Deep Learning, Small Data Pathology Classifiers FID Literature Review and new FID Proposal

Shaylin Chetty
CHTSHA042@myuct.ac.za
University of Cape Town
Cape Town, South Africa

1 INTRODUCTION

Generative Adversarial Networks (GANs) consist of two neural networks called a Generator and Discriminator which compete in an adversarial game. The Generator tries to minimize the loss function while the Discriminator tries to maximize the loss function. However, the loss function is not objective, making it difficult to compare models [5]. As a visual computing method, the manual inspection of images has become a common metric to evaluate the quality of synthetic images [2, 8] however this technique is expensive and cumbersome [1] of which the results are subjective, variant and biased towards models that overfit [1].

Recent literature has proposed Inception Scores. Inception Scores uses an Inception Network trained on the ImageNet dataset whereby the probability that an input image belongs to each class defined in ImageNet is summarized to determine how close an image is to other images in the given class. The higher the score (similarity between images), the better quality of the image. One drawback was that the statistics of real-world samples are not used and compared to the statistics of generated samples [3]. Hence, Inception Scores were supplanted by FID scores. This literature review explores FID scores at a high level and proposes a medical-oriented FID metric using the same computational basis of the original implementation proposed by Heusel et al. in *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium* [3].

2 FID

Generative models aim to produce data that matches the distribution and structure of the original data. To quantify the dissimilarity, Heusel et al. proposed the *Fréchet Inception Distance* (FID)[3] which uses the Wasserstein-2 distance to calculate the distance between the generated data $p(.)$ and real data $p_w(.)$. A perfectly trained GAN should have $p(.) = p_w(.)$.

FID, like Inception Scores, uses InceptionNet [6], which has been trained on the ImageNet dataset. The network is trained as a classification model (see section 2.1). The final output layer is removed and the last coding layer (pool_3) is used to generate vision-relevant features of an image (these features would generally be sent to a softmax layer to classify the image into one of the ImageNet image classes). This layer represents each image as set of 2,048 activation features. We use the mean and covariance of the features to parameterize a multidimensional Gaussian distribution. This process is done for real and synthetic images upon which the Wasserstein-2 distance between these distributions is reported as the FID score (see equation 1). The Gaussian obtained from $p(.)$ has a mean and covariance of m and C and the Gaussian obtained from $p_w(.)$ has a mean and covariance of m_w and C_w. Tr is the trace linear algebra operation.

$$d^2((m, C), (m_w, C_w)) = ||m = m_w||_2^2 + Tr(C + C_w - 2(CC_w)^{1/2}) \quad (1)$$

2.1 What is InceptionNet

InceptionNet is a deep convolutional neural network for classification developed for the ImageNet Large-Scale Visual Recognition Challenge 2014 [6]. The CNN uses *Inception Modules* that use a combination of 1x1, 3x3 and 5x5 convolutions that allow the network to learn feature maps at different scales. These are then concatenated together allowing the network to learn both local and global features. FID scores use InceptionNet V3 which replaces the 5x5 convolution with two 3x3 convolutions, decreasing computation time, and adds in Batch Normalization [7].

3 SOLUTIONS

The drawback of FID scores in medical applications is that ImageNet contains no medical data meaning the feature maps don't capture the semantics of our X-Rays. As such, an FID-like metric for the medical domain is needed. The following options are proposed:

3.1 Train Inception V3 from scratch

This approach involved training an Inception V3 network from scratch using a medical dataset as described in section 3.3. Training would follow the approach of [7]: SGD with momentum = 0.9, auxiliary loss weights of 0.3 with a batch size of 32 for 100 epochs. The learning rate will be set at 0.045 decaying every 2 epochs using an exponential rate of 0.94. However, the resources required would make this method infeasible.

3.2 Transfer learning of Inception V3 trained on ImageNet

To overcome the data requirements and training time needed with training a network from scratch, I will apply transfer learning on a pre-trained Inception V3 network that has been trained on ImageNet (following the computation of FID scores). Transfer learning will be applied to train the network on the labelled neck and elbow data according to its pathologies. FID will then be computed on the new network.

3.3 Datasets

3.3.1 Open-access datasets. A large open-access medical imagery database would be used to create an Inception V3 classifier capable of predicting the labels of the given dataset. Once trained, the final output layer of the InceptionNet will be removed and the last coding layer $pool_3$ will be used to extract feature vectors for the FID calculations. Possible options include:

(1) **MURA-v1.1** [4]. MURA is a collection of 40 561 bone X-rays of the upper extremeties where each X-ray is classified as either *normal* or *abnormal*. Each image is also labelled with the type of body part being studied e.g. finger, elbow, forearm etc. The InceptionNet would be trained to predict the body part being studied with the final coding layer of the InceptionNet generating features for the different body parts. Through this approach, generated images can be tested against similar structures in the human body with the feature vectors highlighting key differences.

(2) **LERA**. LERA (Lower Extremity RAdiographs) is a collection of radiographs of the foot, ankle, hip and knee of 182 patients at the Stanford University Medical Centre whereby each image is labelled as either *normal* or *abnormal*. This will be used in conjunction with the MURA-v1.1 dataset to create a dataset of upper and lower extremeties upon which the Inception Network will be trained.

(3) **MedMNIST** [9]. MedMNIST is a MNIST-like collection of biomedical images with corresponding classification labels. It consists of 12 pre-processed 2D datasets including chest x-rays and colon pathology imagery.

3.3.2 Supervised DEEPPC data. We will leverage the neck and elbow existing data supplied by Dr. Kruger to train the Inception V3 network (specifically under transfer learning). To account for the limited data, traditional geometric data augmentations will be applied including rotations and cropping. The last layer of the InceptionNet will be removed and replaced with a softmax layer to classify images according to their pathology. Once trained, the final output layer of the InceptionNet will be removed and the last coding layer will be used to extract feature vectors for the FID calculations. The drawback is that the metric is not generalizable to other pathologies.

3.4 Testing of FID

FID remained a powerful technique due to its correlation with human perception of images (see figure 1) . To determine the effectiveness of the new metric, this correlation would have to be proven. To do this, an image will undergo an FID calculation before and after various disturbances (including addig blur and noise) have been applied. One expects the FID score to increase as the image quality worsens.

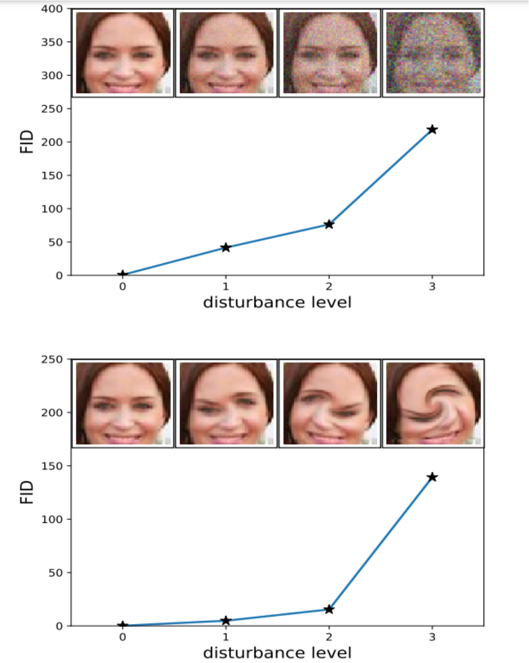

Figure 1: FID scores vs imagery of different quality [3]

REFERENCES

[1] Ali Borji. 2018. Pros and Cons of GAN Evaluation Measures. http://arxiv.org/abs/1802.03446 arXiv:1802.03446 [cs].

[2] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. 2015. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. http://arxiv.org/abs/1506.05751 arXiv:1506.05751 [cs].

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. http://arxiv.org/abs/1706.08500 arXiv:1706.08500 [cs, stat].

[4] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. 2018. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. http://arxiv.org/abs/1712.06957 arXiv:1712.06957 [physics].

[5] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. http://arxiv.org/abs/1606.03498 arXiv:1606.03498 [cs].

[6] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. http://arxiv.org/abs/1409.4842 arXiv:1409.4842 [cs].

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. http://arxiv.org/abs/1512.00567 arXiv:1512.00567 [cs].

[8] Lucas Theis, Aäron van den Oord, and Matthias Bethge. 2016. A note on the evaluation of generative models. http://arxiv.org/abs/1511.01844 arXiv:1511.01844 [cs, stat].

[9] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. MedMNIST v2 – A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data* 10, 1 (Jan. 2023), 41. https://doi.org/10.1038/s41597-022-01721-8 arXiv:2110.14795 [cs, eess].