ISSN 1870-4069

Optimización de la Arquitectura Pix2Pix: Un Estudio de Reducción de Capas y Calidad de Imagen

Alexander Tapia Cortez, José Alejandro Tejeda Sánchez, Yolanda Moyao Martínez, David Eduardo Pinto Avendaño, José de Jesús Lavalle Martínez

Benemérita Universidad Autónoma de Puebla, México

{alexander.tapiaco, jose.tejedasanc}@alumno.buap.mx, {yolanda.moyao, david.pinto, jose.lavalle}@correo.buap.mx

Resumen. En este trabajo en proceso se explora la modificación de la arquitectura Pix2Pix a través de la reducción de capas y cambios de filtros de convolución para encontrar un balance entre la complejidad de la arquitectura y la calidad de las imágenes sintetizadas. Se utilizó un problema clásico en la traducción de imagen a imagen, bocetos a imagen realista aplicado a un conjunto de entrenamiento de flores. El error cuadrático medio (MSE) de la comparación entre las imágenes generadas por la arquitectura original Pix2Pix y la arquitectura propuesta es 0.6718, mientras que el índice de similitud estructural (SSIM) obtenido fue de 0.6347, lo que indica un nivel moderado de similitud perceptual. Estos resultados motivan a seguir explorando con la modificación de capas para mejorar la eficiencia sin comprometer significativamente la calidad visual de las imágenes generadas.

Palabras clave: GAN condicional, arquitectura, filtro.

Pix2Pix Architecture Optimization: A Study of Layer Reduction and Image Quality

Abstract. In this work in progress, we explore the modification of the Pix2Pix architecture through layer reduction and convolution filter modifications to find a balance between architectural complexity and the quality of the synthesized images. A classic problem in image-to-image translation, from sketches to realistic images, was applied to a training set of flowers. The mean square error (MSE) of the comparison between images generated by the original Pix2Pix architecture and the proposed architecture was 0.6718, while the structural similarity index (SSIM) obtained was 0.6347, indicating a moderate level of perceptual similarity. These results motivate further exploration with layer modification to improve efficiency without significantly compromising the visual quality of the generated images.

Keywords: Conditional GAN, architecture, filter.

1. Introducción

En el campo de la inferencia visual, una rama de investigación interesante es la traducción de imagen a imagen, esto se refiere al proceso de introducir una imagen y esta a su vez genera una nueva imagen, este proceso aplica transformaciones mediante convoluciones y convoluciones transpuestas con el objetivo de generar nuevas ilustraciones que mantienen coherencia visual del contenido original. Las aplicaciones de esta técnica van desde la transferencia de estilos y clasificación de objetos hasta la recuperación y edición semántica de imágenes. [2]

Las soluciones a este problema que aplican redes generativas adversarias (GANs) son notables [2] y permiten optimizar el entrenamiento de la red al condicionarla, este condicionamiento introduce en el entrenamiento la imagen, estableciendo así la dirección del aprendizaje. [10] El modelo que se destaca es Pix2Pix. Este estudio está centrado en contrastar diversas variaciones en la arquitectura de la red y analizar su impacto en la calidad de las imágenes generadas, se toma como inspiración de diseño, ejemplos como el de trabajos presentados en el dominio de imágenes médicas, los cuales permitieron una reducción en el coste de recursos. [1]

Uno de los principales desafíos que enfrentan redes como Pix2Pix es encontrar un equilibrio entre la capacidad de representación y la eficiencia computacional. Si el modelo tiene demasiadas capas y filtros, puede volverse demasiado complejo, lo que aumenta la posibilidad de sobreajuste y dificulta la generalización de nuevas imágenes.

Por otro lado, una red con pocas capas o filtros puede tener dificultades para capturar las características necesarias para generar imágenes de alta calidad.

El modelo propuesto aborda estos problemas de manera eficiente al reducir el número de capas y filtros mientras mantiene un balance adecuado entre velocidad y precisión. Al ajustar la arquitectura de manera cuidadosa, el modelo busca preservar las características esenciales necesarias para la generación de imágenes realistas, al mismo tiempo que mejora el tiempo de entrenamiento y reduce los recursos necesarios. Esto no solo hace que el modelo sea más accesible en términos de hardware, sino que también facilita su implementación en escenarios donde los recursos son limitados. La evaluación comparativa de los resultados tras estas modificaciones permite obtener información valiosa sobre cómo simplificar la red sin comprometer de manera significativa la calidad de la salida generada.

El artículo está organizado como se indica a continuación: En la sección 2 se presenta la fundamentación y teoría de las GANs. En la sección 3 se analiza el diseño de la red generativa propuesta. Posteriormente, en la sección 4 se muestran los resultados de las pruebas comparativas entre la arquitectura original y la arquitectura propuesta. Finalmente se presentan las conclusiones del trabajo.

2. Estado del arte

La traducción de imagen a imagen es un proceso en el que se transfiere información desde un dominio fuente a un dominio objetivo, manteniendo el contenido de la representación original. [11] Este proceso tiene una amplia gama de aplicaciones, como

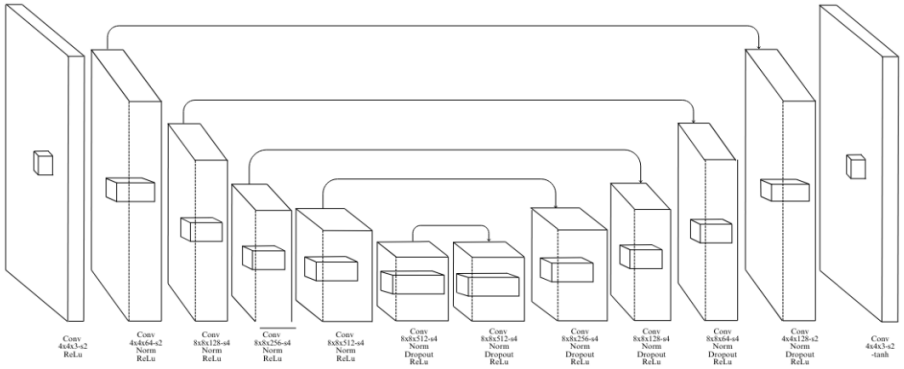

Fig. 1. En la arquitectura de la red generadora propuesta se cuenta con 12 capas convolucionales cada una seguida por un bloque de capas de normalización y ReLu, algunas capas tienen sus excepciones, el input y output de la red son imágenes de 256X256 por 3 filtros 3rgb.

en el trabajo de Kim. S, [5] quien utilizó una red residual con saltos conectados a Pix2Pix para asistir a los arquitectos en la distribución y conectividad de espacios durante el diseño de espacios, su modelo planteado alcanzó un 96% de precisión en el planteamiento de flujos de movilidad. [5] Por otro lado, la investigación de Lui, S. propone una alternativa más económica y accesible para estudios relacionados a la detección de cáncer de mama tras utilizar Pix2Pix para la generación de imágenes IHC que logró superar algoritmos tradicionales. [8] La versatilidad para la adaptación de dominios de Pix2Pix, se puede potencia como muestra el trabajo de Sun J. quien redujo de cinco pares de bloques de convolución y convolución transpuesta a tres, en la red generadora y agregó capas completamente conectadas al final de la red discriminadora, para reducir el nivel de ruido en las SPECT de baja dosis MP. [12] De hecho, investigaciones más recientes han encontrado que las mejoras en los resultados de tareas específicas están fuertemente vinculadas a modificaciones en la arquitectura de las redes, lo que demuestra, cómo la evolución de las técnicas sigue siendo un factor clave para avanzar en este campo [3].

No obstante, a pesar de estos avances, los modelos más populares de GANs continúan enfrentando desafíos importantes. Uno de los problemas más destacados es su incapacidad para generar imágenes de alta resolución [13], además, las GANs siguen siendo propensas a problemas inherentes, como la inestabilidad durante el entrenamiento y el colapso de modo, lo que puede llevar a la generación de resultados inconsistentes o de baja calidad. [6]

Varios estudios han explorado modificaciones en las arquitecturas de Pix2Pix y otros modelos GAN para mejorar su eficiencia y desempeño. Isola [4] presentaron Pix2Pix y optimizaron su arquitectura utilizando redes convolucionales de mapeo directo para tareas de traducción de imágenes, demostrando que la reducción de capas y la simplificación de la red pueden mejorar la eficiencia computacional sin sacrificar demasiado la calidad de la imagen generada [4]. Liu [8] propuso una modificación en la cantidad de filtros en las capas intermedias de una GAN, con el objetivo de reducir

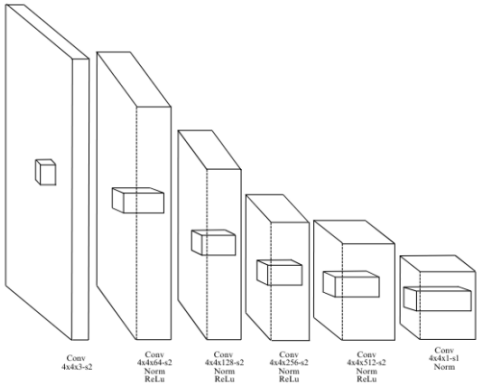

Fig. 2. La arquitectura del del discriminador de tipo *PatchGan* al igual que la del generador contiene capas de convolución, normalización y ReLu, similar a las capas donde se hace la reducción del modelo generador.

el tamaño del modelo y mejorar la velocidad de entrenamiento, sin que esto impactara negativamente en la fidelidad de las imágenes generadas [8]. Estos trabajos resaltan que la simplificación controlada de la arquitectura en las redes GAN es una estrategia efectiva para optimizar el modelo sin comprometer la calidad.

Aunque existen técnicas diseñadas para mitigar estos efectos, estas soluciones aún están en desarrollo y no se abordarán en profundidad. Sin embargo, es importante reconocer que estas estrategias ya están siendo aplicadas en algunos casos con resultados positivos como en el trabajo Liu, M, donde habilitaron saltos de capa para mitigar la pérdida de información durante las transformaciones de convolución [9] técnica que replicamos en nuestro modelo propuesto como se describe a continuación.

3. Metodología

3.1. Redes generativas adversarias

Este modelo comprende un par de redes neuronales interconectadas, denominadas Generador (G) como se muestra en la Figura 1 y Discriminador(D) representado en la Figura 2. El objetivo del generador es producir imágenes sintéticas que resulten lo más realistas posible, mientras que el discriminador recibe pares de imágenes auténticas y sintéticas con fin de clasificar correctamente a cada una según su origen. Ambas redes son entrenadas simultáneamente en un proceso competitivo. [2] Para la implementación de estas redes se utilizaron bloques de capas de normalización, Relu y convolucionales, estas últimas son un pilar del campo de visión por computadora por el alto desempeño que muestran en la comprensión de imágenes. [7]

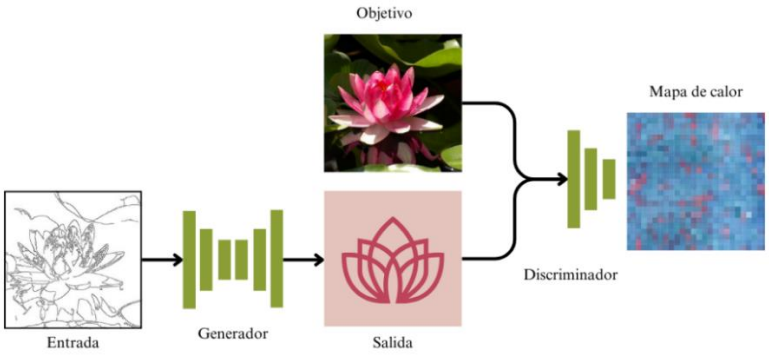

Fig. 3. Proceso de sintetización (generador) y evaluación (discriminador) de la GAN.

Fig. 1. Muestras del grupo objetivo a generar

3.2. Ajustes de la red original

La operación de convolución es fundamental en esta clase de redes por lo que modificar estas operaciones es un enfoque clave para realizar ajustes a la red original. Dentro de los ajustes que se hicieron, se logró una reducción de bloques de capas de tipo Convolución, Normalización, Relu.

Se modificaron los valores de *stride* y *kernel*, a 4 y 8 respectivamente que fueron seleccionados para lograr un equilibrio entre la velocidad de entrenamiento y la calidad de las imágenes generadas. El *stride* de 4 permite una reducción más rápida del tamaño de la imagen, lo que acelera el proceso de entrenamiento al disminuir la cantidad de operaciones requeridas.

Este valor intermedio evita la pérdida excesiva de información mientras mejora la eficiencia computacional. Por otro lado, el *kernel* de 8x8 aumenta el área de cada convolución, lo que permite al modelo capturar patrones más grandes y estructuras más amplias en la imagen. Esto mejora la coherencia y la capacidad del modelo para reconocer formas y características globales. Al combinar un *stride* alto con un *kernel*

Fig. 2. Muestras del conjunto de entrada.

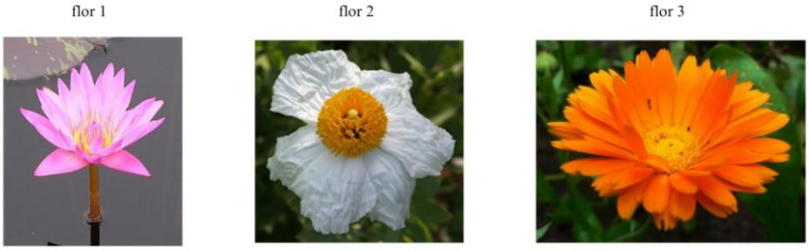

Fig. 6. Objetivo dentro del output de nuestro conjunto de datos.

más grande, se obtiene una mayor reducción de la dimensionalidad sin sacrificar los detalles importantes, lo que permite que el modelo aprenda características generales antes de enfocarse en los detalles en las capas posteriores. Este balance es crucial para mantener una alta calidad de generación sin aumentar innecesariamente el tiempo de entrenamiento o los recursos computacionales. Finalmente se redujo la cantidad de filtros en algunas capas, respetando la filosofía original del modelo, que sigue la estructura de reloj de arena como se muestra en la Figura 1. En cuanto al discriminador, se optó por utilizar el mismo de la arquitectura Pix2Pix como se ve en la Figura 2.

4. Evaluación y entrenamiento de los modelos

Para comparar Pix2Pix con el diseño propuesto, se planteó el problema de convertir una imagen estilo boceto de una flor a una imagen realista como se muestra en la Figura 3, donde los retos para el modelo consisten en generar texturas, colores y contexto lo más fieles posible a la realidad. Ambos modelos fueron entrenados con una taza de aprendizaje de $2e^{-4}$, con un lote de 50 y 80 épocas.

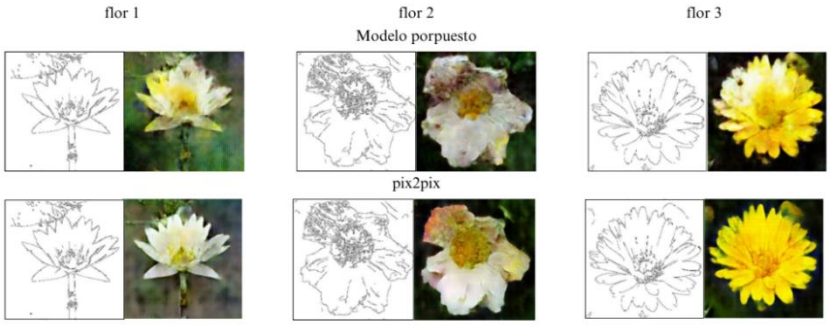

Fig. 7. Comparación de los resultados del modelo propuesto contra pix2pix.

4.1. Conjunto de entrenamiento

El conjunto de entrenamiento propuesto está conformado por 3500 imágenes variadas de flores como se ve en la Figura 4, de las cuales el modelo fue eligiendo el 90% de las imágenes (3150) y el resto (350) se consideraron para la etapa de evaluación, de los modelos. Para generar la entrada que se les dará a ambas redes, se procesaron estas imágenes para conservar únicamente los bordes más significativos como se muestra en la Figura 5. Es importante resaltar que la red generadora no tiene acceso a la imagen objetivo, sino únicamente al input, mientras que el discriminador trabaja con la salida del generador combinándola con la imagen objetivo.

4.2. Métricas de evaluación

El Error Cuadrático Medio (MSE) y la Desviación Estándar son métricas útiles para evaluar la diferencia entre imágenes generadas por GANs. El MSE mide la discrepancia pixel a pixel entre una imagen real y una generada, penalizando fuertemente las diferencias grandes, lo que permite detectar errores significativos en la generación.

Además, es una métrica computacionalmente eficiente y fácil de interpretar. Por otro lado, la desviación estándar evalúa la dispersión de los valores de píxeles en la imagen generada, lo que ayuda a identificar si la distribución de intensidades es consistente con la imagen real.

Una métrica adicional ampliamente utilizada en la evaluación de imágenes generadas es el Índice de Similitud Estructural (SSIM). A diferencia del MSE, que se enfoca en diferencias absolutas, el SSIM considera aspectos perceptuales como la luminancia, el contraste y la estructura de las imágenes, lo que lo hace más representativo de cómo el ojo humano percibe la calidad visual. El SSIM produce un valor entre -1 y 1, donde 1 indica una similitud estructural perfecta. Esta métrica es especialmente valiosa cuando se desea evaluar no solo la precisión numérica, sino también la fidelidad visual de las imágenes generadas.

En el caso de modificaciones en la arquitectura de Pix2Pix, estas métricas permiten cuantificar el impacto de los cambios en la calidad de las imágenes generadas. Por

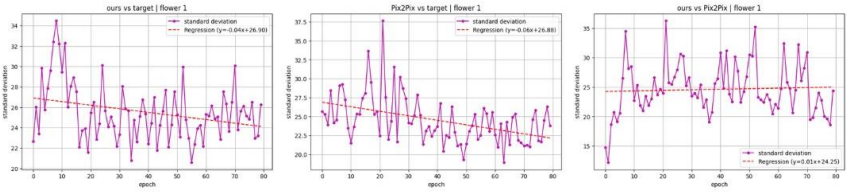

Fig. 8. Gráficas de comparativas de la desviación de la flor 1.

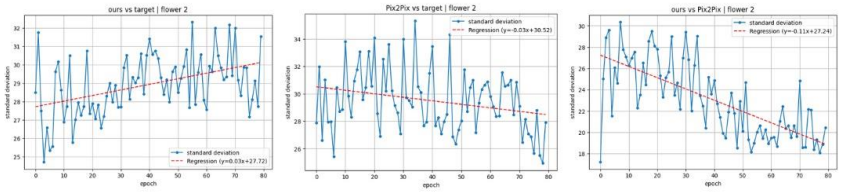

Fig. 9. Gráficas de comparativas de la desviación de la flor 2.

ejemplo, ajustar el tamaño de los filtros en la red puede afectar la precisión, la variabilidad y la percepción visual de la generación. Medir el MSE, la desviación estándar y el SSIM facilita una comparación objetiva y perceptual de los resultados, ayudando a determinar si los ajustes arquitectónicos mejoran o degradan el desempeño del modelo tanto desde una perspectiva técnica como visual.

5. Resultados

Para evaluar el desempeño de ambas arquitecturas (Pix2Pix y la propuesta), se calculó el Error Cuadrático Medio comparando 500 imágenes generadas con sus respectivas imágenes reales de referencia. Para cada par de imágenes, se obtuvo la diferencia pixel a pixel, se elevó al cuadrado y se promedió sobre el total de píxeles de la imagen. Luego, se calculó el promedio de estos errores individuales en las 500 muestras, obteniendo un valor representativo del desempeño del modelo que es de 0.6718. Este enfoque permite cuantificar de manera objetiva qué tan cercanas son las imágenes generadas por cada arquitectura a las originales, proporcionando una base sólida para comparar el impacto de la reducción de capas convolucionales y la modificación de los filtros en la calidad de la generación.

Las imágenes en ambas redes tienen una resolución de 256x256 píxeles, si bien no es perfecto, sin embargo, permite evaluar el desempeño de las arquitecturas. El manejo del color es bastante similar en ambos casos, aunque si hay una diferencia significativa en cuanto a la textura. En todas las imágenes, la silueta de las flores está correctamente

definida, lo que sugiere que el modelo ha logrado abstraer diversas formas de flores como se muestra en la Figura 7.

Para evaluar el desempeño del modelo, se hicieron comparaciones entre las imágenes en escala de grises generadas por los modelos, como se muestra en la Figura 7 con las imágenes objetivo que se muestran en la Figura 6, así como entre sí. Como métrica de comparación, se utilizó la desviación estándar de la siguiente manera:

$$\sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2. \quad (1)$$

En la ecuación 1. *N* es el número total de píxeles y μ representa la media de estos. Esta fórmula se aplicó en cada época del entrenamiento para obtener gráficas que permitan visualizar el desempeño de los modelos a lo largo del proceso de entrenamiento.

Dado que por cada época se obtiene la desviación, se necesita una métrica que muestre la tendencia de la desviación, por lo que se usa la regresión lineal para representar este comportamiento.

Al comparar los resultados del generador con la imagen objetivo, se espera que la desviación siga una tendencia decreciente, lo que indica que las imágenes generadas son cada vez más parecidas. En los casos de la flor 1, mostrada en la Figura 8 y la flor 3, mostrada en la Figura 10, la pendiente de la línea de regresión es negativa, es decir que la desviación estándar disminuye progresivamente. Sin embargo, hay casos donde el modelo propuesto diverge la similitud de resultados, esto es apreciable en la Figura 9 para el caso de la flor 2, donde la desviación estándar aumenta.

Se realizó una comparación entre los resultados del modelo propuesto y el modelo Pix2Pix, donde se espera que la desviación estándar tienda a 0. Esto indicará que ambas imágenes generadas por los modelos son muy parecidas. Una desviación estándar de 0 implicaría que ambas imágenes son idénticas, lo que sugeriría que los ambos modelos se comportan de manera exactamente igual. Además, se espera que la pendiente de la línea de regresión lineal tenga una pendiente negativa, lo que reflejaría la tendencia de la desviación estándar a aproximarse a 0. En este caso, los resultados son más prometedores que en la comparación con la imagen objetivo. El comportamiento de ambos modelos se muestra más estable, con una pendiente cercana a 0 en la regresión lineal.

Además, se observa en las gráficas que el modelo propuesto y el modelo Pix2Pix generan imágenes muy parecidas al inicio del entrenamiento. Sin embargo, alrededor de la época 10, empiezan a divergir para luego volver a asemejarse.

Esto sugiere que el modelo propuesto presenta un comportamiento muy similar al del modelo Pix2Pix, siendo el caso de mayor éxito en la flor 2, como se muestra en la Figura 9.

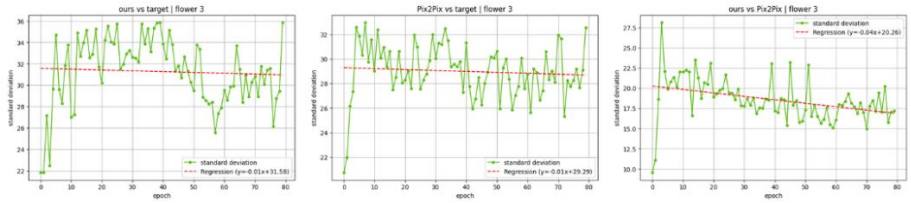

Fig. 10. Gráficas de comparativas de la desviación de la flor 3.

Para evaluar la similitud estructural entre las imágenes generadas por el modelo original Pix2Pix y el modelo propuesto, se utilizó el Índice de Similitud Estructural como métrica de referencia perceptual. El cálculo del SSIM se realizó durante 80 épocas de entrenamiento, comparando en cada una de ellas la imagen generada por la arquitectura modificada contra la correspondiente imagen generada por el modelo base Pix2Pix. Este enfoque permitió analizar cómo evolucionaba la similitud visual entre ambas salidas a lo largo del proceso de aprendizaje. Al finalizar el entrenamiento, se obtuvo un valor promedio de SSIM de 0.6347, lo cual indica un nivel de similitud estructural moderado entre ambas arquitecturas. El SSIM se define como:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (2)$$

En la ecuación 2 μ_x y μ_y representan las medias de las imágenes x e y, σ_x^2 y σ_y^2 son las varianzas, y σ_{xy} la covarianza entre ambas.C_1 y C_2 son constantes de estabilización. Este índice produce un valor entre -1 y 1, donde 1 representa una similitud estructural perfecta. En este contexto, el valor obtenido de 0.6347 sugiere que, a pesar de la simplificación arquitectónica, el modelo propuesto logra mantener una representación visual coherente y comparable a la del modelo original.

6. Conclusión

La reducción en la cantidad de capas convolucionales y la modificación de los filtros en Pix2Pix resultó en un Error Cuadrático Medio de 0.6718, lo que indica que las imágenes generadas por el modelo modificado tienen un nivel de similitud considerable con las generadas por la arquitectura original Pix2Pix. Adicionalmente, se obtuvo un Índice de Similitud Estructural de 0.6347, lo cual sugiere que las imágenes producidas no solo presentan una correspondencia numérica aceptable, sino que también mantienen una coherencia visual perceptible en términos de estructura, forma y contraste. Este valor de SSIM representa una similitud estructural moderada, lo que refuerza la viabilidad de aplicar simplificaciones arquitectónicas sin comprometer significativamente la calidad visual.

Este resultado sugiere que la optimización en la arquitectura no afectó drásticamente la precisión ni la percepción visual de la generación, lo que puede traducirse en un modelo más eficiente computacionalmente y más adecuado para aplicaciones en las que se prioriza la velocidad y el uso reducido de recursos.

Este trabajo es relevante para escenarios donde se busca reducir el consumo de recursos computacionales, como en dispositivos con capacidad de cómputo limitada o en entornos donde la rapidez en la generación de imágenes es prioritaria. A futuro, se podrían explorar otras modificaciones en la arquitectura, como el ajuste fino de hiperparámetros, el uso de estrategias de regularización, o la implementación de técnicas de compresión de redes neuronales, con el fin de continuar optimizando el desempeño del modelo sin comprometer su precisión ni su calidad visual.

Referencias

1. Ali, O., Ali, H., Ali, S.A.: Implementation of a Modified U-Net for Medical Image Segmentation on Edge Devices. IEEE Transactions On Circuits And Systems II: Express Briefs (2022) doi: 10.48550/arXiv.2206.02358.
2. Creswell, A., White, T.: Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine, pp. 53–65 (2018) doi: 10.48550/arXiv.2206.02358.
3. Ikhsan, G., Suciati, N.: The Comparative Study of Adding Edge Information to Pix2pix Architecture for Face Image Generation. In: 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), pp. 160–165 (2022) doi: 10.1109/ICITISEE57756.2022.10057739.
4. Isola, P., Zhu, J.Y., Zhou, T.: Image-to-Image Translation with Conditional Adversarial Networks. CVPR (2017) doi: 10.1109/CVPR.2017.632.
5. Kim, S., Lee, J., Jeong, K.: Automated Door Placement in Architectural Plans Through Combined Deep-Learning Networks of ResNet-50 and Pix2Pix-GAN. Expert Systems with Applications, 244, pp. 122932 (2024) doi: 10.1016/j.eswa.2023.122932.
6. Kossale, Y., Airaj, M., Darouichi, A.: Mode Collapse in Generative Adversarial Networks: An Overview. In: 8th International Conference on Optimization and Applications (ICOA), pp. 1–6 (2022) doi: 10.1109/ICOA55659.2022.9934291.
7. Li, Z., Liu, F., Yang, W.: A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Transactions on Neural Networks and Learning Systems, 33(12), pp. 6999–7019 (2022) doi: 10.1109/TNNLS.2021.3084827.
8. Liu, S., Zhu, C., Xu, F.: Breast Cancer Immunohistochemical Image Generation through Pyramid Pix2pix. ArXiv (2022) doi: 10.48550/arXiv.2204.11425.
9. Liu, M.Y., Tuzel, O.: The Conditional Adversarial Loss For Image-to-Image Translation. CVPR (2019) doi : 10.48550/arXiv.1611.07004.
10. Mirza, M., Osindero, S: Conditional Generative Adversarial Nets. ArXiv (2014)
11. Pang, Y., Lin, J., Qin, T.: Image-to-Image Translation: Methods and Applications. IEEE Transactions on Multimedia, 24, pp. 3859–3881 (2022) doi: 10.1109/TMM.2021.3109419.
12. Sun, J., Du, Y., Li, C.: Pix2Pix Generative Adversarial Network for Low Dose Myocardial Perfusion SPECT Denoising. Quant Imaging Med Surg (2022) doi: 10.21037/qims-21-1042.
13. Wang, L., Sindagi, V.: High-Quality Facial Photo-Sketch Synthesis Using Multi-Adversarial Networks. In: 13th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 83–90 (2018) doi: 10.1109/FG.2018.00022.