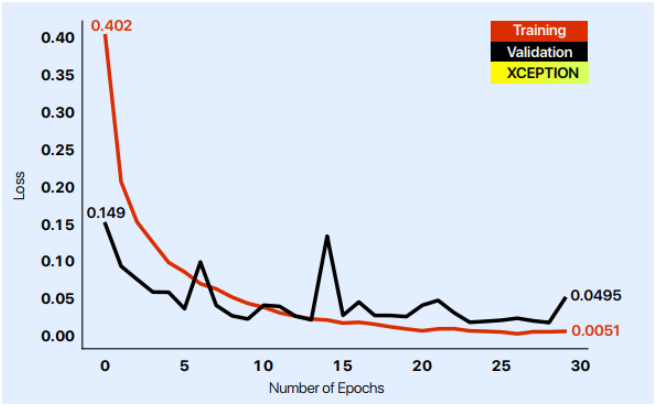
For training, we use the dataset provided by [4], comprising 362K real images extracted from the LSUN dataset and 362K generated images obtained by 20 ProGAN models, each trained on a different LSUN object category. All images have a resolution of 256 256 pixel. A subset of 4K images is used for validation. Available testing datasets are outlined in Figure 2, and we will use at least several of them to perform the testing in both low and high resolution.

We will measure the final performance by the accuracy, and we hope that our combined detector will have a higher detecting accuracy.

5. Preliminary Results

Xception and DenseNet have been done and the performance is just as follows:





For DenseNet, it required large epochs to achieve a good performance and the test accuracy is about 0.82. For Xception, it will achieve a test accuracy of 0.87 at about epoch 9. However, the test error increases for the rest.

6. Detailed Timeline and Roles

Task	Deadline	Who
Implement Learning spatial domain features method	04/18/2022	Zehao Hui
Implement Learning frequency domain features method	04/18/2022	Zhaoyuan Fu
Implement the Ensemble Learning	04/18/2022	Haoxiang Sun
Prepare report and presentation	04/25/2022	all

7. Preliminary Code

GITHUB:
https://github.com/FaustineHui/EC523_finalproject.git

Low Resolution (256×256)		
Name	Content	# Images
Various	ImageNet, COCO, Unpaired-real	11.1k
StyleGAN	Generated objects (LSUN)	6.0k
StyleGAN2	Generated objects (LSUN)	8.0k
BigGAN	Generated objects (ImageNet)	2.0k
CycleGAN	Image-to-image translation	4.0k
StarGAN	Generated faces (CelebA)	2.0k
RelGAN	Generated faces (CelebA)	3.0k
GauGAN	Generated scenes (COCO)	5.0k
High Resolution (1024×1024)		
Name	Content	# Images
RAISE [18]	Central crop of real photos	7.8k
ProGAN	Generated faces (CelebA-HQ)	3.0k
StyleGAN	Generated faces (CelebA-HQ)	3.0k
StyleGAN	Generated faces (FFHQ)	3.0k
StyleGAN2	Generated faces (FFHQ)	3.0k

Figure 2. Datasets used for testing the methods under analysis

References

1) D. Cozzolino, D. Gragnaniello, G. Poggi and L. Verdoliva. Towards Universal GAN Image Detection. 2021 International Conference on Visual Communications and Image Processing, 1-5, 2021.
2) D. Gragnaniello, D. Cozzolino, F. Marra, G.Poggi and L. Verdoliva. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. 2021 IEEE International Conference on Multimedia and Expo, 1-6, 2021.
3) D. Gragnaniello, F. Marra and L. Verdoliva. Detection of AI-Generated Synthetic Faces. Handbook of Digital Face Manipulation and Detection, 191-212, 2022.
4) S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. Efros, “CNN-generated images are surprisingly easy to spot... for now,” in CVPR, 2020.
5) F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, “Detection of GAN-generated fake images over social networks,” in IEEE MIPR, 2018.
6) L. Nataraj et al., “Detecting GAN generated fake images using co-occurrence matrices,” in IS&T EI, Media Watermarking, Security, and Forensics, 2019.
7) X. Zhang, S. Karaman, and S.-F. Chang, “Detecting and Simulating Artifacts in GAN Fake Images,” in IEEE WIFS, 2019, pp. 1–6.
8) X. Xuan, B. Peng, W. Wang, and J. Dong, “On the generalization of GAN image forensics,” in Chinese Conference on Biometric Recognition, 2019, pp. 134–141.
9) S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. Efros, “CNN-generated images are surprisingly easy to spot... for now,” in CVPR, 2020.
10) L. Chai, D. Bau, S.-N. Lim, and P. Isola, “What makes fake images detectable? Understanding properties that generalize,” in ECCV, 2020.