Empirical Study of Quality Image Assessment for Synthesis of Fetal Head Ultrasound Imaging with DCGANs

T. Bautista, J. Matthew, H. Kerdegari, L. Peralta, and M. Xochicale

King's College London,
School of Biomedical Engineering & Imaging Sciences,
London SE1 7EU
{thea.bautista, miguel.xochicale}@kcl.ac.uk

Abstract. In this work, we present an empirical study of DCGANs for synthetic generation of fetal head ultrasound, consisting of hyperparameter heuristics and image quality assessment. We present experiments to show the impact of different image resolutions, epochs, dataset size input, and learning rates for quality image assessment on four metrics: mutual information (MI), Fréchet inception distance (FID), peak-signal-to-noise ratio (PSNR), and local binary pattern vector (LBPv). The results show that FID and LBPv have stronger relationship with clinical image quality scores. The resources to reproduce this work are available at https://github.com/budai4medtech/miua2022.

Keywords: GANs · Ultrasound Fetal Imaging · Medical Image Synthesis

1 Introduction

Synthesis of Ultrasound (US) imaging is growing in medical communities as a way to mitigate the scarcity of datasets due to the cost of data collection, annotation and its ethical policies with a generalisable and reproducible deep learning pipeline [1]. Generative Adversarial Networks (GANs) have the ability to mimic data distributions and synthesise medical images [6, 11]. In the case of US imaging, GANs are able to synthesise 2D images with fetal phantoms [5], intravascular US [12], breast US [2], and fetal head US [8]. Recently, Skandarani *et al.* in 2021 presented an empirical study for GANs in medical imaging on the impact of sensitivity of hyperparameters, dataset and computer scale, FID and image quality, and its clinical usability [11]. However, there are little to no empirical studies on the use of GANs with US fetal imaging but nerve US [7]. Considering that Deep Convolutional GANs (DCGANs) improve image quality generation and training stability of the networks (stride CNNs, extensive use of batchNorm, etc) [9] and its wide application in different medical image modalities [9, 6, 11, 2], we apply DCGAN architecture [9] with different input pixel image size (e.g. DCGAN64 and DCGAN128) (Fig. 1) using an open dataset of 999 real fetal head US images [4].
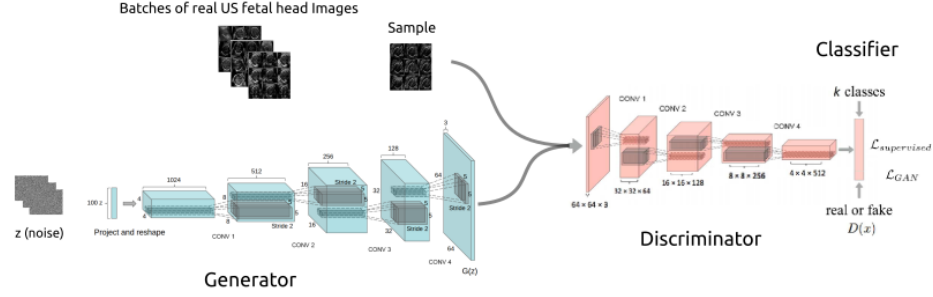
Fig. 1. DCGAN architecture for US imaging. Figure is adapted from the work of Radford et al. 2016 [9].

2   Quality assessment of synthetic fetal head US imaging

For quality assessment of synthetic fetal head US imaging, we consider four metrics for image quality assessment: The Fréchet Inception Distance (FID), peak-signal-to-noise ratio (PSNR), mutual information (MI), and local binary pattern vector (LBPv). FID is one of the most widely used metrics for GANs, where a lower FID indicates better image quality and increased diversity [3]. PSNR is a measurement of the ratio between the maximum power of a signal and the difference between the original and synthetic image, where a higher PSNR indicates closer intensity similarity between the synthetic and original US images. MI is a measure of image similarity and indicates the mutual dependence between two images, where higher MI denotes greater similarity between the images [13]. LBPv are texture descriptors, where the closer these values are from a synthetic image to a reference LBPv calculated from real US images, the higher the image quality is [10].

3   Experiments and results

Fetal head US images were resized to 64x64 (due to training instability at higher resolutions) and augmentation was applied in the form of random horizontal flipping and rotation to mimic variation in US probe location. Then DCGAN model was trained with the original hyperparameters, (learning rate=0.0002; optimiser=Adam optimiser; $\beta 1 = 0.5$; $\beta 2 = 0.999$ and loss function=binary cross entropy loss [9]), demonstrating the most stable training and resulting in the most realistic synthetic images when training the model on augmented data. Hence, the model was trained on 100, 300, 500, 800 and 999 US images, creating 20 sets of 800 synthetic images with 300, 500, 800 and 1000 epochs. FID, MI, PSNR and the LBPv values (sLBPv) were calculated using these sets of synthetic images. A reference LBPv (rLBPv) was calculated from 800 non-augmented original images from which the sLBPv was compared to. To assess
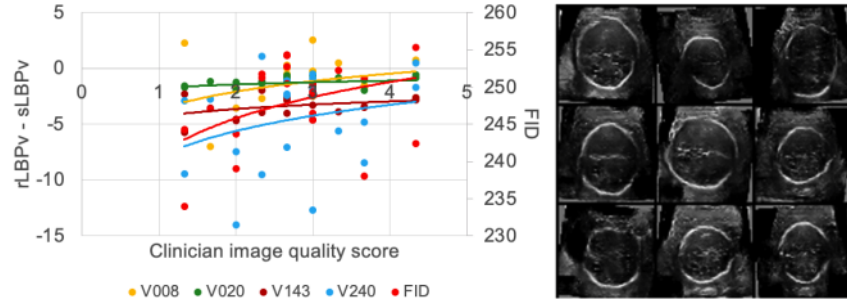
Fig. 2. Regression plots for FID, V8, V20, V143 and V240 with clinician image quality scores (left) and nine synthetic images of 64x64 pixel size from DCGAN64 trained on 300 images after 300 epochs (right).

the usefulness of FID, PSNR, MI and LBPv, a clinician with 10 years' experience with fetal US and two biomedical engineers were asked to differentiate between 68 real and synthetic images and to rate image quality on a scale of 1-5 (1 being poor quality and 5 being high quality). Image quality in the generated images did not vary significantly with increasing training size (MI, PSNR, FID and LBPv did not differ significantly with $p > 0.05$) which could indicate an upper limit in the capability of DCGAN in generating synthetic fetal US images. The biomedical engineers indicated 56.14% of the synthetic images as 'real' which shows that the DCGAN was able to create images that were indifferentiable from the original ones. However, the clinician had marked all images in the survey as 'fake' due to the small resolution and some instances of mode collapse reduced variation in synthetic images; both of which mitigate clinical application. FID and the V8, V20, V143 and V240 elements of the LBPv showed the strongest relationship to clinician image quality scores in regression analysis and highest correlation scores ($r > 0.2$) in comparison to the other metrics (Fig. 2, left); the small correlation was possibly attributed to the low number of data points used for analysis. These LBPv elements also had the smallest difference to the rLBPv thus show promise as indicators for textural similarity.

4 Conclusions and future work

To conclude, DCGANs are capable of satisfactorily learning the data distribution of the training images so as to create synthetic fetal US images indifferentiable from the originals, where FID and LBPv metrics show potential for image quality assessment of synthetic fetal US in comparison to MI and PSNR. However, the synthesised images are limited in clinical application due to their resolution of 64x64 (less than 500x500 pixels) and artefacts such as mode collapse. Hence, to create higher resolution US images for clinical applications, future work might lead to testing other GAN models, architectural alterations, hyperparameter optimization, addresing mode collapse artefacts and using other forms of data augmentation with other fetal US datasets.

Bibliography

[1] Fiorentino, M.C., Villani, F.P., Di Cosmo, M., Frontoni, E., Moccia, S.: A review on deep-learning algorithms for fetal ultrasound-image analysis (2022)

[2] Fujioka, T., Mori, M., Kubota, K., Kikuchi, Y., Katsuta, L., Adachi, M., Oda, G., Nakagawa, T., Kitazume, Y., Tateishi, U.: Breast ultrasound image synthesis using deep convolutional generative adversarial networks. Diagnostics **9**(4) (2019)

[3] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)

[4] van den Heuvel, T.L.A., de Bruijn, D., de Korte, C.L., Ginneken, B.v.: Automated measurement of fetal head circumference using 2d ultrasound images. PLOS ONE **13**(8), 1–20 (08 2018)

[5] Hu, Y., Gibson, E., Lee, L.L., Xie, W., Barratt, D.C., Vercauteren, T., Noble, J.A.: Freehand ultrasound image simulation with spatially-conditioned generative adversarial networks. In: Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment (2017)

[6] Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A.: Gans for medical image analysis. Artificial Intelligence in Medicine **109**, 101938 (2020)

[7] Kumar, D., Mehta, M.A., Chatterjee, I.: Empirical analysis of deep convolutional generative adversarial network for ultrasound image synthesis. The Open Biomedical Engineering Journal **15**(1) (2021)

[8] Lee, L.H., Noble, J.A.: Generating controllable ultrasound images of the fetal head. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 1761–1764 (2020)

[9] Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: International Conference on Learning Representations (2016)

[10] Singh, P., Mukundan, R., Ryke, R.D.: Texture based quality analysis of simulated synthetic ultrasound images using local binary patterns. Journal of Imaging **4** (2018)

[11] Skandarani, Y., Jodoin, P.M., Lalande, A.: Gans for medical image synthesis: An empirical study (2021)

[12] Tom, F., Sheet, D.: Simulating patho-realistic ultrasound images using deep generative networks with adversarial learning. In: 2018 IEEE 15th International Symposium on Biomedical Imaging. pp. 1174–1177 (2018)

[13] Wang, R., Fang, Z., Gu, J., Guo, Y., Zhou, S., Wang, Y., Chang, C., Yu, J.: High-resolution image reconstruction for portable ultrasound imaging devices. EURASIP Journal on Advances in Signal Processing **2019**(1), 56 (Nov 2019)