To FID or not to FID: Applying GANs for MRI Image Generation in HPC

Beatriz Cepa, Cláudia Brito, António Sousa
INESC TEC and University of Minho
Braga, Portugal

***Abstract*—With the rapid growth of Deep Learning models and neural networks, the medical data available for training – which is already significantly less than other types of data – is becoming scarce. For that purpose, Generative Adversarial Networks (GANs) have received increased attention due to their ability to synthesize new realistic images. Our preliminary work shows promising results for brain MRI images; however, there is a need to distribute the workload, which can be supported by High-Performance Computing (HPC) environments. In this paper, we generate 256×256 MRI images of the brain in a distributed setting. We obtained an $FID_{RadImageNet}$ of 10.67 for the DCGAN and 23.54 for the WGAN-GP, which are consistent with results reported in several works published in this scope. This allows us to conclude that distributing the GAN generation process is a viable option to overcome the computational constraints imposed by these models and, therefore, facilitate the generation of new data for training purposes.**

***Index Terms*—MRI Image Generation, DCGAN, WGAN-GP, FID**

I. Introduction

Medical image analysis aims to acquire information about the medical condition of a patient in a non-invasive way [1]. The use of Machine/Deep Learning (ML/DL) models to automatize this task has become increasingly popular since analyzing images manually requires an astounding effort from medical professionals and is very time-consuming [2]. For these models to perform well, they need large amounts of training data, which, in the medical domain, is often difficult to obtain due to privacy issues and time-consuming annotations [2], [3]. In addition, although there are several public medical datasets available for research use, they are smaller than other non-medical datasets [4] – for instance, while *ImageNet* [5] contains more than 14 million images, *RadImageNet* [6] is composed of 5 million images. Thus, the need has emerged to explore new methods of obtaining more data.

For that purpose, Generative Adversarial Networks (GANs) have received increased attention in tasks such as segmentation and classification due to their ability to synthesize new realistic images [1], [4]. Our preliminary work corroborates this claim by showing promising results for brain MRI images; however, we found the model computationally heavy, resulting in high CPU running times [7]. This is a result of the GAN learning process, which might be compared to a two-player game between two models: the Generator, trying to generate new images resembling real ones, and the Discriminator, which discriminates between real and synthetic samples [8]. One key takeaway from the preliminary work was that resorting to High-Performance Computing (HPC) environments would support the need for workload distribution and benefit the model training time and stability [7].

Furthermore, a second takeaway from previous work is that there is no perfect metric to evaluate the output of these generative models (*i.e.,* a performance metric besides human evaluation should be used for scientific evaluations) [8]–[10]. Although works have been proposed showing that Fréchet Inception Distance (FID) can be a possible metric to use, it does not account for the fidelity and diversity of images in a set. With this, better-resolution images may have a lower FID even if the diversity of images in the set is lower than in a set with more diverse images and lower resolution [11]. Moreover, FID is computed with InceptionV3 [12] trained with the ImageNet dataset [5], in which the images do not translate into medical images [1], [3], [4], [9].

Backed by these insights, we re-implement the model used in our previous work (*i.e., DCGAN*), and also re-implement a *WGAN-GP* [13] to generate MRI images of the brain in a larger distributed setting. Both models were implemented in TensorFlow and resorted to the `MultiWorkerMirroredStrategy` module, which splits the model between the several GPU-enabled nodes.

A thorough evaluation was conducted by leveraging the two models and the 2020 and 2021 BraTS datasets [14]–[17] in an HPC infrastructure. Distinct from other works [1]–[4], [18], our models generate 256×256 images and have reached the lowest FID value of approximately 10.67 with the RadImageNet training weights.

II. Background

A. DCGAN

Deep Convolutional GANs (DCGANs) were introduced by Radford et al. [19] and quickly became one of the most used GANs in medical image analysis [4] due to their ability to generate higher-quality images. They rely on fractional-strided convolutions (also named transposed convolutions) on the generator and strided convolutions on the discriminator to generate 64×64 realistic images. In our previous work [7], we implemented this architecture with the Chainer framework [20] and obtained promising results in generating 256×256 MRI images of the brain in a single-node CPU.

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020 (DOI 10.54499/LA/P/0063/2020).

B. WGAN-GP

Conversely, Wasserstein GANs (WGANs) [21] are an alternative solution to DCGANs, where it is introduced a clipping to the weights and the training is asynchronous (*i.e.,* for each training iteration of the generator, the discriminator trains N iterations). Alongside, WGAN-GP [13] was proposed as an improvement of the WGAN. This updated version of the WGAN introduces a penalty term in the discriminator as an alternative to the previous weight clipping (*i.e.,* a gradient penalty (GP)). This aims to "penalize the norm of the gradient of the critic's output with respect to its input" and has shown to improve sample quality and training time [13].

C. GAN evaluation

To use synthetic images in analysis tools and models, it is of high importance that the images are as realistic and similar to the original data as possible [10]. Nonetheless, the human evaluation of such data is time-consuming, subjective, and error-prone, which prompts the need for an evaluation metric to assess image similarity [1], [3]. Despite the increasing use of GANs, evaluating their results remains a difficult task [9]. Several measures and scores have been proposed that attempted to perform a quantitative or qualitative evaluation separately. However, to the best of our knowledge, there is no consensus on which metric is best [8]–[10].

Two of the most commonly used metrics are the Inception Score [22] and the Fréchet Inception Distance (FID) [23], which rely on the Inception V3 network [12] pre-trained on ImageNet [5] to extract the underlying characteristics of the data [4], [9], [10]. The main goal of image generation is to obtain images as similar as possible to the real data, so the original samples should be used for comparison in GAN evaluation. Nevertheless, the Inception Score does not use the original images to evaluate the synthetic ones, which is considered a limitation [23]. The FID, in turn, measures the distance between the real data distribution and the generated data distribution by calculating the mean and covariance of the activations in the final block of the InceptionV3 for both sets of images [1], [9], [23]. Therefore, lower FID scores reveal a smaller distance between the two distributions, with 0.00 being the best value (meaning that the two image sets are identical) [9], [10]. Moreover, FID is more consistent regarding image noise, artifacts, and human judgment than Inception Score and performs well as far as discriminability, computational efficiency, and robustness are concerned [9], [23]. Nonetheless, FID does not consider the fidelity and diversity of data [11], with this last one being one of the metrics used for understanding if the trained model collapsed (*i.e.,* the generated images present low diversity).

III. Methods

A. Datasets

We used three image sets obtained from the Brain Tumor Segmentation (BraTS) 2020 and 2021 datasets [14]–[17]: Set 1 from BraTS 2020 and Sets 2 and 3 from BraTS 2021. Set 1 is the same as in our previous work [7] and presents 720 images of the FLAIR contrast. For Sets 2 and 3, we selected volumes from 48 subjects. Set 2 includes, for each subject, the slices that better represent the brain and tumor structures (central slices, #065 to #094) of the three MRI contrasts (T1, T2, and FLAIR), yielding a total of 4230 images. Finally, Set 3 is composed of slices #052 to #115 (of FLAIR contrast only) of each subject, resulting in a set of 3072 images. This set was created to broaden the cerebral area included in model training.

B. Implementation Details

Both DCGAN and WGAN-GP were implemented using TensorFlow [24] with the Keras API [25], and their architecture follows the one implemented on [7] regarding number and topology of layers. However, we diverge from that initial architecture regarding activation functions, in which we now resort to *Leaky ReLU* with $\alpha = 0.2$ on the Discriminator layers where in [7] *ReLU* had been applied. Regarding hyperparameters, a summary is shown in Table I. In specific, *Lr* represents the learning rate, and the optimizer was Adam [26] with $\beta_1 = 0.5$ on both models. The GP term in WGAN-GP was set to 5.0, and the Discriminator trained 3 iterations for each Generator iteration.

The DCGAN model was trained with the three image sets, while the WGAN-GP was only trained with image Set 1 since it could not handle training with larger image sets.

C. Setup

The models were trained in 3 nodes (of a 4-node cluster with the Simple Linux Utility for Resource Management (SLURM) job management system [27]), each containing one NVIDIA GeForce RTX 2080 Ti GPU. The distribution strategy was TensorFlow's `MultiWorkerMirroredStrategy`, which implements synchronous training where the steps are synced across the workers and replicas. As such, all workers train over different slices of input data and aggregate gradients at each step.

IV. Results and Discussion

Our models generate one 256×256 brain MRI image (axial view) per epoch of training. We measured the FID score for each generated image set with InceptionV3 [12] pre-trained on ImageNet [5] and with RadImageNet [6] weights. The top 3 FID values obtained for each model are presented in Table II, where $\text{FID}_{\text{ImageNet}}$ refers to FID calculated with pre-training on ImageNet and $\text{FID}_{\text{RadImageNet}}$ means FID calculated with RadImageNet weights.

TABLE I
HYPERPARAMETERS USED IN THE TRAINING PROCESS.

Model	Image Set	Epochs	Lr	Batch size	Dropout
DCGAN	Set 1	50	0.0001	8	0.2
	Set 2	100	0.0002	64	0.5
	Set 3	100	0.0002	32	0.5
WGAN-GP	Set 1	50	0.0005	8	0.2

In general, DCGAN FID values are lower than those of WGAN-GP, and evaluating with ImageNet weights scores higher FID than with RadImageNet. This goes in line with the literature [3], [4] and can be explained by the difference in the pre-training datasets: ImageNet is a non-medical dataset, while RadImageNet contains labeled images from several imaging modalities (PET, CT, Ultrasound, and MRI) [6], so applying RadImageNet weights allows the InceptionV3 to grasp features specific to medical images. The $\text{FID}_{\text{ImageNet}}$ values obtained with DCGAN, although significantly high, fall into what other author's findings show (*i.e.,* FID roughly between 50 and 270) [1], [2], so we theorize that our $\text{FID}_{\text{ImageNet}}$ WGAN-GP values are also coherent considering the use of the ImageNet dataset in the medical context. Furthermore, the $\text{FID}_{\text{RadImageNet}}$ scores of both models are consistent with results from other works [4], [8], with the best values ranging from 10.67 to 28.90, representing a decrease of up to $21\times$ when compared to the FID results obtained with ImageNet.

As reported in [1], we found that the larger the image set given to the GAN, the better the FID score of the synthetic images will be (Set 3 has more images than Sets 1 and 2 and gave the best FID score) and, consequently, the better the model will perform. Moreover, we report better FID values with DCGAN trained on BraTS 2021 image sets (Sets 2 and 3) than those obtained in [18] with other GAN architectures. All these findings suggest that our models have a similar performance in a distributed environment to those of others in a non-distributed setting, which supports our goal of applying GANs in an HPC environment to generate training data for other ML/DL models. The best results obtained for Sets 1, 2, and 3 are shown in Figures 1, 2, and 3, respectively.

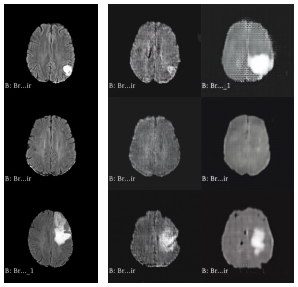

Fig. 1. Best results obtained for Set 1 (by columns, from left to right: original images, DCGAN, WGAN-GP).

As seen in Figures 1-3, the generated images have, overall, less detail than the original samples. Conversely, despite our models' performance in all contrasts, two out of three of our image sets were from FLAIR contrast, so more tests with other contrasts are needed to assess the model's generalization ability. Finally, the WGAN-GP raised issues regarding the training with Sets 2 and 3, which may occur because this model has a slower convergence than DCGAN [13] and training with a larger image set implies learning a higher amount of characteristics. Nevertheless, more tests are needed to pinpoint the exact reason for the WGAN-GP training problems.

TABLE II
TOP 3 FID SCORES FOR THE TWO MODELS.

Model	Image Set	$\text{FID}_{\text{ImageNet}}$	$\text{FID}_{\text{RadImageNet}}$
DCGAN	Set 3	224.43	10.67
	Set 2	232.94	15.01
	Set 1	311.87	16.67
WGAN-GP	Set 1	300.52	23.54
	Set 1	312.06	26.01
	Set 1	316.26	28.90

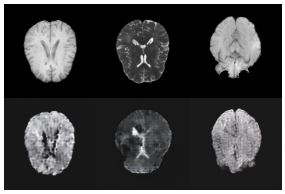

Fig. 2. Best results obtained for Set 2 (upper row contains original images, lower row contains DCGAN images).

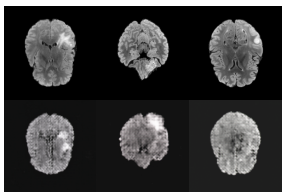

Fig. 3. Best results obtained for Set 3 (upper row contains original images, lower row contains DCGAN images).

V. Conclusions and Future Work

Image generation has turned into a powerful mechanism to overcome the scarcity of medical data for training, with GANs leading in terms of synthetic image quality. Since training these models is computationally demanding and requires a large amount of memory [1], workload distribution becomes the next big step for medical image generation.

In this work, we re-implemented two GANs (a DCGAN and a WGAN-GP) in a distributed setting to generate MRI images of the brain. To evaluate the synthetic images, there is no consensus on which metric is best; each one captures different aspects of the image generation process, so it becomes unlikely that a single measure can encompass all aspects [8]–[10]. We evaluated our results using FID with InceptionV3 pre-trained on ImageNet and RadImageNet weights.

We obtained FID scores consistent with the literature, either with ImageNet or RadImageNet weights, and $\text{FID}_{\text{RadImageNet}}$ was significantly lower than $\text{FID}_{\text{ImageNet}}$. Moreover, we validated that a larger input image set results in better model performance, as the FID scores decreased with Sets 2 and 3. With this, we present evidence that it is possible to obtain similar image quality and model performance using distributed environments, and although we recognize that applying FID to GAN evaluation in medical imaging raises some questions (concerning InceptionV3 being trained with representations of

a non-medical dataset) [1], [3], [4], [9], we believe that FID with RadImageNet weights is a strong metric of image quality, as our findings match those of other works in the same scope.

Future work on this approach includes testing the models in a larger distributed setting, more experiments to fine-tune the hyperparameters (*i.e.,* to improve synthetic image quality), and further investigation on the WGAN-GP training process.

References

[1] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "GANs for medical image synthesis: An empirical study," *Journal of Imaging*, vol. 9, no. 3, 2023. doi: https://doi.org/10.3390/jimaging9030069.

[2] V. Thambawita, P. Salehi, S. A. Sheshkal, S. A. Hicks, H. L. Hammer, S. Parasa, T. d. Lange, P. Halvorsen, and M. A. Riegler, "SinGAN-Seg: Synthetic training data generation for medical image segmentation," *PLOS ONE*, vol. 17, pp. 1–24, 05 2022. doi: https://doi.org/10.1371/journal.pone.0267976.

[3] R. Osuala, G. Skorupko, N. Lazrak, L. Garrucho, E. García, S. Joshi, S. Jouide, M. Rutherford, F. Prior, K. Kushibar, O. Díaz, and K. Lekadir, "medigan: a Python library of pretrained generative models for medical image synthesis," *Journal of Medical Imaging*, vol. 10, no. 6, p. 061403, 2023. doi: https://doi.org/10.1117/1.JMI.10.6.061403.

[4] L. Tronchin, R. Sicilia, E. Cordelli, S. Ramella, and P. Soda, "Evaluating gans in medical imaging," in *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections* (S. Engelhardt *et al.*, eds.), (Cham), pp. 112–121, Springer International Publishing, 2021. doi: https://doi.org/10.1007/978-3-030-88210-5_10.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: https://doi.org/10.1109/CVPR.2009.5206848.

[6] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z. A. Fayad, and Y. Yang, "Radimagenet: An open radiologic deep learning research dataset for effective transfer learning," *Radiology: Artificial Intelligence*, vol. 4, no. 5, p. e210315, 2022. doi: https://doi.org/10.1148/ryai.210315.

[7] B. Cepa, C. Brito, and A. Sousa, "Generative adversarial networks in healthcare: A case study on mri image generation," in *2023 IEEE 7th Portuguese Meeting on Bioengineering (ENBENG)*, pp. 48–51, 2023. doi: https://doi.org/10.1109/ENBENG58165.2023.10175330.

[8] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are gans created equal? a large-scale study," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018. ISBN 9781510884472.

[9] A. Borji, "Pros and cons of gan evaluation measures," *Computer Vision and Image Understanding*, vol. 179, pp. 41–65, 2019. doi: https://doi.org/10.1016/j.cviu.2018.10.009.

[10] T. Nečasová, N. Burgos, and D. Svoboda, "Chapter 25 - validation and evaluation metrics for medical and biomedical image synthesis," in *Biomedical Image Synthesis and Simulation* (N. Burgos and D. Svoboda, eds.), The MICCAI Society book Series, ch. 25, pp. 573–600, Academic Press, 2022. doi: https://doi.org/10.1016/B978-0-12-824349-7.00032-3.

[11] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, "Reliable fidelity and diversity metrics for generative models," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185, PMLR, 13–18 Jul 2020.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, June 2016.

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[14] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, no. 170117, 2017. doi: https://doi.org/10.1038/sdata.2017.117.

[15] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015. doi: https://doi.org/10.1109/TMI.2014.2377694.

[16] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021. doi: https://doi.org/10.48550/arXiv.2107.02314.

[17] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018. doi: https://doi.org/10.48550/arXiv.1811.02629.

[18] M. U. Akbar, M. Larsson, and A. Eklund, "Brain tumor segmentation using synthetic mr images – a comparison of gans and diffusion models," *arXiv preprint arXiv:2306.02986*, 2023. doi: https://doi.org/10.48550/arXiv.2306.02986.

[19] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[20] S. Tokui, R. Okuta, T. Akiba, Y. Niitani, T. Ogawa, S. Saito, S. Suzuki, K. Uenishi, B. Vogel, and H. Yamazaki Vincent, "Chainer: A deep learning framework for accelerating the research cycle," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, (New York, NY, USA), pp. 2002—-2011, Association for Computing Machinery, 2019. doi: https://doi.org/10.1145/3292500.3330756.

[21] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, pp. 214–223, PMLR, 2017.

[22] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016. ISBN 9781510838819.

[23] 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," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017. ISBN 9781510860964.

[24] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016. doi: https://doi.org/10.48550/arXiv.1603.04467.

[25] F. Chollet *et al.*, "Keras." [Online]. Available: https://keras.io/, 2015. Accessed 29 January 2024.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] M. A. Jette and T. Wickberg, "Architecture of the Slurm Workload Manager," in *Job Scheduling Strategies for Parallel Processing* (D. Klusáček, J. Corbalán, and G. P. Rodrigo, eds.), (Cham), pp. 3–23, Springer Nature Switzerland, 2023. doi: https://doi.org/10.1007/978-3-031-43943-8_1.