Resumen Ejecutivo

1. Introducción

Los modelos generativos posibilitan que un sistema computacional adquiera un aprendizaje sobre los patrones inherentes de la distribución de los datos empleados en su proceso de entrenamiento. A través de este aprendizaje, estos modelos tienen la capacidad de generar datos que presentan similitudes notables con aquellos utilizados inicialmente para su capacitación, llegando en ciertos casos a ser virtualmente tan verosímiles como los datos de origen. A través de esta clase de modelos, es posible generar datos inéditos con diversas aplicaciones. Una de ellas consiste en la creación de datos que pueden ser empleados por otros sistemas de aprendizaje para su proceso de entrenamiento, prescindiendo de la necesidad de que estos datos hayan sido previamente existentes. Los avances recientes en técnicas de modelado generativo (generative adversarial networks (GANs), variational auto-encoders (VAEs), autoregressive models (ARMs), energy-based models (EBMs), normalizing flow-based models, diffusion models), redes neuronales profundas para la generación de imágenes, han demostrado un gran éxito en la generación de muestras de alta calidad, pero cada uno tiene algunas limitaciones propias, por ejemplo los modelos GANs son conocidos por un entrenamiento potencialmente inestable y una menor diversidad en la generación debido a su naturaleza de entrenamiento adversario. VAE se basa en una pérdida sustitutiva. Los modelos de flujo tienen que utilizar arquitecturas especializadas para construir transformaciones reversibles (Weng, 2021).

Los modelos de difusión han experimentado un éxito generalizado en la generación de imágenes, prevaleciendo sobre las GANs en fidelidad y diversidad, al tiempo que evitan la inestabilidad en el entrenamiento y los problemas de colapso modal. Los modelos autorregresivos, las GANs, los VQ-VAE, los enfoques basados en transformers y los modelos de difusión han logrado avances significativos en la conversión de texto a imagen, incluido el concurrente DALL-E 2, que utiliza una difusión previa en latentes de texto CLIP y modelos de difusión en cascada para generar imágenes de alta resolución de tamaños de 1024 X 1024.

2. Antecedentes de la difusión

Los modelos de difusión representan un paradigma ejemplar de la inteligencia artificial generativa al iniciar con una entrada inicial x^0 y, de manera sistemática, introducir ruido gaussiano en cada capa t hasta alcanzar una capa final denominada x^T . Este enfoque se fundamenta en los principios de la termodinámica de no equilibrio, que define una cadena de Markov mediante pasos de difusión para incorporar gradualmente ruido aleatorio a los datos. Posteriormente, estos modelos aprenden a invertir el proceso de difusión con el fin de construir muestras de datos deseadas a partir del ruido, como se describe en el trabajo titulado *"Diffusion probabilistic models"* (Sohl-Dickstein et al., 2015), "Noise-conditioned score network" (Yang & Ermon, 2019) y "Denoising diffusion probabilistic models" (Ho et al. 2020). En este contexto, los estados evolucionan a lo largo de períodos prolongados mediante procesos de difusión hacia la homogeneidad. El entrenamiento de los modelos de difusión implica aprender a revertir este proceso, buscando generar la imagen original x^0 a partir de

x^T, siendo x^0 la representación inicial de la imagen (Weng, 2021). La etapa de supresión de ruido en los modelos de difusión, al preservar la integridad de la imagen en cada paso, se establece una conexión estrecha y profunda entre los datos y las predicciones, en contraste con los generadores de texto a imagen que no se fundamentan en el proceso de difusión. Como consecuencia de este enfoque, se logra un resultado más fotorrealista en la generación de imágenes por parte de los modelos basados en la difusión.

3. Generación de Imágenes Realistas de Especies de Aves mediante Modelos Generativos de Difusión para Mejorar la Clasificación Automática

3.1. Planteamiento del problema

En el ámbito del aprendizaje profundo, este proyecto aborda desafíos cruciales con aplicaciones interdisciplinarias, impactando áreas como la ornitología, la conservación de la biodiversidad, el ecoturismo y la monitorización ambiental. La esencia del proyecto radica en la generación de imágenes realistas de especies de aves mediante el uso de modelos generativos de difusión, específicamente el enfoque de Imagen (Saharia et al., 2022), la red neuronal de conversión de texto en imágenes de Google para la generación de imágenes fotorrealistas. El objetivo principal es proporcionar conjuntos de datos de entrenamiento sintéticos de alta calidad para modelos de clasificación de aves.

3.2. Contexto:

En la actualidad, la identificación automática de aves a partir de imágenes capturadas en entornos naturales presenta desafíos considerables. La diversidad de especies, las variaciones en la iluminación y el entorno, así como la presencia de elementos no deseados en las imágenes, complican la tarea de desarrollar modelos de clasificación precisos. Este proyecto busca superar estas limitaciones generando imágenes sintéticas de aves que sean visualmente indistinguibles de las fotografías reales y que puedan ser utilizadas en tareas posteriores (por ejemplo, la clasificación de aves).

3.3. Objetivos:

- **Generación de Imágenes Realistas:** Implementar modelos generativos de difusión, específicamente el enfoque de Imagen (Photorealistic text-to-image diffusion models with deep language understanding), para generar imágenes que reproduzcan con alta fidelidad las características visuales de diversas especies de aves (Limitado de acuerdo con los recursos computacionales).

- **Calidad de Datos de Entrenamiento:** Proporcionar conjuntos de datos de entrenamiento de alta calidad que incluyan imágenes generadas sintéticamente, enriqueciendo así la diversidad y representatividad de los datos utilizados para entrenar modelos de clasificación de aves.

- **Aplicaciones Interdisciplinarias:** Facilitar la identificación automática de aves en imágenes capturadas en la naturaleza y contribuir a la monitorización de

poblaciones de aves en entornos naturales, siendo de particular utilidad para instituciones como El Laboratorio de Ornitología de Cornell.

3.4. Importancia:

Este proyecto aborda la necesidad crítica de mejorar la calidad y diversidad de los conjuntos de datos utilizados en la clasificación automática de aves. La generación de imágenes realistas mediante modelos generativos de difusión no solo beneficia a la investigación en ornitología y conservación de la biodiversidad, sino que también tiene aplicaciones prácticas en el ecoturismo y la monitorización ambiental, contribuyendo al entendimiento y preservación de las poblaciones de aves en entornos naturales.

3.5. Descripción del repositorio del proyecto

- **README.md**: Este archivo contiene información general sobre el proyecto y/o repositorio, cómo utilizarlo, y cualquier otra información relevante para los colaboradores o usuarios.

- Dataset: Este directorio contiene los conjuntos de datos utilizados en el proyecto. Puede incluir datos de entrenamiento, validación o prueba para el modelo generativo de difusión.

- **Modelo_Generativo_de_Difusión_para_Generar_Imágenes_de_Aves.ipynb:** Este archivo es un cuaderno Jupyter (Notebook) que contiene el código fuente para el entrenamiento y/o implementación del modelo generativo de difusión Imagen (Saharia et al., 2022), la red neuronal de conversión de texto en imágenes de Google para la generación de imágenes de aves fotorrealistas.

- **INFORME_PROYECTO.pdf:** Este archivo es el informe detallado del proyecto, que incluye la motivación, la metodología, los resultados y cualquier conclusión obtenida en la implementación del modelo Imagen.

- **ENTREGA1.pdf:** Este archivo es una entrega específica del proyecto un informe asociado con una fase inicial (Propuesta inicial).

- **Datos_adicionales:** Este directorio contiene datos adicionales o complementarios que son relevantes para el proyecto, pero no forman parte directa del conjunto de datos principal (Por ejemplo, los checkpoint.pt, Samples).

- **Estructura:**

```
Generative-Diffusion-Model-for-Bird-Image-Generation
├── README.md
├── Dataset
├── Modelo_Generativo_de_Difusión_para_Generar_Imágenes_de_Aves.ipynb
├── INFORME_PROYECTO.pdf
```

├── ENTREGA1.pdf

└── Datos_adicionales

4. **Optimización del Conjunto de Datos para una Mejor Eficiencia Computacional en la Generación Detallada de imágenes de Aves**

En el contexto del entrenamiento del modelo de difusión para la generación detallada de imágenes de aves, se han identificado desafíos inherentes al tamaño del conjunto de datos NABirds V1. La magnitud del conjunto de datos original es de aproximadamente 9 GB y contiene más de 48,000 fotografías anotadas representando 400 especies de aves en América del Norte, esto plantea limitaciones computacionales notables en el proceso de entrenamiento, especialmente cuando se trabaja con recursos computacionales restringidos (GPUs).

- **Visión general del conjunto de datos:** *NABirds V1*, con más de 550 categorías visuales organizadas taxonómicamente, es una colección exhaustiva. Cada una de las 400 especies cuenta con más de 100 fotografías, proporcionando una representación integral de la diversidad aviar. Las anotaciones detalladas, que incluyen clasificaciones específicas para machos, hembras y juveniles, contribuyen a una colección sofisticada de 700 categorías visuales (Para más información consultar El Laboratorio de Ornitología de Cornell).

- **Estrategia de optimización:** Conscientes de la necesidad de optimizar los recursos computacionales, se ha implementado una estrategia que está centrada en la creación de un subconjunto más manejable (de menor tamaño) del conjunto de datos. El enfoque se ha dirigido específicamente hacia las categorías *"Allen's Hummingbird (Adult Male)"* y *"Allen's Hummingbird (Female, immature)"*, seleccionadas por su relevancia para los objetivos de generación sintética de imágenes detalladas de aves y a su vez el interés en conservar la uniformidad en el conjunto de datos, que permitan que el modelo de red neuronal converja más rápido.

  Nota: En la carpeta dataset se adjunta un archivo llamado classes.txt donde se especifican todas las especies de los subdirectorios del dataset original (enlace de descarga del dataset original https://dl.allaboutbirds.org/nabirds en el caso de que se requieran realizar otros experimentos).

- **Subconjunto:** Para abordar las limitaciones computacionales, se ha implementado un script en Python para extraer y consolidar selectivamente algunas carpetas del conjunto de datos original. Este proceso ha dado lugar a la creación de un nuevo conjunto de datos, reduciendo significativamente su tamaño sin comprometer la integridad de las categorías seleccionadas. Este nuevo conjunto de datos se compone exclusivamente de las categorías especificadas de colibríes, lo que facilita un proceso de entrenamiento más ágil y adaptado a las restricciones computacionales existentes, sin perder de vista los objetivos originales del trabajo aplicado.

5. Arquitectura

- **Codificadores de texto preentrenados:** El modelo de difusión Imagen maximiza el potencial de los codificadores de texto preentrenados (BERT, T5, CLIP) para la síntesis texto-imagen, desviándose de los modelos convencionales entrenados en datos imagen-texto. La congelación de los pesos de los codificadores ofrece ventajas computacionales, y el aumento del tamaño de los codificadores de texto mejora significativamente la calidad de la generación de texto a imagen. Para este caso específico se utilizó la arquitectura de Google T5v1.1 “google/t5-v1_1-large” (Raffel et al., 2020).

  Nota: El modelo y sus pesos se puede encontrar en el siguiente enlace: https://huggingface.co/google/t5-v1_1-large

- **T5 (Text-to-Text Transfer Transformer):** Es una arquitectura basada en Transformer que utiliza un enfoque de texto a texto. Cada tarea, ya sea traducción, respuesta a preguntas o clasificación, se formula como la alimentación del modelo con texto como entrada y entrenándolo para generar algún texto objetivo. Esto permite el uso del mismo modelo, función de pérdida, hiperparámetros, etc., en un conjunto diverso de tareas. Los cambios en comparación con BERT incluyen: i) Adición de un decodificador causal a la arquitectura bidireccional. ii) Reemplazo de la tarea cloze de llenar espacios en blanco con una combinación de tareas de preentrenamiento alternativas (Raffel et al., 2020).

- **Modelos de difusión y guía sin clasificador:** El modelo de difusión Imagen adopta introduce el guiado sin clasificador, evitando los problemas que plantean los pesos de guiado elevados. El umbral dinámico evita activamente la saturación de píxeles, lo que se traduce en un fotorrealismo superior. El modelo logra un equilibrio eficaz, mejorando la alineación imagen-texto y evitando al mismo tiempo la degradación de la calidad asociada a métodos anteriores.

- **Robustos modelos de difusión en cascada:** La robusta arquitectura de Imagen incluye un modelo base de 64 × 64 y dos modelos de difusión de superresolución condicionados al texto. El aumento del acondicionamiento del ruido mejora la fidelidad de la imagen, y la variante Efficient U-Net (Baheti et al., 2020) garantiza una mayor eficiencia de memoria. Los modelos en cascada, conscientes de los niveles de ruido, contribuyen a generar imágenes de alta calidad en todas las resoluciones. Para este caso en específico no se utilizó el modelo de U-net de super resolución debido a los límites en los recursos computacionales.

  A continuación, se detalla el modelo de U-Net instanciado utilizado para el entrenamiento y sus hiperparametros

```
# unets for unconditional imagen
unet = Unet(
    dim = 128,
    dim_mults = (1, 2, 4, 8),
    num_resnet_blocks = (2, 4, 8, 8),
```

```
    layer_attns = (False, False, False, True),
    layer_cross_attns = (False, False, False, True),
    attn_heads = 8
)
```

Se optó por una dimensión base de 128 en lugar de 64 debido a la falta de resultados satisfactorios y a problemas recurrentes al utilizar la dimensión más pequeña (64), con el objetivo de minimizar el uso de memoria. Al finalizar los entrenamientos, se observó la generación de imágenes con niveles significativos de ruido, indicando una falta de convergencia del modelo. Además, la función de pérdida exhibía oscilaciones constantes entre valores similares. Este ajuste en la dimensión base se implementó para abordar estos desafíos y mejorar la estabilidad y la calidad de los resultados obtenidos durante el proceso de entrenamiento.

U-Net (Ronneberger et al., 2015): Es una arquitectura popular para la segmentación semántica, cuya idea principal radica en realizar downsampling progresivo y luego upsampling de la imagen de entrada, añadiendo conexiones de omisión entre capas que tienen la misma resolución. Estas conexiones ayudan con el flujo de gradientes y evitan introducir un cuello de botella en la representación, a diferencia de los autoencoders convencionales. Por lo tanto, los modelos de difusión son autoencoders de eliminación de ruido sin un cuello de botella. La red recibe dos entradas: las imágenes ruidosas y las varianzas de sus componentes de ruido. Se requiere esta última ya que la eliminación de ruido de una señal implica diferentes operaciones en diferentes niveles de ruido. Transformamos las varianzas de ruido utilizando embeddings sinusoidales, de manera similar a las codificaciones posicionales utilizadas tanto en transformers como en NeRF. Esto ayuda a que la red sea altamente sensible al nivel de ruido, lo cual es crucial para un rendimiento óptimo. Implementamos los embeddings sinusoidales mediante una capa Lambda.

6. Descripción de las iteraciones

Se entrenó el modelo de Imagen para generar imágenes de aves fotorrealistas sin texto de tamaño → 64 × 64 sin superresolución (Límite en recursos computacionales). Se utiliza un tamaño de lote de 4 y 20K pasos de entrenamiento para el modelo. Se utilizó una GPU-v4 para el entrenamiento del modelo base de 64 × 64. Todos los entrenamientos fueron realizados en una máquina virtual de Google Colab.

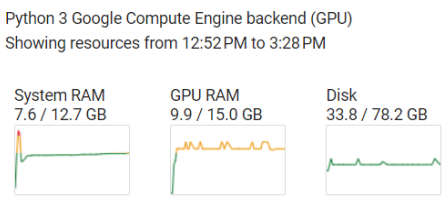

Figura 1. Uso de los recursos de Google Colab

7. Descripción de los resultados.

Kernel Inception Distance (KID): Es una métrica de calidad de imágenes propuesta como alternativa a la popular Distancia de Inception de Frechet (FID). Se prefiere esta métrica KID sobre FID debido a su implementación más sencilla, su capacidad para ser estimada por lote y su menor carga computacional. En este ejemplo, las imágenes se evalúan en la resolución mínima posible de la red Inception (64x64) y la métrica solo se mide en el conjunto de validación por eficiencia computacional. También limitamos el número de pasos de muestreo en la evaluación a 5 por la misma razón. Dado que el conjunto de datos es relativamente pequeño, recorremos las divisiones de entrenamiento y validación varias veces por época. Esto se debe a que la estimación de KID es ruidosa y requiere muchos cálculos.

8. Conclusiones

- **Evolución del Campo y Desafíos en la Evaluación de Modelos Generativos:** Con el avance continuo en el campo y el desarrollo de modelos cada vez más impresionantes y creativos, se observa una creciente limitación en la confiabilidad de los métodos de evaluación actuales discutidos y utilizados en el proyecto. Las métricas seleccionadas, especialmente la fidelidad y la alineación de píxeles en las imágenes, pueden favorecer potencialmente los aspectos fuertes de la imagen generada. Esto destaca la necesidad urgente de adoptar un conjunto más amplio y diverso de criterios de evaluación.

  Es crucial reconocer que las métricas actuales pueden no ser completamente representativas de la calidad y la diversidad de las imágenes generadas, lo que plantea desafíos significativos en la evaluación de modelos generativos. La preferencia por métricas tradicionales como la fidelidad y la alineación de píxeles puede no capturar de manera adecuada la complejidad y la riqueza de las imágenes generadas por modelos avanzados.

- **Limitaciones Recursos Computacionales:** A pesar del reconocimiento de la necesidad de un conjunto más amplio de criterios de evaluación, la implementación de esta idea se ve obstaculizada por la escasez de recursos computacionales. La diversificación y ampliación de las métricas de evaluación requieren un consumo significativo de recursos, lo que puede ser impracticable dado el entorno computacional limitado. En consecuencia, se plantea un dilema en la evaluación de modelos generativos: la necesidad de métricas más abarcadoras versus las limitaciones de recursos disponibles. En este contexto, se podría explorar estrategias alternativas, como la optimización de métricas existentes o la consideración de métodos de evaluación menos intensivos en recursos, sin comprometer la calidad de la evaluación

- **Desafíos para Captar la Complejidad:**

  - **Entrenamientos con el Dataset NABirds V1 y Limitaciones Computacionales:** El proyecto enfrenta desafíos significativos al llevar a cabo entrenamientos con el conjunto de datos NABirds V1. La principal complicación radicó en las limitaciones

de recursos computacionales disponibles. La falta de capacidad computacional dificultó que el modelo de difusión aprendiera de manera efectiva los patrones complejos presentes en los datos. Este obstáculo se manifestó a través de tiempos prolongados de entrenamiento, durante los cuales el modelo no lograba generar imágenes de aves coherentes con las clases observadas en el conjunto de datos.

- **Impacto en la Generación de Imágenes:** La escasez de recursos computacionales influyó negativamente en la capacidad del modelo de difusión para comprender y reproducir las características distintivas de las clases de aves presentes en el conjunto de datos. Como resultado, el tiempo prolongado de entrenamiento no se tradujo en la generación exitosa de imágenes realistas y coherentes con las clases definidas en el NABirds V1.

9. Trabajos futuros

- **Entrenamiento con Dataset NABirds V1 de Forma Condicionada por Texto:** Explorar la posibilidad de entrenar el modelo de difusión utilizando todo el conjunto de datos NABirds V1 de forma condicionada por texto. Esto implica un preprocesamiento de los datos para incorporar incrustaciones de texto, como los nombres científicos y/o comunes de las aves. Investigar técnicas eficientes para la incorporación de información textual en el proceso de generación de imágenes, mejorando así la capacidad del modelo para generar resultados coherentes y específicos en respuesta a las condiciones dadas.

- **Optimización de la Arquitectura del Modelo:** Realizar experimentos con diferentes arquitecturas de modelos de difusión, buscando optimizar la estructura para adaptarse mejor a las características del conjunto de datos y a las limitaciones computacionales. Explorar variantes de modelos generativos que puedan ser más eficientes en términos de consumo de recursos sin comprometer la calidad de generación de imágenes.

- **Transferencia de Aprendizaje:** Investigar y aplicar técnicas de transferencia de aprendizaje para aprovechar conocimientos previos obtenidos en tareas similares o en conjuntos de datos más grandes. Esto podría ayudar a mejorar el rendimiento del modelo con recursos limitados.

- **Exploración de Conjuntos de Datos Reducidos pero Representativos:** Evaluar la viabilidad de trabajar con conjuntos de datos más pequeños pero que conserven la representatividad esencial de las clases de aves. Esto podría facilitar un entrenamiento más efectivo en entornos con recursos computacionales limitados.

- **Incorporación de Otras Métricas de Evaluación:** Ampliar el conjunto de métricas de evaluación para capturar mejor la calidad y la diversidad de las imágenes generadas, superando las limitaciones de métricas tradicionales como la fidelidad y la alineación de píxeles.

Referencias

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, *35*, 36479-36494.

Weng, Lilian. (Jul 2021). What are diffusion models? Lil'Log. https://lilianweng.github.io/posts/2021-07-11-diffusion-models/.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015, June). Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning (pp. 2256-2265). PMLR.

Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems, 32.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, 6840-6851.
Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 35, 36479-36494.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1), 5485-5551.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.

Baheti, B., Innani, S., Gajre, S., & Talbar, S. (2020). Eff-unet: A novel architecture for semantic segmentation in unstructured environment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 358-359).

Bińkowski, M., Sutherland, D. J., Arbel, M., & Gretton, A. (2018). Demystifying mmd gans. arXiv preprint arXiv:1801.01401.

Lucidrains. (2022). imagen-pytorch. GitHub. https://github.com/lucidrains/imagen-pytorch/tree/main