Analysis and Metaphysics
ISSN: 1584-8574, e-ISSN: 2471-0849
Vol 23 (1), 2024
pp. 878–889

Generative adversary neural networks (GANs) in the generation and detection of false faces for biometric security applications

Néstor Sanabria [1], Oscar Gualdron Guerrero [1,3], Marisol Maestre Delgado. [2]

[1]Deparment of Electronic Engineering, University of Pamplona - Colombia.
[2]Faculty of Economics and Business Administration, University of Pamplona, Colombia.
[3]Agroinnova Research Group, Ernesto Schiefelbein Foundation, Bogotá, Colombia.

Email: nestor.sanabria@unipamplona.edu.co, oscar.gualdron@unipamplona.edu.co, marisol.maestre@unipamplona.edu.co

Abstract

This project aims to address the growing issue of identity fraud in digital environments through the research and development of a system based on generative adversarial neural networks for creating and detecting fake faces. Currently, identity fraud poses a significant threat to biometric security in digital settings. Advances in generating fake images, particularly faces, have made secure and reliable authentication increasingly challenging. To tackle this challenge, Generative Adversarial Networks (GANs), a technique used in generating realistic images, will be employed. Research will focus on designing and developing an optimized GAN model for creating and detecting fake faces. The proposed solution involves implementing a comprehensive system that encompasses both the creation and detection of fake faces using artificial intelligence. Machine learning methods and image processing will be employed to develop a robust detection system. Subsequently, testing will be conducted to evaluate the system's performance. Analysis of the results will identify strengths, limitations, and potential areas for improvement, thereby providing recommendations for the effective implementation of the system in biometric security applications.

Resumen

En este artículo se propone abordar el creciente problema del fraude de identidad en entornos digitales mediante la investigación y desarrollo de un sistema basado en redes neuronales generativas adversarias para la creación y detección de rostros falsos. En la actualidad, el fraude de identidad representa una amenaza significativa para la seguridad biométrica en entornos digitales. Los avances en la generación de imágenes falsas, especialmente de rostros, han dificultado la autenticación segura y confiable. Para abordar este desafío, se emplearon las Redes Neuronales Generativas Adversarias (GANs), una técnica utilizada en el ámbito de la generación de imágenes realistas. Se realizó una investigación con un enfoque en el diseño y desarrollo de un modelo GAN optimizado para la creación y detección de rostros falsos. La solución propuesta consistió en implementar un sistema integral que comprenda tanto la creación como la detección de rostros falsos utilizando inteligencia artificial. Se emplearon métodos de aprendizaje automático y procesamiento de imágenes para desarrollar un sistema de detección robusto. Posteriormente, se llevaron a cabo pruebas para evaluar el rendimiento del sistema. El análisis de los resultados obtenidos permitió identificar las fortalezas, limitaciones y posibles áreas de mejora, proporcionando así recomendaciones para la implementación efectiva del sistema en aplicaciones de seguridad biométrica.

Keywords: Generative Adversarial Networks (GANs), digital environments, images, biometric systems, artificial intelligence.

Palabras clave: Redes Neuronales Generativas Adversarias (GANs), entornos digitales, imágenes, sistemas biométricos, inteligencia artificial.

1. Introduction

In the contemporary digital environment, identity fraud represents an increasingly complex and pervasive threat to biometric security. With the proliferation of advanced fake image generation technologies, especially of faces, secure and reliable authentication has become a crucial challenge in many sectors (Sarker et al., 2020). This phenomenon not only compromises the integrity of critical security systems, but also undermines public trust in biometric technologies essential for personal identification and verification (Shein, 2024). The adoption of GAN networks has marked a significant milestone in the ability to generate photorealistic images that can easily fool facial recognition systems. This technological advancement raises the need to develop both innovative and effective methods to counter these counterfeits accurately and effectively. To address this challenge, this work focused on investigating the development of a comprehensive system based on artificial neural networks for the generation and detection of fake faces.

The potential of GANs in the generation of facial images was explored, in addition to their ability to distinguish between genuine and counterfeit faces. In this article, the main objective has been to investigate the design of an architecture that can generate false faces while being able to identify these fakes. This approach not only seeks to advance theoretical-practical understanding, but also to provide an effective practical solution to improve biometric security in digital environments. The proposed methodology includes the investigation of deep learning techniques, the design of a specific architecture for the task of generation and detection of false faces, evaluating the performance of the developed system.

2. Methodology

For the development of this work, an experimental research methodology has been adopted. Through this approach, it was possible to quantitatively evaluate the effectiveness of the designed architecture of the GAN network in the generation and detection of false faces for biometric security applications. Experimental research has focused on designing and carrying out controlled tests that analyze the performance of the model.

The methodology is based on research in the field of GANs networks, the preparation and preprocessing of data, the implementation of the architecture and the evaluation of the system through classification and efficiency metrics. Through these steps, the aim is to obtain results that contribute to the advancement in the field of biometric security.

2.1 Literature review

In the initial stage of the research, a review of the existing literature on generative adversarial networks was conducted. This phase was essential to acquire knowledge of GANs and their application in the generation and detection of false faces in the field of biometric security. The following sources consulted were essential for the development of the architecture:
In the paper (Zumba Narváez & Zumba Narváez, 2022), entitled "Development of a prototype system for generating faces from voice signals using generative adversarial networks", a prototype is presented that uses GANs to generate faces from voice signals. The development of the system is divided into two main parts: the construction of a database for the training of the networks and the development of two generative networks that take noise and voice audio as input parameters.

The article (Atenas et al., n.d.), entitled "Convolutional Adversarial Neural Networks for Image Generation," discusses the use of GANs for image creation, with a specific focus on the generation of faces and other types of images. The study explores the structure and operation of these networks, which are composed of a generator that creates data and a discriminator that distinguishes between real and generated data. This approach is based on game theory, where both components improve their strategies throughout training, making the images generated increasingly realistic. In addition, the implementation and training process using the DCGAN (Deep Convolutional GAN) model is analyzed, including details about the network architecture, activation functions, and the use of normalization by batches to stabilize and accelerate learning. The paper highlights the ability of GANs to produce realistic images and their potential for practical applications, such as generating synthetic data to train other models.

2.2 Generation of a database containing face information to train and evaluate models

In the field of deep learning and GAN networks, the quality and consistency of the dataset are crucial for successful training and realistic image generation (Jordi de la Torre, 2023). To this end, a database of an individual's face was created, composed of 2000 photographs captured with a uniform black background, this decision was made to ensure standardization and consistency in the images, as well as to improve the ability of the GAN model in the detection of false images for applications in biometric security.

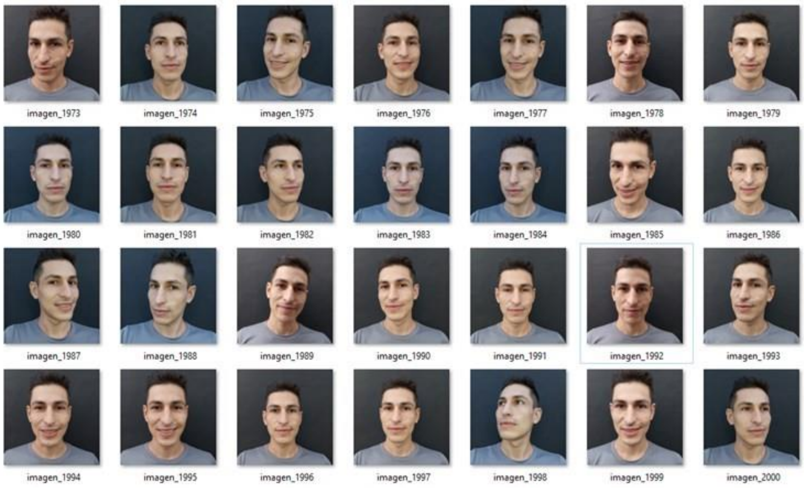

Figure 1. Photographs used to implement the model.
In the original Spanish language

2.3 Design and development of the GAN architecture

An architecture was implemented, DCGAN (Deep Convolutional GAN) (Deep Convolutional Generative Antagonistic *Network | TensorFlow Core*, n.d.). Architecture made up of deep convolutional layers in both the generator and the discriminator, thus allowing the model to learn more complex and subtle features of the images. The design and development of the DCGAN architecture was implemented on the Google Colaboratory platform, to efficiently access and manage datasets stored in the cloud. This integration was also used to load and save the folder of previously resized face images and the generated images. Also, to store checkpoints and results during model training. Within the development of the architecture, the values of the parameters and the seed for the generation of images necessary for the training of the GAN model were established. These parameters control key aspects of the training process and directly affect the performance and results of the model.

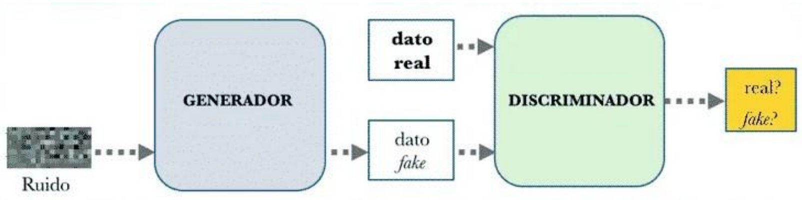

Figure 2. Schematic of a GAN neural network.
In the original Spanish language

2.4 Defining the Generator Model

In this part, the generator model in charge of delivering synthetic faces based on GANs networks was designed, specifically using an architecture known as DCGAN. Generator configuration is a critical part of the system that is responsible for creating synthetic images that mimic the features visual and structural of human faces (April, 2021). The generator configuration, as shown in Figure 3, was implemented using the tools provided by the Keras API. This modular structure makes it easy to define and optimize the model, allowing developers to quickly experiment and adjust the builder architecture to achieve optimal visual results.

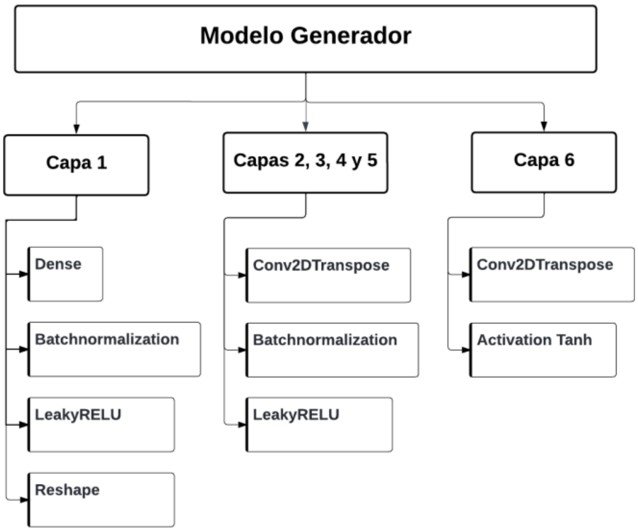

Figure 3 Definition of the generator model.
In the original Spanish language

2.5 Defining the Discriminator Model

The discriminator plays a fundamental role in generative adversarial networks. While the generator is responsible for transforming a noise vector into a coherent image, the discriminator acts as a critic that evaluates the authenticity of these generated images (*What is a GAN?*, n.d.). Its function is to distinguish between the images generated by the generator and the actual images of the training dataset. In essence, the discriminator learns to discern the distinguishing features of real images and to identify those that are artificially created by the generator. This dynamic of competition between the generator and the discriminator drives the adversarial learning process, where both models are trained simultaneously to improve the quality and authenticity of the images generated (Atenas et al., n.d.). This competitive interaction not only determines the success of the GAN network in the generator's ability to create convincing visual representations from random input data, but also in the discriminator's learning to distinguish between synthetically generated and real images.

Figure 4 presents the architecture of the discriminator developed within the context of a GAN network. The code details how the discriminator layers are structured using the tools provided by the Keras API, focusing on its ability to discern between synthetic and real images during the adversarial training process.

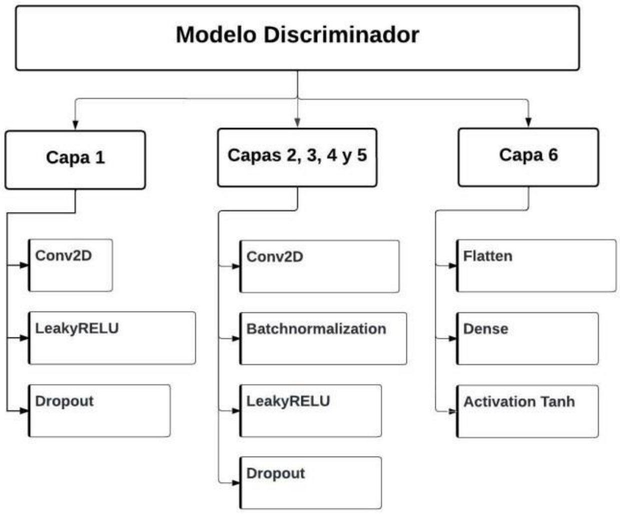

Fig.4: Definition of the discriminating model.
In the original Spanish language

The discriminator uses a DCGAN deep convolutional architecture optimized for precise feature discrimination in images. Convolutional layers are critical for extracting and learning discriminative representations that allow the model to differentiate between synthetic and real images (Wang et al., 2024). This structure is based on its ability to capture local and global details of the images, thus contributing to improving the quality and realism of the images generated by the network generator.

Stall functions are critical in training and are used to assess how well both the generator and the discriminator are performing (Dai et al., 2024). In the GAN model developed, binary cross-entropy was used, which measures the difference between two probability distributions, the true distribution and the predicted distribution. It is used to evaluate the veracity of the discriminator's predictions and the generator's ability to produce realistic images (Nadal Comparini, 2022). During the training phase the process is started using the dataset by performing a specific number of epochs. In summary, it initiates the training process of the generative adversarial neural network GAN, where the generators and discriminators are iteratively optimized through multiple epochs using the provided dataset (Benítez Iglésias, n.d.).

3 Results

This section presents the results obtained from the training and evaluation of the developed generator model. The images produced are analyzed and the impact of hyperparameters and training strategies on the overall performance of the GAN model is examined. The architecture takes advantage of deep convolutional layers to improve the quality and realism of the images generated. The image in the first period of the first training delivered by the generator in this DCGAN architecture corresponds to a clean image in gray tones where the noise is barely noticeable.

The results obtained in the first epoch are in accordance with a first training, where the generator has not learned the details and shapes of the person's face, and the discriminator is unable to differentiate between a real face and a generated one, this can be seen in figure 5, where the sequence of images produced in the 100 epochs is also observed. 200 and 1000 respectively. After 100 periods of training, it can be observed that the model has begun to generate images with discernible characteristics. This indicates that the generator is learning to produce more complex and realistic patterns from the input noise.

Fig.5: Images generated at times 1, 100, 200 and 1000 of the first training.

The image generated after 200 epochs of the first training of the implemented model shows a marked improvement compared to a basic DCGAN architecture. Although the image still exhibits some blur, features that resemble a human face can be distinguished, indicating a significant advancement in the generator's ability to produce coherent facial details. By configuring the hyperparameters and proceeding to train the model again, the results improve significantly in the third training compared to the first architecture used, and in a shorter execution time.

Fig .6: Image generated at the time 1000, third training.

As shown in Figure 6, the model has learned to capture and replicate fundamental facial features, such as the shape of the face, eyes, and mouth, although not with perfect clarity. This improvement in image quality suggests that the modifications implemented in the architecture and training of the model are allowing for more realistic image generation. This enhancement reflects the generator's ability to capture finer details and fool the discriminator more effectively. Although the images generated do not reach the quality of a realistic photograph, the GANs show promising results for the generation of fake faces.

To assess the quality of the images generated by the model, a metric was used to compare them with real images. The Structural Similarity Index (SSIM) was selected due to its ability to assess the structural similarity between two images, considering aspects such as luminance, contrast, and structure (Díaz-Ramírez, 2021). This metric provides an objective assessment that allows determining how close the generated images are to the actual reference images (Prieto Renieblas et al., 2016).

The SSIM is a metric used to measure the perceptual similarity between two images (Chevalier del Río et al., 2009). An SSIM value close to 1 indicates a high similarity between the images compared, while a value close to 0 indicates low similarity. In the network training for 1000 epochs, an average SSIM of 0.4278 was obtained. This result reveals that the images generated by the GAN network have a moderate perceptual similarity with respect to the original images. The average SSIM of 0.4278 suggests that the model has managed to capture some structural features of the original images, but there is still room to improve visual quality and perceptual similarity. This output can guide further adjustments to the GAN architecture or training parameters to achieve greater similarity to the reference images.

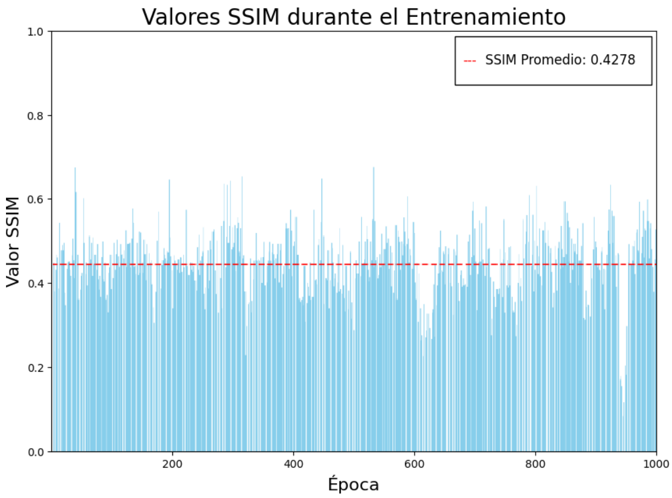

Fig.7: Average value of the SSIM evaluation between generated images and dataset images in 1000 iterations. In the original Spanish language

The process of evaluating the discriminating model is crucial to understand its ability to distinguish between real and model-generated images. This analysis was carried out after 1000 iterations of the fourth training of the DCGAN model, where the discriminator evaluations for real and generated images were recorded and analyzed. During training, the DCGAN model was tuned using a specific dataset and imager. Discriminator evaluations were performed by comparing how the model ranks the actual images of the original dataset against the images generated by the model. To evaluate the effectiveness of the discriminator, the results obtained after the training iterations were analyzed. These results provide a clear view of how the discriminator has learned to distinguish between the two classes of images throughout the training process.

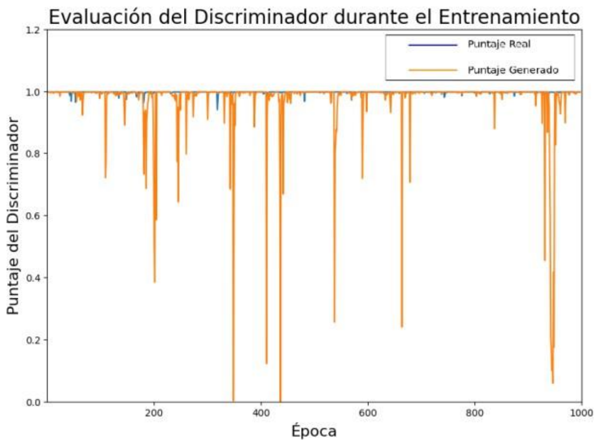

Fig.8: Evaluations of the discriminator in 1000 iterations, fourth training. In the original Spanish language

Figure 8 presents the evaluations of the discriminator for real images and generated throughout the fourth training of the model. The blue line represents evaluations for actual images, while the orange line shows evaluations for generated images. It is observed that the evaluations for real images are consistently high, close to 1, indicating a high certainty in the correct classification. On the other hand, evaluations for generated images vary widely, showing that the discriminator encounters more challenges in distinguishing between generated and real images.

These values are crucial for assessing the discriminator's ability to discern between real and generated images. In this specific context, the high number of classifications of the generated images as real suggests that the discriminator may be having difficulty fully distinguishing between the real and the generated images. This could indicate that the images generated by the generating network are quite convincing or that the discriminator needs more training to improve its accuracy.

3.1 Discriminator evaluation using confusion matrix and F1-score metric

In the context of the development and evaluation of machine learning models, the F1-score metric plays a critical role in measuring the accuracy and completeness of predictions, especially in binary classification tasks (Díaz, 2020). This metric combines accuracy and sensitivity (Recall) into a single measure, thus providing a balance between a model's ability to correctly predict positive classes and its ability to capture all instances of the actual positive class (Borja-Robalino et al., 2020). When evaluating, both the discriminator's ability to correctly identify real images (PV) and its ability to avoid misclassifying the generated images as real (FP) are considered (Rocha, n.d.). This provides a comprehensive assessment of the discriminator's performance in the task of discerning between real and synthetic data, thus ensuring the quality and consistency of GAN model generations (Pineda, 2022).

Using the F1-score metric, the evaluation of the discriminating model is performed on the results of the 1000 iterations. Table 1 presents the detailed results of this evaluation, calculated using the f1_score function of the sklearn.metrics library.

Table 1 Evaluations and results of the discriminator.

Total, of true positives (PV)	1000
Total, of true negatives (VN)	23
Total false positives (FP)	977
Total, false negatives (FN)	0
Precision	0.5058
Recall	1.00
F1 Scores	0.6718
Average SSIM value	0.4278

Where:

- True Positives (PV): 1000 real images classified as real.
- True Negatives (VN): 23 generated images classified as generated.
- False Positives (FP): 977 generated images classified as real.
- False Negatives (FN): 0 real images classified as generated.

Table 2: Confusion matrix for discriminator results.

	Correct predictions	Incorrect predictions
Real	1000	0
Generated	23	977

In this study, a threshold of 0.8 was used to assess the ability of a discriminating model in classifying generated and actual images. When the discriminator output for a real image exceeds this threshold, the model is considered to correctly classify the image as real and is recorded as True Positive (PV). In contrast, if the output for a generated image exceeds the threshold, it is incorrectly classified as real, being recorded as False Positive (FP). Accuracy is the measure that indicates how many of the positive cases were correctly identified in relation to the total positive results. In other words, it represents the proportion of positive cases that were correctly detected. This metric is calculated as follows:

$$\text{Accuracy} = \frac{\text{VP}}{\text{VP} + \text{FP}} = \frac{1000}{1000 + 977} = 0.5058$$

Equation 1: Precision of the discriminating model.

In this case, the model is correct approximately 50.58% of the time in predicting positive instances, which suggests that there are a significant number of false positives in the model's predictions.

Recall, also known as sensitivity, indicates the proportion of positive cases that were correctly identified by the algorithm. In the context of the developed GAN model, this metric is critical to assess how many of the actual positive cases were correctly detected by the discriminating model. It is calculated using the following equation:

$$\text{Sensitivity} = \frac{\text{VP}}{\text{VP} + \text{FN}} = \frac{1000}{1000 + 0} = 1$$

Equation 2: Sensitivity of the discriminating model.

A sensitivity of 1 means that the model correctly identified all the actual images. This is an ideal result and shows that the discriminator is very effective at detecting real images.

Accuracy is defined as the proportion of correct predictions made by the model, represented by the sum of the true results (positive and negative) divided by the total number of cases evaluated. This metric is calculated using the following formula:

$$\text{Model accuracy} = 2 * \frac{\text{Accuracy} * \text{Sensitivity}}{\text{Accuracy} + \text{Sensitivity}} = 2 * \frac{0.5058 * 1}{0.5058 + 1} \approx 0.6718$$

Equation 3: Accuracy of the discriminating model.

With a value of 0.6718, this model shows reasonably good performance in terms of accuracy and recall. This value is particularly useful when seeking a balance between correctly identifying positive instances and minimizing false positives.

4 Conclusions

This study represents a significant advance in the development and evaluation of GANs for the generation and detection of false faces in biometric security applications. Although the discriminator demonstrated good performance when classifying real images, it faced difficulties in distinguishing generated images as fake. This finding underscores the complexity and continued need to improve the generator's ability to create images that are indistinguishable from the real thing.

The SSIM obtained of 0.4278 by the generator indicates that while it has captured some structural features of the original images, there is still room to improve visual quality and perceptual similarity with respect to the authentic images. This aspect is crucial to strengthen the robustness of the system in the detection of more sophisticated digital manipulations.

In summary, this work provides a solid foundation for future research aimed at optimizing both the generation and detection of fake biometric images. By addressing these areas of improvement, it promotes progress towards more reliable and effective systems in the field of digital security, driven by innovative technologies such as Generative Adversarial Neural Networks. During the conduct of this study on GANs in the generation and detection of false faces for biometric security applications, several observations have been identified and key recommendations for future research and practical applications are formulated.

The performance of the GAN model could be affected by variations in the input datasets, such as different lighting conditions, viewing angles, and backgrounds. Increasing the number of images available in the dataset can improve the model's ability to learn general and specific features, thus making it easier to generalize to new conditions and scenarios. It is recommended to investigate robust data pre-processing methods and regularization techniques to handle these variations and strengthen the model's ability to produce consistent, high-quality results in different biometric contexts.

The evaluation of the model should include both objective metrics and visual evaluations for a complete understanding of its performance. Regular validations and ongoing evaluations are critical to monitor and improve system performance over time. It is suggested to explore training optimization techniques and the use of advanced hardware platforms to improve efficiency and reduce training times. Based on research and implementation, Deep Convolutional GANs (DCGANs) stand out as an ideal choice for imaging. These architectures employ deep convolutions in the generator and discriminator, allowing them to capture and learn complex features from images. The structure of DCGANs contributes significantly to improving the visual quality and stability of training, making them a powerful tool for the synthesis of realistic and detailed images.

In addition, it is recommended to research and implement Progressive Growing GANs (PGGANs) (Brownlee, 2019), which represent another prominent choice in advancing imaging capabilities. These architectures allow for progressive training that starts from a low resolution and gradually increases the resolution during the process. This scalable approach facilitates the generation of detailed and realistic images, improving both the stability of the model and the final quality of the images generated.

5 References

April, R. R. (2021, May 26). Generative Adversarial Networks • LMO article. *The Oracle Machine*. https://lamaquinaoraculo.com/deep-learning/redes-generativas-adversarias/

Atenas, F., Sanhueza, F., & Valenzuela, C. (n.d.). *Convolutional Adversarial Neural Networks for Imaging*.

Benítez Iglésias, R. E. B., Gerard; Kanaan Izquierdo, Samir; Masip Rodó, David. (n.d.). *Advanced artificial intelligence*. Open University of Catalonia; Digitalia. https:www.digitaliapublishing.com/a/79414

Borja-Robalino, R., Monleon-Getino, A., & Benedé, J. (2020). *Standardization of Performance Metrics for Machine Classifiers and Deep Learning*.

Brownlee, J. (2019, August 11). A Gentle Introduction to the Progressive Growing GAN. *MachineLearningMastery.Com*. https://machinelearningmastery.com/introduction-to-progressive-growing-generative-adversarial-networks/

Chevalier del Río, M., Guibelalde del Castillo, E., & Prieto Renieblas, G. (2009). *Use of SSIM as an index of medical image quality*. WorldCat. https://hdl.handle.net/20.500.14352/46231

Dai, Q., Yan, Q., Zou, Q., Li, Y., & Yan, J. (2024). Generative adversarial network with the discriminator using measurements as an auxiliary input for single-pixel imaging. *Optics Communications*, *560*, 130485. https://doi.org/10.1016/j.optcom.2024.130485

Diaz, R. (2020, May 8). ▹ Ranking Metrics—Learn how to EVALUATE your model. *The Machine Learners*. https:www.themachinelearners.com/metricas-de-clasificacion/

Díaz-Ramírez, J. (2021). Machine Learning and Deep Learning. *INGENIARE - Revista Chilena de Ingeniería*, *29*(2), 180-181. Engineering Source.

Jordi de la Torre. (2023). *GENERATIVE ADVERSARIAL NETWORKS (GANS) THEORETICAL FOUNDATIONS AND APPLICATIONS.* https://arxiv.org/ftp/arxiv/papers/2302/2302.09346.pdf

Nadal Comparini. (2022, July 29). *Generative adversarial neural networks (GANs).* https://blog.damavis.com/redes-neuronales-generativas-adversarias-gans/

Pineda, J. M. (2022). Predictive models in health based on machine learning. *CENTRAL TOPIC: CLINICAL DATA SCIENCE*, *33*(6), 583-590. https://doi.org/10.1016/j.rmclc.2022.11.002

Prieto Renieblas, G., Guibelalde del Castillo, E., & Universidad Complutense de Madrid Faculty of Medicine Department of Radiology and Physical Medicine. (2016). *Analysis and proposal of medical image quality metrics that mimic the human observer*. WorldCat.

What is a GAN? - Understanding Generative Adversarial Networks - AWS. (n.d.). Amazon Web Services, Inc. Retrieved April 17, 2024, from https://aws.amazon.com/es/what-is/gan/

Deep Convolutional Generative Antagonistic Network | TensorFlow Core. (n.d.). TensorFlow. Retrieved April 17, 2024, from https:// www.tensorflow.org/tutorials/generative/dcgan?hl=es-419

Rocha, Á. (n.d.). *Information Technology and Systems: ICITS 2024, Volume 1*. Springer. https://books.google.com.co/books?id=uWL8EAAAQBAJ

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: An overview from machine learning perspective. *Journal of Big Data*, *7*(1), 41. https://doi.org/10.1186/s40537-020-00318-5

Shein, E. (2024). Governments Setting Limits on AI. *Communications of the ACM*, *67*(4), 12-14. Engineering Source.

Wang, L., Li, Y., Liu, J., Peng, J., Zhang, Q., & Fu, W. (2024). Research on fault diagnosis of industrial robots based on generative adversarial network. *Physical Communication*, 102355. https://doi.org/10.1016/j.phycom.2024.102355

Zumba Narváez, E. A., & Zumba Narváez, F. P. (2022). *Development of a prototype system for generating faces from voice signals using generative adversarial networks* [bachelorThesis]. http://dspace.ups.edu.ec/handle/123456789/21758