UNIVERSIDAD NACIONAL DE LA PATAGONIA SAN JUAN BOSCO

FACULTAD DE INGENIERÍA

TESINA PARA OBTENER EL TÍTULO DE:
Licenciatura en Informática

Personalización de un modelo text-to-image para la generación de imágenes con Inteligencia Artificial.

ALUMNO: Pablo Dibez[1]

TUTORA: Mg.Viviana Mercado
CO-TUTOR: Lic.Daniel Ormachea

Comodoro Rivadavia (Chubut) - Argentina

Junio de 2024

[1]pdibez@gmail.com

Resumen

En los últimos años, el ámbito de la Inteligencia Artificial ha experimentado un crecimiento significativo, destacándose en la generación de imágenes. Estos avances han incrementado el interés en las oportunidades de automatización en diversos campos y las restricciones asociadas.

La técnica de generación de imágenes a partir de texto, conocida como texto a imagen, ha emergido como un área de investigación prometedora con aplicaciones en entretenimiento, diseño gráfico, publicidad y creación de contenido visual. Esta tecnología permite a los usuarios convertir descripciones textuales en imágenes pertinentes y realistas.

Este estudio explora a fondo el estado del arte de los modelos generativos de texto a imagen, resaltando sus capacidades y limitaciones. Se han investigado técnicas especializadas para la personalización de modelos orientada a sujetos específicos, facilitando una adaptación más precisa del aprendizaje.

El marco teórico aborda la clasificación de los modelos generativos, los modelos de difusión y los de texto a imagen, detallando sus mecanismos fundamentales y aplicaciones impactantes en áreas como el arte, la medicina y los medios de comunicación. Esta base teórica proporciona una comprensión profunda del potencial transformador de estas tecnologías.

En la parte experimental, se ha demostrado la aplicación práctica de estas teorías a través del reentrenamiento del modelo Stable Diffusion con la técnica de ajuste fino, utilizando Dreambooth y personalizándolo para generar imágenes de Lilo, una gata. Este proceso ha confirmado la eficacia del ajuste fino con un número reducido de imágenes y ha proporcionado un modelo robusto para generar imágenes de gran calidad y fidelidad.

Palabras claves: Inteligencia Artificial, Generación de Imágenes, Texto a Imagen, Modelos Generativos, Ajuste Fino

Abstract

In recent years, the field of artificial intelligence has experienced significant growth, particularly in image generation. These advancements have increased interest in automation opportunities across various fields and associated constraints.

The technique of generating images from text, known as de texto a imagen, has emerged as a promising research area with applications in entertainment, graphic design, advertising, and visual content creation. This technology enables users to convert textual descriptions into relevant and realistic images.

This study thoroughly explores the state-of-the-art in generative models of de texto a imagen, highlighting their capabilities and limitations. Specialized techniques for customizing models aimed at subject-driven have been investigated, facilitating a more precise adaptation of learning.

The theoretical framework addresses the classification of generative models, diffusion models, and de texto a imagen models, detailing their fundamental mechanisms and impactful applications in areas such as art, medicine, and media. This theoretical foundation provides a deep understanding of the transformative potential of these technologies.

In the experimental part, the practical application of these theories has been demonstrated through the retraining of the Stable Diffusion model with fine-tuning technique, using Dreambooth and customizing it to generate images of Lilo, a cat. This process has confirmed the effectiveness of fine-tuning with a reduced number of images and has provided a robust model for generating high-quality and faithful images.

Keywords: Artificial Intelligence, Image Generation, Text to Image, Generative Models, Fine-Tuning

Agradecimientos

A Dios, quien me dió todo y me sorprende cada día de mi vida con tantas cosas grandiosas y que nunca hubiera imaginado.

A la Patria y su fuente de oportunidades únicas; como la Universidad pública y gratuita que me posibilitó el acceso al estudio y a seguir creciendo en el mundo de conocimiento. Considero hoy mas que nunca que es necesario defenderla para mantenerla como baluarte para las próximas generaciones.

A mi familia: a mi mamá Norma y mi papá José, que con gran esfuerzo y aún tiempos difíciles, cultivaron mi opción por el estudio y me apoyaron a continuar en tal camino. Y a mis hermanos Fede e Igna quienes también fueron pilares fundamentales en continuar hacia este objetivo.

A mis compañeros que con el transcurso de las cursadas se convirtieron en grandes amigos: Maca, Martín y Santiago; con quienes compartimos días y noches de estudio. Mas allá de lo académico supimos forjar un gran vínculo que hasta el día de hoy mantenemos.

A Claudia, quien siempre fue generosa y flexible al momento de solicitar tiempo y días en el trabajo, para asistir a las cursadas y rendir exámenes.

A mi tutora Viviana y mi cotutor Daniel, quienes me guiaron y ayudaron a la largo de este trabajo.

Índice general

Parte I

Introducción

Capítulo 1

Introducción

En los últimos años, el ámbito de la *Inteligencia Artificial* (AI) ha experimentado un crecimiento significativo en múltiples áreas, abarcando también la generación de imágenes. Estos avances han despertado un interés cada vez mayor en las oportunidades de automatización en diversos campos de aplicación, así como en las limitaciones y restricciones asociadas.

Específicamente, el aprovechamiento de la AI para crear imágenes a partir de texto implica la habilidad de generar imágenes que concuerden con la descripción ingresada, utilizando un modelo que comprenda la información proporcionada. Este procedimiento se fundamenta en el entrenamiento previo del modelo, utilizando conjuntos de datos específicos diseñados con ese propósito.

La generación de imágenes a partir de texto (también conocida como *text-to-image* en inglés), ha surgido como un campo de investigación prometedor. Esta técnica halla aplicaciones en diversos ámbitos, como la industria del entretenimiento, el diseño gráfico, la publicidad y la creación de contenido visual, entre otros (Yang et al., 2023); permitiendo a los usuarios expresar ideas visuales a través de texto y obtener imágenes pertinentes y realistas como respuesta.

En el transcurso de este trabajo, se tratarán los desafíos y las limitaciones que se presentan al generar imágenes a partir de texto, además de las soluciones propuestas en la literatura científica. Por un lado, se llevará a cabo un estudio exhaustivo de los métodos y las aplicaciones de los *Modelos de Difusión* (DM) (Yang et al., 2023), mientras que por otro lado, se analizará

la generación de imágenes de alta resolución utilizando *Modelos de Difusión Latentes* (LDM) (Rombach et al., 2022)).

En el plano aplicado, esta tesina adoptará un enfoque centrado en la personalización de los DM de texto a imagen basados en un sujeto específico mediante la técnica de *Ajuste Fino* (FT) (Ruiz et al., 2023).

En resumen, esta tesina se concentra en el estudio de la generación de imágenes a partir de texto, utilizando la AI como herramienta principal y enfocándose en la personalización de un modelo de difusión con dicho propósito. A continuación, se presentará la motivación que impulsa esta investigación y los objetivos específicos que se buscarán alcanzar.

1.1. Motivación

Tomando en cuenta lo expresado anteriormente, antes de entrar en los desafíos e iniciativas sobre el uso de AI para generar imágenes, es necesario comprender cómo funcionan realmente estos modelos en su interior llevando a cabo una investigación de lo que se conoce hasta el momento.

Seguidamente es menester apropiarse de estas técnicas y evaluar problemas prácticos concretos con una implementación de estas tecnologías.

Un aspecto a tener en cuenta es que generalmente, los resultados de las AI están restringidos a las entradas en las que fueron entrenadas, y las correspondientes a generación de imágenes no escapan de esta limitación.

Sin embargo, es posible tomar un enfoque para personalizar estos modelos e incorporar nuevos sujetos al conocimiento original, obteniendo un nuevo modelo y seguir ampliando las posibilidades de creación. Esto significa que se pueden crear nuevas imágenes cómo lo hacían originalmente y también de los sujetos que se hayan incorporado al modelo en el proceso aplicado (conocido como FT). Esto último es lo que también se pretende llevar a cabo en el presente proyecto de tesina.

1.2. Objetivos

1.2.1. Objetivo General

- Realizar una personalización de un modelo de texto a imagen con un *conjunto de datos* (conocido en inglés como *dataset*) propio para generar imágenes con AI.

1.2.2. Objetivos Específicos

- Relevar y comparar las principales técnicas de los modelos de texto a imagen basados en DM.
- Realizar una revisión sistemática del estado del arte de los modelos de texto a imagen basados en DM.
- Identificar el impacto de la utilización de modelos de texto a imagen para la generación de imágenes.
- Realizar una personalización de un modelo de texto a imagen para introducir un sujeto nuevo en su conocimiento aplicando la técnica FT.
- Generar imágenes del nuevo sujeto introducido en el modelo de texto a imagen personalizado.

1.3. Desarrollos Propuestos

Los desarrollos propuestos para el presente proyecto son:

- Estudiar los modelos de generación de imágenes de texto a imagen basados en difusión para analizar y evaluar sus capacidades.
- Elaborar un conjunto de datos consistente de imágenes de un el nuevo sujeto a agregar en el modelo.
- Aplicar la técnica de FT para personalizar el modelo *Stable Diffusion* (SD) con el conjunto de datos previamente creado.

- Desplegar una aplicación (de manera temporal por cuestiones de consumo de servicios), a fin de poder probar el modelo que se generará en el proyecto.
- Dejar a disposición el conjunto de datos y el modelo personalizado como contribución científica al Departamento de Informática.

1.4. Resultados Esperados

Los resultados esperados para el presente proyecto son:

- Generar un nuevo modelo de texto a imagen. El mismo estará basado en SD e incorporará un nuevo sujeto al modelo original.
- A partir de las imágenes resultantes, desarrollar diversas situaciones y contextos que se puedan imaginar ingresándolas como prompt.
- Experimentar y evaluar los modelos obtenidos previamente, analizando su viabilidad y efectividad.

1.5. Metodología Propuesta

Este trabajo explorará a fondo el estado del arte de los modelos generativos de texto a imagen y se investigarán técnicas especializadas para la personalización de modelos orientada a sujetos específicos.

Desde el marco teórico se abordará la clasificación de los modelos generativos, los DM y los modelos de texto a imagen, explicando sus mecanismos fundamentales y sus aplicaciones más impactantes en campos tan diversos como el arte, la medicina y los medios de comunicación.

En la fase experimental, se aplicarán prácticamente estas teorías mediante el reentrenamiento del modelo SD. Se empleará la técnica de FT junto con Dreambooth, y se utilizará Google Colab como plataforma para ejecutar los scripts necesarios.

Posteriormente, el modelo se desplegará en una plataforma para realizar pruebas interactivas que evalúen su capacidad de generar representaciones

de alta fidelidad. Finalmente, se medirá el desempeño del modelo utilizando métricas específicas de fidelidad de imagen. Los resultados obtenidos servirán para determinar la validez de las técnicas utilizadas tanto para la personalización como para la evaluación del modelo.

1.6. Estructura del Desarollo

La estructura y organización del desarrollo se definirá de la siguiente manera:

Parte I - Introducción

- *Capítulo 1 - Introducción:* Este capítulo define el tema a investigar e implementar. Se expone la motivación detrás del estudio, se resaltan las problemáticas a abordar, los objetivos, los desarrollos propuestos, los resultados esperados y la metodología que se seguirá.
- *Capítulo 2 - Estado del Arte:* Se describirán y analizarán trabajos previos que hayan realizado aportes significativos en el ámbito de los modelos generativos de texto a imagen y modelos dirigidos por sujetos.

Parte II - Marco Teórico

- *Capítulo 3 - Marco Teórico:* Se analizarán y describirán los conceptos clave de los modelos generativos, incluyendo la clasificación de estos modelos, los DM y los modelos de texto a imagen. También se explicarán sus mecanismos fundamentales y sus aplicaciones más impactantes.

Parte III - Desarrollo y Experimentos

- *Capítulo 4 - Experimentación:* Este capítulo describirá las herramientas, modelos y librerías que se utilizarán; y detallará el proceso de experimentación del estudio, incluyendo la preparación del conjunto de datos y el FT realizado para generar el nuevo modelo, así como la evaluación posterior del mismo.

Parte IV - Conclusiones y Trabajo Futuro

- *Capítulo 5 - Conclusiones y Trabajo Futuro:* Este capítulo resume los principales resultados obtenidos en este trabajo. Además, se sugieren posibles líneas de investigación y desarrollo a futuro que podrían expandir y mejorar los hallazgos actuales.

- *Anexos A:* Los anexos proporcionan detalles adicionales sobre las abreviaturas y un diccionario de términos con breves explicaciones de los mismos. La lectura de estos anexos no es necesaria para comprender los conceptos principales abordados en esta tesina, pero pueden ser útiles para aquellos que deseen profundizar en los términos específicos utilizados.

Capítulo 2

Estado del Arte

En este capítulo, se explorarán minuciosamente las investigaciones más destacadas y los desarrollos seleccionados como casos de estudio en el campo de la generación de imágenes con AI. Se pondrá un énfasis especial en el uso de DM y en la aplicación de diversas técnicas innovadoras en este ámbito en constante evolución.

Inicialmente se llevará a cabo una exhaustiva introducción a los modelos de texto a imagen, donde se analizará detalladamente su funcionamiento y arquitectura subyacente. Se discutirá cómo estos modelos están revolucionando la forma en que se generan imágenes a partir de descripciones textuales, anticipando su amplia aplicación en campos como la publicidad, el entretenimiento y la educación.

Más adelante, se realizará un análisis profundo de tres casos de estudio representativos en el ámbito de los modelos de texto a imagen: *Imagen*, *DALL-E 2* y SD. Se examinarán sus características clave, su rendimiento en tareas específicas y su relevancia en el panorama actual de la investigación en AI aplicada a la generación de imágenes.

Posteriormente, se ofrecerá una breve introducción al concepto de FT en modelos de texto a imagen. Se explorarán las posibilidades que ofrece esta técnica para adaptar los modelos preentrenados a dominios específicos o tareas particulares, lo que permitirá mejorar su rendimiento y adaptabilidad en escenarios reales.

Por último, se analizarán los casos de DB y *SuTi*, dos ejemplos destacados de cómo el FT puede potenciar aún más la capacidad de los modelos de texto a imagen para generar contenido visual de alta calidad y relevancia.

Con esta amplia revisión del estado del arte en la generación de imágenes con AI, se pretende ofrecer una visión panorámica de las últimas tendencias y avances en este emocionante campo de investigación, así como identificar posibles áreas de desarrollo futuro y aplicación práctica.

2.1. Modelos de Texto a Imagen

Los modelos de texto a imagen son sistemas de AI que generan imágenes a partir de descripciones textuales (Foster, 2023). Estos modelos son una aplicación específica de la generación de imágenes por parte de sistemas de AI, y han ganado interés en campos como la *Visión por Computadora* (CV) y el *Procesamiento del Lenguaje Natural* (NLP).

Las características principales de los modelos de texto a imagen son:

- *Generación de imágenes basada en texto:* Los modelos de texto a imagen toman como entrada descripciones textuales, como frases o párrafos, y generan imágenes correspondientes a esas descripciones.
- *Uso de redes neuronales:* Estos modelos generalmente se basan en arquitecturas de redes neuronales profundas, que pueden ser modelos generativos como *Redes Generativas Adversarias* (GAN) o modelos de atención como *Transformers*[1] (T) (Elgendy, 2020).
- *Transferencia de conocimiento:* Algunos modelos de texto a imagen pueden ser preentrenados en grandes conjuntos de datos para aprender representaciones de alto nivel de las relaciones entre texto e imágenes, lo que les permite generar imágenes realistas y detalladas a partir de descripciones textuales.
- *Adaptabilidad:* Los modelos de texto a imagen suelen ser adaptables a diferentes dominios y tareas mediante técnicas como el FT, que les permite ajustarse a conjuntos de datos específicos o a requisitos de generación de imágenes particulares (Babcock & Bali, 2021).

[1] Un transformer es una red neuronal que aprende contexto y, por lo tanto, significado mediante el seguimiento de relaciones en datos secuenciales como las palabras de una oración.

- *Aplicaciones prácticas:* Estos modelos tienen una amplia gama de aplicaciones prácticas, incluyendo la generación automática de contenido multimedia, la creación de arte digital, la asistencia en el diseño y la producción de contenido para la web y las redes sociales.

2.1.1. Imagen

Ante todo, es relevante mencionar el trabajo titulado *"Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding"* de Chitwan et al. (2022), que dio origen a *Imagen*, un DM [2] de texto a imagen reconocido por su notable realismo fotográfico y su profundo entendimiento del lenguaje.

Pueden apreciarse ejemplos de los resultados obtenidos con *Imagen* en la Figura 2.1.

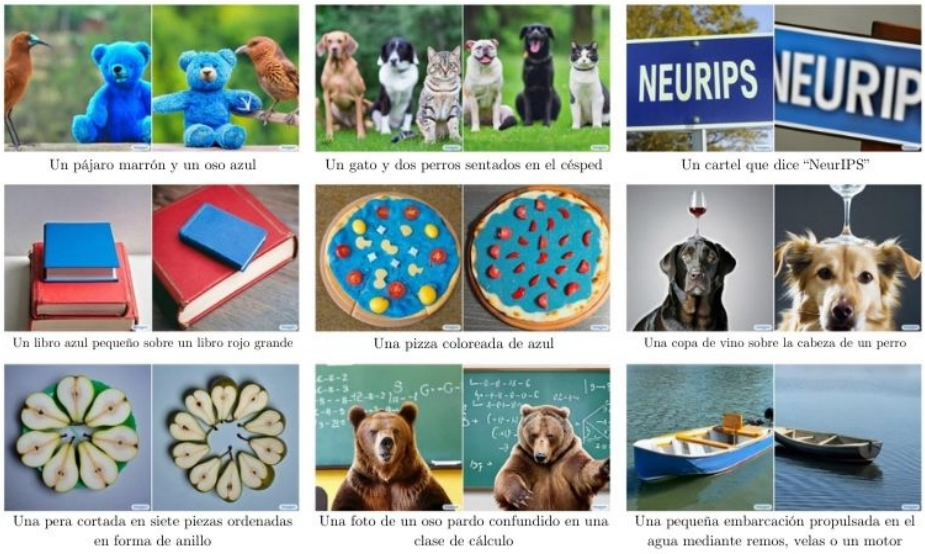

Figura 2.1. Comparación entre imágenes de conjunto de validación de DrawBench y las generadas por Imagen. Gráfico extraído de Chitwan et al. (2022)

El modelo *Imagen* se compone de un codificador de texto que transforma el texto en una secuencia de representaciones y una serie de DM condicional

[2]Los DM son una familia de modelos generativos probabilísticos que destruyen progresivamente los datos mediante la inyección de ruido, para luego aprender a revertir este proceso para la generación de muestras. Para obtener más detalles, consultar la sección 3.5

que convierten estas representaciones en imágenes de mayor resolución. Es decir, este enfoque capitaliza la capacidad de los grandes modelos de lenguaje tipo T para comprender el texto y se basa en la eficacia de los DM para generar imágenes altamente fidedignas.

Lo más revelador de *Imagen* es la eficacia de los modelos de lenguaje genéricos de gran tamaño, como T5 [3], que están entrenados exclusivamente en texto; en la codificación de texto para la generación de imágenes. Esto implica que al aumentar el tamaño del modelo de lenguaje se evidencia un impacto mucho mayor en la calidad de las muestras y por otro lado en la coherencia entre imágenes y texto; que simplemente aumentar el tamaño del modelo de difusión de imágenes.

Imagen también se destaca por su énfasis en la orientación sin clasificador y presenta el concepto de umbral dinámico, que permite utilizar pesos de orientación muy grandes. Estos elementos innovadores permiten que *Imagen* genere muestras de 1024x1024 con un asombroso realismo fotográfico y una fuerte correspondencia con el texto.

Para evaluar el rendimiento de *Imagen*, se utilizaron varios conjuntos de datos, como el conjunto de validación de COCO [4] y DrawBench [5], propuesto como alternativa a COCO.

En cuanto a las métricas de evaluación, *Imagen* se sometió a pruebas de FID[6] para medir la *fidelidad* de la imagen (Figura 2.2), en particular, con FID-30K[7] en un escenario de cero disparos [8].

[3]T5 es un modelo Transformer de Google que utiliza la estructura codificador-decodificador. Se caracteriza por redefinir una variedad de tareas en un marco de trabajo de texto a texto, que incluye traducción, aceptabilidad lingüística, similitud de oraciones y resumen de documentos (Raffel et al., 2019).

[4]COCO es un estándar en la evaluación de modelos de texto a imagen

[5]DrawBench comprende 11 categorías que ponen a prueba diversas capacidades de los modelos, incluida la representación precisa de colores, números de objetos, relaciones espaciales, texto en la escena e interacciones inusuales entre objetos.

[6]FID (del inglés Frechet Inception Distance) es una métrica popular utilizada para evaluar la calidad de las imágenes generadas por redes GAN. Mide la distancia entre las distribuciones gaussianas multivariadas del conjunto de datos de imágenes generadas y los datos reales que el GAN trata de replicar (de Deijn et al., 2024)

[7]FID que utiliza 30.000 imágenes

[8]escenario de aprendizaje automático en el que un modelo de AI se entrena para reconocer y categorizar objetos o conceptos sin haber visto ejemplos previos de esas categorías o conceptos.

Modelo	FID-30K	Tiro Cero FID-30K
AttnGAN	35.49	
DM-GAN	32.64	
DF-GAN	21.42	
DM-GAN + CL	20.79	
XMC-GAN	9.33	
LAFITE	8.12	
Make-A-Scene	7.55	
DALL-E		17.89
LAFITE		26.94
GLIDE		12.24
DALL-E 2		10.39
Imagen		**7.27**

Figura 2.2. Resultados de la evaluación Imagen con respecto a otros modelos con la métrica FID utilizando COCO como conjunto de datos. Entre los modelos analizados figuran GLIDE y DALL-E. Tabla extraída de Chitwan et al. (2022)

También se empleó la puntuación *CLIP-I*[9] para evaluar la coherencia entre imagen y texto, además de una evaluación humana (Figura 2.3) para contrarrestar las limitaciones de las métricas anteriores. Los evaluadores humanos compararon las imágenes generadas por *Imagen* con las imágenes de COCO y calificaron la *calidad* de la imagen y la *similitud del texto*, con resultados que respaldan la *excelencia* de *Imagen* en términos de alineación imagen-texto.

Modelo	***Fotorealismo*** ↑	***Alineación*** ↑
Original		
Original	50.0%	91.9 ± 0.42
Imagen	$39.5 \pm 0.75\%$	91.4 ± 0.44
Sin Personas		
Original	50.0%	92.2 ± 0.54
Imagen	$43.9 \pm 1.01\%$	92.1 ± 0.55

Figura 2.3. Resultados de la evaluación humana de imágenes generadas con Imagen, separando el conjunto de datos sin personas y completo, analizando las métricas de alineación con el texto y fotorrealismo. Tabla extraída de Chitwan et al. (2022)

Finalmente, en las comparaciones directas con otros modelos recientes (Figura 2.4) utilizando Drawbench, se incluyeron VQ-GAN+CLIP, LDM y

[9]CLIP-I representa la similitud promedio de cosenos entre las incrustaciones de imágenes generadas por CLIP y las imágenes reales

DALL-E 2; *Imagen* se destacó positivamente en términos de calidad de las muestras y alineación entre imagen y texto, según la preferencia de los evaluadores humanos.

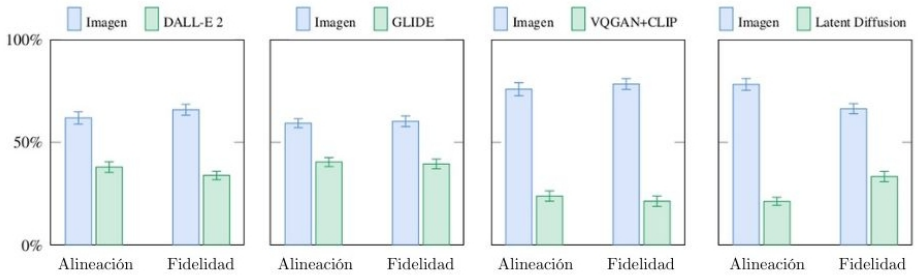

Figura 2.4. Resultados con respecto a la comparación de Imagen con otros modelos utilizando Drawbench, notando la superioridad en cuanto a fidelidad y alineación. Gráfico extraído de Chitwan et al. (2022)

2.1.2. DALL-E 2

En el trabajo titulado *"Hierarchical Text-Conditional Image Generation with CLIP Latents"* de Aditya et al. (2022), se introduce un modelo llamado unCLIP [10], también conocido en la industria como DALL-E 2, que revoluciona la generación de imágenes.

Este enfoque consta de dos etapas fundamentales: un modelo previo denominado *prior* que genera una representación de imagen CLIP [11] basada en una descripción de texto y un *decodificador* que crea una imagen vinculada a esta representación de imagen, logrando resultados excepcionales con una pérdida mínima de fotorrealismo y similitud de descripción.

Por un lado, se debe destacar el papel crucial de los modelos contrastivos como CLIP, que han demostrado la capacidad de aprender representaciones sólidas de imágenes, capturando tanto la semántica como el estilo de un

[10]unCLIP propone un modelo de dos etapas: un *prior* que genera una incrustación de imagen CLIP dado un texto, y un *decodificador* que genera una imagen condicionada a la incrustación de imagen. Se lo llama así porque representa una versión "des-CLIPada" (es decir que deshace) del modelo CLIP original

[11]CLIP es una red neuronal que aprende eficientemente conceptos visuales a partir de supervisión en lenguaje natural. CLIP se puede aplicar a cualquier prueba de clasificación visual simplemente proporcionando los nombres de las categorías visuales a reconocer (Radford et al., 2021)

texto. Las incrustaciones de CLIP presentan propiedades deseables, como su resistencia a cambios en la distribución de imágenes y su capacidad de realizar tareas de tiro cero de manera impresionante, destacándose en una amplia variedad de desafíos relacionados con visión y lenguaje.

En contraste, se exploran los DM en el *decodificador*, comparando modelos autoregresivos ARM con DM para el modelo previo. Los resultados indican que estos últimos son computacionalmente más eficientes y generan muestras de mayor calidad. Es decir, los DM han emergido como una base prometedora en la creación de modelos generativos, liderando en el estado del arte en la creación de imágenes y videos.

La fusión de estos dos enfoques resuelve eficazmente el desafío de la generación de imágenes basada en texto. Inicialmente se capacita un decodificador de difusión para invertir la codificación de imágenes de CLIP. Este inversor no determinista, puede generar múltiples imágenes correspondientes a una representación de una imagen dada. Esta capacidad de codificar y decodificar imágenes permite una exploración más allá de la traducción de texto a imagen, revelando las características reconocidas o ignoradas por CLIP.

La estructura generativa de unCLIP se compone de dos componentes clave: un *prior* P que produce incrustaciones de imagen de CLIP condicionadas por descripciones y un *decodificador* D que genera imágenes basadas en estas incrustaciones de imagen de CLIP (y, opcionalmente, descripciones de texto).

En cuanto a la evaluación, unCLIP se compara con otros sistemas, como DALL-E y GLIDE. Se encuentra que las muestras de unCLIP son comparables en calidad a GLIDE, pero con una mayor diversidad en sus generaciones. Las métricas más destacadas en este estudio incluyen la importancia del *prior*, evaluaciones humanas, el equilibrio entre diversidad y fidelidad con guía y la comparación con MS-COCO (Figuras 2.5, 2.6, 2.7 y 2.8).

Figura 2.5. Comparación entre DALL-E 2 con y sin uso de prior. Gráfico extraído de Aditya et al. (2022)

Modelo	FID	Tiro Cero FID	Tiro Cero FID (filt)
AttnGAN (Xu et al., 2017)	35.49		
DM-GAN (Zhu et al., 2019)	32.64		
DF-GAN (Tao et al., 2020)	21.42		
DM-GAN + CL (Ye et al., 2021)	20.79		
XMC-GAN (Zhang et al., 2021)	9.33		
LAFITE (Zhou et al., 2021)	8.12		
Make-A-Scene (Gafni et al., 2022)	**7.55**		
DALL-E (Ramesh et al., 2021)		~ 28	
LAFITE (Zhou et al., 2021)		26.94	
GLIDE (Nichol et al., 2021)		12.24	12.89
Make-A-Scene (Gafni et al., 2022)			11.84
unCLIP (AR prior)		10.63	11.08
unCLIP (Diffusion prior)		**10.39**	**10.87**

Figura 2.6. Comparación de DALL-E 2 contra otros modelos evaluando la métrica FID con el conjunto de datos COCO. Gráfico extraído de Aditya et al. (2022)

Figura 2.7. Comparación con de DALL-E contra otros modelos utilizando el MS-COCO. Gráfico extraído de Aditya et al. (2022)

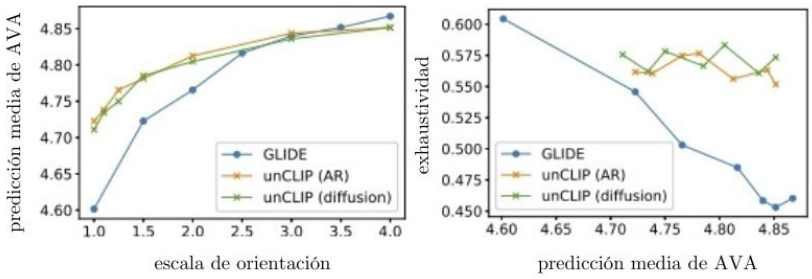

Figura 2.8. Comparación de calidad estética entre GLIDE y unCLIP (DALL-E 2). Gráfico extraído de Aditya et al. (2022)

La propuesta de DALL-E 2 consiste un modelo innovador que fusiona conceptos contrastivos y DM para la generación de imágenes basada en texto,

demostrando su eficacia a través de una serie de métricas y evaluaciones rigurosas. Este enfoque promete avances significativos en la generación de contenido visual y su correspondencia con el lenguaje natural.

2.1.3. Stable Diffusion

Uno de los aportes más significativos en el campo de generación de imágenes de alta resolución es el trabajo titulado *"High-Resolution Image Synthesis with Latent Diffusion Models"* de Rombach et al. (2022), que dio origen a uno de los modelos más destacados en la síntesis de texto a imagen hasta la fecha denominado SD.

Los modelos tradicionales operan en el espacio de píxeles, lo que requiere una optimización computacionalmente costosa y evaluaciones secuenciales que consumen una cantidad considerable de recursos GPU. Teniendo esto en cuenta y con el objetivo de permitir el entrenamiento de DM en entornos con recursos limitados sin sacrificar su calidad y flexibilidad, la solución propuesta son los LDM, que mejoran significativamente tanto la generación como la eficiencia de muestreo. Estos modelos operan en el espacio latente de potentes codificadores automáticos previamente entrenados aplicando y eliminando ruido de las imágenes sin degradar la calidad.

Una característica clave de este trabajo es la introducción de capas de atención cruzada en la arquitectura de los LDM. Estas capas mejoran la robustez y la capacidad de generación de imágenes de alta resolución, lo que permite que los modelos acepten diferentes tipos de entradas, ya sea texto o cuadros.

En términos de resultados, los LDM logran puntuaciones de última generación en la generación de imágenes y la síntesis de imágenes condicionales de clase. También demuestran un rendimiento altamente competitivo en diversas tareas, como la síntesis de texto a imagen y la generación incondicional[12] de imágenes y superresolución. Lo más importante es que logran estos resultados con requisitos computacionales significativamente reducidos en comparación con los enfoques basados en píxeles, un logro clave de este estudio.

Las evaluaciones realizadas sobre SD incluyen la compensación de la compresión de percepción, que analiza el comportamiento de los LDM con diferentes factores de reducción de resolución, y la comparación de la generación de imágenes. En estas pruebas, los LDM demuestran su superioridad

[12]Capacidad de un modelo generativo para producir datos sin restricciones específicas

en términos de calidad y capacidad de generación en comparación con otros modelos, incluidos los basados en GAN (Figuras 2.9, 2.10, 2.11, 2.12).

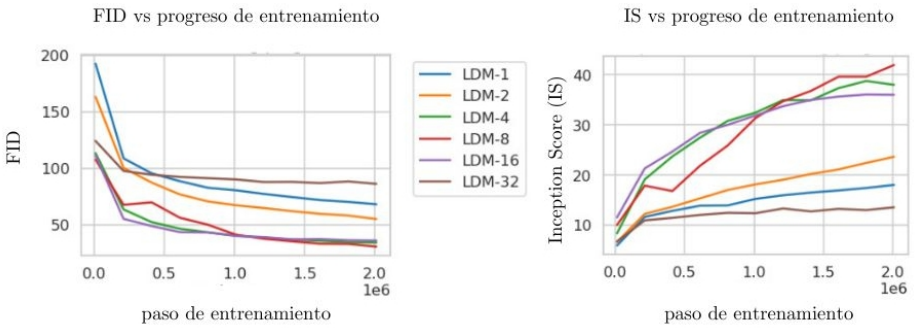

Figura 2.9. Análisis de entrenamiento de LDM condicionales con diferentes factores de reducción. Gráfico extraído de Rombach et al. (2022)

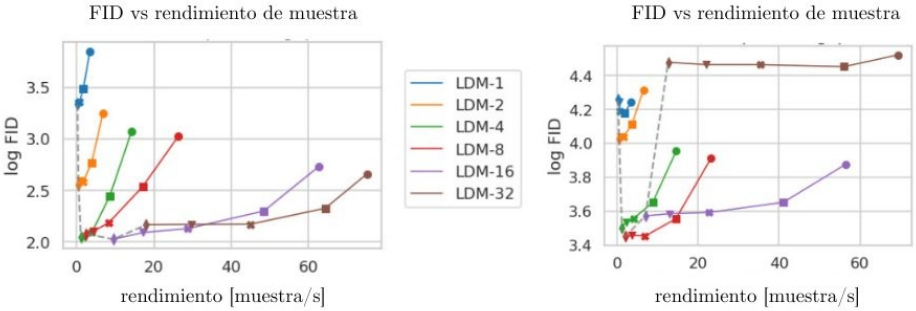

Figura 2.10. Comparación de LDM con compresión variable en el conjunto de datos. Gráfico extraído de Rombach et al. (2022)

CelebA-HQ 256 × 256				FFHQ 256 × 256			
Método	FID↓	Prec. ↑	Exhaust.↑	Método	FID ↓	Prec.↑	Exhaust.↑
DC-VAE [63]	15.8	-	-	ImageBART [21]	9.57	-	-
VQGAN+T. [23] (k=400)	10.2	-	-	U-Net GAN (+aug) [77]	10.9 (7.6)	-	-
PGGAN [39]	8.0	-	-	UDM [43]	5.54	-	-
LSGM [93]	7.22	-	-	StyleGAN [41]	4.16	0.71	0.46
UDM [43]	7.16	-	-	ProjectedGAN [76]	**3.08**	0.65	0.46
LDM-4	**5.11**	0.72	0.49	LDM-4	4.98	**0.73**	**0.50**

LSUN-Churches 256 × 256				LSUN-Bedrooms 256 × 256			
Método	FID ↓	Prec. ↑	Exhaust.↑	Método	FID↓	Prec.↑	Exhaust.↑
DDPM [30]	7.89	-	-	ImageBART [21]	5.51	-	-
ImageBART [21]	7.32	-	-	DDPM [30]	4.9	-	-
PGGAN [39]	6.42	-	-	UDM [43]	4.57	-	-
StyleGAN [41]	4.21	-	-	StyleGAN [41]	2.35	0.59	0.48
StyleGAN2 [42]	3.86	-	-	ADM [15]	1.90	**0.66**	**0.51**
ProjectedGAN [76]	**1.59**	0.61	0.44	ProjectedGAN [76]	**1.52**	0.61	0.34
LDM-8	4.02	**0.64**	**0.52**	LDM-4	2.95	**0.66**	0.48

Figura 2.11. Métricas de evaluación para la síntesis de imágenes incondicional. Tabla extraída de Rombach et al. (2022)

Síntesis de Imágenes Condicionada por Texto				
Método	FID↓	IS ↑	N parametros	
CogView†	27.10	18.20	4B	autoclasificación, tasa de rechazo 0.017
LAFITE†	26.94	26.02	75M	
GLIDE*	12.24	-	6B	277 pasos DDIM c.f.g. s = 3
Make-A-Scene*	**11.84**	-	4B	c.f.g para modelos AR s = 5
LDM-KL-8	23.31	$20.03_{\pm 0.33}$	1.45B	250 pasos DDIM
*LDM-KL-8-G**	12.63	$\mathbf{30.29}_{\pm \mathbf{0.42}}$	1.45B	250 pasos DDIM c.f.g. s = 1.5

Figura 2.12. Métricas de evaluación para la síntesis de imágenes condicionales de texto. Tabla extraída de Rombach et al. (2022)

Finalmente, este trabajo destaca la eficacia de los LDM y la atención cruzada en la generación de imágenes de alta resolución de manera eficiente. Estos enfoques tienen el potencial de superar a los métodos actuales en una amplia gama de tareas de síntesis de imágenes condicionales sin requerir arquitecturas específicas para cada tarea.

2.2. Ajuste Fino

El FT se erige como una estrategia omnipresente en el *Aprendizaje por Transferencia* (de inglés *Transfer Learning* TL), donde se emplea una red

neuronal preentrenada y se ajustan sus pesos mediante ejemplos del nuevo conjunto de datos o tarea específica (Howard & Ruder, 2018). Este enfoque, altamente efectivo en la adaptación de modelos preentrenados a contextos particulares, encuentra su aplicación de manera destacada en los modelos de texto a imagen, donde la personalización es fundamental para lograr resultados óptimos.

Para comenzar, se detallará el trabajo realizado en DB, explorando cómo este enfoque de FT se ha utilizado para mejorar la capacidad de generación de imágenes a partir de texto (Ruiz et al., 2023). Asimismo, se examinará el caso de estudio de *SuTi*, evidenciando cómo la aplicación de técnicas de FT ha permitido la personalización de modelos de texto a imagen para satisfacer las demandas específicas de diferentes dominios o aplicaciones (Wenhu et al., 2023).

Este análisis exhaustivo de trabajos previos brindará una comprensión más profunda de cómo el FT ha contribuido al avance de los modelos de texto a imagen, sentando las bases para nuestra investigación y exploración en este campo.

2.2.1. DreamBooth

Después de investigar los modelos más significativos en la generación de imágenes, es fundamental analizar cómo personalizarlos para satisfacer necesidades específicas mediante tecnologías de este tipo. En este contexto, se destaca el trabajo (Ruiz et al., 2023) en el cual se introduce DB, un marco de trabajo que permite ajustar un modelo de texto a imagen previamente entrenado, (como por ejemplo SD), utilizando solo unas pocas imágenes de un tema específico. El objetivo principal es aprender a vincular un identificador único con un sujeto particular.

Una vez que el sujeto está incorporado en el dominio de salida del modelo, el identificador único se utiliza para generar imágenes fotorrealistas novedosas del sujeto en diferentes contextos. Es relevante destacar que este proceso de FT puede funcionar con tan solo 3-5 imágenes de sujetos, lo que lo convierte en una técnica altamente accesible y utilizable.

Para las pruebas de personalización con DB, se utilizaron 30 sujetos, divididos en dos categorías: objetos y sujetos vivos/mascotas (21 objetos y 9 sujetos vivos/mascotas). Se generaron cuatro imágenes por tema y por

mensaje en el conjunto de evaluación, lo que resultó en un total de 3000 imágenes.

Un aspecto crítico a evaluar es la *fidelidad* del sujeto, es decir, la preservación de los detalles del sujeto en las imágenes generadas. Para esto, se calcularon dos métricas: CLIP-I y DINO [13].

En las pruebas, se compararon los resultados de DB utilizando *Imagen*, DB utilizando SD y *Textual Inversión* utilizando SD. Los cálculos de CLIP-I y DINO (Figura 2.13) llevaron a la conclusión de que DB (sobre el modelo *Imagen*) logra puntuaciones más altas tanto en *fidelidad* de sujeto como en alineación del texto que DB (sobre el modelo SD), acercándose al límite superior de fidelidad de sujeto para imágenes reales.

Método	DINO ↑	CLIP-I ↑	CLIP-T ↑
Imagenes reales	0.774	0.885	N/A
DreamBooth (Imagen)	**0.696**	**0.812**	**0.306**
DreamBooth (Stable Diffusion)	0.668	0.803	0.305
Textual Inversion (Stable Diffusion)	0.569	0.780	0.255

Figura 2.13. Comparación de métricas CLIP-I y DINO para fidelidad del sujeto y CLIP-T para alineación del texto. Tabla extraída de Ruiz et al. (2023)

Además, se compararon Textual Inversión (con SD) y DB (también con SD) mediante un estudio de usuarios (ver Figura 2.14). Se solicitó a 72 usuarios que respondieran cuestionarios con 25 preguntas comparativas (3 usuarios por cuestionario), lo que resultó en 1800 respuestas. Las muestras se seleccionaron aleatoriamente de un gran conjunto de datos. Cada pregunta mostraba un conjunto de imágenes reales de un sujeto y una imagen generada de ese sujeto por cada método (con un mensaje aleatorio). Los usuarios debían responder preguntas sobre la reproducción de la identidad del sujeto y la adecuación al texto de referencia.

[13]DINO es una métrica propuesta por los autores y mide la similitud promedio de coseno entre las incrustaciones ViT-S/16 DINO de imágenes generadas y reales

Método	Fidelidad del sujeto ↑	Fidelidad del texto ↑
DreamBooth (Stable Diffusion)	**68%**	**81%**
Textual Inversion (Stable Diffusion)	22%	12%
Indeciso	10%	7%

Figura 2.14. Resultados de evaluación humana en Dreambooth. Tabla extraída de Ruiz et al. (2023)

Como conclusión de esta prueba, se observó una preferencia abrumadora por DB, tanto en términos de *fidelidad del sujeto* como en *fidelidad de las descripciones ingresadas* [14].

2.2.2. SuTI

Para finalizar con el enfoque de personalización de modelos de texto a imagen, un grupo de investigadores propuso *SuTI* (Wenhu et al., 2023), un generador de texto a imagen centrado en temas que reemplaza el FT específico de un tema con el aprendizaje en contexto.

SuTI permite generar representaciones instantáneas y novedosas de un tema en diversas escenas sin necesidad de una optimización específica del tema. Esto se logra a través del aprendizaje por imitación, donde se entrena un único modelo de aprendiz utilizando datos generados por una amplia variedad de modelos expertos. Además, *SuTI* es el primer generador de texto a imagen basado en temas que funciona completamente en contexto, lo que significa que puede generalizarse a través de múltiples dominios visuales.

La creación de *SuTI* involucró la extracción de millones de conjuntos de imágenes de Internet, cada uno centrado en un tema visual específico. Estos conjuntos se utilizaron para entrenar numerosos modelos expertos, cada uno especializado en un tema particular. Luego, el modelo aprendiz SuTI aprende a imitar el comportamiento de estos expertos afinados (Figura 2.15).

[14]Los autores calculan las correlaciones entre los puntajes DINO/CLIP-I y los puntajes normalizados de preferencia humana. DINO tiene un coeficiente de correlación de Pearson de 0.32 con la preferencia humana (en comparación con 0.27 para la métrica CLIP-I), con un valor de p muy bajo de 9.44×10^{-30}. *NOTA:* Se utilizará como notación de separador decimal al punto (.)

Figura 2.15. Diagrama de aprendizaje de SuTI. Gráfico extraído de Wenhu et al. (2023)

En cuanto a los resultados, *SuTI* puede generar imágenes personalizadas y de alta calidad de temas específicos hasta 20 veces más rápido que los métodos de vanguardia basados en optimización. En los desafíos DreamBench y DreamBench-v2, la evaluación humana demuestra que *SuTI* supera significativamente a los modelos existentes hasta la fecha, como *InstructPix2Pix*, *Textual Inversion*, *Imagic*, *Prompt2Prompt*, *Re-Imagen* y DB, especialmente en términos de fidelidad y alineación del texto.

Para evaluar *SuTI*, se definieron los siguientes elementos:

- *Modelo experto*: Se entrenó el modelo de difusión texto → 64x64, conservando la superresolución original de 256x256 y 1024x1024, similar al modelo *Imagen*. Se utilizó una guía sin clasificador para muestrear nuevas imágenes, con un peso de guía establecido en 30. Para evitar problemas de memoria, se usaron expertos optimizados para muestrear imágenes pseudoobjetivo y luego se almacenaron las muestras como archivos separados.
- *Modelo aprendiz*: Este modelo contiene 2.5 mil millones de parámetros, lo que representa un aumento de 400 millones de parámetros en comparación con el modelo *Imagen* 64x64 original de 2.1 mil millones. Los parámetros adicionales se agregaron a través de capas de atención adicionales en las entradas de imagen y texto.
- *Conjunto de datos*: Se empleó el conjunto de datos DreamBench propuesto por DB y se creó un conjunto de datos DreamBench-v2 para aumentar aún más la dificultad y diversidad de las pruebas.
- *Métricas de evaluación*: Se utilizaron las mismas métricas de evaluación empleadas en DreamBooth, incluyendo DINO, CLIP-I para evaluar la *fidelidad del sujeto*, CLIP-T para evaluar la *fidelidad del texto* y la evaluación humana para complementar estas métricas.

Los resultados de estas pruebas confirman la superioridad de SuTI sobre DB, tanto en términos de métricas (Figura 2.16) como en evaluaciones realizadas por seres humanos (Figura 2.17).

Metodos	Modelo	DINO ↑	CLIP-I ↑	CLIP-T ↑
Real Image (Oracle)	-	0.774	0.885	-
DreamBooth [27]	Imagen [28]	0.696	0.812	**0.306**
DreamBooth [27]	SD [25]	0.668	0.803	0.305
Textual Inversion [10]	SD [25]	0.569	0.780	0.255
Re-Imagen [6]	Imagen [28]	0.600	0.740	0.270
SuTI	Imagen [28]	**0.741**	**0.819**	0.304

Figura 2.16. Evaluación de SuTI en las métricas DINO, CLIP-I y CLIP-T. Tabla extraída de Wenhu et al. (2023)

Metodo	Modelo	Espacio	Tiempo	Objeto	Texto	Fotorealismo	General
Modelos que requieren ajuste en el momento de la prueba							
Textual Inversion [10]	SD [25]	$	30 mins	0.22	0.64	0.90	0.14
Null-Text Inversion [19]	Imagen [28]	$$	5 mins	0.20	0.46	0.70	0.10
Imagic [15]	Imagen [28]	$$$$	70 mins	0.78	0.34	0.68	0.28
DreamBooth [27]	SD [25]	$$$	6 mins	0.74	0.53	0.85	0.47
DreamBooth [27]	Imagen [28]	$$$	10 mins	0.88	0.82	**0.98**	0.77
InstructPix2Pix [4]	SD [25]	-	10 secs	0.14	0.46	0.42	0.10
Re-Imagen [6]	Imagen [28]	-	20 secs	0.70	0.65	0.64	0.42
SuTI	Imagen [28]	-	30 secs	**0.90**	**0.90**	0.92	**0.82**

Figura 2.17. Evaluación humana de SuTI. Tabla extraída de Wenhu et al. (2023)

En resumen, *SuTI* logra superar a DB en un 5 por ciento en la puntuación general, lo que destaca su capacidad mejorada en la generación de imágenes personalizadas basadas en temas.

2.3. Resumen

En este capítulo se examinaron los modelos de texto a imagen más destacados en la actualidad, así como las tendencias para personalizarlos. Se complementó el análisis de los modelos con tablas, gráficos y métricas.

Entre ellos se pudo destacar que *Imagen* (Chitwan et al., 2022) logra una alta fidelidad de imagen basada en modelos de lenguaje genéricos, mientras que *DALL-E 2* (Aditya et al., 2022) propone un enfoque de dos etapas para la generación de imágenes con un equilibrio entre diversidad y realismo. Por su parte, SD (Rombach et al., 2022) propone la utilización de LDM para mejorar la fidelidad visual y permitir la generación de alta resolución.

En la personalización de estos modelos, DB (Ruiz et al., 2023) ajusta modelos preentrenados basándose en unas pocas imágenes de un sujeto, permitiendo la generación de imágenes personalizadas. *SuTI* (Wenhu et al., 2023), por otro lado, elimina la necesidad de FT específico del sujeto al aprender de múltiples modelos expertos para generar imágenes de alta calidad y personalizadas de un sujeto de manera eficiente.

Parte II

Marco Teórico

Capítulo 3

Marco Teórico

3.1. Introducción

El paradigma de la AI se refiere a la capacidad intrínseca de un sistema para realizar tareas que normalmente requieren facultades intelectuales humanas (Norvig & Russell, 2021). En este contexto, se destaca el enfoque pragmático de la AI, que se centra en obtener resultados efectivos mediante el uso de algoritmos y técnicas de *Aprendizaje Automático* (del inglés *Machine Learning* (ML)).

La aplicación de modelos de texto a imagen, en particular, representa un avance significativo en la intersección entre el lenguaje y la representación visual. Estos modelos facilitan la creación de contenido visual a partir de descripciones. Desde la generación de contenido multimedia hasta la mejora de la comunicación, los modelos de texto a imagen desempeñan un papel crucial en la expansión de las aplicaciones de la AI. Su capacidad para traducir eficientemente entre el lenguaje humano y la representación visual potencia significativamente la utilidad y el impacto de la AI en diversos sectores y disciplinas.

Inicialmente, se expondrán los conceptos claves necesarios para comprender a fondo el ámbito de la AI y los modelos generativos.

A continuación, se proporcionará una explicación detallada de qué son los modelos generativos, incluyendo su arquitectura básica, su taxonomía y una descripción minuciosa de las diferentes familias.

Posteriormente, se abordará la temática de los DM, detallando su definición, clasificación y una descripción pormenorizada de los diversos tipos de DM existentes, así como sus aplicaciones más comunes.

Finalmente, se conceptualizarán y ampliarán los modelos generativos, centrándose en su aplicación para la generación de imágenes a partir de texto.

3.2. Conceptos y Terminología Generales

Para comprender plenamente los modelos generativos en AI, y en particular los de texto a imagen, es esencial tener una comprensión sólida de los siguientes conceptos:

- *Aprendizaje Automático (Machine Learning (ML))*

  El ML es una rama de la AI que se enfoca en desarrollar algoritmos y modelos que permiten a las computadoras aprender patrones y tomar decisiones a partir de datos sin necesidad de ser programadas explícitamente para cada tarea (Alpaydin, 2014).

- *Aprendizaje Profundo (Deep Learning (DL))*

  El DL es una categoría de algoritmos de ML que emplea múltiples capas de unidades de procesamiento para adquirir representaciones de alto nivel a partir de *datos no estructurados* [1] (Foster, 2023).

  Cuando se trabaja con datos no estructurados, la información contenida en píxeles individuales, frecuencias o caracteres es mínima. Por ejemplo, conocer el color de un píxel específico en una imagen o el carácter en una posición determinada de un texto generalmente no proporciona información útil para identificar el contenido de la imagen o el tema del texto.

[1] Los datos no estructurados se refieren a cualquier tipo de información que no esté naturalmente organizada en columnas de características, como imágenes, audio y texto. Si bien es cierto que las imágenes tienen una estructura espacial, las grabaciones de audio tienen una estructura temporal y los pasajes de texto pueden tener tanto estructura espacial como temporal, estos datos no están dispuestos en columnas de características, lo que los califica como no estructurados.

Por otro lado, un modelo de DL tiene la capacidad de extraer automáticamente características relevantes y de alto nivel directamente de los datos no estructurados.

La mayoría de los sistemas de aprendizaje profundo se basan en *Redes Neuronales Artificiales* (del inglés de *Artificial Neural Networks* (ANN)), las cuales constan de múltiples capas ocultas apiladas.

- *Aprendizaje por Transferencia (Transfer learning (TL))*

  El TL es una técnica que implica aprovechar una red entrenada (Perez-Aguilar et al., 2021). En otras palabras, se reutiliza la arquitectura y los pesos de un modelo entrenado previamente con grandes volúmenes de datos de entrada, y se aplican a diferentes escenarios con otros conjuntos de datos. El propósito principal es lograr clasificaciones más rápidas y reducir la carga computacional.

- *Ajuste Fino (Fine Tuning (FT))*

  El FT es una estrategia común en el TL. En este proceso, se toma una red entrenada y se ajustan sus pesos utilizando ejemplos del nuevo conjunto de datos o tarea específica. Durante esta fase, los pesos se modifican para adaptar la red a las características particulares del nuevo conjunto de datos, lo que permite que la red se especialice en la tarea específica que se está abordando. El FT resulta especialmente útil cuando se busca mejorar el rendimiento de la red en una tarea específica sin tener que entrenarla completamente desde cero, aprovechando así el conocimiento previo adquirido durante el entrenamiento inicial en un dominio más general. Este enfoque puede acelerar el proceso de entrenamiento y mejorar la capacidad de la red para realizar tareas específicas.

- *Aprendizaje de Representación y Espacio Latente*

  El *aprendizaje de representación* y el *espacio latente* son conceptos fundamentales en los modelos de AI. Tal como lo define Foster (Foster, 2023) en lugar de intentar modelar directamente el espacio de muestras de alta dimensión, cada observación en el conjunto de entrenamiento se describe utilizando un espacio de menor dimensión llamado *espacio latente*. Luego, se aprende una función de mapeo que puede tomar un punto en el espacio latente y mapearlo a un punto en el dominio original. En otras palabras, cada punto en el espacio latente representa alguna observación de alta dimensión.

En la práctica, esto significa que se pueden representar las observaciones complejas de manera más simple y compacta en el espacio latente. Por ejemplo, si se tiene un conjunto de imágenes de latas de galletas, cada imagen se puede convertir en un punto en un espacio latente de dimensiones reducidas, como la altura y el ancho de la lata. Esta representación simplificada en el espacio latente permite producir imágenes de latas que no existen en el conjunto de entrenamiento mediante una función de mapeo adecuada.

La abstracción de que el conjunto de datos original puede ser descrito por un espacio latente más simple es una tarea fundamental del ML y, más específicamente, del DL. Este enfoque de codificar el conjunto de datos en un espacio latente y luego decodificarlo de vuelta al dominio original es común en muchas técnicas de modelado generativo.

- *Redes Neuronales Artificiales (Artificial Neural Networks (ANN))*

  Las ANN son modelos computacionales que se inspiran en la estructura y el funcionamiento del cerebro humano (Goodfellow et al., 2016). Estas redes están compuestas por capas de nodos interconectados, conocidos como neuronas, donde cada conexión tiene un peso que regula la influencia de una neurona en otra (representado en la Figura 3.1). A través del ajuste de estos pesos, estas redes tienen la capacidad de aprender y generalizar a partir de conjuntos de datos.

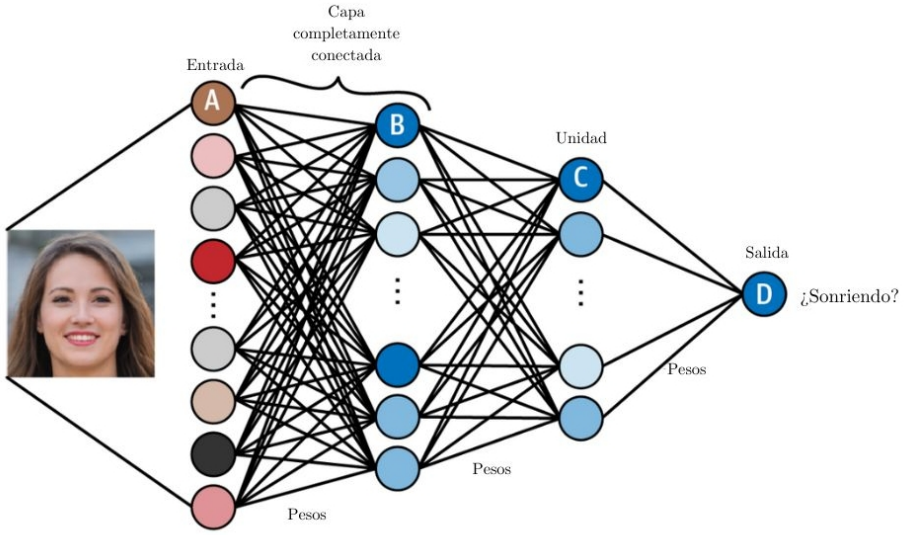

Figura 3.1. Diagrama de una red neuronal básica. Gráfico extraído de Foster (2023)

Como señala el autor Foster (Foster, 2023), si bien existen diversos tipos de capas, una de las más comunes es la *capa completamente conectada* (también conocida como capa densa), que establece conexiones directas entre todas las unidades de la capa actual y las de la capa anterior.

- *Red Neuronal Convolucional (Convolutional Neural Networks (CNN))*

  Las CNN son aquellas que emplean una capa conocida como *capa convolucional* (Foster, 2023). Esta capa aplica filtros para capturar la información estructural de los datos no estructurados procesados por la red, como la disposición espacial en las imágenes de entrada.

  En esencia, una capa convolucional se compone de una colección de filtros cuyos pesos son aprendidos por la red neuronal durante el entrenamiento (representado en la Figura 3.2). Inicialmente, estos pesos son aleatorios, pero se ajustan gradualmente para resaltar características relevantes.

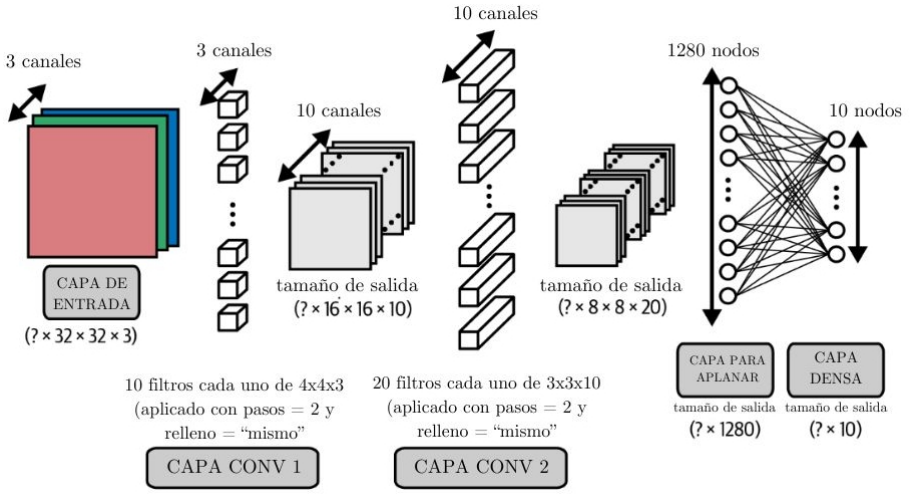

Figura 3.2. Diagrama de red neuronal convolucional. Gráfico extraído de Foster (2023)

Por ejemplo, al convolucionar dos secciones distintas de una imagen en escala de grises de tamaño $3 \times 3 \times 1$ con un filtro (o núcleo) de $3 \times 3 \times 1$, se realiza una multiplicación píxel a píxel entre el filtro y la sección de la imagen, seguida de una suma de los resultados. La salida es más positiva cuando la sección de la imagen coincide con el filtro y más negativa cuando es inversa. Si se desplaza el filtro sobre toda la imagen, se obtiene una nueva matriz que resalta una característica específica de la entrada.

Las *capas convolucionales* son fundamentales para extraer características de las imágenes mediante la aplicación de múltiples filtros. La salida de la capa en un píxel determinado es una suma ponderada de los pesos de los filtros multiplicados por los valores de la capa anterior en una región pequeña centrada en el píxel. Este enfoque es efectivo para detectar bordes, texturas y, en capas más profundas, formas y características más complejas.

- *Espacio Muestral*

  Según lo señalado por Foster (Foster, 2023), el *espacio muestral* es el conjunto completo de todos los valores que una observación x puede tomar.

- *Función de Densidad de Probabilidad*

  De acuerdo con Foster (Foster, 2023), una *función de densidad de probabilidad* (o simplemente función de densidad) es una función $p(x)$ que asigna un punto x en el espacio muestral a un número entre 0 y 1. Para que sea una distribución de probabilidad bien definida, la integral de la función de densidad sobre todos los puntos en el espacio muestral debe ser igual a 1.

- *Verosimilitud (Likelihood)*

  Según Foster (2023), la *verosimilitud* $L(\theta \mid x)$ de un conjunto de parámetros θ es una función que evalúa la plausibilidad de θ dado algún punto observado x. En esencia, la verosimilitud de un conjunto de parámetros θ se define como la probabilidad de observar los datos si la verdadera distribución generadora de datos fuera el modelo parametrizado por θ y su fórmula podría escribirse como figura en la siguiente Ecuación 3.1

$$\mathcal{L}(\theta \mid \mathbf{X}) = \prod_{\mathbf{x} \in \mathbf{X}} p_\theta(\mathbf{x}) \tag{3.1}$$

- *Incrustación (Embbeding)*

  La *incrustación* (z) es la compresión de un dato original, como una imagen, en un espacio latente de menor dimensionalidad. Esta técnica permite generar datos novedosos al seleccionar puntos en el espacio latente y pasarlos a través de la red diseñada con ese propósito (Foster, 2023).

- *Tokenización*

  La *tokenización* es el proceso de dividir el texto en unidades individuales, como palabras o caracteres (Foster, 2023). La elección de cómo tokenizar un texto dependerá de los objetivos del modelo generativo.

  Si se elige la *tokenización por palabras*, se sugiere convertir todo el texto a minúsculas para mantener la coherencia en la tokenización, aunque esto puede afectar a los nombres propios y lugares. Además, es útil manejar las palabras poco frecuentes mediante un token para "palabra desconocida" y considerar el uso de *stemming*[2] para reducir la complejidad del vocabulario. Sin embargo, es importante tener en cuenta que el modelo estará limitado a predecir palabras dentro del vocabulario de entrenamiento y será necesario abordar la tokenización de la puntuación.

  Por otro lado, si se opta por la *tokenización por caracteres*, el modelo puede generar nuevas palabras fuera del vocabulario de entrenamiento, lo cual puede ser deseable en ciertos casos. Las letras mayúsculas pueden tratarse como caracteres separados o convertirse a minúsculas según la preferencia. Además, el vocabulario resultante suele ser más reducido, lo que beneficia la velocidad de entrenamiento al haber menos pesos que aprender en la capa de salida final.

3.3. Modelos Generativos

En el universo de los modelos de texto a imagen, los modelos generativos son un pilar fundamental. En este contexto, el enfoque no se limita a la transformación directa de las entradas en salidas, sino que su característica distintiva es la introducción de un componente probabilístico en el resultado final.

[2]Cuando se habla *stemming* significa que las palabras pueden ser llevadas a su raíz, es decir reducirlas a su forma más simple, de modo que diferentes tiempos de un verbo permanecen agrupados en una misma tokenización. Por ejemplo, "browse" (navegar), "browsing" (navegando), "browses" (navega) y "browsed" (navegó) serían todos reducidos a la raíz "brows".(Foster, 2023)

3.3.1. Definición de los Modelos Generativos

Los modelos generativos, dentro del ámbito de la AI, representan una categoría avanzada de algoritmos con el propósito fundamental de comprender y reproducir la complejidad presente en diversos conjuntos de datos. Según el autor Goodfellow (Goodfellow et al., 2014), estos modelos van más allá de la clasificación o predicción convencional al aprender la distribución de probabilidad conjunta entre datos observados y latentes, permitiendo así la generación de nuevas instancias de datos con similitudes estructurales con las muestras originales.

En contraste con los *modelos discriminativos*, que se centran en clasificar o predecir, los *modelos generativos*, como indica el autor Bishop (Bishop, 2006), tienen la capacidad única de generar instancias de datos que siguen la distribución de los datos de entrenamiento. Esta habilidad para sintetizar datos realistas los convierte en herramientas esenciales para diversas aplicaciones prácticas, revolucionando campos como la CV y el NLP (LeCun et al., 2015).

En el ámbito médico, por ejemplo, se utilizan para generar imágenes médicas sintéticas para entrenar modelos de diagnóstico por imágenes (Frid-Adar et al., 2018), mientras que en la música se emplean para crear composiciones originales basadas en estilos musicales existentes (Gaetan et al., 2017).

El modelo GPT (Generative Pre-trained Transformer) de OpenAI, descrito por el autor Brown (Brown et al., 2020), ejemplifica el potencial de los modelos generativos en el NLP, mostrando destreza en la generación de texto coherente y contextual para aplicaciones como la redacción automatizada y los asistentes virtuales.

En resumen, los modelos generativos no solo representan una innovación técnica en AI, sino que también están transformando profundamente la manera en que se interactúa con datos complejos y abren nuevas posibilidades en múltiples campos.

Tal como explica Foster (Foster, 2023), los modelos generativos son una rama del ML que se enfoca en crear modelos capaces de generar nuevos puntos de datos que se asemejen a los datos de entrenamiento. Por ejemplo, si se tiene un conjunto de datos con imágenes de caballos, es posible construir un modelo que genere una imagen completamente nueva de un caballo, incluso si nunca ha existido en la realidad. La particularidad radica en que esta imagen generada parece auténtica, ya que el modelo ha internalizado las

reglas generales que determinan la apariencia de un caballo. Este tipo de desafío es precisamente el que aborda la modelación generativa.

Un resumen de un proceso típico de modelado generativo se muestra en la Figura 3.3

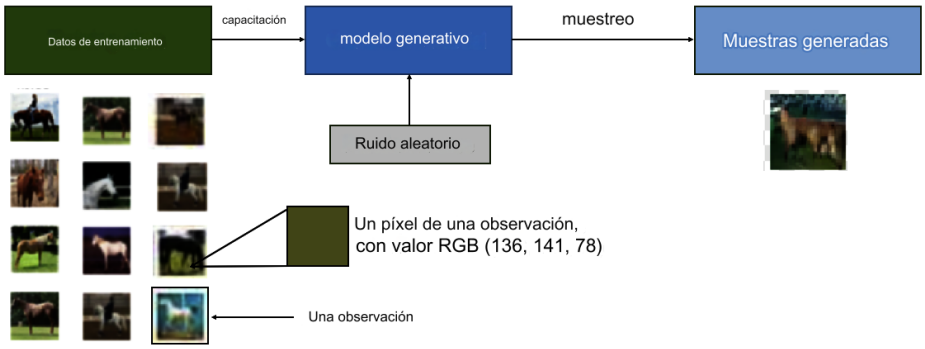

Figura 3.3. Proceso de modelos generativos. Gráfico extraído de Foster (2023)

Al analizar un modelo generativo típico, es fundamental comenzar con un *conjunto de datos* (del inglés dataset) que contenga numerosos ejemplos de la entidad que se desea generar. Estos datos contienen observaciones, cada una abarcando múltiples características. En el contexto de la generación de imágenes, estas características suelen representar los valores de píxeles individuales. El objetivo central radica en construir un modelo capaz de generar nuevas combinaciones de características que se asemejen a las que se encuentran en los datos originales. Es importante destacar que esta tarea es excepcionalmente desafiante debido a la inmensa variedad de formas en que los valores de píxeles pueden combinarse, en contraste con el número relativamente limitado de arreglos que constituyen una imagen de la entidad que pretendemos simular.

Además para que un modelo sea considerado generativo, debe ser *probabilístico* en lugar de ser determinista. Si simplemente realizara un cálculo fijo, como tomar el valor promedio de cada píxel en el conjunto de datos, no sería generativo, ya que produciría la misma salida en cada ocasión. El componente clave es la inclusión de un elemento estocástico o aleatorio que afecta las muestras individuales generadas por el modelo.

En términos más simples, se podría imaginar la existencia de una distribución de probabilidad desconocida que explica por qué ciertas imágenes están presentes en el conjunto de datos de entrenamiento y por qué otras no

lo están. Por lo tanto, la tarea central consiste en construir un modelo que pueda imitar esta distribución de la manera más precisa posible y, luego, extraer muestras de esta distribución para generar nuevas observaciones únicas que parezcan pertenecer al conjunto de entrenamiento original.

3.3.2. Diferencia entre Modelado Generativo y Modelado Discriminativo

De acuerdo con Foster (2023), para comprender plenamente el propósito y la importancia del *modelado generativo*, es esencial contrastarlo con su contraparte, el *modelado discriminativo*.

Si se dispone de un conjunto de datos que contiene pinturas, algunas creadas por Van Gogh y otros artistas; con suficientes datos, se podría entrenar un modelo discriminativo para determinar si una pintura en particular fue realizada por Van Gogh. Este modelo aprendería a reconocer ciertos colores, formas y texturas que son indicativos de las obras del pintor holandés. Cuando se le presenta una pintura con estas características, el modelo aumentaría su confianza en su predicción de que la obra es de Van Gogh.

Este enfoque, conocido como *modelado discriminativo*, se centra en la diferenciación y clasificación. Su objetivo principal es discernir entre diferentes categorías o clases predefinidas. En este caso, las categorías son "pinturas de Van Gogh" y "pinturas de otros artistas". El modelo utiliza las características distintivas que ha aprendido de las pinturas de Van Gogh para realizar estas distinciones. La siguiente Figura 3.4 representa un esquema típico de un modelo discriminativo.

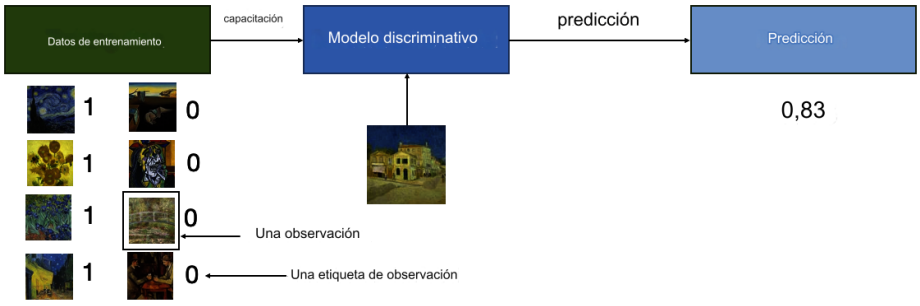

Figura 3.4. Proceso de modelos discriminativos. Gráfico extraído de Foster (2023)

En contraste, el *modelado generativo* se enfoca en crear nuevos ejemplos que sean similares a los datos de entrenamiento, sin preocuparse por clasificarlos en categorías preexistentes. En lugar de determinar si una pintura es de Van Gogh o de otro artista, un *modelo generativo* podría producir una nueva pintura que tenga las características estilísticas de Van Gogh, incluso si nunca antes existió una obra específica con esas características.

Por lo tanto, mientras que el *modelado discriminativo* se utiliza para la clasificación y la identificación, el *modelado generativo* se emplea para la creación de datos nuevos y auténticos que sigan las pautas aprendidas de los datos de entrenamiento, lo que lo hace fundamental en áreas como la generación de imágenes, el procesamiento de lenguaje natural y más.

Otra diferencia clave es que, al realizar el modelado discriminativo, cada observación en los datos de entrenamiento está etiquetada. En un problema de clasificación binaria como el discriminador de obras de artistas, las pinturas de Van Gogh estarían etiquetadas como 1, mientras que las que no son de Van Gogh estarían etiquetadas como 0. El modelo aprende, por lo tanto, a discriminar entre estos dos grupos y a emitir una probabilidad de que una nueva observación tenga la etiqueta 1, lo que indica que fue pintada por Van Gogh.

Debido a esto, el modelado discriminativo se asocia con el *aprendizaje supervisado*[3]. Por otro lado, el modelado generativo generalmente se lleva a cabo con un conjunto de datos no etiquetado, lo que se conoce como *aprendizaje no supervisado*. No obstante, también es posible aplicar el modelado generativo a un conjunto de datos etiquetado para aprender a generar observaciones que pertenezcan a cada clase distintiva.

3.3.3. Marco de un Modelo Generativo Básico

Tal como se define en Foster (2023), el desarrollo de un modelo generativo básico debe resultar en la obtención de un modelo p_{model} que imite p_{data}, y cuya ejecución permita realizar muestras de p_{model} para generar observaciones que parezcan haber sido extraídas de p_{data}.

Por ello, se necesita definir un conjunto de datos de observaciones X y suponer que las observaciones han sido generadas de acuerdo a una distribución desconocida denominada p_{data}.

[3]Donde se aprende una función que relaciona una entrada con una salida utilizando un conjunto de datos etiquetado

Las propiedades deseables de p_{model} son:

- *Precisión*: Si $p_{\text{model}}(x)$ es alto, x debe parecer haber sido extraído de p_{data}. Si $p_{\text{model}}(x)$ es bajo, x no debe parecer haber sido extraído de p_{data}.
- *Generación*: Debería ser posible muestrear fácilmente una nueva observación x de $p_{\text{model}}(x)$.
- *Representación*: Debería ser posible comprender cómo p_{model} representa diferentes características de alto nivel.

3.3.4. Alucinaciones en Modelos Generativos

La alucinación en AI es un fenómeno en el cual un *Modelo de Lenguaje de Gran Escala* (del inglés Large Language Model (LLM)), a menudo un chatbot generativo de AI o una herramienta de CV, percibe patrones u objetos que no existen o que son imperceptibles para los observadores humanos, creando resultados que son absurdos o completamente inexactos[4].

Generalmente, cuando un usuario hace una solicitud a una herramienta generativa de AI, desea un resultado que aborde adecuadamente la solicitud (es decir, una respuesta correcta a una pregunta). Sin embargo, a veces los algoritmos no producen resultados esperados; en otras palabras, "alucinan" la respuesta.

Las alucinaciones son causadas principalmente por datos de entrenamiento sesgados, indicaciones ambiguas y parámetros inexactos, y ocurren principalmente al combinar hechos matemáticos con contexto basado en el lenguaje (Roychowdhury, 2023).

Con la creciente presencia de la AI en diversos ámbitos, han surgido preocupaciones respecto a esta problemática. De acuerdo a algunos estudios existe una falta de consistencia en cómo se utiliza el término alucinación en AI y si es realmente el mismo significado para cada caso (Maleki & Padmanabhan, 2024). Por otro lado, algunos estudios proponen soluciones para minimizar las alucinaciones, por ejemplo en aplicaciones de chatbot para la toma de decisiones en el ámbito financiero (Roychowdhury, 2023).

[4]https://www.ibm.com/topics/ai-hallucinations

3.4. Clases de Modelos Generativos

Según los trabajos de Yang (Yang et al., 2023) y Foster (Foster, 2023), dentro de la familia de modelos generativos, se pueden distinguir seis clases principales: Autocodificadores Variacionales (del inglés Variational Autoencodres (VAE)), GAN, Flujos Normalizadores (del inglés Normalization Flows (NF)), ARM, Modelos basados en energía (del inglés Energy Based Models (EBM)) y DM.

A continuación, se analizarán brevemente las características principales, ventajas y limitaciones de las primeras cinco clases. Además, se estudiarán con más detalle los DM como una sexta clase importante en la familia de modelos generativos.

3.4.1. Autocodificadores Variacionales (VAE)

Como explica Foster (Foster, 2023), un *autocodificador* (del inglés autoencoder (AE)) es una ANN que se entrena para realizar la tarea de codificar (es decir, representar un elemento del dominio original en el espacio latente) y decodificar (es decir, recrear un elemento a partir de su ubicación en el espacio) un elemento a representar, de modo que la salida de este proceso sea lo más cercana posible al artículo original (Figura 3.5).

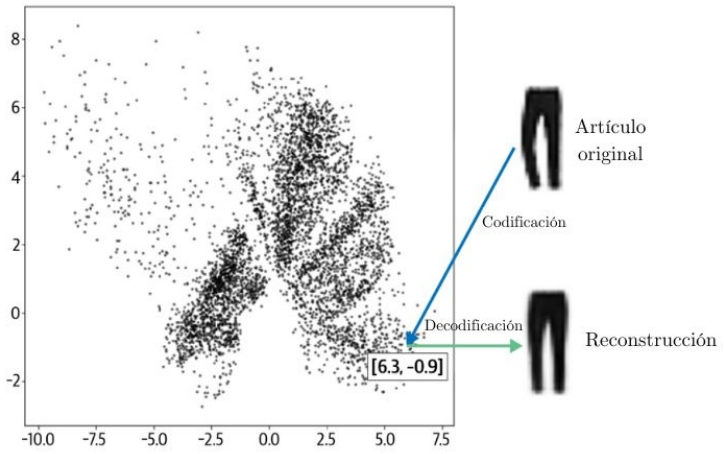

Figura 3.5. Codificación y decodificación de autoencoder básico en el espacio latente. Gráfico extraído de Foster (2023)

Un AE consta de dos partes principales (Figura 3.6): una red codificadora que comprime datos de entrada de alta dimensionalidad, como una imagen, en un *vector de incrustación* de dimensionalidad inferior, y una red decodificadora que descomprime un vector de incrustación dado de vuelta al dominio original (por ejemplo invierta una imagen).

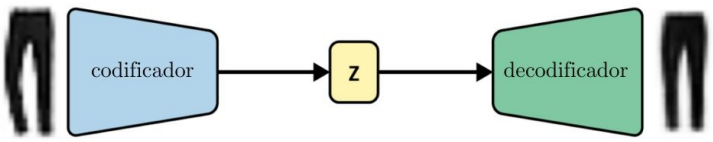

Figura 3.6. Arquitectura de un AE. Gráfico extraído de Foster (2023)

En la práctica, el espacio latente de un AE generalmente tendrá más de dos dimensiones para capturar mayor sutileza de las imágenes.

Los AE clásicos presentan algunos problemas, como la elección uniforme de puntos en un espacio limitado, la falta de claridad en la elección de puntos aleatorios en el espacio latente y la presencia de agujeros en el espacio latente donde ninguna de las imágenes originales está codificada.

Para abordar estos problemas, se desarrollaron los VAE, donde en un VAE, cada elemento se mapea a una distribución normal multivariada alrededor de un punto en el espacio latente. El codificador mapea cada entrada a un vector de medias y un vector de logaritmos de varianzas, y el decodificador es idéntico al de un AE simple. Esta variación garantiza que incluso cuando se elige un punto en el espacio latente que nunca ha sido visto por el decodificador, es probable que se decodifique a una imagen bien formada.

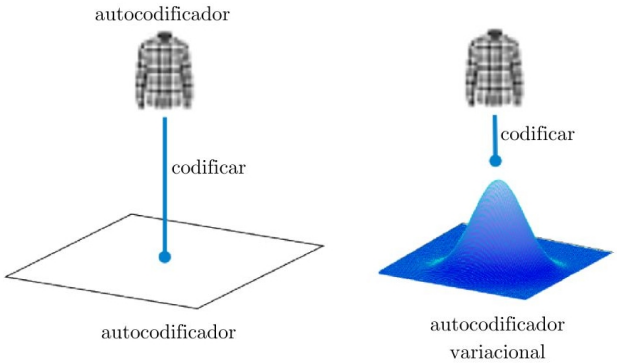

Figura 3.7. Diferencias entre autoencoders y autoencoders variacionales. Gráfico extraído de Foster (2023)

3.4.2. Redes Generativas Adversariales (GAN)

Las GAN constituyen un tipo específico de modelo generativo que se compone de dos redes neuronales que operan en competencia: el *generador* y el *discriminador* (Goodfellow et al., 2014). El *generador* produce muestras de datos sintéticos, mientras que el *discriminador* evalúa la autenticidad de estas muestras. Este proceso de confrontación entre ambas redes permite mejorar constantemente la calidad de las muestras generadas.

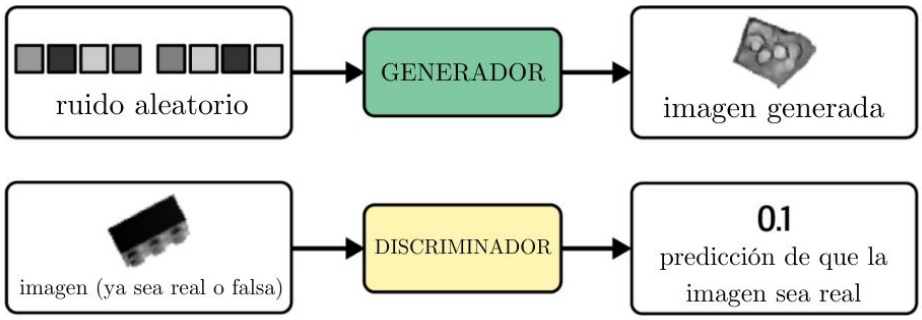

Figura 3.8. Arquitectura básica de una red GAN. Gráfico extraído de (Foster, 2023)

Como describe (Foster, 2023), una GAN se puede visualizar como una batalla entre dos adversarios: el *generador* y el *discriminador*. El generador intenta transformar ruido aleatorio en observaciones que se asemejen a las muestras del conjunto de datos original, mientras que el discriminador trata de discernir si una observación proviene del conjunto de datos original o es una creación del generador.

Inicialmente, el *generador* produce imágenes de baja calidad y el *discriminador* realiza predicciones aleatorias. La esencia de las GAN radica en la alternancia entre el entrenamiento de ambas redes: a medida que el generador mejora en su capacidad para engañar al *discriminador*, este último se adapta para mantener su habilidad de identificar correctamente las observaciones falsas. Este proceso impulsa al generador a buscar nuevas estrategias para engañar al *discriminador*, lo que alimenta un ciclo de mejora continua.

La función principal del *discriminador* en una GAN es determinar si una imagen es real o falsa. Este proceso se asemeja a un problema de clasificación de imágenes supervisado, donde se puede emplear una arquitectura similar a la de capas convolucionales apiladas, con un único nodo de salida que indica la autenticidad de la imagen.

En contraposición, el *generador* toma como entrada un vector extraído de una distribución normal estándar multivariante y produce como salida una imagen del mismo tamaño que las imágenes originales del conjunto de datos de entrenamiento.

El *generador* en una GAN tiene una función similar al decodificador en un VAE: convierte un vector en el espacio latente en una representación visual, es decir, en una imagen. Este concepto de mapeo desde el espacio latente de vuelta al dominio original es fundamental en la modelización generativa, ya que permite manipular los vectores en el espacio latente para alterar características de alto nivel de las imágenes en el dominio original.

Tal como añade Yang et al. (2023), las GAN suelen construirse mediante ANN, aunque podrían implementarse en cualquier sistema diferenciable capaz de mapear los datos de entrada de un espacio a otro. La optimización de las GAN se asemeja a un problema de *minimax*[5], donde se busca una función de valor $V(G, D)$.

El proceso de optimización conduce a un punto de equilibrio en el que se alcanza un mínimo para el generador y un máximo para el discriminador, conocido como equilibrio de Nash[6].

Las GAN, a pesar de sus ventajas, enfrentan una serie de desafíos en su entrenamiento, lo que las hace notoriamente difíciles de manejar. Algunos de los problemas más comunes son:

- *Dominio del Discriminador*: Si el *discriminador* se vuelve demasiado efectivo en distinguir entre imágenes reales y generadas, la señal de la función de pérdida se debilita, lo que dificulta las mejoras significativas en el *generador*. En el peor de los casos, el *discriminador* puede aprender a separar perfectamente las imágenes reales de las generadas, lo que resulta en gradientes casi inexistentes y un estancamiento del entrenamiento.

[5]El algoritmo Minimax es un método de toma de decisiones utilizado en juegos de dos jugadores de suma cero, donde un jugador busca maximizar su ganancia mientras que el otro intenta minimizarla. Este algoritmo evalúa posibles movimientos y selecciona la mejor opción basándose en el supuesto de que el oponente también juega de manera óptima. (Norvig & Russell, 2021)

[6]El equilibrio de Nash, nombrado así por el matemático John Nash, es un concepto fundamental en la teoría de juegos; y se define como una situación en la que, dentro de un juego con dos o más jugadores, ninguno puede mejorar su resultado eligiendo una estrategia diferente, siempre y cuando los otros jugadores mantengan sus estrategias constantes (Myerson, 1991)

- *Dominio del Generador*: Por otro lado, si el *discriminador* no es lo suficientemente poderoso, el *generador* puede engañarlo fácilmente con muestras de baja calidad. Esto conduce a un fenómeno conocido como *colapso de modo*, donde el *generador* produce solo un conjunto limitado de resultados.
- *Pérdida Poco Informativa*: La función de pérdida del *generador* puede no ser un indicador confiable de la calidad de las imágenes producidas. Esto se debe a que la evaluación del *generador* depende del *discriminador* actual, que está en constante mejora. Por lo tanto, una pérdida baja no siempre garantiza una buena calidad visual.
- *Ajuste de Hiperparámetros*: Las GAN tienen una gran cantidad de hiperparámetros que requieren ajuste. La sensibilidad a pequeños cambios en estos parámetros hace que encontrar una configuración efectiva sea más un proceso de prueba y error que seguir pautas establecidas.

Las GAN condicionales (CGAN) representan un avance significativo en los modelos generativos al permitir el control sobre la salida generada. Esta idea, introducida por primera en Mehdi & Simon (2014), es una extensión relativamente simple de la arquitectura de GAN (Figura 3.9).

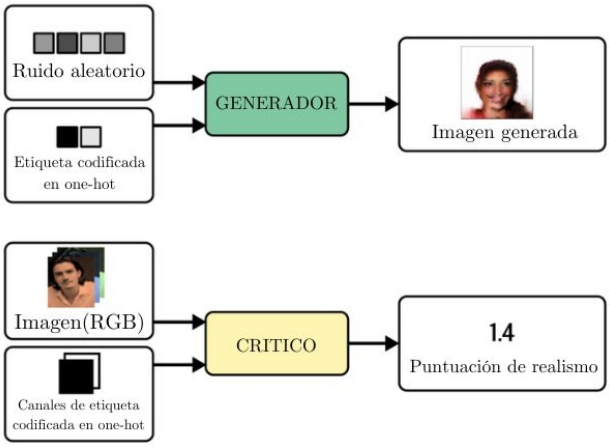

Figura 3.9. Arquitectura de una red CGAN. Gráfico extraído de Foster (2023)

La distinción principal entre una GAN estándar y una CGAN radica en que en esta última se proporciona información adicional relacionada con la etiqueta tanto al *generador* como al *discriminador*. En el generador, esta

información se añade como un *vector codificado en caliente* [7] que se concatena con la muestra del espacio latente. Mientras tanto, en el discriminador, la información de la etiqueta se agrega como canales adicionales a la imagen RGB, repitiendo el vector codificado en caliente para llenar la misma forma que las imágenes de entrada.

Las CGAN son efectivas porque el *discriminador* tiene acceso a información adicional sobre el contenido de la imagen. Por lo tanto, el *generador* debe asegurarse de que su salida concuerde con la etiqueta proporcionada para seguir engañando al *discriminador*. Si el *generador* produjera imágenes que no coincidieran con la etiqueta de la imagen, el *discriminador* podría identificarlas como falsas simplemente porque las imágenes y las etiquetas no coinciden. Este enfoque condicional amplía significativamente las capacidades de las GAN al permitir la generación de imágenes específicas según ciertas condiciones proporcionadas.

3.4.3. Flujos Normalizadores (Normalization Flow (NF))

Los NF, como señala Yang (Yang et al., 2023), son modelos generativos que poseen la capacidad de generar distribuciones manejables para la representación de datos de alta dimensionalidad. Estos flujos tienen la habilidad de transformar una distribución de probabilidad simple en una distribución de probabilidad altamente compleja, lo que resulta beneficioso en diversas aplicaciones, como modelos generativos, aprendizaje por refuerzo, inferencia variacional y otros campos relacionados.

Tal como aclara Foster (2023), los NF comparten similitudes con los ARM y los VAE en su intento de modelar explícita y factiblemente la distribución generadora de datos p_x. Al igual que los VAE, los NF buscan mapear los datos hacia una distribución más simple, como una *distribución gaussiana*[8]. La diferencia clave radica en que los NF imponen una restricción en la forma de la función de mapeo, garantizando que sea invertible y, por lo tanto, utilizable para generar nuevos puntos de datos.

[7] Un "vector codificado en caliente" (o en inglés "one-hot encoded vector") se refiere a una representación de datos categóricos donde cada categoría se representa como un vector binario, donde un único bit está activado (establecido en 1) para indicar la pertenencia a una categoría específica (Goodfellow et al., 2016)

[8] La distribución gaussiana, también conocida como distribución normal, es una de las distribuciones de probabilidad más comunes en estadística. Se caracteriza por ser una distribución simétrica en forma de campana, donde la mayoría de los datos se concentran cerca de la media y disminuyen gradualmente hacia los extremos.

En un modelo de NF, la función de decodificación está diseñada para ser la inversa exacta de la función de codificación y, además, de rápida computación, lo que confiere a los NF la propiedad de la factibilidad. Sin embargo, las ANN no son, por defecto, funciones invertibles.

Para abordar este desafío, los NF utilizan una técnica conocida como *cambio de variables*. En el caso general, la ecuación del cambio de variables requiere el cálculo de un *determinante de Jacobiano* [9] altamente complejo, lo cual es impráctico para todos excepto los ejemplos más simples.

Sin embargo, al aplicarlo en la práctica, surgen dos problemas principales:

- En primer lugar, calcular el determinante de una matriz de alta dimensión es *computacionalmente costoso*, específicamente, es $O(n^3)$; lo que en otras palabras significa que es completamente impráctico de implementar.
- En segundo lugar, no es inmediatamente obvio cómo se debería calcular la función inversa f_x.

Para resolver estos dos problemas, es necesario utilizar una arquitectura especial de ANN que asegure que la función de cambio de variables f sea invertible y tenga un determinante fácil de calcular. Precisamente se utiliza una técnica llamada transformaciones de preservación de volumen no reales (RealNVP). Este modelo introduce un nuevo tipo de capa, llamada *capa de acoplamiento*.

Las *capas de acoplamiento* son aquellas capas que producen un factor de escala y una traslación para cada elemento de su entrada. En otras palabras, generan dos tensores del mismo tamaño que la entrada, uno para el factor de escala y otro para la traslación.

Es importante destacar que estas capas enmascaran los datos mientras fluyen a través de la red, de modo que el Jacobiano resulta ser triangular

[9] El determinante Jacobiano es una medida de cuánto cambian las variables de salida de una transformación respecto a las variables de entrada. En el contexto del cálculo multivariable y la optimización, el determinante Jacobiano se utiliza para calcular la tasa de cambio de una transformación en un punto dado del espacio. Es especialmente útil en problemas de optimización y en la transformación de coordenadas entre sistemas de referencia. (Strang, 2009)

inferior y, por lo tanto, tiene un determinante fácil de calcular. La visibilidad completa de los datos de entrada se logra mediante la inversión de las máscaras en cada capa.

Por diseño, las operaciones de escala y traslación pueden invertirse fácilmente, lo que permite ejecutar los datos a través de la red en reversa una vez que el modelo está entrenado. Esto significa que se puede dirigir el proceso de transformación hacia adelante hacia una distribución gaussiana estándar, de la cual es fácil muestrear y luego ejecutar los puntos muestreados hacia atrás a través de la red para generar nuevas observaciones.

3.4.4. Modelos Autoregresivos (Autoregresives Model (ARM))

Según Foster (Foster, 2023), los ARM representan una familia de modelos que simplifican el problema de modelado generativo al tratarlo como un proceso secuencial. Los ARM condicionan las predicciones en valores anteriores en la secuencia, en lugar de en una variable aleatoria latente. Por lo tanto, intentan modelar explícitamente la distribución que genera los datos en lugar de una aproximación de ella.

Los ARM, como explica Yang (Yang et al., 2023), operan dividiendo la distribución conjunta de los datos en una serie de distribuciones condicionales utilizando la regla de la cadena de probabilidad. Esta descomposición se realiza desglosando la probabilidad conjunta de los datos en una secuencia de probabilidades condicionales, donde cada una depende de los datos anteriores hasta cierto punto.

Los avances recientes en el campo del DL han impulsado un progreso sustancial en una amplia variedad de tipos de datos, incluyendo imágenes, audio y texto. Los ARM emergieron como poderosas herramientas generativas, ya que emplean una única red neuronal para esta tarea. Sin embargo, el proceso de muestreo en los ARM es intensivo en términos computacionales, ya que requiere un número de llamadas a la red igual a la dimensionalidad de los datos. Esto significa que, aunque los ARM son eficaces como estimadores de densidad, el proceso de muestreo puede ser lento y continuo, especialmente cuando se trata de datos de alta dimensionalidad.

Entre los ARM más conocidos se encuentran *Long Short-Term Memory* (LSTM) y T para la generación de texto, así como *PixelCNN* para imágenes.

Los LSTM representan un tipo particular de Red Neuronal Recurrente (del inglés *Recurrent Neural Network* (RNN))[10].

Sin embargo, hay varias diferencias significativas entre los *datos de texto* y los *datos de imagen* que implican que muchos de los métodos exitosos para los datos de imagen no son tan fácilmente aplicables a los datos de texto. A continuación, se mencionan algunas diferencias:

- Los *datos de texto* están compuestos por fragmentos discretos, ya sean caracteres o palabras, mientras que los píxeles en una imagen son puntos en un espectro continuo de colores. Es sencillo modificar un píxel verde para que sea más azul, pero no es tan evidente cómo deberíamos modificar la palabra "gato" para que sea más similar a la palabra "perro", por ejemplo.
- Los *datos de texto* tienen una dimensión temporal pero no una dimensión espacial, mientras que los *datos de imagen* tienen dos dimensiones espaciales pero no una dimensión temporal. El orden de las palabras es crítico en los *datos de texto* y las palabras no tendrían sentido si se invierte su orden, mientras que las imágenes generalmente pueden ser volteadas sin afectar su contenido.
- Los *datos de texto* son altamente sensibles a pequeños cambios en las unidades individuales (palabras o caracteres), mientras que los *datos de imagen* son generalmente menos sensibles a los cambios en las unidades de píxeles individuales.
- Los *datos de texto* tienen una estructura gramatical basada en reglas, mientras que los *datos de imagen* no siguen reglas establecidas sobre cómo deben asignarse los valores de los píxeles.

Para procesar textos con ARM como LSTM, el primer paso suele ser limpiar y tokenizar el texto, convirtiéndolo en una secuencia de tokens comprensibles para la red neuronal. Esto permite que la red aprenda a generar texto secuencialmente, teniendo en cuenta el contexto de las palabras anteriores.

[10]Las RNN contienen una capa recurrente (o celda) que puede manejar datos secuenciales, haciendo que su propia salida en un momento dado forme parte de la entrada al siguiente paso de tiempo

La arquitectura del modelo LSTM en general consiste en una entrada al modelo que es una secuencia de tokens enteros y una salida que es la probabilidad de que cada palabra en el vocabulario de 10000 palabras aparezca a continuación en la secuencia. Para comprender cómo funciona esto en detalle, es necesario describir dos capas: *Incrustación* (Embedding) y *LSTM*.

a) *La Capa de Incrustación (Embedding):* La *capa de Incrustación* funciona esencialmente como una tabla de búsqueda que convierte cada token entero en un vector de longitud tamaño incrustación (Figura 3.10). Los vectores de búsqueda son aprendidos por el modelo como pesos, lo que significa que el número total de pesos aprendidos por esta capa es igual al tamaño del vocabulario multiplicado por la dimensión del vector de incrustación (por ejemplo, $10000 \times 100 = 1000000$).

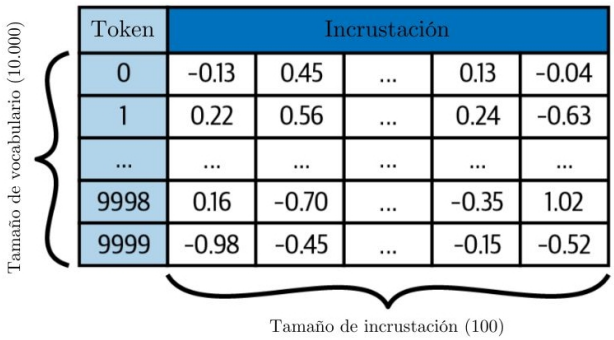

Figura 3.10. Capa de incrustación. Gráfico extraído de Foster (2023)

Cada token entero se convierte en un vector continuo mediante la incrustación en una capa de incrustación. Esto permite al modelo aprender una representación para cada palabra que puede ser actualizada a través de la retropropagación. El uso de esta capa de incrustación es preferible porque hace que la incrustación sea entrenable en sí mismo, lo que otorga al modelo más flexibilidad para determinar cómo incrustar cada token y mejorar su rendimiento.

b) *La Capa LSTM:* La capa LSTM se basa en una estructura recurrente general (Figura 3.11). Esta capa tiene la capacidad especial de procesar datos de entrada secuenciales $x_1, \ldots, x_n$. Consiste en una celda que actualiza su estado oculto, h_t, a medida que cada elemento de la secuencia x_t se pasa a través de ella, paso a paso.

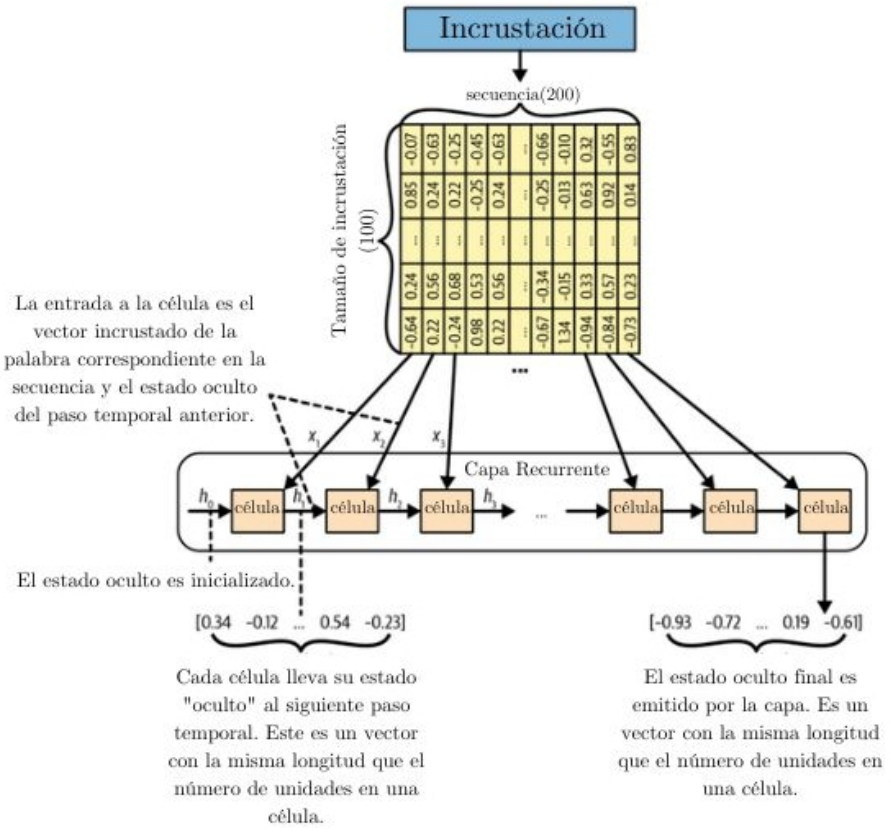

Figura 3.11. Funcionamiento de una RNN. Gráfico extraído de Foster (2023)

El *estado oculto* es un vector con longitud igual al número de unidades en la celda y representa la comprensión actual de la secuencia por parte de la celda. En el paso de tiempo t, la celda utiliza el valor anterior del estado oculto, h_{t-1}, junto con los datos del paso de tiempo actual x_t, para generar un vector de estado oculto actualizado, h_t. Este proceso recurrente continúa hasta el final de la secuencia. Una vez que la secuencia ha concluido, la capa produce el estado oculto final de la celda, h_n, que luego se pasa a la siguiente capa de la red.

La función principal de la celda LSTM es generar un *nuevo estado oculto*, h_t, dados su *estado oculto anterior*, h_{t-1}, y la incrustación de palabras actual, x_t. Es importante destacar que la longitud de h_t es igual al número de unidades en la LSTM, un parámetro que se define al configurar la capa y no está relacionado con la longitud de la secuencia.

Además del *estado oculto*, una celda LSTM mantiene un estado de celda, C_t, que se puede entender como las creencias internas de la celda

sobre el estado actual de la secuencia. A diferencia del estado oculto, h_t, que se produce al final del proceso, el estado de la celda tiene la misma longitud que el estado oculto (es decir, el número de unidades en la celda).

Una aplicación de modelos autoregresivos fue introducida por van de Oord (van den Oord et al., 2016) con *PixelCNN*, el cual se trata de un modelo que genera imágenes píxel por píxel al predecir la probabilidad del siguiente píxel basándose en los píxeles anteriores. Este enfoque permite la generación de imágenes de forma autoregresiva. PixelCNN se apoya en dos conceptos clave: *capas convolucionales enmascaradas* y *bloques residuales*.

- Para aplicar *capas convolucionales* a la generación de imágenes de manera autoregresiva, es necesario ordenar los píxeles y asegurarse de que los filtros solo puedan ver los píxeles anteriores al píxel en cuestión. Luego, se pueden generar imágenes un píxel a la vez, aplicando filtros convolucionales a la imagen actual para predecir el valor del próximo píxel a partir de todos los píxeles anteriores.
- Un *bloque residual* comprende un conjunto de capas donde la salida se suma a la entrada antes de pasar al resto de la red. Esto se logra mediante una conexión de salto, donde la entrada tiene una ruta directa hacia la salida sin pasar por las capas intermedias. La inclusión de esta conexión de salto permite a la red aprender transformaciones más complejas al evitar la necesidad de encontrar un mapeo de identidad a través de las capas intermedias.

Para evitar que el proceso de entrenamiento sea lento debido a la gran cantidad de valores de píxeles independientes (256 para una imagen en escala de grises), *PixelCNN* utiliza una distribución de mezcla en la salida en lugar de un *softmax*[11] sobre los 256 valores de píxeles discretos. Esta distribución de mezcla consiste en una mezcla de varias distribuciones de probabilidad, como distribuciones logísticas con diferentes parámetros. Además, se necesita una distribución categórica discreta que denote la probabilidad de elegir cada una de las distribuciones incluidas en la mezcla.

[11]La función softmax es una función de activación utilizada comúnmente en redes neuronales para convertir un vector de números reales en un vector de probabilidades. La salida de la función softmax es una distribución de probabilidad que asigna una probabilidad a cada posible clase o categoría. Esta función es especialmente útil en la capa de salida de redes neuronales utilizadas para clasificación multiclase, ya que garantiza que las probabilidades sumen uno y permite interpretar la salida como una distribución de probabilidad.(Goodfellow et al., 2016)

3.4.5. Modelos Basados en Energía (Energy Based Models (EBM))

Según explica Foster (Foster, 2023), los EBM son una amplia clase de modelos generativos que adoptan una idea fundamental de la modelización de sistemas físicos: la probabilidad de un evento puede expresarse utilizando una *distribución de Boltzmann* [12], una función que normaliza una función de energía entre 0 y 1.

Los EBM, como menciona Yang (Yang et al., 2023), representan una variante generativa de los discriminadores y tienen la capacidad de aprender a partir de datos de entrada no etiquetados.

Para comprender su funcionamiento, considerar una muestra de entrenamiento, denotada como x, que sigue una distribución de probabilidad $p_{\text{data}}(x)$. El objetivo de un modelo basado en energía es aproximar esta distribución $p_{\text{data}}(x)$ mediante una función de densidad de probabilidad $p_0(x)$, donde θ representa los parámetros del modelo.

La expresión que define un modelo basado en energía es bastante compleja, y la parte fundamental es la función de partición Z_0, que se calcula como figura en la siguiente Ecuación 3.2:

$$\int e^{f_0(x)}\, dx \tag{3.2}$$

Esta función de partición es especialmente desafiante de calcular cuando se trabaja con datos de alta dimensionalidad, como imágenes.

La función $f_0(x)$ se parametriza típicamente mediante una CNN que produce un valor escalar. Esto implica entrenar una ANN E_{X} para asignar puntajes bajos a observaciones probables (para que p_{X} esté cerca de 1) y puntajes altos a observaciones improbables (para que p_{X} esté cerca de 0).

La idea clave detrás de un EBM es que se puedan usar técnicas de aproximación para evitar la necesidad de calcular el denominador intratable.

Los EBM evitan este problema mediante dos estrategias: la divergencia contrastiva para entrenar la red y la dinámica de Langevin para muestrear

[12] La distribución de Boltzmann fue propuesta originalmente por Ludwig Boltzmann en 1868 para describir gases en equilibrio térmico

nuevas observaciones, basándose en las ideas presentadas en el artículo de los autores Du y Mordatch (Du & Mordatch, 2019).

La *función de energía* $E_\theta(x)$ es una ANN con parámetros θ que transforma una imagen de entrada x en un valor escalar. A lo largo de esta red, se utiliza una función de activación llamada *Swish*[13], la cual puede calcularse como figura en la Ecuación 3.3

$$\text{swish}(x) = x \cdot \text{sigmoid}(x) = \frac{x}{e^{-x} + 1} \tag{3.3}$$

En la práctica, se aproxima el proceso de muestreo de las muestras falsas para garantizar la eficiencia del algoritmo. El muestreo de EBM profundos se logra mediante la *dinámica de Langevin*[14], una técnica que utiliza el gradiente de la puntuación con respecto a la imagen de entrada para transformar gradualmente el ruido aleatorio en una observación plausible mediante la actualización de la entrada en pequeños pasos, siguiendo el gradiente cuesta abajo.

3.5. Modelos de Difusión

Los DM, según Yang et al. (2023), constituyen una familia de modelos generativos probabilísticos que introducen ruido de manera progresiva en los datos y luego aprenden a revertir este proceso para la generación de muestras.

Estos modelos han surgido como la nueva metodología en la generación de modelos generativos, desafiando la larga hegemonía de las GAN en la compleja tarea de sintetizar imágenes. Además, han demostrado un gran potencial en una amplia gama de dominios, desde la CV, el NLP y el modelado de datos temporales, hasta la modelización multimodal, el aprendizaje robusto

[13]Swish es una alternativa a ReLU introducida por Google en 2017. Visualmente, Swish es similar a ReLU, pero es suave, lo que ayuda a mitigar el problema del gradiente desvaneciente.

[14]La dinámica de Langevin es un enfoque utilizado en física estadística y simulaciones computacionales para describir el comportamiento de partículas en medios con influencias estocásticas, como la fricción y la agitación térmica. Se basa en la ecuación de Langevin, que tiene en cuenta tanto las fuerzas deterministas como las fluctuaciones aleatorias, modeladas usualmente como ruido blanco gaussiano, para predecir la evolución temporal de un sistema(Frenkel & Smit, 2002).

y aplicaciones interdisciplinarias en campos como la química computacional o bien en la reconstrucción de imágenes médicas.

Según Foster (2023), los fundamentos de los DM comparten similitudes con tipos anteriores de modelos generativos, como los AE de eliminación de ruido y los EBM.

En 2015, se estableció un vínculo entre la difusión termodinámica y el DL, lo que inspiró el nombre "difusión" para estos modelos generativos (Sohl-Dickstein et al., 2015). Yang Song y Stefano Ermon desarrollaron una *Red de Puntuación Condicional al Ruido* (del inglés *Noise Conditional Score Network* (NCSN)), que empleaba perturbaciones de ruido múltiples para mejorar la eficiencia en regiones de baja densidad de datos. Sin embargo, el punto de inflexión en el campo de los DM ocurrió en 2020 (Song & Ermon, 2020), cuando se descubrió una conexión profunda entre estos modelos y los modelos generativos basados en puntajes (del inglés *Score-Based Generative Models* (SGM)). Este hallazgo condujo al desarrollo del Modelo Probabilístico de Difusión de Eliminación de Ruido (del inglés *Diffusion Probabilistic Model* (DDPM)), capaz de rivalizar con las GAN en múltiples conjuntos de datos.

3.6. Tipos de Modelos de Difusión

Según Yang (Yang et al., 2023) la investigación actual sobre DM se basa principalmente en tres formulaciones predominantes: DDPM, SGM, y *Ecuaciones Diferenciales Estocásticas de Puntuación* (del inglés *Score-Based Stochastic Differential Equations* (Score SDE)).

3.6.1. Modelo Probabilístico de Difusión por Eliminación de Ruido (DDPMs)

Según autores como Yang (Yang et al., 2023) y Ho (Ho et al., 2020), un DDPM hace uso de dos *cadenas de Markov*[15] es decir,una cadena hacia

[15]Una *cadena de Markov* es un modelo estocástico que describe una secuencia de eventos donde la probabilidad de que ocurra un evento futuro depende únicamente del evento actual y no de los eventos anteriores. Se caracteriza por tener la propiedad de Markov, que implica que la distribución de probabilidad de transición entre estados solo depende del estado actual y no de cómo se llegó a ese estado. Las cadenas de Markov se utilizan en una amplia variedad de campos, como procesos estocásticos, modelado de sistemas dinámicos, procesamiento de señales y aprendizaje automático.(Durrett, 2016)

adelante que perturba los datos con ruido, y una cadena hacia atrás que convierte el ruido de vuelta en datos.

La primera suele diseñarse con el objetivo de transformar cualquier distribución de datos en una distribución previa simple (por ejemplo, una distribución Gaussiana estándar), mientras que la última cadena de Markov revierte la primera mediante el aprendizaje de núcleos de transición parametrizados por redes neuronales profundas. Los nuevos puntos de datos se generan posteriormente primero muestreando un vector aleatorio de la distribución previa, seguido por un *muestreo ancestral*[16] a través de la cadena de Markov inversa.

Por otro lado, tal como explica Foster, la idea principal detrás de un DDPM es simple: entrenar un modelo de DL para eliminar el ruido de una imagen a lo largo de una serie de pasos muy pequeños. Si se parte de un ruido puro y aleatorio, en teoría se debería poder seguir aplicando el modelo hasta obtener una imagen que parezca haber sido extraída del conjunto de entrenamiento (Foster, 2023).

Para agregar ruido a las imágenes, se utiliza un proceso denominado *difusión hacia adelante*, el cual comienza con una imagen inicial x_0 que se corrompe gradualmente a lo largo de múltiples pasos (por ejemplo, $T = 1000$), con el objetivo final de que sea indistinguible del ruido gaussiano estándar. Para lograr esto, se define una función q que agrega una pequeña cantidad de ruido gaussiano con varianza β_t a una imagen x_{t-1}, generando así una nueva imagen c_t. Al seguir aplicando esta función, se obtiene una secuencia de imágenes con cada vez más ruido $(x_0, ..., x_T)$.

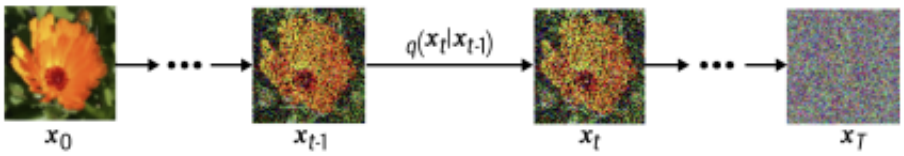

Figura 3.12. Proceso de Difusión Hacia Adelante. Gráfico extraído de Foster (2023)

El uso de un truco de reparametrización permite calcular las imágenes con ruido en cualquier paso del proceso hacia adelante sin la necesidad de

[16]El *muestreo ancestral* es una técnica utilizada en estadísticas y genética para generar muestras de una distribución de probabilidad posterior. En el contexto de la inferencia bayesiana, el muestreo ancestral es una forma de aproximar la distribución posterior al muestrear directamente de la distribución conjunta de las variables de interés y los parámetros del modelo.(Lai, 2016)

realizar múltiples pasos de ruido. Esto puede mejorar significativamente la eficiencia computacional del proceso.

Es crucial elegir un cronograma de parámetros adecuado para agregar ruido a los datos, ya que esto afecta la calidad y la naturaleza de la corrupción gradual de la imagen. Ajustar la varianza del ruido (β_t) a lo largo del tiempo es especialmente importante, ya que controla la velocidad a la que la imagen se corrompe gradualmente y garantiza que, al final del proceso, sea indistinguible del ruido gaussiano estándar. La selección cuidadosa de estos parámetros es fundamental para el éxito general del DM hacia adelante.

Por otro lado, el *proceso de difusión inversa* tiene como objetivo desarrollar una ANN que cumpla con la Ecuación 3.4:

$$p_\theta(x_{t-1} \mid x_t) \tag{3.4}$$

y que sea capaz de deshacer el proceso de corrupción de ruido, es decir, aproximando la distribución inversa como figura en la Ecuación 3.5:

$$q(x_{t-1} \mid x_t) \tag{3.5}$$

Si se logra este objetivo, es posible muestrear ruido aleatorio de una distribución inicial p_0 y luego aplicar el proceso de difusión inversa varias veces para generar una imagen nueva y original.

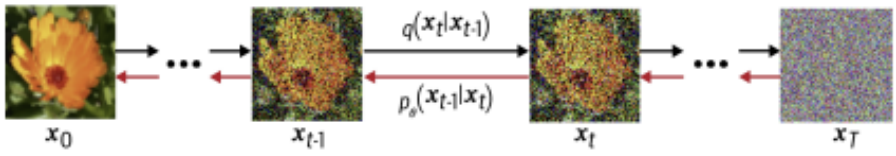

Figura 3.13. Proceso de Difusión Inversa. Gráfico extraído de Foster (2023)

Existen similitudes entre el proceso de difusión inversa y el decodificador de un VAE. En ambos casos, se busca transformar ruido aleatorio en una salida significativa utilizando una red neuronal. La diferencia principal radica en que en un VAE, el proceso hacia adelante (es decir, convertir imágenes en ruido) es parte del modelo y se aprende durante el entrenamiento, mientras que en un DM, este proceso no está parametrizado. Por lo tanto, tiene sentido

aplicar la misma función de pérdida que en un autocodificador variacional. En la propuesta original de DDPM (Ho et al., 2020) se derivó la forma exacta de esta función de pérdida y se demostró que se puede optimizar entrenando una red $\mathcal{U}_\theta$ para predecir el ruido $\mathcal{U}$ que se agregó a una imagen $\tilde{x}_0$ en el paso de tiempo t.

El *proceso de difusión hacia atrás* está parametrizado por una arquitectura de ANN conocida como *U-Net*, que intenta predecir el ruido en cada paso de tiempo dada la imagen corrompida con ruido y la tasa de ruido en ese paso.

El *proceso de difusión inversa* está dirigido por una *U-Net*, la cual busca anticipar el nivel de ruido en cada etapa temporal, partiendo de la imagen con ruido y la cantidad de ruido presente en dicha etapa.

Los autores del artículo que dió origen a DDPM (Ho et al., 2020) emplearon una arquitectura conocida como *Modelo de quita de ruido U-Net*. Esta arquitectura consta de bloques de muestreo descendente que aumentan el número de canales mientras reducen el tamaño de la imagen, y bloques de muestreo ascendente que hacen lo opuesto, es decir, disminuyen el número de canales mientras aumentan el tamaño de la imagen. La tasa de ruido se codifica utilizando incrustaciones sinusoidales.

Similar a un VAE, una U-Net se compone de dos partes:

a) una mitad de *muestreo descendente*, donde las imágenes de entrada se comprimen espacialmente pero se expanden en número de canales,

b) y una mitad de *muestreo ascendente*, donde las representaciones se expanden espacialmente mientras que el número de canales se reduce.

Lo distintivo de una U-Net radica en sus conexiones de omisión, que permiten que la información evite ciertas partes de la red y fluya hacia capas posteriores.

Este tipo de arquitectura resulta especialmente útil cuando se busca que la salida mantenga la misma forma que la entrada. Por ejemplo, en el caso de un DM que busca predecir el ruido añadido a una imagen, que debería tener la misma forma que la imagen original, una U-Net se presenta como la elección natural para la arquitectura de la red.

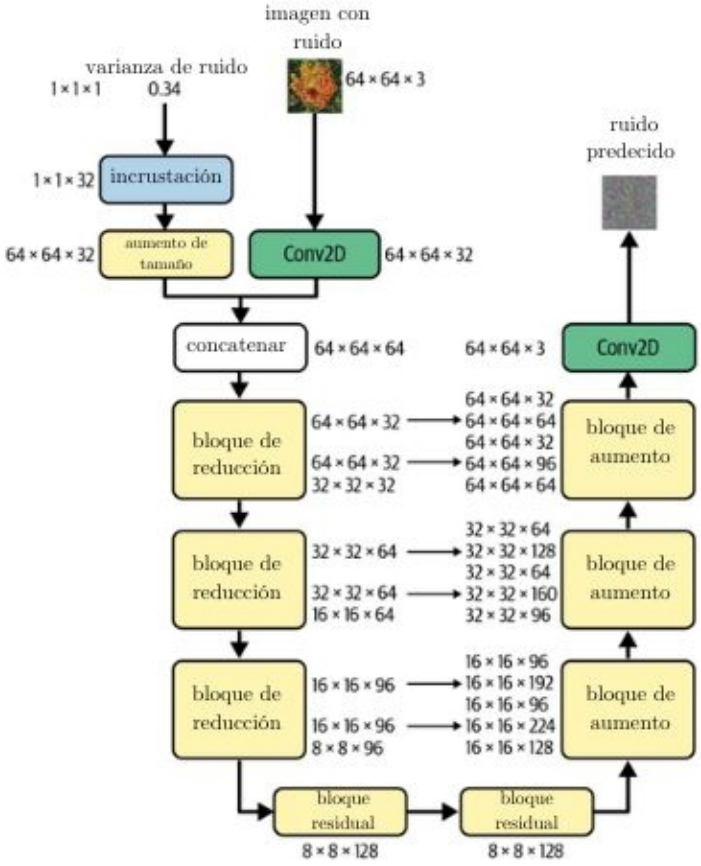

Figura 3.14. Arquitectura U-Net. Gráfico extraído de Foster (2023)

3.6.2. Modelos Generativos Basados en Puntuación (SGM)

La idea clave de los SGM, de acuerdo a Song y Ermon, es perturbar los datos con una secuencia de ruido gaussiano intenso y estimar conjuntamente las funciones de puntuación para todas las distribuciones de datos ruidosas mediante el entrenamiento de un modelo NCSN (Song & Ermon, 2019).

En el núcleo de los SGM (Song & Ermon, 2019, 2020) se encuentra el concepto de la *función de puntuación* (también conocida como función de score de Stein[17]).

[17]La *función de puntuación de Stein* es una herramienta estadística utilizada en estimación paramétrica, que se utiliza para calcular la puntuación de un estimador de máxima verosimilitud en modelos estadísticos donde los parámetros no pueden estimarse directamente a través de métodos de diferenciación estándar.(Geyer, 2010)

Dada una función de densidad de probabilidad $p(x)$, su función de puntuación se define como el gradiente del logaritmo de la densidad de probabilidad, tal como se ve en la Ecuación 3.6:

$$\nabla_x \log p(x) \tag{3.6}$$

A diferencia de la *puntuación de Fisher*[18] comúnmente utilizada en estadísticas con la Ecuacion 3.7 :

$$\nabla_\theta \log p_\theta(x) \tag{3.7}$$

La puntuación de Stein considerada aquí es una función de los datos x en lugar del parámetro del modelo θ. Es un campo vectorial que señala direcciones a lo largo de las cuales la función de densidad de probabilidad tiene la mayor tasa de crecimiento.

Las muestras se generan encadenando las funciones de puntuación en niveles de ruido decrecientes con enfoques de muestreo basados en puntuación, incluyendo Monte Carlo Langevin[19] ((Jolicoeur-Martineau et al., 2021b), (Song & Ermon, 2019),(Song & Ermon, 2020)), SDE ((Jolicoeur-Martineau et al., 2021a), (Song & Ermon, 2020)), ecuaciones diferenciales ordinarias ((Karras et al., 2022), (Lu et al., 2022), (Song et al., 2021), (Song & Ermon, 2020)), y sus diversas combinaciones (Song & Ermon, 2020).

El entrenamiento y el muestreo están completamente desacoplados en la formulación de los SGM, por lo que se puede utilizar una multitud de técnicas de muestreo después de la estimación de las funciones de puntuación.

3.6.3. Ecuaciones Diferenciales Estocásticas (SDE de Puntuación)

Los DDPM y los SGM pueden generalizarse aún más al caso de un número infinito de pasos de tiempo o niveles de ruido, donde los procesos de

[18]La *función de puntuación de Fisher* es una medida utilizada en estadística para estimar los parámetros de un modelo de manera iterativa. Esta función se utiliza en el contexto de la estimación de máxima verosimilitud para calcular los gradientes de la función de log-verosimilitud con respecto a los parámetros del modelo.(Rice, 2006)

[19]La *función de Monte Carlo Langevin* es un algoritmo utilizado en simulaciones computacionales para generar muestras de una distribución de probabilidad deseada. Este algoritmo combina métodos de Monte Carlo con la dinámica de Langevin(Liang & Liu, 2009)

perturbación y desruido son soluciones a SDE. Esta formulación, conocida como SDE de Puntuación (Song & Ermon, 2020), aprovecha las SDE para la perturbación del ruido y la generación de muestras, mientras que el proceso de desruido implica la estimación de las funciones de puntuación de las distribuciones de datos ruidosos.

3.7. Modelos de Difusión Condicionales

Según la definición de Yang (Yang et al., 2023), existen dos paradigmas de aplicación básicos de los DM: DM incondicionales y DM condicionales.

La evolución de los DM sigue un patrón similar al de otros modelos generativos, como VAE, GAN y NF. Inicialmente, todos estos modelos se enfocaron en la generación incondicional, y posteriormente se desarrolló la generación condicional.

La *generación incondicional* se utiliza frecuentemente para evaluar el rendimiento máximo del modelo generativo. Por otro lado la *generación condicional* se centra más en el contenido específico para aplicaciones, ya que permite controlar los resultados de la generación de acuerdo con las intenciones pretendidas.

A continuación, se explicarán las diferentes formas de condicionamiento que existen para los DM:

- *Mecanismos de Condicionamiento en Modelos de Difusión*

  Los DM destacan por su *capacidad de controlabilidad*. Hacen uso extensivo de diversas formas de condicionamiento para dirigir la generación, incluyendo etiquetas, clasificadores, texto, imágenes, mapas semánticos, gráficos, entre otros. Sin embargo, algunas de estas condiciones son estructuralmente complejas.

  Principalmente, existen cuatro tipos de mecanismos de condicionamiento: *concatenación*, *basado en gradientes*, *atención cruzada* y *normalización adaptativa de capas*.

  a) *Concatenación:* significa que los DM concatenan orientación informativa con objetivos de eliminación de ruido intermedios en el proceso de difusión, como la incrustación de etiquetas y los mapas de características semánticas.

b) *Basado en gradientes:* incorpora el gradiente relacionado con la tarea en el proceso de muestreo de difusión para la generación controlable.

c) *Atención cruzada:* realiza un paso de mensajes atencional entre la orientación y los objetivos de difusión, lo cual suele llevarse a cabo de manera por capas en redes de eliminación de ruido.

d) *Normalización adaptativa de capas:* exploran la sustitución de capas normales estándar en los esqueletos de difusión basados en transformadores con normalización adaptativa de capas

- *Condicionamiento de la Difusión en Etiquetas y Clasificadores*

  El condicionamiento del proceso de difusión *en función de las etiquetas* es una forma directa de incorporar propiedades deseadas en las muestras generadas. Sin embargo, cuando las etiquetas son limitadas, resulta difícil para los DM capturar completamente toda la distribución de datos.

- *Condicionamiento de la Difusión en Textos, Imágenes y Mapas Semánticos*

  Las investigaciones recientes han comenzado a condicionar el proceso de difusión *en función de semánticas más ricas*, como textos, imágenes y mapas semánticos, con el fin de expresar mejor las complejas características semánticas en las muestras generadas.

- *Condicionamiento de la Difusión en Grafos*

  Los *datos estructurados en forma de grafos* suelen exhibir relaciones complejas entre los nodos, lo que hace extremadamente difícil el condicionamiento sobre grafos para los DM.

3.8. Aplicaciones de los Modelos de Difusión

En los últimos años, las aplicaciones de los DM se extendieron a diversas áreas, incluyendo: CV, NLP,generación multimodal, modelado de datos temporales, aprendizaje robusto y aplicaciones interdisciplinarias.

3.8.1. Visión por Computadora

Los modelos generativos han demostrado ser efectivos en una variedad de tareas de restauración de imágenes en el campo de la CV.

Las tareas que se incluyen son:

- *Super resolución*: para generar imágenes condicionales de alta resolución (por ejemplo SR3 (Saharia et al., 2021b)).
- *Relleno de vacíos*: Métodos como RePaint (Lugmayr et al., 2022) utilizan estrategias mejoradas de reducción de ruido con iteraciones de remuestreo para lograr una mejor condición de la imagen. Palette (Saharia et al., 2021a) emplea DM condicionales para abordar varias tareas de generación de imágenes en un marco unificado.
- *Traducción de imágenes*: SDEdit (Meng et al., 2021) utiliza ecuaciones diferenciales estocásticas previas para mejorar la fidelidad en la traducción de imágenes con estilos específicos.
- *Resolución de problemas inversos*: DDRM (Kawar et al., 2022) aprovecha modelos generativos de difusión de eliminación de ruido preentrenados para resolver problemas inversos lineales en diversas tareas de restauración de imágenes.

Además existen otros enfoques dentro de la CV que aprovechan tanto la potencia de los modelos generativos como la capacidad de los DM para capturar y modelar la complejidad de los datos visuales. Entre ellos se encuentran:

- La *segmentación semántica*[20] es una tarea crucial en el campo de la visión por computadora. Métodos recientes han explorado el potencial de las representaciones aprendidas a través de DDPM en esta área específica.

  Por ejemplo, DDeP (Asiedu et al., 2022) integra DM con AE de eliminación de ruido, lo que resulta en un enfoque prometedor para la segmentación semántica eficiente en etiquetas.

[20]La *segmentación semántica* implica etiquetar cada píxel de una imagen según categorías de objetos establecidas

- La *generación de videos* de alta calidad sigue siendo un desafío debido a la complejidad y la continuidad espacio-temporal de los fotogramas de video. Sin embargo, la investigación reciente ha demostrado que los DM pueden ser una herramienta efectiva para mejorar la calidad de los videos generados.

  Por ejemplo, el modelo FDM (Harvey et al., 2022) permite el muestreo de cualquier subconjunto arbitrario de fotogramas de video, lo que brinda flexibilidad en la generación de videos. Además, el modelo RVD (Yang et al., 2022c) utiliza un enfoque autoregresivo basado en un DM de video, lo que permite generar videos de alta calidad manteniendo la coherencia temporal entre los fotogramas.

  Por lo tanto, estos enfoques muestran el potencial de los DM en la generación de videos realistas y de alta fidelidad.

- Los DM también demostraron ser eficaces para abordar el desafío de la *completitud de nubes de puntos* 3D, utilizando técnicas que les permiten inferir partes faltantes y reconstruir formas completas.

  Por ejemplo, tanto el modelo PVD (Zhou et al., 2021) como el modelo PDR (Lyu et al., 2021) han sido utilizados con éxito en la reconstrucción de formas 3D completas a partir de observaciones parciales. Estos métodos aprovechan la capacidad de los DM para aprender representaciones robustas de datos tridimensionales y completar información faltante en las nubes de puntos, lo que resulta en una reconstrucción más precisa y detallada de las formas.

- La *detección de anomalías* es un desafío crítico en el campo del ML y la CV. En este sentido, modelos generativos como AnoDDPM (Wyatt et al., 2022) y DDPM-CD (Bandara et al., 2022) han demostrado ser prometedores, haciendo uso de técnicas de difusión para corromper imágenes de entrada y luego reconstruir aproximaciones saludables, lo que se ha mostrado como una estrategia potencialmente más efectiva que las alternativas basadas en el entrenamiento adversarial.

3.8.2. Generación de Lenguaje Natural (NLP)

La generación de texto se ha convertido en una tarea crítica y desafiante en el NLP. Busca componer texto plausible y legible en lenguaje humano a partir de diferentes fuentes, como texto, audio o incluso ruido aleatorio. Para abordar esta tarea, se han desarrollado numerosos enfoques basados en DM.

Por ejemplo, D3PM (Austin et al., 2021) introduce modelos generativos similares a la difusión para la generación de texto a nivel de caracteres. Amplía el DM multinomial a través de procesos de corrupción con probabilidades de transición uniformes. Además, los grandes ARM pueden generar texto de alta calidad. Sin embargo, para desplegarlos de manera confiable en aplicaciones del mundo real, es crucial que el proceso de generación de texto sea controlable.

El control del comportamiento de los modelos de lenguaje sin necesidad de volver a entrenarlos es un problema importante en la generación de texto. Por ejemplo, Analog Bits (Chen et al., 2022) genera los bits analógicos para representar las variables discretas y mejora aún más la calidad de la muestra con autocondicionamiento e intervalos de tiempo asimétricos. Este enfoque permite ajustar y dirigir la generación de texto según los requisitos deseados, como tema y estructura sintáctica, lo que mejora significativamente la capacidad de estos modelos para adaptarse a diferentes escenarios de aplicación.

3.8.3. Generación Multimodal

El aprendizaje generativo tradicionalmente se ha centrado en una sola modalidad de datos, como texto, imágenes o música. Sin embargo, los humanos son capaces de cruzar fácilmente entre diferentes modalidades, como escribir una descripción de una imagen. Para las computadoras, esto representa un desafío, ya que requiere un enfoque diferente.

El entrenamiento de modelos generativos para convertir entre dos o más tipos diferentes de datos se conoce como *aprendizaje multimodal* (Foster, 2023).

En los últimos años, han surgido modelos generativos multimodales impresionantes que pueden manejar múltiples tipos de datos simultáneamente. Estos modelos abren nuevas oportunidades para aplicaciones que requieren comprensión y generación de datos en diferentes modalidades, como traducción de imágenes a texto, descripción de vídeos o síntesis de música basada en imágenes. Entre estos tipos de modelos multimodales se destacan:

- *Generación de Imagen desde Grafos de Escena:* A pesar de los avances en la generación de texto a imagen, la reproducción fiel de oraciones

complejas con muchos objetos y relaciones sigue siendo un desafío. SG-Diff (Yang et al., 2022b) propone abordar esta dificultad mediante un DM diseñado específicamente para la *generación de imágenes a partir de gráfos de escena*. Este modelo aprende una incrustación de gráfico de escena continua para condicionar el espacio latente del DM. Este enfoque permite capturar y representar de manera más efectiva las complejas relaciones y estructuras presentes en las escenas, lo que potencialmente mejora la calidad y la fidelidad de las imágenes generadas.

- La *generación de contenido 3D a partir de texto* ha ganado interés en diversas aplicaciones. Ejemplos de enfoques en esta área incluyen DreamFusion (Poole et al., 2022) y Magic3D (Lin et al., 2022), ambos utilizan DM para sintetizar contenido tridimensional a partir de descripciones de texto. Estos modelos permiten la creación de mundos virtuales y escenarios tridimensionales basados en las especificaciones proporcionadas en texto, lo que amplía las posibilidades de creación y diseño en entornos virtuales y de realidad aumentada.

- La *generación de movimiento humano* es fundamental en la animación por computadora, con aplicaciones que van desde juegos hasta robótica. Ejemplos de modelos adaptados para esta tarea son MDM (Tevet et al., 2022) y FLAME (Kim et al., 2022), los cuales utilizan técnicas basadas en T para manejar datos de movimiento y texto. Estos modelos permiten la creación de animaciones realistas y fluidas, lo que contribuye significativamente a la creación de contenido virtual y al desarrollo de aplicaciones interactivas y de entretenimiento.

- *Generación de Texto a Video:* El reciente progreso en la generación basada en difusión de texto a imagen motivó el desarrollo de la *generación de texto a video*. Ejemplos como Imagen Video (Ho et al., 2022) y FateZero (Qi et al., 2023) son enfoques que extienden modelos de texto a imagen para generar videos. Estos modelos aprovechan las capacidades de los DM para crear videos realistas y coherentes a partir de descripciones textuales, abriendo nuevas posibilidades en áreas como la producción de contenido multimedia y la generación de material audiovisual automatizado.

- *Generación de Texto a Audio:* La *generación de texto a audio* es la tarea de transformar textos del lenguaje natural en salidas de voz. Ejemplos como Diffsound (Yang et al., 2022a) y EdiTTS (Tae et al., 2021) son modelos que utilizan técnicas basadas en difusión para esta tarea. Estos enfoques aprovechan los modelos generativos para crear voces sintéticas

que suenan naturales y expresivas, lo que tiene aplicaciones importantes en tecnologías de asistencia, narración de contenido y generación de contenido multimedia.

- También se incluyen en esta categoría la *Generación de Texto a Imagen* (que se profundizará en la Sección 3.10)

3.8.4. Modelado de Datos Temporales

El *modelado de datos temporales* en AI constituye un área crucial de investigación que busca capturar y comprender la dinámica temporal inherente a numerosos fenómenos del mundo real. Entre sus aplicaciones se encuentran:

- *Imputación de Series Temporales:* Las *series temporales* se utilizan ampliamente en muchas aplicaciones del mundo real y suelen contener valores faltantes por diversas razones. Para abordar este problema, se han desarrollado métodos de imputación, incluyendo enfoques basados en difusión. Un ejemplo destacado es el modelo de Difusión Basado en Puntaje Condicional para Imputación (CSDI) (Tashiro et al., 2021), que introduce un enfoque novedoso aprovechando los SGM para la imputación de series temporales. Además, otros métodos como CSDE (Park et al., 2021) y SSSD (Lopez Alcaraz & Strodthof, 2022) se han propuesto para capturar dependencias a largo plazo en series temporales, mostrando un buen rendimiento en tareas de imputación y pronóstico. Estos enfoques ofrecen soluciones efectivas para manejar valores faltantes en series temporales, lo que es crucial para muchas aplicaciones prácticas, como finanzas, predicción del clima y monitoreo de la salud.

- *Pronóstico de Series Temporales:* El *pronóstico de series temporales* es fundamental para predecir los valores futuros a lo largo de un período de tiempo. En este contexto, los métodos neurales han ganado una amplia aceptación, ya sea en métodos de pronóstico univariados o probabilísticos, tanto en series univariadas como multivariadas. TimeGrad (Rasul et al., 2021) introduce un enfoque autorregresivo para el pronóstico de series temporales multivariadas, utilizando DM y optimizando una variante del límite variacional sobre la probabilidad de los datos. Este enfoque ofrece una metodología avanzada para el pronóstico preciso y probabilístico de series temporales, lo que es fundamental para aplicaciones como finanzas, climatología y análisis de datos industriales.

- *Procesamiento de Señales de Forma de Onda:* En disciplinas como la electrónica y la acústica, la *representación de una señal* se realiza mediante su forma de onda, que describe cómo varía la señal en función del tiempo. WaveGrad (Chen et al., 2020) y DiffWave (Kong et al., 2020) son ejemplos destacados de DM probabilística diseñados para generar formas de onda, utilizando métodos basados en gradientes para mejorar la calidad de las muestras. Estos modelos producen audio de alta fidelidad y se utilizan en una variedad de aplicaciones, desde síntesis de voz hasta generación de música, donde la calidad del audio es crucial.

3.8.5. Aprendizaje Robusto

El *aprendizaje robusto* constituye una categoría de técnicas defensivas diseñadas para conferir robustez a las redes de aprendizaje frente a perturbaciones adversariales o ruido. Aunque el entrenamiento adversarial se ha establecido como un método de defensa estándar contra ataques adversariales en clasificadores de imágenes, la purificación adversarial ha demostrado un rendimiento considerable como alternativa. Este enfoque transforma las imágenes atacadas en imágenes limpias mediante un proceso de purificación independiente. Diversas estrategias, como DiffPure (Nie et al., 2022) y *Purificación de Eliminación de Ruido Adaptativo* (del inglés *Adaptative Denoising Purification* ADP) (Yoon et al., 2021), emplean procesos de difusión para restaurar imágenes limpias a partir de ejemplos adversariales. Otros métodos, como *Descenso de Gradiente Proyectado* (del inglés *Projected Gradent Descent* (PGD)) (Blau et al., 2022), incorporan preprocesamiento robusto basado en difusión estocástica para defensa adversarial. Además, algunos enfoques proponen la aplicación de difusión guiada para una purificación adversarial más avanzada.

3.8.6. Aplicaciones Interdisciplinarias

La sinergia entre la AI y una amplia gama de disciplinas ha propiciado un extenso repertorio de aplicaciones interdisciplinarias, que van desde el diseño de fármacos y biología hasta la reconstrucción de imágenes médicas. Entre ellas se destacan:

- *Diseño de Medicamentos y Ciencias de la Vida:* En el campo del *diseño de fármacos y las ciencias de la vida*, las redes neuronales gráficas y las

técnicas de aprendizaje de representaciones han tenido un gran éxito en la modelización de moléculas y proteínas para diversas aplicaciones, que van desde la predicción de propiedades hasta la generación de moléculas y proteínas completas.

Torsional diffusion (Jing et al., 2022) introduce un novedoso marco de difusión que opera en el espacio de ángulos de torsión, utilizando un proceso de difusión en el hiperspacio y un modelo de puntuación extrínseco a intrínseco. Por otro lado, GeoDiff (Xu et al., 2022) demuestra que las cadenas de Markov que evolucionan con núcleos de Markov equivariantes pueden producir una distribución invariante, utilizando bloques diseñados para preservar la propiedad de equivariancia deseada. Además, se han desarrollado otros métodos que incorporan la propiedad de equivariancia en la generación de moléculas 3D y la generación de proteínas. Por ejemplo, ConfGF (Shi et al., 2021) estima directamente los campos de gradiente de la densidad logarítmica de las coordenadas atómicas en la generación de conformaciones moleculares.

- *Diseño de Drogas y Proteínas:* Recientemente, el *diseño de moléculas de drogas pequeñas en 3D* ha ganado impulso gracias a los DM. Por ejemplo, TargetDiff (Guan et al., 2023) emplea proteínas objetivo como información de orientación y genera moléculas paso a paso, modelando explícitamente la interacción entre las proteínas y las moléculas en el espacio tridimensional. Además, estudios como DiffAb (Luo et al., 2022) aprovechan DM para la generación de proteínas, proponiendo un marco de diseño de anticuerpos basado en difusión.

- *Diseño de Materiales:* El diseño y fabricación de materiales sólidos constituyen la base crítica de numerosas tecnologías en industrias claves. Autocodificador Variacional de Difusión Cristalina (del inglés *Crystal Diffusion Variational AutoEncoder* CDVAE) (Xie et al., 2021), por ejemplo, aborda este campo al incorporar la estabilidad como un sesgo inductivo. Propone una red de puntuación condicional al ruido, la cual utiliza simultáneamente propiedades de permutación, traslación, rotación e invariancia periódica.

- *Reconstrucción de Imágenes Médicas:* En la *reconstrucción de imágenes médicas*, los problemas inversos son fundamentales en técnicas como la tomografía computarizada y la resonancia magnética. Aquí, los SGM juegan un papel crucial al reconstruir imágenes coherentes tanto con la prioridad establecida como con las mediciones observadas. En el caso de la reconstrucción de imágenes de resonancia magnética, estos métodos

guían gradualmente el proceso de difusión inversa, dados los datos de señal observados en el espacio k, y proponen un algoritmo de muestreo de grueso a fino para asegurar un muestreo eficiente.

3.9. Modelos de Difusión de Espacio Latente

Los LDM constituyen una clase de modelos que se entrenan en un espacio de representación de baja dimensión y eficiente, en contraposición al espacio de píxeles de alta dimensión (Rombach et al., 2022). Estos modelos se basan en la idea de compresión perceptual, donde se eliminan los detalles de alta frecuencia pero se conserva la variación semántica relevante. La distinción fundamental de los LDM respecto a otros enfoques radica en su independencia de una compresión espacial excesiva, ya que se entrenan en un espacio latente aprendido que presenta mejores propiedades de escalado en términos de dimensionalidad espacial. Esto posibilita una generación eficiente de imágenes desde el espacio latente con una sola pasada de ANN.

En síntesis, los LDM representan modelos que operan en un espacio de representación latente, lo que facilita una síntesis de alta calidad con una carga computacional reducida.

Las ventajas que ofrecen estos modelos en la síntesis de imágenes son diversas:

1. En comparación con los enfoques puramente basados en T ((Campbell et al., 2022) y (Fefferman et al., 2016)), los LDM *escalan de manera más fácil a datos de mayor dimensionalidad*, lo que les permite trabajar con un nivel de compresión que proporciona reconstrucciones más fieles. Esto facilita su aplicación eficiente en la síntesis de imágenes de alta resolución de megapíxeles.

2. Ofrecen un *rendimiento competitivo en múltiples tareas*, como la síntesis de imágenes incondicionales, el inpainting y la super resolución estocástica, mientras reducen significativamente los costos computacionales. En comparación con los enfoques de difusión basados en píxeles, también disminuyen de manera notable los costos de inferencia.

3. Al no requerir el aprendizaje simultáneo de una arquitectura de codificador/decodificador y un prior basado en puntajes, estos modelos

eliminan la necesidad de una ponderación delicada de las habilidades de reconstrucción y generativas. Esto garantiza *reconstrucciones extremadamente fieles* y requiere poca regularización del espacio latente.

4. Pueden aplicarse de manera convolucional y renderizar imágenes grandes y consistentes de hasta 1024x1024 píxeles para tareas densamente condicionadas, como la super resolución, la restauración de imágenes y la síntesis semántica.

5. Por último, ofrecen un mecanismo de *condicionamiento de propósito general basado en atención cruzada*, lo que permite el entrenamiento multimodal. Esto es fundamental para entrenar modelos de clase condicional, texto a imagen y diseño a imagen.

3.10. Modelos de Texto a Imagen

La *generación de texto a imagen* es una tarea que tiene como objetivo producir una imagen correspondiente a un texto descriptivo (conocido por su nombre en inglés *prompt*).

Los modelos responsables de llevar a cabo esta tarea son conocidos como *modelos texto a imagen* y se clasifican dentro de la categoría de *modelos multimodales*. Estos modelos se encargan de realizar tareas que abarcan desde la comprensión de texto hasta la generación de imágenes que representen el texto proporcionado.

Por ejemplo, dado el texto de entrada *"Un gato jugando al tenis"*, esperamos que el modelo pueda generar una imagen que refleje exactamente esa descripción textual representado gráficamente en la Figura 3.15.

Figura 3.15. Ejemplo de una imagen generada con el prompt "Un gato jugando al tenis" en la herramienta Leonardo AI.

Para abordar esta tarea, se han desarrollado varios enfoques basados en DM como Imagen (Chitwan et al., 2022)) DALL-E 2 (Aditya et al., 2022) y SD (Rombach et al., 2022), los cuales fueron analizados con mayor detalle en el Capítulo 2 de este trabajo.

3.10.1. Modelos de Texto a Imagen Impulsados por el Sujeto

La *generación de imágenes impulsada por el sujeto* tiene como objetivo la tarea de generar imágenes altamente personalizadas con respecto a un sujeto específico (Wenhu et al., 2023). Por lo tanto, este tipo de modelos está vinculado directamente con los modelos de generación de imágenes impulsada por el texto.

Los modelos de generación impulsada por el sujeto a menudo FT sobre un DM preentrenado de texto a imagen en un conjunto de demostraciones proporcionadas C_s sobre un sujeto específico s. Formalmente, dicha demostración contiene un conjunto de pares de texto e imagen centrados en el sujeto s y expresados en la Ecuación 3.8 :

$$C_s = \left(\mathbf{x}_k, \quad \mathbf{c}_k\right) K_{ks} \tag{3.8}$$

Las imágenes x_k contienen imágenes del mismo sujeto s, mientras que c_s es una breve descripción de las imágenes x_k.

Para abordar esta tarea, se desarrollaron varias propuestas basadas en DM y aplicando FT como DB (Ruiz et al., 2023) y *SuTI* (Wenhu et al., 2023), los cuales fueron analizados con mayor detalle en el Capítulo 2 de este trabajo.

3.11. Resumen

En este capítulo se proporcionó una comprensión sólida de los conceptos que son esenciales para el diseño y la implementación de modelos generativos en AI, especialmente en el contexto de los modelos de texto a imagen.

En primer lugar se dieron definiciones tales como Aprendizaje Automático (Machine Learning), Aprendizaje Profundo (Deep Learning), Aprendizaje por Transferencia (Transfer Learning), Ajuste Fino (Fine Tuning), Aprendizaje de Representación y Espacio Latente, Redes Neuronales Convolucionales (CNN), Espacio Muestral, Función de Densidad de Probabilidad y Verosimilitud, Incrustación (Embedding) y Tokenización.

Seguidamente se abordó el tema de los modelos generativos en el contexto de la AI, destacando su importancia, clasificación y aplicaciones en diversos campos. Los modelos generativos son algoritmos avanzados que van más allá de la simple clasificación o predicción al aprender la distribución de probabilidad entre datos observados y latentes.

Se contrastaron los *modelos generativos* contra los *modelos discriminativos*, ya que pueden generar instancias de datos que siguen la distribución de los datos de entrenamiento, lo que los hace útiles para aplicaciones como CV, NLP, medicina y música.

También se describió el proceso para desarrollar un modelo generativo básico que imite la distribución de probabilidad de los datos de entrenamiento y se definieron las propiedades deseables de un modelo generativo, como *precisión*, *capacidad* de generación y *representación* de características de alto nivel.

Además, se abordó el fenómeno de la *alucinación en modelos generativos*. Donde una alucinación se produce cuando un modelo de lenguaje o una herramienta de visión por computadora generan resultados absurdos o incorrectos, percibiendo patrones u objetos que no existen para los humanos. Por un lado, se destacó la preocupación debido a sus implicaciones prácticas y se mencionaron algunos estudios que proponen soluciones para esta problemática. Finalmente, una vez definidos los conceptos de AI y dado el marco teórico sobre modelos generativos se explicó en detalle las *clases de modelos generativos*.

Se presentó a los VAE, redes neuronales que transforman datos de entrada en valores en un espacio latente continuo, permitiendo la generación de nuevos datos. Se destacan por su capacidad para generar distribuciones de probabilidad alrededor de puntos en el espacio latente, lo que facilita la generación de datos realistas y coherentes.

Cabe destacar, las GAN se componen de dos redes neuronales en competencia: el *generador* y el *discriminador*. Estas redes se entrenan en un proceso

de confrontación, donde el generador produce muestras sintéticas y el discriminador evalúa su autenticidad. Este enfoque ha demostrado ser efectivo para generar datos de alta calidad, aunque enfrenta desafíos como el colapso de modo y la optimización inestable.

Los NF son modelos generativos que transforman una distribución de probabilidad simple en una distribución compleja, facilitando la generación de datos de alta dimensionalidad. Estos flujos utilizan técnicas como el cambio de variables y las capas de acoplamiento para garantizar la invertibilidad de la función de mapeo, permitiendo la generación de nuevos puntos de datos de manera eficiente.

Los ARM son una familia de modelos que simplifican el problema de la generación de datos al tratarlo como un proceso secuencial. Estos modelos condicionan las predicciones en valores anteriores en la secuencia, lo que les permite modelar explícitamente la distribución que genera los datos. Aunque son eficaces para datos secuenciales como texto y audio, el proceso de muestreo puede ser computacionalmente intensivo.

Los EBM adoptan una idea fundamental de la física, expresando la probabilidad de un evento mediante una distribución de Boltzmann. Estos modelos son una variante generativa de los discriminadores y pueden aprender a partir de datos no etiquetados. Utilizan estrategias como la divergencia contrastiva y la dinámica de Langevin para entrenar y muestrear nuevas observaciones, respectivamente.

Dada la importancia para el fin de este trabajo se analizaron con gran detalle los DM. Los DM son una familia de modelos generativos probabilísticos que introducen ruido de manera progresiva en los datos para luego aprender a revertir este proceso y generar muestras. Estos modelos han emergido como una nueva metodología en la generación de modelos generativos, desafiando el predominio de las GAN.

Los DM se dividen en tres formulaciones principales: DDPM, SGM y Score SDE.

Los DDPM hacen uso de dos cadenas de Markov; es decir, una cadena hacia adelante que perturba los datos con ruido y una cadena hacia atrás que convierte el ruido de vuelta en datos. El *proceso de difusión hacia adelante* comienza con una imagen inicial que se corrompe gradualmente con ruido, mientras que el *proceso de difusión inversa* intenta desarrollar una red neuronal capaz de deshacer este proceso de corrupción de ruido.

Los SGM perturban los datos con una secuencia de ruido gaussiano intenso y estiman conjuntamente las funciones de puntuación para todas las distribuciones de datos ruidosas mediante el entrenamiento de una red neuronal profunda condicionada a los niveles de ruido.

Los Score SDE son una generalizacion de los modelos DDPM y SGM para caso de un número infinito de pasos de tiempo o niveles de ruido, donde los procesos de perturbación y desruido son soluciones a ecuaciones diferenciales estocásticas.

Los DM se pueden aplicar de manera *incondicional* o *condicional*. La generación incondicional se utiliza para evaluar el rendimiento máximo del modelo, mientras que la generación condicional se centra en el contenido específico para aplicaciones particulares, permitiendo controlar los resultados de la generación de acuerdo con las intenciones del usuario.

Los DM hacen uso extensivo de *diversos mecanismos de condicionamiento* para dirigir la generación, incluyendo etiquetas, clasificadores, texto, imágenes, mapas semánticos y gráficos. Estos mecanismos proporcionan controlabilidad sobre el proceso de generación y permiten la expresión de características semánticas complejas en las muestras generadas.

Un enfoque interesante dentro de los DM son los LDM, los cuales se entrenan en un espacio de representación de baja dimensión en contraposición al espacio de píxeles de alta dimensión. Estos modelos se basan en la compresión perceptual, conservando la variación semántica relevante mientras eliminan los detalles de alta frecuencia. La principal ventaja de los LDM es su independencia de una compresión espacial excesiva, lo que facilita la generación eficiente de imágenes desde el espacio latente con una sola pasada de red neuronal.

También se mencionó que existen diversas aplicaciones de los DM en campos como visión por computadora, generación de lenguaje natural, generación multimodal, modelado de datos temporales, aprendizaje robusto y aplicaciones interdisciplinarias.

Además, se trataron los *modelos de texto a imagen*, que son modelos multimodales encargados de generar imágenes correspondientes a descripciones de texto proporcionadas como entrada. Estos modelos abordan tareas que van desde la comprensión del texto hasta la generación de imágenes que representen la descripción textual dada. Se mencionan varios enfoques basados en DM, como *Imagen*, *DALL-E 2* y SD.

Por último, se introdujo el concepto de *modelos de texto a imagen impulsados por el sujeto*, que tienen como objetivo generar imágenes altamente personalizadas para un sujeto específico. Estos modelos aplican FT sobre un DM de texto a imagen preentrenado en un conjunto de demostraciones proporcionadas sobre un sujeto particular. Estas demostraciones consisten en pares de texto e imagen centrados en el sujeto, lo que permite generar imágenes personalizadas basadas en descripciones específicas del sujeto. Se mencionan ejemplos de modelos de generación impulsada por el sujeto, como DB y *SuTI*.

Parte III

Desarrollo y Experimentos

Capítulo 4

Experimentación

En este capítulo se detallan las *herramientas y librerías* específicas para la generación de modelos de texto a imagen que serán utilizadas en el desarrollo de un prototipo capaz de convertir descripciones textuales en representaciones visuales detalladas.

Además se explicará el proceso empleado para reentrenar un modelo de generación de imágenes basado en DM, utilizando específicamente como SD base.

Se recurrirá a DB para personalizar el modelo a fin de que pueda generar objetos del conjunto de datos específico seleccionado. Luego, el modelo personalizado será desplegado en la plataforma de *Hugging Face* para realizar pruebas exhaustivas.

Finalmente, se evaluará el desempeño del modelo utilizando diversos métodos y la métrica de *fidelidad*, con el objetivo de medir con precisión su rendimiento y su capacidad para generar imágenes que sean fieles y de alta calidad.

4.1. Herramientas y Librerías

El proceso requerido para lograr los objetivos de este trabajo implica el uso de tecnologías avanzadas en el campo del DL y la AI, adecuadas

para ser ejecutadas en entornos de alto rendimiento como Google Colab[1]. Las herramientas seleccionadas incluyen DB, que permite personalizaciones específicas y generaciones creativas basadas en descripciones textuales de un sujeto, así como el uso de la plataforma *Hugging Face*[2], que facilita el acceso a modelos preentrenados y la implementación de estos en un entorno interactivo y accesible.

Estas herramientas son fundamentales para explorar y aprovechar el potencial de los modelos generativos de texto a imagen, permitiendo la creación de imágenes únicas a partir de simples descripciones textuales.

4.1.1. Google Colab

Google Colab, o también conocido como "Colaboratory", es un entorno de Jupyter Notebook[3] gratuito que se ejecuta completamente en la nube. Está desarrollado por Google y permite a los usuarios escribir y ejecutar código de Python en un navegador web sin necesidad de configuración previa. *Google Colab* está especialmente diseñado para facilitar la colaboración en tiempo real, el aprendizaje de la programación, la investigación científica, y los proyectos de ML y análisis de datos, ofreciendo una plataforma accesible y altamente versátil para una amplia gama de usuarios, desde estudiantes hasta investigadores y desarrolladores[4].

Las principales características de *Google Colab* son las siguientes:

- *Acceso gratuito a recursos computacionales:* Colab ofrece acceso gratuito a hardware potente, incluyendo GPU[5] y TPU[6], que pueden acelerar significativamente los cálculos, especialmente útil para el entrenamiento de modelos de ML[4].
- *Sin configuración:* Colab permite a los usuarios comenzar a trabajar sin pasar por complicadas configuraciones de entorno. Esto elimina las barreras para muchos usuarios que quieren empezar rápidamente con Python y la ciencia de datos[4].

[1]https://colab.google
[2]https://huggingface.co
[3]https://jupyter.org
[4]https://research.google.com/colaboratory/faq.html
[5]Unidades de Procesamiento Gráfico
[6]Unidades de Procesamiento Tensorial

- *Compatibilidad con bibliotecas populares:* Se integra sin problemas con bibliotecas y frameworks populares en el ámbito de la ciencia de datos y ML como *TensorFlow*[7], *PyTorch*[8], *Keras*[9], y *OpenCV*[10], permitiendo importar y utilizar estas bibliotecas directamente en los notebooks[11].

- *Colaboración en tiempo real:* Similar a *Google Docs*, *Google Colab* permite a los usuarios colaborar en un mismo documento en tiempo real. Los usuarios pueden compartir sus notebooks con otros, quienes pueden ver y editar el código de forma simultánea[4].

- *Integración con Google Drive:* está integrado con *Google Drive*, lo que facilita guardar y compartir notebooks y acceder a archivos desde Drive. Esto también permite una fácil colaboración y gestión de archivos[4].

- *Ambiente basado en Jupyter Notebook:* Ofrece todas las funcionalidades de *Jupyter*, incluyendo *Markdown*, carga de imágenes, visualización de datos, y mucho más, lo que lo hace adecuado para la enseñanza y la presentación de proyectos de análisis de datos[3].

- *Uso común:* se utiliza ampliamente en educación para enseñar programación y ciencia de datos, en investigación para prototipos rápidos y experimentación con nuevos modelos de ML, y en la industria para análisis de datos y desarrollo de modelos predictivos. Su capacidad para manejar computación intensiva sin necesidad de una infraestructura local lo hace especialmente valioso para estudiantes y profesionales que no tienen acceso a recursos computacionales de alto rendimiento[4].

4.1.2. DreamBooth

DB es un modelo generativo de AI que permite a los usuarios personalizar modelos de generación de imágenes con unos pocos ejemplos de imágenes específicas. Desarrollado inicialmente por investigadores de *Google* (Ruiz et al., 2023), DB utiliza técnicas de DL para adaptar modelos preentrenados, como SD, permitiendo la generación de imágenes que mantienen características específicas identificadas en las imágenes de ejemplo proporcionadas por el usuario.

[7]https://www.tensorflow.org/?hl=es-419
[8]https://pytorch.org
[9]https://keras.io
[10]https://opencv.org
[11]https://colab.research.google.com/notebooks/intro.ipynb

Para ver en más detalle, volver a la Sección 2.2.1 en donde se analizó DB como parte del estado del arte de la personalización de modelos de texto a imagen.

4.1.3. Hugging Face

Hugging Face es una plataforma de código abierto que se centra en el desarrollo de herramientas y modelos de AI [12].

Sus principales características incluyen:

- *Modelos Preentrenados:* Hugging Face ofrece una amplia variedad de modelos preentrenados en NLP, que van desde modelos para tareas específicas hasta modelos de lenguaje generales como BERT, GPT, y más.
- *Biblioteca:* Esta biblioteca proporciona una interfaz unificada y fácil de usar para trabajar con modelos de múltiples tipos, lo que facilita la carga, el uso y el FT de modelos preentrenados.
- *Hub de Modelos:* Hugging Face cuenta con un hub de modelos donde los usuarios pueden compartir, descubrir y descargar modelos preentrenados y sus pesos para una amplia gama de tareas.
- *Pipeline de Procesamiento de Texto:* La plataforma ofrece pipelines preconfigurados para realizar diversas tareas de procesamiento de texto, como clasificación de texto, generación de texto, extracción de información, traducción y más, de manera sencilla y rápida.
- *Comunidad Activa:* Hugging Face cuenta con una comunidad activa de desarrolladores, investigadores y entusiastas que contribuyen con código, modelos y recursos para mejorar la plataforma.

4.2. Conjunto de Datos

El *conjunto de datos* seleccionado consiste en imágenes de *Lilo*, una gata de color blanco y manchas grises, destinadas a incorporar un nuevo estilo

[12] https://huggingface.co/about

felino al modelo. Las imágenes fueron tomadas con la cámara de un teléfono celular y posteriormente redimensionadas a un tamaño de 1200px x 1200px para homogeneizar la entrada durante el entrenamiento.

Figura 4.1. Conjunto de datos de imágenes utilizadas para entrenar el modelo Lilo. Gráfico de elaboración propia.

Este enfoque de reentrenamiento con DB ha demostrado ser eficaz con un número mínimo de imágenes, típicamente entre 3 y 5, para lograr resultados significativos (Ruiz et al., 2023).

4.3. Modelo Stable Diffusion

Se seleccionó el modelo de SD en su versión 2.1 como base para la generación del modelo texto a imagen. Dicho modelo se encuentra disponible en el siguiente repositorio: `https://huggingface.co/stabilityai/stable-diffusion-2-1`.

4.4. Ajuste Fino con Dreambooth

Una vez definido el conjunto de datos y el modelo base, se procedió a aplicar el FT con DB. El código fue adaptado del repositorio FastDB, disponible en `https://github.com/TheLastBen/fast-stable-diffusion`.

El proceso se llevó a cabo ejecutando el código en *Google Colab*, utilizando un entorno con GPU T4.

Inicialmente, se habilitó la carga desde Google Drive para guardar tanto el modelo generado como el conjunto de datos empleado. Posteriormente, se instalaron las librerías necesarias y se descargó el modelo SD 2.1 desde el repositorio de *Hugging Face*.

Las imágenes definidas previamente fueron cargadas en el entorno para ser utilizadas en el entrenamiento.

Se definieron los parámetros de entrenamiento como se muestra en la Tabla 4.1.

Parámetro	**Valor**
UNet Training Steps	1500
UNet Learning Rate	2e-6
Text Encoder Training Steps	350
Text Encoder Learning Rate	1e-6
Offset Noise	Deshabilitado
External Captions	Deshabilitado
Resolution	512
Save Checkpoint Every n Steps	Deshabilitado

Cuadro 4.1: *Valores de los parámetros utilizados en el entrenamiento con DB. Tabla de elaboración propia.*

Como último paso, tras completar el entrenamiento, se generó el modelo *Lilo*, cuyo archivo binario se guardó como *lilo_01.ckpt*.

4.5. Despliegue del Modelo en Hugging Face

Con el modelo ya configurado, se procedió a ejecutar el código para desplegar el modelo en un repositorio de Hugging Face. Esta plataforma propor-

ciona una interfaz interactiva (Figura 4.2) que permite introducir *prompts*[13] para probar el modelo, en este caso, generando imágenes.

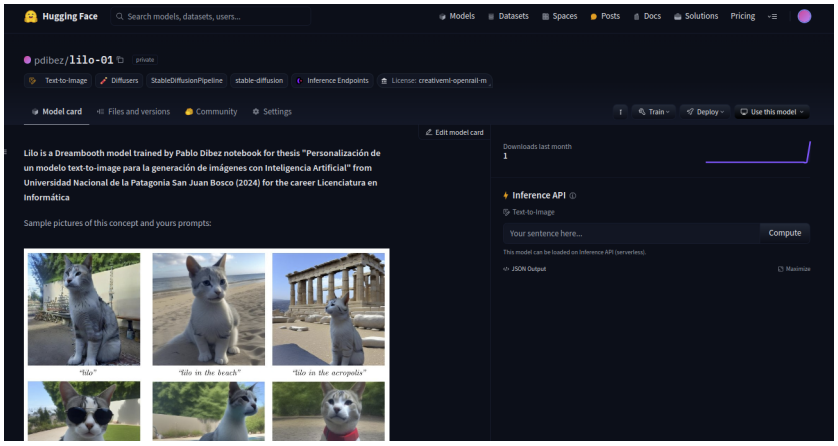

Figura 4.2. Modelo Lilo desplegado en Hugging Face. Gráfico de elaboración propia.

Una vez cargado el modelo en Hugging Face, se realizaron diversas pruebas para generar imágenes con el modelo *Lilo*, percibiendo la gran similitud en la generación de imágenes a las caracteristicas del sujeto original. Se pueden visualizar en la Figura 4.3 algunos ejemplos de las imágenes generadas.

Figura 4.3. Ejemplos de imágenes generadas con el modelo Lilo y sus prompts. Gráfico de elaboración propia.

[13]Se denomina *prompt* al texto o información que se le proporciona a un modelo de AI para generar su salida

4.6. Evaluación

En esta sección, se explicará cómo se procedió a evaluar la efectividad del modelo de generación de imágenes Lilo utilizando la métrica *fidelidad del sujeto*.

Para llevar a cabo esta evaluación, se diseñó un conjunto de pruebas que utilizan diferentes tipos de indicaciones (*prompts*) para guiar la generación de imágenes del modelo. Estos *prompts* se han categorizado en tres tipos: 5 imágenes *básicas* del sujeto sin modificaciones; 5 de *recontextualización*, donde el entorno o escenario del sujeto cambia; y 5 de *accessorización*, que incorporan accesorios o elementos adicionales al sujeto. El sujeto central de nuestras pruebas es *Lilo*, una figura representativa que ha sido continuamente empleada en experimentaciones previas.

Se definieron los *prompts* y posteriormente se generaron las imágenes para evaluación (Figura 4.4) con los métodos seleccionados.

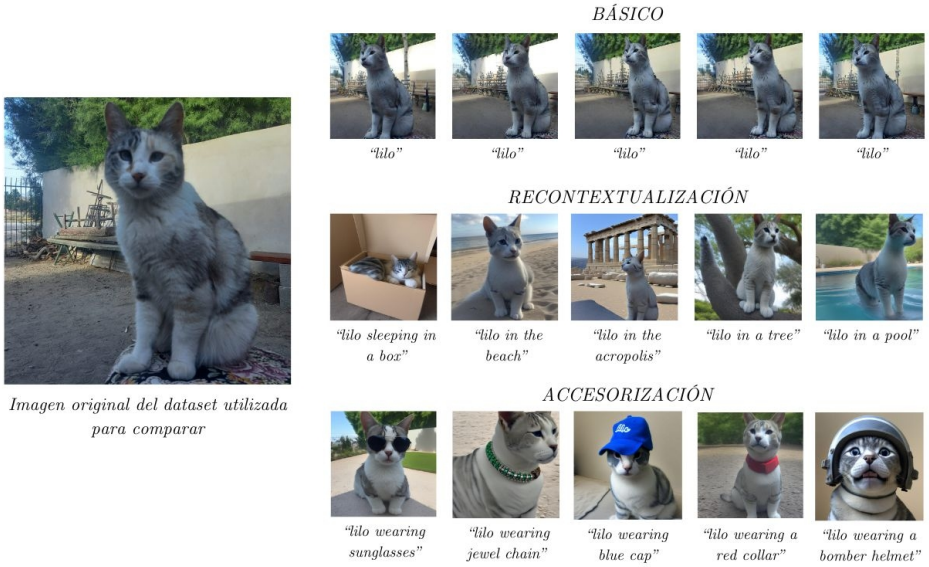

Figura 4.4. Prompts e imágenes generadas para la evaluación. Gráfico de elaboración propia.

A partir de la aplicación de estos métodos, se pretende establecer un marco robusto para evaluar y mejorar continuamente la capacidad del modelo para generar imágenes que no solo sean visualmente atractivas, sino también fieles al sujeto original en diversos contextos y modificaciones.

- *Análisis de la métrica: Fidelidad del Sujeto*

  Tal como se explica en el trabajo de DB (Ruiz et al., 2023), la *fidelidad del sujeto* indica cuán probable es que se preserven los detalles del sujeto en las imágenes generadas con respecto al conjunto de datos original.

4.7. Resultados de la Evaluación

Para cuantificar la *fidelidad del sujeto* en las imágenes generadas, se aplicaron dos métodos de evaluación distintos: *CLIP-I* y *DINO*. Por lo tanto se seleccionó la imagen más representativa del conjunto de datos con el que fue entrenado el modelo *Lilo* y se la comparó con cada una de las imágenes generadas por cada categoría aplicando el cálculo de *CLIP-I* y *DINO*.

4.7.1. CLIP I

Los resultados obtenidos al calcular la *fidelidad del sujeto* con CLIP-I se visualizan en la Tabla 4.2.

De acuerdo a los resultados se puede resaltar que con una probabilidad promedio de *0.901526755094528*, se concluye que la métrica *CLIP-I* demuestra que el modelo *Lilo* genera representaciones altamente fieles a las imágenes originales. Además, se observa que la generación de imágenes básicas presenta resultados superiores en comparación con aquellas pruebas donde se incorporan accesorios o se recontextualiza al sujeto.

4.7.2. DINO

Los resultados obtenidos al calcular la *fidelidad del sujeto* con *DINO* se visualizan en la Tabla 4.3

De acuerdo a los resultados se puede resaltar que con una probabilidad promedio de *0.8368679692348*, se concluye que la métrica *DINO* también confirma que el modelo *Lilo* genera representaciones fieles a las imágenes originales. Sin embargo, se observa una diferencia significativa en la fidelidad de las imágenes básicas en comparación con aquellas a las que se les incorporan accesorios o se recontextualiza el sujeto.

Categoría	Prompt	Resultado CLIP-I
básico	lilo	0.964837908744812
básico	lilo	0.968794882297516
básico	lilo	0.941496670246124
básico	lilo	0.962206780910492
básico	lilo	0.944729715585709
recontextualización	lilo sleeping in a box	0.8622825443744659
recontextualización	lilo in the beach	0.91531702876091
recontextualización	lilo in the Acropolis	0.8383466303348541
recontextualización	lilo in a tree	0.8984368145465851
recontextualización	lilo in a pool	0.9161779284477234
accesorización	lilo wearing sunglasses	0.9086306095123291
accesorización	lilo wearing jewel chain	0.8487137258052826
accesorización	lilo wearing blue cap	0.8387371897697449
accesorización	lilo wearing a red collar	0.8988291621208191
accesorización	lilo wearing a bomber helmet	0.815363734960556

Cuadro 4.2: *Resultados de evaluación con la métrica CLIP-I. Tabla de elaboración propia.*

Categoría	Prompt	Resultado DINO
básico	lilo	0.947639167308807
básico	lilo	0.95164030790329
básico	lilo	0.936329036951065
básico	lilo	0.939815372228622
básico	lilo	0.931950896978378
recontextualización	lilo sleeping in a box	0.692872837185860
recontextualización	lilo in the beach	0.810725927352905
recontextualización	lilo in the Acropolis	0.637392327189446
recontextualización	lilo in a tree	0.794460475444794
recontextualización	lilo in a pool	0.818364739418030
accesorización	lilo wearing sunglasses	0.881823062896729
accesorización	lilo wearing jewel chain	0.798172414302826
accesorización	lilo wearing blue cap	0.827085822820663
accesorización	lilo wearing a red collar	0.872227400541305
accesorización	lilo wearing a bomber helmet	0.712519749999046

Cuadro 4.3: *Resultados de evaluación con la métrica DINO. Tabla de elaboración propia.*

4.8. Resumen

En este capítulo se exploró el proceso de reentrenamiento de un modelo de generación de imágenes utilizando la técnica FT sobre el modelo SD, con personalización a través de DB para adaptarse a un conjunto de datos específico, y su posterior despliegue en la plataforma *Hugging Face*.

Se abordaron las herramientas claves en el ámbito de la AI y la personalización de modelos de texto a imagen para este trabajo: *Google Colab*, DB y *Hugging Face*.

El conjunto de datos utilizado se constituyó por imágenes de *Lilo*, una gata de color blanco y manchas grises; las cuales fueron capturadas con un teléfono móvil y redimensionadas a 1200px x 1200px. Este enfoque ha demostrado ser efectivo para el reentrenamiento con un mínimo de imágenes.

La personalización del modelo se realizó en *Google Colab* utilizando un entorno GPU T4 y el modelo SD 2.1 de *Hugging Face*. Las imágenes de entrenamiento se cargaron en el entorno y, tras el FT, se generó el archivo binario del modelo *Lilo*.

Posteriormente, el modelo se desplegó en *Hugging Face*, proporcionando una interfaz para introducir prompts y generar imágenes, donde varios ejemplos se ilustraron en las figuras correspondientes.

Para *evaluar la fidelidad del sujeto*, se diseñaron pruebas utilizando diferentes tipos de prompts: imágenes básicas, recontextualización y accessorización, con el objetivo de verificar la capacidad del modelo para mantener la fidelidad del sujeto en diversas situaciones. Las pruebas emplearon las métricas *CLIP-I* y *DINO*, mostrando altos niveles de fidelidad en imágenes básicas, aunque se notó una reducción al incorporar accesorios o cambiar el contexto.

Parte IV

Conclusiones y Trabajo Futuro

Capítulo 5

Conclusiones y Trabajo Futuro

5.1. Conclusiones

Este trabajo ha explorado a fondo el estado del arte de los modelos generativos de texto a imagen como *Imagen*, *DALL-E 2* y SD, destacando sus capacidades y limitaciones. Además, se investigaron técnicas especializadas para la personalización de modelos orientada a sujetos específicos, tales como DB y *SuTi*, que permiten una adaptación más precisa y focalizada del aprendizaje.

El marco teórico abordó la clasificación de los modelos generativos, los DM y los modelos de texto a imagen, explicando sus mecanismos fundamentales y sus aplicaciones más impactantes en campos tan diversos como el arte, la medicina y los medios de comunicación. Este fundamento teórico estableció una base sólida para comprender la profundidad y el potencial transformador de estas tecnologías.

La parte experimental del estudio demostró la aplicación práctica de estas teorías mediante el reentrenamiento del modelo SD con la técnica de FT utilizando DB, personalizado específicamente para generar imágenes de *Lilo*, una gata. Este proceso no solo confirmó la eficacia del FT con un número reducido de imágenes, sino que también proporcionó una plataforma robusta para evaluar la calidad y fidelidad de las imágenes generadas.

Al desplegar el modelo en *Hugging Face*, se pudieron realizar pruebas interactivas que confirmaron la capacidad del modelo para mantener una

alta fidelidad en representaciones básicas, aunque con una ligera disminución en escenarios de recontextualización o accessorización. Esto fue cuantificado a través de las métricas *CLIP-I* y *DINO*, cuyos resultados reforzaron la validez de las técnicas empleadas para la personalización y evaluación del modelo.

Personalmente, este proyecto ha sido una inmersión profunda en el campo de la AI, brindándome un aprendizaje significativo sobre las capacidades de los modelos generativos. La experiencia ha reforzado mi comprensión de que estos modelos no solo poseen un gran potencial para la creación artística y la innovación tecnológica, sino que también plantean desafíos éticos y técnicos que deben abordarse con cuidado y precisión.

En conclusión, este estudio ha demostrado que con las técnicas adecuadas, los modelos de generación de imágenes pueden ser personalizados efectivamente para tareas específicas, ofreciendo resultados de alta fidelidad y abriendo puertas a futuras investigaciones y aplicaciones en la generación de contenido personalizado y relevante.

5.2. Trabajo Futuro

A lo largo de este estudio, se han identificado diversas oportunidades para expandir y profundizar el trabajo realizado. A continuación, se delinean varias áreas clave donde futuras investigaciones podrían aportar significativamente al desarrollo y la optimización de modelos generativos personalizados para la generación de imágenes.

- *Ampliación del Dataset:* Una de las limitaciones del presente estudio es el uso de un conjunto de datos relativamente pequeño y centrado en contextos limitados. Para futuras investigaciones, se propone ampliar el dataset no solo en cantidad sino también en diversidad de contextos y escenarios en los que el sujeto, en este caso Lilo, aparece. Esto permitiría evaluar la robustez del modelo bajo una variedad más amplia de situaciones y mejorar su capacidad de generalización, asegurando que las representaciones generadas mantengan la fidelidad incluso en condiciones variadas.
- *Evaluación Humana de las Imágenes Generadas:* Complementar las métricas automáticas como CLIP-I y DINO con evaluaciones humanas proporcionaría una dimensión adicional de validación para la calidad

de las imágenes. Al incorporar juicios humanos, se puede obtener una medida más directa de la percepción de la fidelidad y la creatividad de las imágenes, lo que podría ayudar a refinar aún más los modelos para que sus resultados sean tanto técnicamente sólidos como visualmente atractivos para los usuarios finales.

- *Experimentación con Diferentes Arquitecturas de Modelos Generativos:* Mientras que este trabajo se centró en Stable Diffusion, futuras investigaciones podrían explorar y comparar otras arquitecturas, como DALL-E 2 o modelos basados en técnicas emergentes. El análisis comparativo podría revelar fortalezas particulares y limitaciones de cada modelo, ofreciendo una perspectiva más rica sobre cómo las diferencias en las arquitecturas afectan los resultados de personalización.

- *Implementación de Mejoras en la Personalización del Sujeto:* Explorar métodos avanzados para la personalización de modelos, tal como mejorar la técnica de DreamBooth con adaptaciones que permitan una menor dependencia de la cantidad de datos o una mayor resistencia a la sobreajuste, podría conducir a mejoras significativas en la eficiencia y efectividad del entrenamiento de modelos.

- *Evaluación de la Ética y los Aspectos Legales:* Con la creciente capacidad de los modelos generativos para producir contenidos realistas, es crucial abordar las implicaciones éticas y legales de su uso. Investigaciones futuras podrían centrarse en desarrollar directrices y protocolos que aseguren el uso responsable de la tecnología generativa, especialmente en contextos sensibles como la representación de individuos o la creación de contenido multimedia.

Estas líneas de trabajo no solo prometen mejorar la calidad y eficacia de los modelos generativos de imágenes, sino que también ofrecen la posibilidad de explorar nuevos horizontes en la intersección de la tecnología, el arte y la ética.

Apéndice A

Anexos

A.1. Siglas y Abreviaturas

ADP Adaptative Denoising Purification, (Purificación de Eliminación de Ruido Adaptativo)

AE Autoencoders, (Autocodificadores)

AI Artificial Intelligence, (Inteligencia Artificial)

ANN Artificial Neural Networks, (Redes Neuronales Artificiales)

ARM Autoregressive Models, (Modelos Autoregresivos)

BERT Bidirectional Encoder Representations from Transformers, (Representación de Codificador Bidireccional de Transformadores)

CDVAE Crystal Diffusion Variational AutoEncoder, (Autocodificador Variacional de Difusión Cristalina)

CGAN Conditional Generative Adversarial Networks, (Redes Generativas Adversariales Condicionales)

CLIP Contrastive Language–Image Pre-training, (Preentrenamiento de Contraste Lenguaje-Imagen)

CNN Convolutional Neural Network, (Red Neuronal Convolucional)

CSDI Conditional Score-based Diffusion Model for Imputation, (Modelo de Difusión Basado en Puntuación Condicional para Imputación)

CT Computed Tomography, (Tomografía Computarizada)

CV Computer Vision, (Visión por Computadora)

DB DreamBooth

DDPM Denoising Diffusion Probabilistic Models, (Modelo Probabilístico de Difusión de Eliminación de Ruido)

DL Deep Learning, (Aprendizaje Profundo)

DM Diffusion Model, (Modelos de Difusión)

EBM Energy-Based Models, (Modelos Basados en Energía)

FID rechet Inception Distance, (Distancia de Frechet Inception)

FT Fine Tuning, (Ajuste Fino)

GAN Generative Adversarial Networks, (Redes Generativas Adversariales)

GPT Generative Pre-trained Transformer, (Transformador Generativo Preentrenado)

GPU Graphics Processing Unit, (Unidad de Procesamiento Gráfico)

LDM Latent Diffusion Models, (Modelos de Difusión Latente)

LLM Large Language Model, (Modelo de Lenguaje de Gran Escala)

LSTM Long Short-Term Memory, (Redes Neuronales de Memoria a Largo Plazo y Corto Plazo)

ML Machine Learning, (Aprendizaje Automático)

MR Magnetic Resonance, (Resonancia Magnética)

NCSN Noise Conditional Score Network, (Red de Puntuación Condicional al Ruido)

NF Normalization Flow, (Flujo Normalizador)

NLP Natural Language Processing, (Procesamiento de Lenguaje Natural)

PGD Projected Gradent Descent, (Descenso de Gradiente Proyectado)

RealNVP Real-valued Non-Volume Preserving, (Transformaciones de Preservación de Volumen No Reales)

RNN Recurrent Neural Network, (Red Neuronal Recurrente)

Score SDE Score Stochastic Differential Equations, (Ecuaciones Diferenciales Estocásticas de Puntuación)

SD Stable Diffusion, (Difusión Estable)

SDE Stochastic Differential Equations, (Ecuaciones Diferenciales Estocásticas)

SGM Score-based Generative Models, (Modelos Generativos Basados en Puntuación)

T Transformer, (Transformador)

TL Transfer Learning, (Aprendizaje por Transferencia)

VAE Variational Autoencoders, (Autocodificadores Variacionales)

A.2. Glosario

- *Aprendizaje de cero ejemplos:* escenario de aprendizaje automático en el que un modelo de AI se entrena para reconocer y categorizar objetos o conceptos sin haber visto ejemplos previos de esas categorías o conceptos.
- *Aprendizaje Supervisado:* escenario de aprendizaje donde se aprende una función que relaciona una entrada con una salida utilizando un conjunto de datos etiquetado
- *Cadena de Markov:* modelo matemático que describe un sistema en el que la probabilidad de transición a un estado futuro depende solo del estado actual y no de cómo se llegó a él.
- *CLIP:* red neuronal que aprende eficientemente conceptos visuales a partir de supervisión en lenguaje natural. CLIP se puede aplicar a cualquier prueba de clasificación visual simplemente proporcionando los nombres de las categorías visuales a reconocer.

- *CLIP-I:* métrica que representa la similitud promedio de cosenos entre las incrustaciones de imágenes generadas por CLIP y las imágenes reales.
- *COCO:* estándar en la evaluación de modelos de texto a imagen.
- *Dataset:* colección de datos utilizados para entrenar y validar un modelo.
- *Datos No Estructurados:* Los datos no estructurados se refieren a cualquier tipo de información que no esté naturalmente organizada en columnas de características, como imágenes, audio y texto. Si bien es cierto que las imágenes tienen una estructura espacial, las grabaciones de audio tienen una estructura temporal y los pasajes de texto pueden tener tanto estructura espacial como temporal, estos datos no están dispuestos en columnas de características, lo que los califica como no estructurados.
- *Determinante de Jacobiano:* medida de cuánto cambian las variables de salida de una transformación respecto a las variables de entrada. En el contexto del cálculo multivariable y la optimización, el determinante Jacobiano se utiliza para calcular la tasa de cambio de una transformación en un punto dado del espacio. Es especialmente útil en problemas de optimización y en la transformación de coordenadas entre sistemas de referencia. (Strang, 2009)
- *Dinámica de Langevin:* enfoque utilizado en física estadística y simulaciones computacionales para describir el comportamiento de partículas en medios con influencias estocásticas, como la fricción y la agitación térmica. Se basa en la ecuación de Langevin, que tiene en cuenta tanto las fuerzas deterministas como las fluctuaciones aleatorias, modeladas usualmente como ruido blanco gaussiano, para predecir la evolución temporal de un sistema.(Frenkel & Smit, 2002)
- *DINO:* métrica propuesta por los autores de Dreambooth (Ruiz et al., 2023) y mide la similitud promedio de coseno entre las incrustaciones ViT-S/16 DINO de imágenes generadas y reales.
- *Distribución de Boltzmann* distribución propuesta originalmente por Ludwig Boltzmann en 1868 para describir gases en equilibrio térmico
- *Distribución Gaussiana:* también conocida como distribución normal, es una de las distribuciones de probabilidad más comunes en estadística.

Se caracteriza por ser una distribución simétrica en forma de campana, donde la mayoría de los datos se concentran cerca de la media y disminuyen gradualmente hacia los extremos.

- *DrawBench:* comprende 11 categorías que ponen a prueba diversas capacidades de los modelos, incluida la representación precisa de colores, números de objetos, relaciones espaciales, texto en la escena e interacciones inusuales entre objetos.

- *Equilibrio de Nash:* nombrado así por el matemático John Nash, es un concepto fundamental en la teoría de juegos; y se define como una situación en la que, dentro de un juego con dos o más jugadores, ninguno puede mejorar su resultado eligiendo una estrategia diferente, siempre y cuando los otros jugadores mantengan sus estrategias constantes (Myerson, 1991)

- *FID:* métrica popular utilizada para evaluar la calidad de las imágenes generadas por redes GAN. Mide la distancia entre las distribuciones gaussianas multivariadas del conjunto de datos de imágenes generadas y los datos reales que el GAN trata de replicar (de Deijn et al., 2024).

- *FID-30K:* FID que utiliza 30000 imágenes.

- *Función de Monte Carlo Langevin:* es un algoritmo utilizado en simulaciones computacionales para generar muestras de una distribución de probabilidad deseada. Este algoritmo combina métodos de Monte Carlo con la dinámica de Langevin(Liang & Liu, 2009).

- *Función de puntuación de Fisher:* medida utilizada en estadística para estimar los parámetros de un modelo de manera iterativa. Esta función se utiliza en el contexto de la estimación de máxima verosimilitud para calcular los gradientes de la función de log-verosimilitud con respecto a los parámetros del modelo(Rice, 2006).

- *Función de puntuación de Stein:* herramienta estadística utilizada en estimación paramétrica, que se utiliza para calcular la puntuación de un estimador de máxima verosimilitud en modelos estadísticos donde los parámetros no pueden estimarse directamente a través de métodos de diferenciación estándar(Geyer, 2010).

- *Generación incondicional:* capacidad de un modelo generativo para producir datos sin restricciones específicas.

- *GPU:* Unidad de Procesamiento Gráfico.

- *Minimax:* es un método de toma de decisiones utilizado en juegos de dos jugadores de suma cero, donde un jugador busca maximizar su ganancia mientras que el otro intenta minimizarla. Este algoritmo evalúa posibles movimientos y selecciona la mejor opción basándose en el supuesto de que el oponente también juega de manera óptima (Norvig & Russell, 2021).

- *Modelos de difusión:* familia de modelos generativos probabilísticos que destruyen progresivamente los datos mediante la inyección de ruido, para luego aprender a revertir este proceso para la generación de muestras.

- *Muestreo Ancestral:* técnica utilizada en estadísticas y genética para generar muestras de una distribución de probabilidad posterior. En el contexto de la inferencia bayesiana, el muestreo ancestral es una forma de aproximar la distribución posterior al muestrear directamente de la distribución conjunta de las variables de interés y los parámetros del modelo(Lai, 2016).

- *Prompt:* texto o información que se le proporciona a un modelo de AI para generar su salida.

- *Redes Neuronales Recurrentes (RNN):* contienen una capa recurrente (o celda) que puede manejar datos secuenciales, haciendo que su propia salida en un momento dado forme parte de la entrada al siguiente paso de tiempo.

- *Segmentación Semántica:* implica etiquetar cada píxel de una imagen según categorías de objetos establecidas.

- *Softmax:* función de activación utilizada comúnmente en redes neuronales para convertir un vector de números reales en un vector deprobabilidades. La salida de la función softmax es una distribución de probabilidad que asigna una probabilidad a cada posible clase o categoría. Esta función es especialmente útil en la capa de salida de redes neuronales utilizadas para clasificación multiclase, ya que garantiza que las probabilidades sumen uno y permite interpretar la salida como una distribución de probabilidad(Goodfellow et al., 2016).

- *Stemming:* significa que las palabras pueden ser llevadas a su raíz, es decir reducirlas a su forma más simple, de modo que diferentes tiempos de un verbo permanecen agrupados en una misma tokenización.

- *Swish:* alternativa a ReLU introducida por Google en 2017. Visualmente, Swish es similar a ReLU, pero es suave, lo que ayuda a mitigar el problema del gradiente desvaneciente.
- *T5:* modelo Transformer de Google que utiliza la estructura codificador-decodificador. Se caracteriza por redefinir una variedad de tareas en un marco de trabajo de texto a texto, que incluye traducción, aceptabilidad lingüística, similitud de oraciones y resumen de documentos.
- *TPU:* Unidad de Procesamiento Tensorial.
- *Transformer:* red neuronal que aprende contexto y, por lo tanto, significado mediante el seguimiento de relaciones en datos secuenciales como las palabras de una oración.
- *unCLIP:* modelo de dos etapas: un prior que genera una incrustación de imagen CLIP dado un texto, y un decodificador que genera una imagen condicionada a la incrustación de imagen. Se lo llama así porque representa una versión "des-CLIPada" (es decir que deshace) del modelo CLIP original.
- *Vector codificado en caliente:* (o en inglés "one-hot encoded vector") se refiere a una representación de datos categóricos donde cada categoría se representa como un vector binario, donde un único bit está activado (establecido en 1) para indicar la pertenencia a una categoría específica (Goodfellow et al., 2016).

A.3. Repositorio

Toda la programación de este proyecto se realizó con el lenguaje Python y está disponible en Hugging Face en el siguiente enlace:

https://huggingface.co/pdibez/lilo-01

En el repositorio se incluye: el conjunto de datos y de evaluación, el código de implementación, el modelo generado, las salidas de ejecución y un entorno interactivo para probar el modelo.

Parte V

Bibliografía

Bibliografía

Ramesh Aditya, Dhariwal Prafulla, Nichol Alex, Chu Casey, & Chen Mark. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Ethem Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2014.

Anónimo. Zero-shot learning, enero 2024. URL `https://www.ibm.com/topics/zero-shot-learning`. Consultado el 18 de marzo de 2024.

Emmanuel Brempong Asiedu, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, & Mohammad Norouzi. Decoder denoising pretraining for semantic segmentation. *arXiv preprint arXiv:2205.11423*, 2022. URL `https://arxiv.org/abs/2205.11423`.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, & Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *arXiv preprint arXiv:2107.03006*, 2021. URL `https://arxiv.org/abs/2107.03006`.

Joseph Babcock & Raghav Bali. *Hands-On Generative AI with Python and TensorFlow 2*. Packt, 2021.

Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, & Vishal M. Patel. Ddpm-cd: Denoising diffusion probabilistic models as feature extractors for change detection. *arXiv preprint arXiv:2206.11892*, 2022. URL `https://arxiv.org/abs/2206.11892`.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Tsachi Blau, Roy Ganz, Bahjat Kawar, Alex Bronstein, & Michael Elad. Threat model-agnostic adversarial defense using diffusion models. 2022. URL `https://arxiv.org/abs/2207.08089`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, & Dario Amodei. Language models are few-shot learners. 2020. URL `https://arxiv.org/abs/2005.14165`.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, & Arnaud Doucet. A continuous time framework for discrete denoising models. 2022. URL `https://arxiv.org/abs/2205.14987`.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, & William Chan. Wavegrad: Estimating gradients for waveform generation. 2020. URL `https://arxiv.org/abs/2009.00713`.

Ting Chen, Ruixiang Zhang, & Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022. URL `https://arxiv.org/abs/2208.04202`.

Saharia Chitwan, Chan William, Saxena Saurabh, Li Lala, Whang Jay, Denton Emily, Seyed Ghasemipour Seyed Kamyar, Karagol Ayan Burcu, Mahdavi S. Sara, Gontijo Lopes Rapha, Salimans Tim, Ho Jonathan, J Fleet David, & Norouzi Mohammad. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, & Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. URL `https://arxiv.org/abs/1406.1078`.

Ricardo de Deijn, Aishwarya Batra, Brandon Koch, Naseef Mansoor, & Hema Makkena. Reviewing fid and sid metrics on generative adversarial networks. 2024.

Laurent Dinh, Jascha Sohl-Dickstein, & Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803v3*, 2016. URL `https://arxiv.org/abs/1605.08803v3`.

Yilun Du & Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019. URL https://arxiv.org/abs/1903.08689.

Vincent Dumoulin & Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2018. URL https://arxiv.org/abs/1603.07285.

Rick Durrett. *Essentials of Stochastic Processes*. Springer, 2016.

Mohamed Elgendy. *Deep Learning for Vision Systems*. Manning, 2020.

Hugging Face. About us, n.d. URL https://huggingface.co/about.

Charles Fefferman, Sanjoy Mitter, & Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4): 983–1049, 2016. URL https://arxiv.org/abs/1310.0425.

David Foster. *Generative Deep Learning*. Editorial O'Reilly, 2023.

Daan Frenkel & Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 2002.

Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, & Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. 2018. URL https://arxiv.org/abs/1801.02385.

Hadjeres Gaetan, Pachet François, & Frank Nielsen. Deepbach: A steerable model for bach chorales generation. *arXiv preprint arXiv:1612.01010*, 2017.

Charles J. Geyer. *Introduction to Markov Chain Monte Carlo*. Institute for Mathematics and Its Applications, 2010.

Ian Goodfellow, Yoshua Bengio, & Aaron Courville. *Deep Learning*. The MIT Press, 2016.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, & Yoshua Bengio. Generative adversarial networks. Preprint available at arXiv, 2014. URL https://arxiv.org/abs/1406.2661.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, & Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. 2023. URL https://arxiv.org/abs/2303.03543.

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, & Frank Wood. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022. URL https://arxiv.org/abs/2205.11495.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. URL https://arxiv.org/abs/1512.03385.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, & Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. URL https://arxiv.org/abs/2210.02303.

Jonathon Ho, Ajay Jain, & Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020. URL https://arxiv.org/abs/2006.11239.

Sepp Hochreiter & Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997. URL https://www.bioinf.jku.at/publications/older/2604.pdf.

Jeremy Howard & Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics, 2018.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, & Tommi Jaakkola. Torsional diffusion for molecular conformer generation. 2022. URL https://arxiv.org/abs/2206.01729.

Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, & Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. 2021a. URL https://arxiv.org/abs/2011.13456.

Alexia Jolicoeur-Martineau, Remi Piche-Taillefer, Rémi Tachet des Combes, & Ioannis Mitliagkas. Adversarial score matching and improved sampling for image generation. 2021b. URL https://arxiv.org/abs/2009.05475.

Tero Karras, Miika Aittala, Timo Aila, & Samuli Laine. Elucidating the design space of diffusion-based generative models. 2022. URL https://arxiv.org/abs/2206.00364.

Bahjat Kawar, Michael Elad, Stefano Ermon, & Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022. URL https://arxiv.org/abs/2201.11793.

Jihoon Kim, Jiseob Kim, & Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. *arXiv preprint arXiv:2209.00349*, 2022. URL https://arxiv.org/abs/2209.00349.

Diederik P. Kingma & Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. URL https://arxiv.org/abs/1312.6114.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, & Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. 2020. URL https://arxiv.org/abs/2009.09761.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report*, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

K. K. Lai. Ancestral sampling for particle gibbs. *IEEE Transactions on Signal Processing*, 64(20):5230–5241, 2016.

Yann LeCun, Yoshua Bengio, & Geoffrey Hinton. Deep learning. *Nature*, 521 (7553):436–444, 2015.

Faming Liang & Chuanhai Liu. The langevin and hamiltonian monte carlo methods in bayesian computation. *Journal of the American Statistical Association*, 103(482):653–664, 2009.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, & Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022. URL https://arxiv.org/abs/2211.10440.

Juan Miguel Lopez Alcaraz & Nils Strodthof. Diffusion-based time series imputation and forecasting with structured state space models. 2022. URL https://arxiv.org/pdf/2208.09399.pdf.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, & Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. 2022. URL https://arxiv.org/abs/2206.00927.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, & Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models, 2022. URL https://arxiv.org/abs/2201.09865.

Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, & Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *bioRxiv*, 2022. doi: 10.1101/2022.07.10.499510. URL https://www.biorxiv.org/content/10.1101/2022.07.10.499510v5.full.

Zhaoyang Lyu, Zhifeng Kong, Xudong Xu, Liang Pan, & Dahua Lin. A conditional point diffusion-refinement paradigm for 3d point cloud completion. *arXiv preprint arXiv:2112.03530*, 2021. URL https://arxiv.org/abs/2112.03530.

Negar Maleki & Balaji Padmanabhan. Ai hallucinations: A misnomer worth clarifying, 2024.

Mirza Mehdi & Osindero Simon. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, & Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2021. URL https://arxiv.org/abs/2108.01073.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

Roger B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1991.

Alex Nichol, Prafulla Dhariwal, Alistair Muldal, & John Schulman. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021. URL https://arxiv.org/abs/2102.09672.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, & Anima Anandkumar. Diffusion models for adversarial purification. 2022. URL https://arxiv.org/abs/2205.07460.

J. R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, UK, 1998.

Peter Norvig & Stuart Russell. *Artificial Intelligence: A Modern Approach*. Pearson, 4th edition, 2021.

Sung Woo Park, Kyungjae Lee, & Junseok Kwon. Neural markov controlled sde: Stochastic optimization for continuous-time data. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/pdf?id=7DI6op61AY.

Daniel Perez-Aguilar, Redy Risco-Ramos, & Luis Casaverde-Pacherrez. Transfer learning en la clasificación binaria de imágenes térmicas. *Ingenius. Revista de Ciencia y Tecnología*, (26):71–86, 2021. DOI: https://doi.org/10.17163/ings.n26.2021.07.

Ben Poole, Ajay Jain, Jonathan T. Barron, & Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. URL https://arxiv.org/abs/2209.14988.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, & Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. URL https://arxiv.org/abs/2303.09535.

Alec Radford, Luke Metz, & Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2016. URL https://arxiv.org/abs/1511.06434.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, & Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, & Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. URL https://arxiv.org/abs/1910.10683.

Kashif Rasul, Calvin Seward, Ingmar Schuster, & Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. 2021. URL https://arxiv.org/abs/2101.12072.

John A. Rice. *Mathematical Statistics and Data Analysis*. Thomson Brooks/Cole, 2006.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, & Bjorn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752v2*, 2022.

Sohini Roychowdhury. Journey of hallucination-minimized generative ai solutions for financial decision makers, 2023.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, & Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2023.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, & Mohammad Norouzi. Palette: Image-to-image diffusion models, 2021a. URL https://arxiv.org/abs/2111.05826.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, & Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021b. URL https://arxiv.org/abs/2104.07636.

Chence Shi, Shitong Luo, Minkai Xu, & Jian Tang. Learning gradient fields for molecular conformation generation. 2021. URL https://arxiv.org/abs/2105.03902.

Jascha Sohl-Dickstein, Eric A. Weiss, & Naveen Goyal. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015. URL https://arxiv.org/abs/1503.03585.

Yang Song & Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019. URL https://arxiv.org/abs/1907.05600.

Yang Song & Stefano Ermon. Improved techniques for training score-based generative models. *arXiv preprint arXiv:2006.09011*, 2020. URL https://arxiv.org/abs/2006.09011.

Yang Song, Conor Durkan, Iain Murray, & Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, pages 1415–1428, 2021. URL https://arxiv.org/abs/2101.09258.

StabilityAI. Stable diffusion 2.1. https://huggingface.co/stabilityai/stable-diffusion-2-1, n.d. Fecha de acceso: 19 de Abril de 2024.

Gilbert Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 2009.

Jaesung Tae, Hyeongju Kim, & Taesu Kim. Editts: Score-based editing for controllable text-to-speech. *arXiv preprint arXiv:2110.02584*, 2021. URL https://arxiv.org/abs/2110.02584.

Yusuke Tashiro, Jiaming Song, Yang Song, & Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *arXiv preprint arXiv:2107.03502*, 2021. URL https://arxiv.org/abs/2107.03502.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, & Amit H. Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. URL https://arxiv.org/abs/2209.14916.

TheLastBen. Fast stable diffusion. https://github.com/TheLastBen/fast-stable-diffusion, n.d. Fecha de acceso: 19 de Abril de 2024.

Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, & Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016. URL https://arxiv.org/abs/1601.06759.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, & Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. URL https://arxiv.org/abs/1706.03762.

Chen Wenhu, Hu Hexiang, Li Yandong, Ruiz Nataniel, Jia Xuhui, Chang Ming-Wei, & W. Cohen William. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023.

Julian Wyatt, Adam Leach, Sebastian M. Schmon, & Chris G. Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 650–656. IEEE/CVF, June 2022. doi: 10.1109/CVPRW52783.2022.00082. URL https://openaccess.thecvf.com/content/CVPR2022W/NTIRE/html/Wyatt_AnoDDPM_Anomaly_Detection_With_Denoising_Diffusion_Probabilistic_Models_Using_Simplex_CVPRW_2022_paper.html.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, & Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021. URL https://arxiv.org/abs/2110.06197.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, & Jian Tang. Geodiff: a geometric diffusion model for molecular conformation generation. 2022. URL https://arxiv.org/abs/2203.02923.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, & Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983*, 2022a. URL https://arxiv.org/abs/2207.09983.

Ling Yang, Zhilin Huang, Yang Song, Shenda Hong, Guohao Li, Wentao Zhang, Bin Cui, Bernard Ghanem, & Ming-Hsuan Yang. Diffusion-based scene graph to image generation with masked contrastive pre-training. *arXiv preprint arXiv:2211.11138*, 2022b. URL https://arxiv.org/abs/2211.11138.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, & Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796v10*, 2023.

Ruihan Yang, Prakhar Srivastava, & Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022c. URL https://arxiv.org/abs/2203.09481.

Jongmin Yoon, Sung Ju Hwang, & Juho Lee. Adversarial purification with score-based generative models. 2021. URL https://arxiv.org/abs/2106.06041.

Qinsheng Zhang & Yongxin Chen. Fast sampling of diffusion models with exponential integrator. 2022. URL https://arxiv.org/abs/2204.13902.

Linqi Zhou, Yilun Du, & Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. *arXiv preprint arXiv:2104.03670*, 2021. URL https://arxiv.org/abs/2104.03670.