Does a GAN leave distinct model-specific fingerprints? (Supplemental Material)

Yuzhen Ding
Yuzhen.Ding@asu.edu
Nupur Thakur
nsthaku1@asu.edu
Baoxin Li
Baoxin.Li@asu.edu

Ira A. Fulton Schools of Engineering
Arizona State University
699 S Mill Ave, Tempe, AZ 85281

1 FID score for GAN images

Frechet Inception Distance (FID) is a metric that is used to determine the quality of GAN generated images. It calculates the distance between the Inception feature vectors obtained from the generated and original data. Lower the value of FID, better the quality of generated images. Figure 1 shows the sample real and GAN-generated images that we use for the experiments and their corresponding FID scores. As these scores are comparable to the ones achieved by [1, 2], it is clear that we use high-quality GAN-generated input images for our experiments.

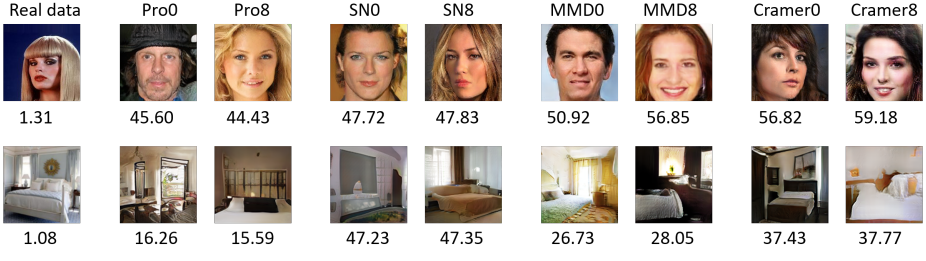


Figure 1: Sample images of celebA (upper row) and LSUN (bottom row) dataset and their corresponding FID scores.

2 More Experiments

In this section, we compare BFR-VAE with baselines and show a further application of BFR-VAE: attributing single image.

© 2021. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

GAN model	Pro0	Pro8	SN0	SN8
Pro0	3.73 ± 2.25	**3.61 ± 2.21**	31.00 ± 3.23	31.09 ± 3.23
Pro8	3.73 ± 2.23	**3.62 ± 2.19**	30.91 ± 3.22	31.00 ± 3.22
SN0	31.28 ± 3.29	31.01 ± 3.28	**3.65 ± 2.20**	3.70 ± 2.22
SN8	31.28 ± 3.31	31.00 ± 3.31	**3.64 ± 2.18**	3.69 ± 2.20
MMD0	32.77 ± 4.42	32.56 ± 4.42	32.74 ± 4.31	32.83 ± 4.31
MMD8	34.65 ± 3.80	34.44 ± 3.80	34.53 ± 3.85	34.62 ± 3.85
Cramer0	36.37 ± 3.41	36.15 ± 3.41	35.94 ± 3.50	36.08 ± 3.51
Cramer8	36.75 ± 3.39	36.52 ± 3.39	36.36 ± 3.40	36.49 ± 3.40

GAN model	MMD0	MMD8	Cramer0	Cramer8
Pro0	33.45 ± 2.85	35.91 ± 2.81	37.58 ± 3.07	38.06 ± 3.07
Pro8	33.45 ± 2.84	35.90 ± 2.79	37.56 ± 3.05	38.04 ± 3.05
SN0	33.85 ± 2.75	36.08 ± 2.71	37.39 ± 3.08	37.91 ± 3.08
SN8	33.84 ± 2.74	36.07 ± 2.70	37.42 ± 3.05	37.95 ± 3.06
MMD0	**4.56 ± 3.03**	5.18 ± 3.27	31.60 ± 4.13	32.18 ± 4.13
MMD8	**4.37 ± 2.70**	4.40 ± 2.77	33.23 ± 3.60	33.25 ± 3.61
Cramer0	30.42 ± 3.23	32.94 ± 3.23	**3.58 ± 2.36**	3.95 ± 2.49
Cramer8	30.74 ± 3.24	32.71 ± 3.25	**3.75 ± 2.35**	3.78 ± 2.38

Table 1: Mean and standard deviation of JSD between fingerprint of a single test image and the GAN fingerprints. BFR-VAE is trained on celebA images generated by ProGANx (Prox), SNGANx (SNx), MMDGANx (MMDx) and CramerGANx (Cramerx) where x=$\{0,8\}$ is the initialization seed used for training GAN. Largely speaking, the images from the same GAN show lower JSD than those from different GANs. *Note all the values have 1e-02 as a multiplication factor. The most similar fingerprints are **highlighted**.

As mentioned in the main paper, we can use the extracted parameters of a single image as the fingerprint to attribute it to its source. We use our model trained on celebA images generated using 4 GAN models (i.e., ProGAN, SNGAN, MMDGAN and CramerGAN) as the base model. The base fingerprints are calculated as the average of all the fingerprints extracted from training images generated by the respective GAN. Next, we extract the fingerprint of every single test image from a testset of 4 GAN models, each with two initializations. Table 1 and Table 2 show the mean and standard deviation of JSD and correlation coefficient values between a single test image and the base fingerprint of GAN models, respectively. We observe that a single image tends to have a lower JSD and a higher correlation with its source GAN (regardless of the initializations) than other GAN models. This further verifies the conclusion that each GAN has its own fingerprint.

We also calculate the classification accuracy, where the predicted GAN label is the one corresponding to the lowest JSD or highest correlation coefficient values. The results are shown in Table 3. We observe that BFR-VAE achieves higher accuracy than [4] (99.43%) and [3] (86.61%), which indicates BFR-VAE can attribute a GAN image to its source successfully.

It is worth mentioning that, the accuracy reported by [4] is obtained using their classifier. If we take their fingerprint (a $512 \times 1 \times 1$ tensor) and predict the label based on JSD or correlation coefficient values, it is observed that [4] fails to acquire a decent classification accuracy (only 25.52% and 32.54% accuracy, respectively). This implies the fingerprints extracted by [4] may not be able to generalize to other scenarios as the classification performance highly depends on a classifier trained with the same data.

GAN model	Pro0	Pro8	SN0	SN8
Pro0	**77.70 ± 4.06**	77.69 ± 4.05	17.49 ± 4.75	17.48 ± 4.75
Pro8	77.73 ± 4.03	**77.74 ± 4.01**	17.65 ± 4.74	17.65 ± 4.74
SN0	17.30 ± 4.91	17.46 ± 4.91	**77.11 ± 3.66**	77.11 ± 3.66
SN8	17.29 ± 4.90	17.45 ± 4.90	**77.10 ± 3.61**	77.10 ± 3.61
MMD0	15.65 ± 4.86	15.68 ± 4.86	15.53 ± 4.49	15.54 ± 4.49
MMD8	13.35 ± 4.07	13.38 ± 4.08	13.38 ± 4.13	13.40 ± 4.12
Cramer0	11.77 ± 4.21	11.81 ± 4.22	12.26 ± 4.35	12.21 ± 4.35
Cramer8	11.92 ± 4.38	11.98 ± 4.38	12.38 ± 4.42	12.32 ± 4.42
GAN model	**MMD0**	**MMD8**	**Cramer0**	**Cramer8**
Pro0	16.05 ± 4.24	12.77 ± 4.03	11.64 ± 4.43	11.69 ± 4.44
Pro8	16.09 ± 4.21	12.81 ± 3.97	11.69 ± 4.38	11.75 ± 4.40
SN0	15.76 ± 4.05	12.66 ± 3.82	11.99 ± 4.43	12.01 ± 4.45
SN8	15.78 ± 4.05	12.68 ± 3.84	11.94 ± 4.39	11.95 ± 4.42
MMD0	**75.37 ± 5.17**	74.57 ± 5.64	17.64 ± 4.75	17.55 ± 4.80
MMD8	79.99 ± 3.85	**80.80 ± 4.27**	15.41 ± 4.14	16.31 ± 4.22
Cramer0	18.31 ± 4.88	14.93 ± 4.77	**78.53 ± 3.93**	77.98 ± 3.85
Cramer8	18.37 ± 4.65	15.92 ± 4.44	78.55 ± 4.01	**79.11 ± 3.95**

Table 2: Mean and standard deviation of Correlation coefficient between the fingerprint of a single test image and base GAN fingerprints. BFR-VAE is trained on celebA images generated by ProGANx (Prox), SNGANx (SNx), MMDGANx (MMDx) and CramerGANx (Cramerx) where x=$\{0,8\}$ is the initialization seed used for training. Largely speaking, the same GAN shows a higher correlation than different GANs. *Note all the values have 1e-02 as a multiplication factor. The most similar fingerprints are **highlighted**.

3 Discussion

BFR-VAE has two layers in the latent space and is trained on a combination of losses, each contributing to the final GAN fingerprint. In this section, we study how each component affects the final GAN fingerprint extraction process.

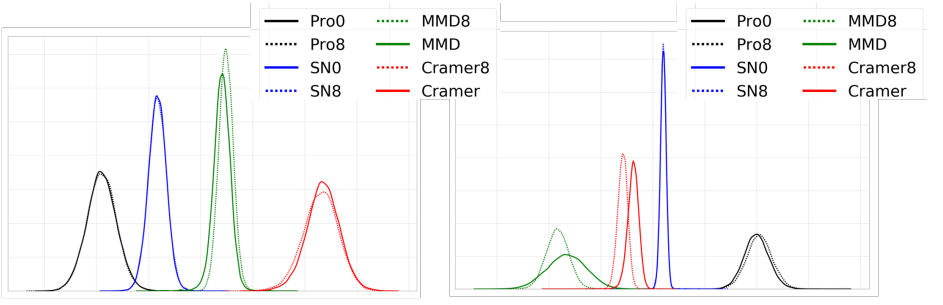
3.1 Latent Space

VAE has shown its capability in extracting the highly condensed features of the input and recovering it. Since the "fingerprint" we defined is also a condensed representation of GAN images, thus we carried out an experiment to extract possible fingerprints using VAE with triplet loss added to the latent representation. Figure 2 (a) and (b) compare the extracted fingerprint in 1-D using BFR-VAE and VAE, respectively. Although VAE can distinguish different GANs, it seems difficult in handling different initializations (red and green curves in Figure 2 (b)), even for the data that are used in the training. This indicates that VAE lacks the capacity to discriminate the features that are linked to each GAN and thus a richer latent space is required.

3.2 Ablation Study

We use celebA dataset to evaluate each loss component in BRF-VAE.

Reconstruction Loss: The reconstruction loss is Mean Squared Error (MSE) between the input and the reconstructed output. By including this loss in the final training loss, BFR-



(a) The fingerprints of multiple GANs (model trained on celebA with BFR-VAE)

(b) The fingerprints of multiple GANs (model trained on celebA with VAE)

Figure 2: Visualizing the distribution of the extracted fingerprint in 1-D for different GANs. The same color represents the same GAN while different line patterns represent different initializations of a GAN. (a) the data is trained with BRF-VAE; (b) the data is trained with VAE.

GAN model	Pro	SN	MMD	Cramer	Average
BFR-VAE	99.99 / 99.99	99.97 / 99.97	99.84 / 99.92	99.97 / 99.93	99.94 / 99.95

Table 3: Classification Accuracy (%). In each cell, we present two values as A/B, where the predicted GAN label is the one corresponding to the lowest JSD value in A and the highest correlation coefficient value in B.

VAE is trained to learn a compact representation of an image that captures the important and distinct features.

BFR-VAE produces poor quality reconstructed images if it is trained with only the triplet loss. The correlation coefficients of the extracted fingerprints are around 0.6 for the same as well as different GANs. Moreover, the fingerprints of different GANs even have smaller JSD values than that of the same GANs which is anti-intuitive. This observation shows the BFR-VAE fails to learn a meaningful fingerprint and indicates that the fingerprint extraction process also relies on the features learned for image reconstruction.

Triplet Loss: The triplet loss is imposed on the latent space parameters of BFR-VAE to make sure that fingerprint of images generated from the same GAN model is as close as possible whereas they are as distinct as possible for the images generated from different GAN models.

For the case where the BFR-VAE is trained without triplet loss, the extracted fingerprints have correlation coefficients that are greater than 0.9 and JSD values being around $6.33e-04$ for the same as well as different GANs. This implies that different GANs cannot be identified using the extracted fingerprints. Though the trained model learns significant features as a result of the reconstruction loss producing good quality output images, the fingerprints extracted using this BFR-VAE are no longer distinguishable for different GANs.

From the above two experiments, it is clear that both the loss components are essential for training BFR-VAE and to ensure a good quality of the extracted fingerprints.

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[3] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511. IEEE, 2019.

[4] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7556–7566, 2019.