Pre-trained vs. Random Weights for Calculating Fréchet Inception Distance in Medical Imaging

2021 13th Biomedical Engineering International Conference (BMEiCON) | 978-1-6654-2627-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BMEiCON53485.2021.9745214

Jamie A. O'Reilly and Fawad Asadi
College of Biomedical Engineering
Rangsit University
Pathumthani, Thailand
jamie.o@rsu.ac.th

***Abstract*— Fréchet Inception Distance (FID) is an evaluation metric for assessing the quality of images synthesized by generative models. Conventionally this involves using an Inception v3 convolutional neural network that has been pre-trained to classify everyday color images with the ImageNet dataset. The final classification section of this network is omitted, leaving an efficient feature extractor that outputs an encoded representation of each input image in the form of a 2048 element vector. Difference or similarity between samples of images can then be compared by measuring the distance between the distributions of their corresponding feature representations. Researchers have raised concerns about the utility of FID for evaluating unorthodox images (e.g. medical images) that are unlike those used for model training; suggesting that randomly initialized convolutional neural networks may be more appropriate. The aim of this study was to compare pre-trained and random approaches for evaluating medical images. Robustness to synthetic image distortions (Gaussian noise, blurring, swirl, and impulse noise) and different image types (X-ray, CT, fundus, and everyday images) was addressed. Feature representations were converted into two-dimensional space and visualized using t-distributed stochastic neighbor embedding (tSNE) and principal component analysis (PCA). Normalized FID between image classes was substantially larger and more consistent for the pre-trained model. Overall, this suggests that the pre-trained model is preferable to the randomly initialized model for evaluating medical images.**

***Index Terms*— Generative adversarial network, synthesized images, evaluation metric, deep learning, biomedical image processing**

I. Introduction

Generative adversarial networks (GANs) have brought tremendous advances in image synthesis [1]. In the domain of medical images, GANs have been applied for synthesizing computed tomography (CT), fundus (photographs of the retina), and X-ray images, among others [2]. Motivations for generating realistic medical images primarily stem from desires to improve other image processing applications, such as image classification or segmentation, by enhancing data augmentation [3], [4]. Quantifying whether synthetic medical images are indeed realistic can be tackled in broadly two ways: human expert [5], which is slow and resource-intensive, or automated computational assessment. This paper concerns the latter, computational approach for evaluating similarity between different types of images.

One of the most prominent synthetic image evaluation metrics to emerge in recent years is Fréchet Inception Distance (FID), which conventionally uses a pre-trained Inception v3 image classification network [6]. By replacing its top fully-connected and softmax layers with average pooling, this network is repurposed as an image encoder, which outputs 2048-element vector representations of input images, also known as embeddings. FID measures the distance between distributions of these embeddings for different types of images (e.g. real and artificial). The Inception model used to calculate FID has typically been trained with everyday color images from a subsample of the ImageNet Large Scale Visual Image Recognition Challenge (ILSVRC) dataset [7]. It has been argued that using these pre-trained weights may be suboptimal for evaluating types of images that were not included in the training distribution, such as medical images, and that a randomly initialized model may perform better by removing this dataset bias [8]. To determine whether this is the case, an Inception model with randomly initialized weights may be created for extracting image features.

In the present study, we evaluate the use of pre-trained and random weights for computing FID within and between different types of medical (X-ray, CT, and fundus) and everyday color images. We begin by partially replicating the work of [6] using chest X-ray films to examine the influence of image distortions on FID. In addition to comparing identical images (as did Heusel *et al.*), we split samples in half for comparing zero-distortion images, providing a more suitable benchmark for synthetic images, which should not be a perfect copy of the training dataset. This is followed up by computing the FID between different types of images using pre-trained and randomly initialized models, and subsequently visualizing the distributions of embeddings derived from the two models.

II. Methods

A. Data

One-hundred chest X-ray [9], thoracic contrast-enhanced CT [10], fundus [11], and golden retriever [12] images were sampled. Each image was resized to 299 by 299 RGB pixels with bi-cubic interpolation and stored in unsigned 8-bit integer

978-1-6654-2627-5/21/$31.00 ©2021 IEEE

(.jpg) format. Examples are shown in Fig. 1. Images were scaled to the range $[-1, 1]$ for input to the Inception model.

B. Image Distortions

To assess within-image-type FID differences caused by distortions to medical images, the work of [6] was replicated using chest X-ray images. The 100 unaltered X-ray images were considered to have distortion level 0, and these were compared with increasing distortion levels 1, 2, and 3 for each of the following types of image distortion. Additionally, FID between each half of the X-ray images (50:50) was calculated to provide a more realistic analysis of zero distortion on the same types of images.

1) Gaussian Noise

Random noise from a Gaussian distribution with zero mean was added to each image. Distortion level was controlled by setting the standard deviation of the noise distribution to 10, 20, and 40.

2) Blur

Average filter kernels of size 5, 9, and 15 were used to blur images to varying degrees.

3) Swirl

Swirl transformations were applied with a radius of 200 px and strengths of 2, 4, and 6 to control the amount of swirl.

4) Impulse Noise

Impulse or "salt and pepper" noise was simulated by selecting the proportion of pixels that were randomly converted to black or white: 2%, 10%, or 20%.

C. Fréchet Inception Distance

The Inception v3 model [7] is a deep convolutional neural network architecture that won the ILSVRC 2014. This image classification network was repurposed as a feature extractor by replacing the top fully-connected and output layers with average pooling, combining feature maps into 2048-length vectors. Using Tensorflow, this model was initialized with pre-trained (ImageNet) and random weights.

As shown in equation (1), FID [6] measures the distance between two multidimensional Gaussian distributions of image feature representations, denoted here as image *set a* (μ_a, σ_a) and image *set b* (μ_b, σ_b). As described above, image sets were either original and distorted images (Fig. 2), or different types of images (Fig. 3).

$$FID = |\mu_a - \mu_b|^2 + tr\left(\sigma_a + \sigma_b - 2(\sigma_a \sigma_b)^{1/2}\right) \tag{1}$$

D. Dimensionality Reduction

Image feature representations were transformed from 2048 to 2 elements. This was required to visualize image representations in two-dimensional space (Fig. 4). Two complimentary approaches were used to avoid bias that may be associated with either one, described as follows.

1) t-Distributed Stochastic Neighbor Embedding

The method of tSNE maps data from a higher dimension space into a lower dimension space while minimizing the Kullback-Leibler divergence between their corresponding probability distributions [13]. This approach maintains comparable distances between points in higher and lower dimension spaces.

2) Principal Component Analysis

PCA is an established method of transforming data into a set of orthogonal variables, known as principal components, which are ordered according to the proportion of variance in the original data that they account for. This was used to reduce image feature representations into two-dimensional space by selecting only the first two principal components to represent each image.

E. Software

Python 3 in Google Colaboratory was used with and Tensorflow 2.4.1, Scikit-learn 0.22.2, Scikit-image 0.16.2, Scipy 1.4.1, Numpy 1.19.5, Matplotlib 3.2.2, and OpenCV-Python 4.1.2.30.

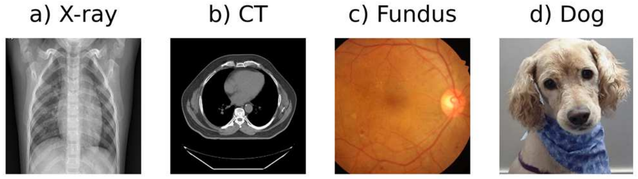

Fig. 1 Different types of images used in this study. Medical images included greyscale X-ray (a) and CT (b), color fundus photographs (c), and everyday color images of dogs (d) were also used for comparison.

III. Results and Discussion

FID measurements between chest X-ray images with and without distortion are plotted in Fig. 2. Greater levels of distortion increase FID, in agreement with [6]. When zero-distortion images were split in half, FID between these halves had similar FID to some of the level one distortion conditions. While this partly reflects sample size, it provides a more realistic measurement than comparing identical images without any distortion, which produces zero FID. There is a notable difference in scale between pre-trained and random models. The pre-trained model produced relatively consistent values of FID across image distortions, while the random model was notably more sensitive to Gaussian and impulse noise. Moreover, the random model was comparably insensitive to swirl distortions.

FID within (50:50) and between (100:100) chest X-ray, CT, eye, and dog images are shown in Fig. 3. Normalized values are also shown for comparing the results of pre-trained and random models despite their differences in scale. The random model produced lower FID within the same types of images, which is a desirable property, but also between image types, which is undesirable, relative to the pre-trained model. In contrast, the pre-trained model produced larger FID between different types of images, whereas within the same image type was slightly larger than the random model. Overall, the pre-trained model appears to do a better job of separating the different types of images.

Image feature representations transformed into two-dimensional space using tSNE and PCA are plotted in Fig. 4. Both of these dimensionality reduction techniques produced

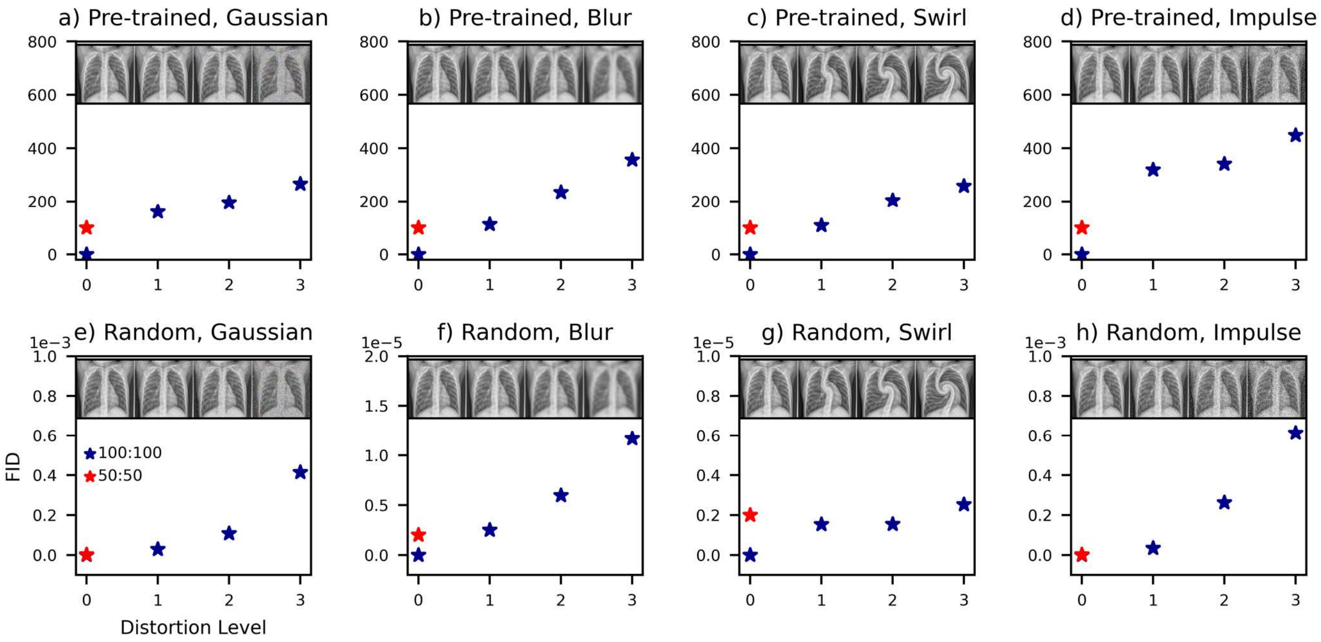

Fig. 2 Effect of distortions on FID measurements between chest X-ray images calculated using pre-trained (top) and random weights (bottom). Gaussian noise (a,e), mean blur (b,f), swirl (c,g), and impulse noise (d,h) distortions were applied in different levels, as explained in the methods section. Blue stars represent comparisons between unchanged 100 images and the same images with distortion. At distortion level 0 this produces zero FID, because the images were identical. Red stars represent FID between two different samples of 50 unchanged X-ray images. This analysis was inspired by [6].

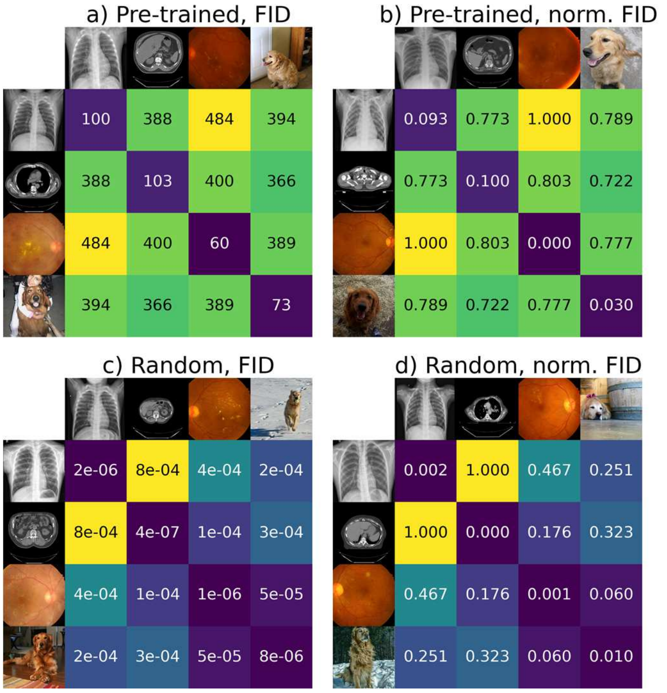

Fig. 3 Analysis of FID between different types of images using pre-trained (top) and randomized (bottom) approaches. The raw FID distances presented in (a) and (c) demonstrate different scales, so these were normalized to values between [0, 1] to improve their comparison as shown in (b) and (d). The randomized approach appeared to reduce the distance between images of the same type, while the pre-trained model increased the distance between different types of images.

linearly separable feature representations from embeddings produced by the pre-trained model. Reduce dimension feature representations from the randomly initialized model could not be linearly separated in the same manner. Taken together, these analyses show that the pre-trained model has more consistent and preferable output across different types of images than the randomly initialized model. This result appears to contradict [8], who proposed that dataset bias might make the (ImageNet) pre-trained model worse for types of images that were not included in the training distribution.

There are a few differences between our approach and that of M. F. Naeem *et al.* [8] that could explain this apparent contradiction. We used Inception v3 whereas they used the VGG16 [14], although this should not implicitly account for differences in the efficacy of image feature representations produced by pre-trained and randomly initialized models. In addition, they supported their statements with data from MNIST and voice spectrogram datasets, which are both relatively simple, greyscale images. In contrast, we included more complex medical grayscale (chest X-ray and CT) with medical (fundus) and everyday (dog) color images. The randomly initialized model separated greyscale images fairly well, as shown in Fig. 3; however, it did not separate greyscale from color images as well. We also used a smaller sample size, but consider that increasing the number of images would not drastically alter the overall pattern observed. Finally, M. F. Naeem *et al.* introduced their own density and coverage metrics, whereas here we used FID to evaluate the similarity and difference between embeddings for different types of images.

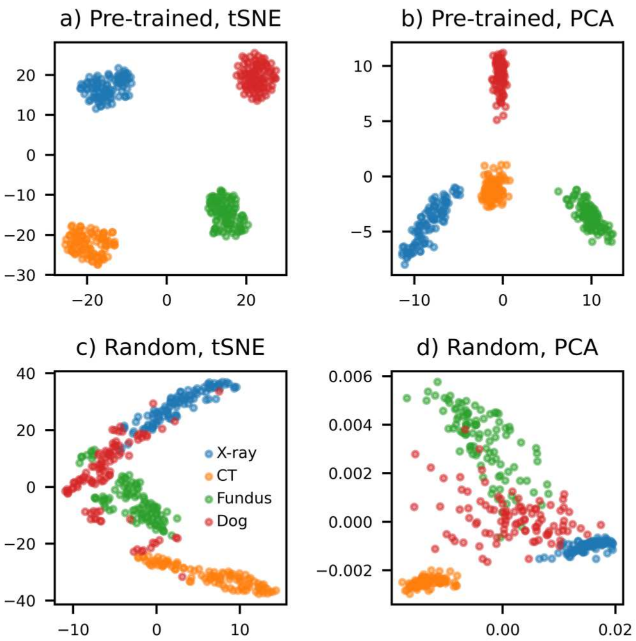

Fig. 4 Visualization of encodings transformed into two-dimensional space using tSNE (left) and PCA (right). This analysis illustrates that embeddings generated using the pre-trained model (top) have a more uniform distribution for the same types of images and more distance between different types of images, resulting in distinct. In contrast, the random model (bottom) produced embedding distributions that are more diffuse and not linearly separable.

Future work may seek to determine the suitability of FID for identifying more subtle differences between medical images of the same modality, for example between chest X-ray and abdominal X-ray images.

IV. Conclusions

Using an Inception model that has been pre-trained on the ImageNet database is preferable to using one with randomly initialized weights for extracting medical image feature representations. Although the training dataset does not include medical images, the pre-trained model nevertheless outperforms a randomly initialized model for representing different types of medical images. This is also beneficial for ensuring more reliable comparisons between studies, partly for the reason that random weights produced embeddings with a much smaller scale, and also given the inherent variability of randomly initialized weights.

Acknowledgment

We wish to thank M. F. Naeem and colleagues for stimulating this research and sharing their thoughts on YouTube.

References

[1] I. Goodfellow *et al.*, “Generative adversarial nets,” *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[2] X. Yi, E. Walia, and P. Babyn, “Generative adversarial network in medical imaging: A review,” *Med. Image Anal.*, vol. 58, p. 101552, Dec. 2019.

[3] A. P. Sagar and K. Venu, “Evaluation of Deep Convolutional Generative Adversarial Networks for data augmentation of chest X-ray images,” 2020.

[4] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, “Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Dec. 2019.

[5] M. J. M. Chuquicusma, S. Hussein, J. Burt, and U. Bagci, “How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis,” in *Proceedings - International Symposium on Biomedical Imaging*, 2018, vol. 2018-April, pp. 240–244.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, pp. 6627–6638, Jun. 2017.

[7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 2818–2826.

[8] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, “Reliable fidelity and diversity metrics for generative models,” in *37th International Conference on Machine Learning, ICML 2020*, 2020, vol. PartF16814, pp. 7133–7142.

[9] D. S. Kermany *et al.*, “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning,” *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, Feb. 2018.

[10] J. Yang *et al.*, “Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017,” *Med. Phys.*, vol. 45, no. 10, pp. 4568–4581, Oct. 2018.

[11] P. Porwal *et al.*, “Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research,” *Data*, vol. 3, no. 3, p. 25, Jul. 2018.

[12] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, “Novel dataset for fine-grained image categorization: Stanford dogs,” in *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011, vol. 2, no. 1.

[13] L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE,” 2008.

[14] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, Sep. 2014.