
AQUI FALTA EL TITULO

TITLE MISSING



Trabajo de Fin de Grado
Curso 2023–2024

Autor

Sergio Llorente Hernando
Isabella Romano Ramos

Director

Gonzalo Mendez Pozo
Pablo Gérvias Gómez-Navarro

Grado en **Ingeniería Informática**

Facultad de Informática
Universidad Complutense de Madrid

AQUI FALTA EL TITULO

TITLE MISSING

Trabajo de Fin de Grado en **Ingeniería Informática**

Autor

Sergio Llorente Hernando
Isabella Romano Ramos

Director

Gonzalo Mendez Pozo
Pablo Gérvias Gómez-Navarro

Convocatoria: *Junio 2024*

Grado en **Ingeniería Informática**

Facultad de Informática
Universidad Complutense de Madrid

14 de mayo de 2024

Dedicatoria

*A Gonzalo por guiarnos en todo momento y a
nuestras familias por hacer esto posible.*

Agradecimientos

A todas las personas que han apoyado este proyecto y durante nuestra etapa universitaria.

Resumen

El mundo tal y como lo conocemos está avanzando cada vez más y en todos los aspectos que conocemos, entre estos avances podemos destacar la esperanza de vida y el uso de las nuevas tecnologías. Aunque en principio no veamos mucha relación entre ambos conceptos, están mucho más unidos de lo que pensamos. El aumento de la esperanza de vida se debe a las mejoras de los procesos y condiciones médicas y sanitarias que han sido de revolución en los últimos 100 años, gracias a ello podemos vivir mucho más tiempo y disfrutar de la vida unos años más. Sin embargo, hay consecuencias que sufren las personas que alcanzan los años más avanzados de edad, entre tantas, las más importantes son el deterioro de algunas capacidades sensoriales, como lo pueden ser la vista o el oído; y capacidades cognitivas como la orientación o la memoria. En esta última es en la que se basa la principal enfermedad que afecta a las personas de mayor edad, el Alzheimer, que ataca directamente a la memoria destruyendo lentamente su capacidad hasta llegar a las funciones más básicas. Aquí es cuando entran las últimas innovaciones tecnológicas realizadas, en dónde introducimos el concepto de Inteligencia Artificial para la realización de este trabajo. En concreto, la generación de imágenes a través de esta. Nuestro objetivo es crear recuerdos representados por imágenes generadas automáticamente y al momento, a través de palabras del propio afectado por el Alzheimer, para poder construir un libro de vida, que le ayudará a través de terapias de reminiscencia que favorecen que la enfermedad avance paulatinamente.

Palabras clave

Inteligencia artificial, generación de imágenes, prompt, Alzheimer, terapias de reminiscencia, redes neuronales.

Abstract

The world as we know it is advancing more and more and in all the aspects that we know, among these advances we can highlight life expectancy and the use of new technologies. Although at first sight we do not see much relationship between both concepts, they are much closer than we think. The increase in life expectancy is due to improvements in medical and health processes and conditions that have been revolutionary in the last 100 years. Thanks to this, we can live much longer and enjoy life for a few more years. However, there are consequences suffered by people who reach the most advanced years of age, among many, the most important are the deterioration of some sensory abilities, such as sight or hearing; and cognitive abilities such as orientation or memory. The latter is the basis of the main disease that affects older people, Alzheimer's, which directly attacks memory, slowly destroying its capacity until it reaches the most basic functions. This is when the latest technological innovations come in, where we introduce the concept of Artificial Intelligence. Specifically, the generation of images through it. Our objective is to create memories represented by images generated automatically and at the moment, through the words of the person affected by Alzheimer's, in order to build a life book, which will help them through reminiscence therapies that encourage the disease to progress gradually.

Keywords

Artificial intelligence, image generation, prompt, Alzheimer's, reminiscence therapies.

Índice

1. Introducción	1
1.1. Motivos	2
1.2. Objetivos	2
1.3. Plan de trabajo	2
2. Estado de la Cuestión	5
2.1. ¿Qué es el Alzheimer?	5
2.2. La terapia de reminiscencia	6
2.3. ¿Qué es la Inteligencia Artificial?	7
2.4. Redes neuronales	9
2.4.1. Redes Neuronales Feedforward (FNN)	10
2.4.2. Redes Neuronales Recurrentes (RNN)	11
2.4.3. Redes Neuronales LSTM	12
2.4.4. Redes Neuronales Convolutivas (CNN)	14
2.4.5. Redes Generativas Adversarias (GAN)	16
2.4.6. Redes Neuronales Transformer	17
2.5. Stable Diffusion	19
2.5.1. Funcionamiento interno de Stable Diffusion	19
2.5.2. Parámetros adicionales	22
3. Descripción del Trabajo	25
3.1. Elección del modelo	25

3.2. Entrenamiento con Stable Diffusion	26
3.3. Generación de imágenes con la interfaz de Stable Diffusion	33
3.3.1. Requisitos de instalación	33
3.3.2. Funcionamiento y detalles de la interfaz	33
3.3.3. Resultados de generación de imágenes	33
4. Conclusiones y Trabajo Futuro	39
Introduction	41
Conclusions and Future Work	43
Contribuciones Personales	45
Bibliografia	49

Índice de figuras

2.1.	Obstrucción de neuronas por la vitamina tau	6
2.2.	Estructura del perceptrón	9
2.3.	Red neuronal feedforward	11
2.4.	Red neuronal recurrente	12
2.5.	Red neuronal Long Short Term Memory	14
2.6.	Red neuronal convolucional	15
2.7.	Red neuronal generativa antagónica	17
2.8.	Red neuronal Transformer	18
2.9.	difusión directa de una imagen de un cachorro de Shiba Inu comiendo	20
2.10.	difusión inversa de una imagen de un cachorro de Shiba Inu comiendo	20
2.11.	Representación gráfica del proceso de conversión al espacio latente . .	21
2.12.	Arquitectura de una red U-Net	22
2.13.	Representación gráfica del funcionamiento de Stable Diffusion	23
3.1.	Dataset seleccionado para el entrenamiento de personas con Lora . .	29
3.2.	Procedimiento del entrenamiento mediante Dreambooth	30
3.3.	Dataset seleccionado para el entrenamiento con lugares	31
3.4.	Dataset seleccionado para el entrenamiento con animales	32
3.5.	Mismo prompt con 1000, 2000 y 10.000 pasos de entrenamiento . . .	34
3.6.	Imagen generada con el modelo Stable Diffusion 1.5, Lora Hachiko . .	34
3.7.	Resultados de entrenamiento de una persona con Dreambooth	35
3.8.	Resultados de entrenamiento de una persona con diferentes estilos . .	35

Índice de tablas

3.1. Tabla de resultados obtenidos de entrenamiento	30
---------------------------------------------------------------	----

Capítulo **1**

Introducción

“Todo se hunde en la niebla del olvido pero cuando la niebla se despeja, el olvido está lleno de memoria”

— Mario Benedetti

Nuestra vida se compone desde los primeros instantes en los tomamos nuestro primer aliento hasta que respiramos el último. Sin embargo, a la hora de relatar nuestra vida, esta historia estaría compuesta solamente de recuerdos. Desde nuestro primer recuerdo, que normalmente está comprendido entre los 2 y 4 años, debido a que la formación de nuevas neuronas impide que la corteza almacene recuerdos hasta esa edad, hasta el último, el cual puede darse el caso de ser distorsionado o borrado completamente debido al deterioro de nuestro cerebro y la imposibilidad de nuestras neuronas a realizar las conexiones posibles para ello. Este último caso es una enfermedad degenerativa comúnmente conocida como Alzheimer, que actualmente afecta a más de 55 millones de personas en todo el mundo. Sin embargo, no son sólo estas personas quiénes la sufren, sino todos sus allegados también.

Para ayudar a combatir esta enfermedad existen métodos farmacológicos y no farmacológicos, en este proyecto nos centraremos en el segundo grupo, más concretamente, en la terapia de reminiscencia. En cuanto a los métodos que existen actualmente para realizar esta terapia se ha comprobado que la ayuda de apoyo visual es mucho más efectiva a entrevistas con muchas preguntas seguidas, ya que esta última puede provocar que el paciente se vea agobiado y abrumado. Es por este motivo, entre muchos otros, que nuestro proyecto está orientado a ser una herramienta de apoyo visual a la narración de libros de vida. Para ello haremos uso de Inteligencias artificiales que convierten entradas de texto a imágenes y de esta manera, podremos transformar recuerdos narrados por el paciente y convertirlos en imágenes. Lo interesante es poder dar la posibilidad de hacer estas imágenes más personales de forma que se puedan añadir una serie de imágenes propias, de familiares o de eventos importantes de la vida del paciente, para entrenar al modelo y que el resultado final sea único, especial y de gran ayuda para la persona que sufre de esta enfermedad.

1.1. Motivos

1.2. Objetivos

- Utilizar la inteligencia artificial generativa para crear imágenes personales que ayuden a los pacientes a evocar momentos y experiencias emocionales e integrarlos en el presente.
- Generar fotografías que respalden historias de los pacientes.
- Brindar material de apoyo para la terapia de reminiscencia.
- Elaborar un programa mediante el cual, un ayudante pueda incorporar imágenes propias al modelo.

Por ejemplo: un paciente recuerda cuando vio el mar por primera vez, pero no conoce detalles suficientes como para tener integrada una historia que contar en la mente, ni tomó ninguna fotografía en aquel entonces. El modelo puede generar una foto del paciente en el mar, y el hecho de evocar ese recuerdo, le provoca bienestar y felicidad.

De esta manera, podemos aportar un material muy valioso para la terapia de reminiscencia, ya que se necesita material visual que permita crear una conexión con la vida del paciente, y este material en ocasiones puede ser muy limitado.

1.3. Plan de trabajo

Una vez definidos los objetivos, se debe establecer un método para tratar de llegar a los resultados esperados. En primer lugar, se debe realizar una amplia investigación acerca de los tipos de inteligencia artificial que existen y cuál de todas es la que mejor se adapta a nuestro objeto de estudio. Pero antes de profundizar en las diferentes técnicas y modelos, debemos ser conscientes del motivo por el que realizamos este trabajo, es decir, quién es el destinatario y qué espera del producto final. Por tanto, necesitamos indagar en el foco del problema y saber cómo la inteligencia artificial puede ayudar a resolverlo, o bien a mitigarlo.

Sabiendo que para llegar a los resultados deseados necesitamos una inteligencia artificial generativa de imágenes a partir de texto, necesitamos conocer cuáles son las mejores que están a nuestra disposición, y si las podemos utilizar y trabajar sobre ellas. Por tanto, una parte importante de nuestro proyecto consiste en realizar pruebas de cada una de las inteligencias artificiales generativas y valorar cuál genera imágenes de mayor calidad y cuál la genera en un tiempo aceptable. Es necesario probarlas todas y cada una de las que estén disponibles, y saber cuál va a ser nuestro entorno de ejecución.

Una vez se haya decidido cuál va a ser el o los modelos que elijamos para desa-

rrollar nuestro proyecto, llegará el momento de saber cómo vamos a desplegar las diferentes tecnologías y qué vamos a añadir para que sea algo útil y completamente novedoso para nuestros destinatarios. La idea es crear un programa, que contenga el modelo de inteligencia artificial elegido y una interfaz sencilla y eficaz para utilizarla en las terapias ocupacionales. Además, estos modelos deberán tener la posibilidad de ser personalizables, de manera que se puedan crear imágenes con los elementos o personas que se deseen. Para ello, se investigará sobre los distintos modos de entrenamiento, y realizando múltiples pruebas y analizando los diferentes resultados, se razonará cuál será el entrenamiento óptimo para nuestro proyecto.

Cuando ya disponemos de un modelo de inteligencia artificial generativa de imágenes que otorgue buenos resultados, simularemos casos de uso que ejemplifiquen cómo los usuarios pueden tener una experiencia plenamente satisfactoria. A raíz de esto, podremos obtener una serie de conclusiones y anotar si se han cumplido las expectativas y si en un futuro los pacientes pueden ver su calidad de vida incrementada gracias a nuestra iniciativa y a nuestro trabajo.

Capítulo 2

Estado de la Cuestión

A continuación, se abordarán diferentes aspectos referentes al proyecto y, que son de suma importancia para la comprensión de éste. Es preciso saber cómo está actualmente el campo de estudio que trata el proyecto para que al final, se pueda sacar en claro conclusiones de lo que hemos aportado y aprendido realizándolo.

En primer lugar, trataremos en líneas generales lo que es el Alzheimer y en específico, la terapia de reminiscencia y cómo ayuda a paliar esta enfermedad. Una vez ya puestos en contexto, se explicará lo que es la Inteligencia Artificial y cómo funciona exactamente, en concreto, la Inteligencia Artificial generativa de imágenes que funciona a través de redes neuronales. Lo que nos llevará al siguiente punto a tratar, qué son las redes neuronales, las diferentes variedades que existen, qué puntos positivos presenta cada una y cuál hemos elegido y por qué para implementar el modelo base que se utilizará en nuestro trabajo.

2.1. ¿Qué es el Alzheimer?

El Alzheimer es una enfermedad que destruye lentamente la memoria y que, además, también va deteriorando los pensamientos y la conducta, hasta que poco a poco se ven afectadas las funciones más básicas. El Alzheimer es la principal causa de la demencia.

El cerebro envía estímulos químicos a través de las neuronas creando conexiones cerebrales, y mediante miles de millones de estas conexiones, se obtienen nuestros recuerdos, sentimientos, pensamientos y capacidades locomotoras. Aunque todavía no se sabe con certeza el motivo que causa esta enfermedad en el cerebro, se ha investigado que hay dos proteínas en el cerebro que con el tiempo se vuelven tóxicas, tau y beta-amiloide, que se acumulan hasta obstruir la conexión entre las neuronas

y provocar que estas mueran, como se puede ver en ?? Con la destrucción de las neuronas, el cerebro se va encogiendo y con él también severamente, el hipocampo, que es una parte clave fundamental en nuestro cerebro a la hora de formar nuevos recuerdos y para el aprendizaje, lo que causa que nuestra memoria, nuestra capacidad para tomar decisiones y el habla, fallen.



Figura 2.1: Obstrucción de neuronas por la vitamina tau

2.2. La terapia de reminiscencia

Afortunadamente, existen terapias que ayudan a mejorar la calidad de vida de las personas que padecen esta enfermedad. Entre ellas la estimulación cognitiva, orientación a la realidad, ejercicio terapéutico, musicoterapia, y estimulación multisensorial, entre otras.

Nuestro proyecto está ligado a la terapia de reminiscencia, que es un proceso que ayuda a la persona a evocar momentos y experiencias emocionales e integrarlos en el presente, lo cual puede mejorar la autoestima y la calidad de vida.

En concreto, la técnica consiste en mostrar a la persona una herramienta o material visual, musical o incluso, olfativo, vinculado a su propia experiencia o hechos históricos. De esta manera se promoverá una revisión de la vida de la persona, de modo que se consiga una conexión con sus vivencias, con el fin de reconstruir un libro de vida y reforzar su identidad como persona. Esta terapia se puede clasificar como un tipo de ensueño que los lleva a su pasado, lo que les permite entrar en un estado de concentración con el fin de poder reforzar su memoria general, lo cual fortalece al cerebro y desarrolla sus capacidades sociales, estimula recuerdos a través de órganos sensoriales y activa su sentido de la identidad.

Algunos de los beneficios de la creación del libro de vida y la terapia ocupacional de reminiscencia, entre muchos otros, son:

Bienestar emocional: Se permite que las personas puedan recordar experiencias positivas y significativas de sus vidas, lo cual puede generar sensaciones satisfactorias, de alegría y felicidad.

Autoconocimiento: Cuando una persona revisa y reflexiona sobre eventos pasados y logros personales, puede obtener una comprensión de sí misma. Mediante la interacción de sus recuerdos, ayudado de la guía de cuidadores y terapeutas, puede evocar aspectos de sus valores, fortalezas y debilidades que de otra manera no serían accesibles.

Sentido y propósito personal: Mediante la rememoración de escenas significativas, las personas pueden encontrar una guía emocional que permite dar sentido y dirección a sus vidas, especialmente en momentos de confusión o desorientación.

Reducción del estrés: Al enfocarse en recuerdos positivos, las personas experimentan una sensación de calma, de tranquilidad y de bienestar emocional, y encuentran un espacio para escapar del estrés y tensiones del presente. Favorecer las relaciones sociales: El hecho de compartir recuerdos y experiencias con otras personas, puede fortalecer los lazos sociales y fomentar una mayor conexión con cuidadores, familia y amigos, lo cual es altamente satisfactorio para el paciente.

Aumentar el desarrollo del lenguaje y la persona: Al relatar experiencias pasadas, las personas mejoran su capacidad para comunicarse de manera efectiva, así como su expresión verbal, lo cual impulsa un crecimiento personal significativo a nivel emocional e intelectual.

Prevenir la incapacidad: El hecho de potenciar las habilidades lingüísticas y de mantener la mente activa, puede ayudar a preservar la función cognitiva a lo largo del tiempo.

Existen numerosos estudios y tesis que afirman que las capacidades cognitivas se mantienen y se consigue frenar en cierta medida el deterioro, además de reducir la ansiedad y la depresión, por lo que se hace una recomendación y un llamamiento para que se realicen estas actividades de terapia ocupacional.

2.3. ¿Qué es la Inteligencia Artificial?

La Inteligencia Artificial es un campo de la Informática que trata la creación de herramientas, procedimientos, máquinas y computadores que simulen el proceso de inteligencia humana siendo capaces de realizar tareas como el aprendizaje, el razonamiento, la resolución de problemas, el reconocimiento de patrones, la comprensión del lenguaje natural, la percepción visual e incluso la creatividad. Este proceso se lleva a cabo utilizando principalmente el análisis de datos y las estadísticas, la inge-

niería de hardware y software, aunque también abarca disciplinas como la lingüística, la neurociencia y hasta la filosofía y la psicología.

Ahora, existen dos tipos principales de inteligencia artificial: La IA débil que se centra en tareas específicas y está diseñada para realizar funciones específicas sin poseer una inteligencia general. Y, por otro lado, la IA fuerte que busca replicar la inteligencia humana de manera más completa, con capacidad para comprender, aprender y adaptarse a una amplia variedad de tareas, con el fin de percibir su entorno y aplicando los conocimientos y datos almacenados, tener la capacidad de tomar decisiones para lograr la consecución de objetivos proporcionados.

Al hablar sobre inteligencia artificial, es muy importante mencionar los conceptos de aprendizaje automático (machine learning) y aprendizaje profundo (deep learning), que son subconjuntos de la inteligencia artificial que comparten la meta común de permitir a las máquinas aprender y mejorar su rendimiento en tareas específicas sin intervención humana directa.

El aprendizaje automático es un enfoque que abarca diferentes técnicas en las que se crean modelos de entrenamiento a partir de datos para aplicar los conocimientos adquiridos con el fin de realizar predicciones o tomar decisiones en función de nuevos datos. Es decir, se utiliza para mejorar el rendimiento de la inteligencia artificial en tareas específicas a medida que se expone a mayor cantidad de datos, que en general requiere intervención humana de forma manual.

Algunas categorías dentro del aprendizaje automático son:

- Aprendizaje supervisado: Un algoritmo se entrena con un conjunto de datos etiquetado, donde se le proporcionan ejemplos de entrada y la salida esperada. El modelo aprende a realizar predicciones o tomar decisiones basándose en estos ejemplos.

- Aprendizaje no supervisado: El algoritmo se enfrenta a datos no etiquetados y debe encontrar patrones o estructuras por sí mismo. Esto se utiliza comúnmente para la clasificación o agrupación de datos.

- Aprendizaje por refuerzo: El modelo aprende a través de la interacción con un entorno. Recibe recompensas o castigos según las acciones que realiza, lo que le ayuda a aprender qué comportamientos son más beneficiosos.

El aprendizaje profundo, a lo que respecta, es una disciplina dentro del aprendizaje automático que utiliza redes neuronales profundas con múltiples capas, que contienen datos de entrada y salida, de manera que cada capa puede aprender de la anterior a partir de los datos transformados para realizar predicciones y aprender de manera automatizada, reduciendo gran parte de la intervención humana y per-

mitiendo aumentar el conjunto de los datos a tratar.

Posibilidad de añadir más de información acerca de Deep Learning

2.4. Redes neuronales

Las redes neuronales fueron creadas con el objetivo de simular el comportamiento del cerebro humano para que las máquinas pudieran ser capaces de aprender tareas computacionales de forma similar a las neuronas de nuestro sistema nervioso. La forma más básica de una neurona artificial se puede definir a través del concepto de perceptrón. El concepto de perceptrón fue visto por primera vez en 1956 por Frank Rosenblatt, psicólogo estadounidense. Un perceptrón es una unidad de red neuronal que funciona a través de un algoritmo de clasificación que realiza determinados cálculos en el que se ponderan las entradas, sumándolas y generando así una salida binaria, de manera que permite separar los datos en dos categorías. Para ello, cada entrada tiene asignada un peso que determina la relevancia de esta para su cálculo posterior en la suma ponderada, es decir, cada entrada se multiplica por su peso correspondiente para determinar su importancia relativa. La salida, además, se calcula aplicando una función de activación al resultado de la suma ponderada previamente calculada para así, dictaminar la salida final.

El perceptrón puede constar de una sola capa, que se compone de entrada y salida, o puede constar de múltiples capas, más conocidos como MLP (Multilayer Perceptrons). A diferencia de los perceptrones simples, los MLP constan de unas capas ocultas adicionales, componiéndose así de múltiples neuronas conectadas que se dividen típicamente en tres capas: capa de entrada, capa oculta y capa de salida. En resumidas cuentas, los perceptrones son la forma más simple de una red neuronal.

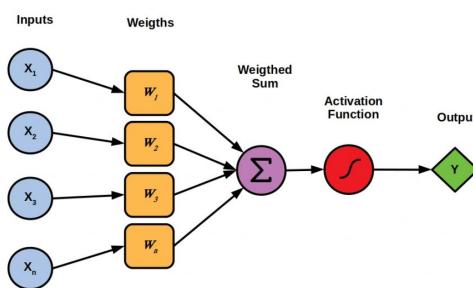


Figura 2.2: Estructura del perceptrón

La primera imagen generada por Inteligencia Artificial fue en 1957, por el propio creador del perceptrón, Frank Rosenblatt, quien entrenó al propio perceptrón con

una serie de imágenes de rostros humanos para que éste aprendiera e identificara un patrón y reprodujera una nueva imagen. La imagen fue generada a través de una matriz de puntos de luz que aunque no se asemejara a lo que realmente era una fotografía de una persona real, hizo que este fenómeno marcará un antes y un después en el desarrollo de la Inteligencia artificial generativa.

¿Qué tipos de redes neuronales existen?

La variedad de redes neuronales es considerablemente grande, y cada una de ellas se han ido desarrollando y diseñando para elaborar tareas específicas. De esta manera, las diferentes redes neuronales han ido adoptando diferentes arquitecturas para tratar diferentes tipos de datos y problemas.

Entre ellas, mencionaremos las más relevantes hoy en día y, analizaremos el funcionamiento de cada una para tener claro cuál es la más adecuada para este modelo, haciendo una profunda comparación entre unas y otras sobre todo, nos centraremos en la red neuronal convolucional, ya que es la implementada para este trabajo.

2.4.1. Redes Neuronales Feedforward (FNN)

Son un tipo de redes multicapa, que como hemos visto, están formadas por conjuntos de neuronas agrupadas en varios niveles o capas, en los que cada neurona está conectada y recibe señales de otras neuronas pertenecientes a la capa anterior, que a su vez, se encargan de transmitir información por señales a las neuronas de la capa posterior, en dirección a la salida de la red. De esta forma, las salidas de cada capa constituyen la entrada a la capa inmediatamente posterior. En todo caso, las conexiones de la red fluyen exclusivamente en una sola y única dirección, de ahí el nombre feedforward, que traducido es hacia delante. Por norma general, la arquitectura típica que sigue una red neuronal multicapa consta de tres capas: capa de entrada, capa oculta y capa de salida.

Sin embargo, cabe la posibilidad de que haya más de una capa de cada tipo, y cuando se da el caso de que la red consta de más de una capa oculta, la red se califica como profunda, traducida al inglés como deep neural network. Concretamente, añadir más de una capa oculta a la red permite crear un modelo interno que reconoce patrones y proporciona un mayor rendimiento en la interpretación y estructuración de diferentes propiedades de objetos. Por ello, estas redes fueron diseñadas específicamente para resolver problemas de clasificación, regresión y por supuesto, como hemos visto, realizar reconocimiento de patrones.

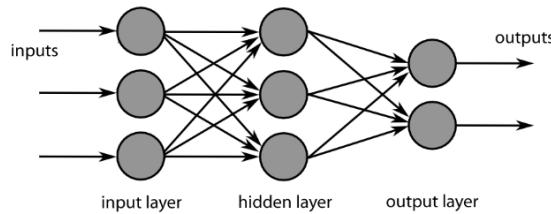


Figura 2.3: Red neuronal feedforward

2.4.2. Redes Neuronales Recurrentes (RNN)

Como hemos visto, las redes neuronales feedforward están habilitadas para que la información fluya en una sola dirección. Sin embargo, cuando se les proporciona una memoria, el resultado que se obtiene son las redes neuronales recurrentes, o Recurrent Neural Networks (RNN) en inglés. El origen de estas redes se popularizó en 1982 por el físico americano John Hopfield, en las que se destacan por su comportamiento dinámico y estable.

¿Y cómo es posible añadir memoria a las propias conexiones? Simplemente generalizando sus conexiones, es decir, alimentar a sus propias entradas o inputs con las salidas o outputs generados previamente por las conexiones anteriores provocando que el modo en el que fluyen las conexiones sea bidireccional. Es decir, incluyendo conexiones hacia atrás con las que se trabaja en una serie de pasos de tiempo, conocidos como timesteps, donde se procesan los elementos de la secuencia uno por uno, manteniendo una memoria de los estados anteriores a medida que avanzan en la secuencia.

El entrenamiento de redes neuronales de estas características se realiza contando con diferentes algoritmos y técnicas, la más básica y conocida en este tipo de red neuronal es el algoritmo de propagación de errores (back-propagation en inglés) formalizado en 1986 por Rumelhart, Hinton y Williams. Este método, en concreto, consiste en aplicar un patrón a la primera capa de la red, el cual se va propagando hacia las capas superiores con el objetivo de generar una salida que se pueda comparar con la salida deseada y así poder calcular el error para cada neurona de la salida obtenida en función de los diferentes parámetros de la red. En otras palabras, cómo varía el error en relación con la variación de los parámetros de la red neuronal. Si a este fenómeno le sumamos la variable del tiempo, se obtiene la técnica de retropropagación a través del tiempo (backpropagation through time, BPTT), en la que después de analizarse la secuencia completa, calcular el error y cambiar los parámetros para minimizar el mismo, se propaga el error a través del tiempo desde la última capa hasta la superior en cada paso de tiempo con el fin de que la red aprenda de las secuencias y mejore su predicción futura. Este método combinado con la técnica del gradiente descendiente encargada de la optimización del error buscando los parámetros adecuados para poder reducirlo al mínimo es el modo de aprendizaje supervisado más popular que existe. Aunque, como hemos visto anteriormente, no

es el único, ya que coexiste con técnicas de entrenamiento como el aprendizaje no supervisado y el aprendizaje por refuerzo.

Así que, podemos concluir en que las redes neuronales son idóneas para tareas de modelado de datos secuenciales y con dependencias temporales, como lo pueden ser el procesamiento de lenguaje natural, la generación de texto o la traducción automática, entre muchas otras.

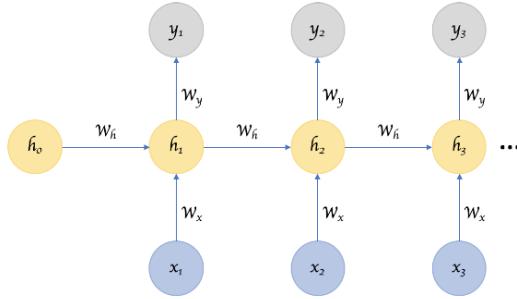


Figura 2.4: Red neuronal recurrente

2.4.3. Redes Neuronales LSTM

Las redes neuronales Long Short-Term Memory surgen a partir de las redes neuronales recurrentes cuyo principal objetivo y propósito es ratificar el problema del desvanecimiento del gradiente.

En primer lugar y para entender en qué consiste este problema, es necesario saber qué es exactamente un gradiente y a qué nos referimos cuando hablamos del desvanecimiento del mismo. Un gradiente es una medida que determina la variación de una función con respecto al cambio que se produce en sus variables, ya sea en términos de optimización, rapidez o maximización de la función objetivo.

Matemáticamente, se puede entender como un vector multivariable cuyas variables son las derivadas parciales de dicha función. Se puede ver en el ejemplo de la figura. Geométricamente hablando, el gradiente indica la dirección en la que la función crece a mayor velocidad y, de esta manera, podemos concluir que el uso de los gradientes se hace con el fin de ajustar los parámetros para reducir al mínimo la diferencia entre los valores obtenidos y los deseados. Un claro ejemplo de ello son los modelos de predicción, en los que se busca reducir dichas predicciones con los valores reales.

Ahora bien, cuando hablamos del desvanecimiento del gradiente nos referimos al fenómeno que ocurre en la propagación hacia atrás, cuando a medida que se va

produciendo la propagación en capas cada vez más profundas, el gradiente va disminuyendo hasta llegar a ser tan pequeño que las capas superiores sean incapaces de aprender de manera efectiva. Este tipo de problema es común en redes conformadas por una gran cantidad de capas, dado que se disminuye exponencialmente a medida que va recorriendo cada capa hasta impedir que los pesos se actualicen de manera correcta y que puedan tener un aprendizaje lo suficientemente eficiente para ser capaces de resolver patrones complejos. La solución a este problema se abordó introduciendo nuevos componentes a la red que ayudan y controlan el flujo de información de la red, consiguiendo aumentar la memoria del conjunto de la red al almacenar la información durante períodos más largos de tiempo. Los componentes principales son los siguientes: Memory cells o celdas de memoria: contienen la información relevante y actualizada a lo largo del tiempo.

Input gates o puertas de entrada: son las encargadas de controlar la cantidad de información que entra a las celdas de memoria.

Forget gates o puertas de olvido: seleccionan la información que debe ser eliminada de la celda, al no ser de utilidad.

Output gates o puertas de salida: su tarea radica en seleccionar la información de la celda que va a pasar a la siguiente capa de la red, basándose en el estado oculto actual.

Para entender cómo funcionan y se relacionan estos componentes en la estructura de la red, lo veremos con un ejemplo hipotético en el que la celda de memoria representa una caja fuerte y las diferentes puertas son componentes de una cinta transportadora.

La caja fuerte contiene información valiosa que se transporta en la cinta, el la puerta de entrada actúa como una máquina que trabaja en la cinta que se encarga de introducir nueva información a la caja fuerte y regular la información contenida en esta; la puerta de olvido actúa como una máquina que descarta fragmentos de información contenidas en la caja fuerte; y por último, el componente que representa la puerta de salida, se encarga de elegir si la información contenida en la caja fuerte es apta para salir de la cinta transportadora.

De este modo, se consigue erradicar el problema del desvanecimiento del gradiente y se consigue procesar grandes secuencias de datos durante largos períodos de tiempo. Es especialmente útil en tareas que requieren capturar grandes dependencias de datos temporales, ejemplo de ello son el procesado y/o reconocimiento de lenguaje natural, en la traducción automática y en la generación de texto.

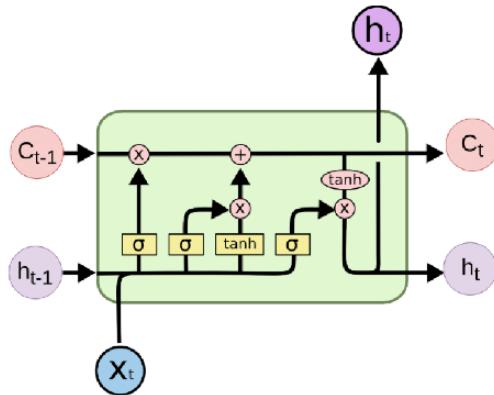


Figura 2.5: Red neuronal Long Short Term Memory

2.4.4. Redes Neuronales Convolutivas (CNN)

Las redes neuronales convolucionales se introdujeron por primera vez en la década de los 50 por David Hubel y Torsten Wiesel cuando experimentaron con las neuronas biológicas, lo que le sirvió de inspiración a Kunihiko Fukushima en la década de los 80 para desarrollar el Neocognitron, una red neuronal que se conoce como la primera CNN. Este concepto fue cobrando forma con el paso de los años y el modelo tal y como lo conocemos hoy en día, fue obra de Yann LeCun en 1998 al introducir el aprendizaje a través de la técnica de backforwarding.

Las redes neuronales convolutivas están formadas por una secuencia de capas que podemos clasificar en tres tipos: capas convolutivas, capas de pooling y capas completamente conectadas.

Capas convolutivas: suponen la capa principal de la red, y el papel que desempeña en el procesamiento de imágenes, es realizar la gran parte de los cálculos que se necesitan para extraer características de las imágenes. Para ello, hace falta la intervención de tres elementos principales: datos de entrada, un filtro y un mapa de características.

Los datos de entrada son el elemento que se trata de analizar, por ejemplo, en el caso de una imagen de color que estuviera compuesta por una matriz de píxeles en 3D, las dimensiones serían la altura, la anchura y la profundidad de la misma.

Por otro lado, el filtro o kernel, se trata de un detector de características, que va pasando por cada área de la imagen para identificar diferentes características, este proceso se denomina convolución.

El siguiente paso es representar mediante una matriz bidimensional de pesos, que al aplicarse en cada área calcula un producto escalar a partir de los píxeles de los datos de entrada y del filtro. El producto escalar se utiliza como input en la matriz de salida para que el filtro sea capaz de repetir el proceso por toda la imagen. La finalidad del filtro es ser capaz de distinguir diferentes patrones, como lo pueden ser

texturas, bordes o figuras.

Finalmente, la suma de los diferentes productos escalares y el o los filtros utilizados, da como resultado final lo que se conoce como mapa de características.

Después de cada capa de convolución, se aplica una función de activación no lineal al mapa de características, a través de la ReLU (Rectified Linear Unit). El fin de esta función es introducir no linealidad a la red, lo que le permite mejorar la complejidad entre las diferentes características.

Capa de agrupación o pooling: son capas dedicadas a reducir la dimensión del mapa a través de la disminución del número de parámetros de entrada.

Comúnmente, se utilizan dos técnicas conocidas como max pooling y average pooling. Max pooling consiste en seleccionar el pixel con mayor valor a medida que el filtro recorre la imagen para enviar el máximo a la matriz de salida, en cambio, average pooling lo que busca es calcular el valor medio del campo. El inconveniente que presenta esta capa es la gran pérdida de información que existe, sin embargo, presenta una gran ventaja a la hora de reducir la complejidad computacional y concentrarse en las características más importantes, lo que claramente, mejora el rendimiento del modelo y evita el riesgo de que se produzca un sobreajuste.

Capa totalmente conectada: en las últimas capas de la red los nodos están conectados con los de la capa anterior para producir la salida final incorporando las características extraídas y aprendidas de los procesos realizados en las capas anteriores. Las capas totalmente conectadas se preocupan de realizar las funciones de clasificación de la imagen y de regresión para producir el resultado deseado.

Las principales funciones y tareas que abarcan las redes neuronales convolucionales son el reconocimiento de imágenes identificando y detectando objetos, personas o animales; análisis de imágenes para diferentes propósitos, por ejemplo, médicos; reconocimiento facial; segmentación semántica; y por supuesto, generación de imágenes.

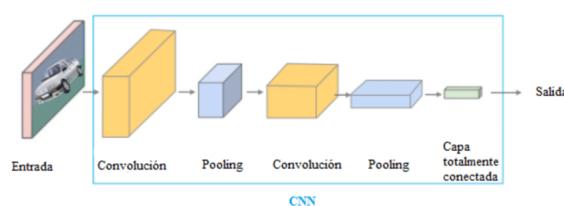


Figura 2.6: Red neuronal convolucional

2.4.5. Redes Generativas Adversarias (GAN)

Las redes GAN, del inglés, Generative Adversarial Networks son un tipo de redes neuronales profundas revolucionarias, que son relativamente novedosas, ya que surgieron en el año 2014 por primera vez por Ian Goodfellow y sus compañeros en la Universidad de Montréal.

El funcionamiento de este tipo de redes involucra a dos redes neuronales diferentes, en las que cada una se encarga de realizar una tarea específica y “competir” contra la otra, para así, cada vez ir mejorando más los resultados obtenidos.

La primera red neuronal de este sistema se conoce como “Generador” y se encarga de producir y crear datos totalmente nuevos basándose en los datos aportados en el entrenamiento de la red. Se le asigna como entrada un vector de ruido y es responsable de crear datos que se asimilen a los datos originales. El entrenamiento del Generador es constante y siempre busca mejorar los resultados a medida que los va produciendo para alcanzar el máximo realismo posible.

La segunda red neuronal de la que se compone este sistema es el “Discriminador”, se dedica a analizar los resultados producidos por el Generador e identificar si son los reales o los creados por la red. A medida que va avanzando su entrenamiento, la capacidad del discriminador en distinguir entre un resultado real o falso dada una cierta entrada va mejorando cada vez más, haciendo que su predicción sea más certera. Un clásico ejemplo llevado a la vida real de esto puede ser el caso de un falsificador de billetes y un detective, en el que el primero tiene como muestra cierta cantidad de billetes e intenta replicarlos para que más adelante el segundo agente intente detectar la copia del original. A medida que pasa el tiempo, cada uno de los dos individuos van mejorando en su tarea llegando a un nivel de equilibrio.

Matemáticamente, el discriminador tiene que generar una salida, expresada como $D(x)$, basándose en la probabilidad de que la entrada sea sintética o real, suponiendo que en cuanto más cercana a 1 sea, la entrada es original. Y por el contrario y dada una muestra aleatoria z en función de cierta distribución de probabilidad, el generador tiene que producir una muestra, expresada como $G(z)$, que el discriminador tiene que clasificar como cercana a 0, produciendo una salida del tipo $D(G(z))$ en la que el generador tiene que intentar que su probabilidad se aproxime a 1, justo al contrario que el discriminador. Suponiendo que entre todas las muestras del modelo, una mitad son auténticas y la otra mitad son falsas, se tiene que alcanzar el conocido como equilibrio de Nash en el que las muestras del modelo son igual a los datos y en las que la probabilidad del discriminador es $D(x) = 0,5$ para todo x .

Por último, cabe destacar que para el entrenamiento de este tipo de red se utilizan los métodos del gradiente descendente y backpropagation, vistos en redes neuronales

anteriores.

Este tipo de redes son ideales para tareas que requieren creación de datos realistas y artísticos, como lo pueden ser la generación de imágenes, de música o incluso, síntesis de voz.

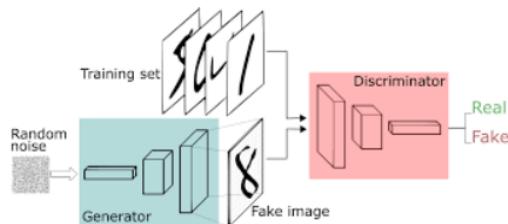


Figura 2.7: Red neuronal generativa antagónica

2.4.6. Redes Neuronales Transformer

Las redes neuronales Transformer son las más actuales tratadas en este trabajo, surgieron en 2017 a través del artículo “Attention is all you need” elaborado por Vaswani et al.. El artículo trata una mejora de las redes recurrentes y convolucionales conocidas introduciendo un mecanismo basado sólidamente en la atención, es decir, realizar paralelismo entre las diferentes tareas que realiza la red. Para ello, se parte de una entrada mediante embedding y se hace uso de dos componentes: codificador y decodificador.

Embedding: es el primer bloque por el que está compuesto la red neuronal, su función es primordial para el manejo de los datos, ya que transforma el texto de entrada en unos determinados vectores o tokens, los cuales son la representación numérica del valor inicial.

Codificador o encoder: después del embedding de la entrada, el siguiente bloque es el codificador posicional cuya tarea es indicar a la red el orden de los diferentes elementos del vector, es decir, de las palabras en el texto. Esta función es esencial ya que la secuencia se procesa en paralelo, y del contrario, no se podrían concretar las posiciones de cada uno.

A continuación, se encuentran conectados en secuencia los codificadores. Cada codificador se compone de 4 elementos: el bloque residual, una red neuronal, otro bloque residual, y por último, un bloque atencional siendo el más importante de todos al encargarse de determinar la relevancia de cada uno de los tokens para la frase junto a su asociación. Los restantes elementos se encargan de normalizar la entrada y la salida para poder seguir entrenando la red de forma productiva.

Decodificador o decoder: cada codificador está conectado a un decodificador, que cumplen una composición similar, sino igual, a la explicada en los codificadores, es decir, cuatro elementos: dos bloques residuales, una red neuronal y un bloque atencional. A los que se le añaden dos elementos más en la estructura: un tercer bloque residual y un bloque atencional con enmascaramiento.

Sin embargo, el funcionamiento difiere al de la estructura anterior. Se comienza con el bloque atencional con enmascaramiento que codifica las relaciones entre elementos atendiendo únicamente a palabras actuales y pasadas. Por otro lado, el bloque atencional del decodificador se conecta con el del codificador para establecer el orden de prioridad al que se debe prestar atención en la secuencia, sus valores son probabilidades entre 0 y 1 siendo el valor más alto el seleccionado. Por último, el comportamiento de los demás bloques cumplen la mismas funciones que en el codificador, incluyendo entre ellos, los bloques residuales, el bloque de codificación posicional y la salida en función del embedding.

Por lo tanto, podemos afirmar que estas redes neuronales aprenden contexto y significado mediante el seguimiento de relaciones en datos secuenciales. Todo tipo de organizaciones utilizan estos modelos para conversión de secuencias, incluidas las de reconocimiento de voz y la traducción automática. Esta red neuronal procesa secuencias largas con cálculo paralelo, con el objetivo de reducir significativamente el tiempo de entrenamiento y de procesamiento. A raíz de este modelo, surgen técnicas innovadoras como el aprendizaje por transferencia y la generación aumentada de recuperación (RAG). El objetivo principal es entrenar inicialmente los modelos de conjuntos de datos amplios y después refinarlos de manera precisa, utilizando conjuntos de datos más específicos. De este modo, se maximiza su utilidad y relevancia dentro de cualquier contexto.

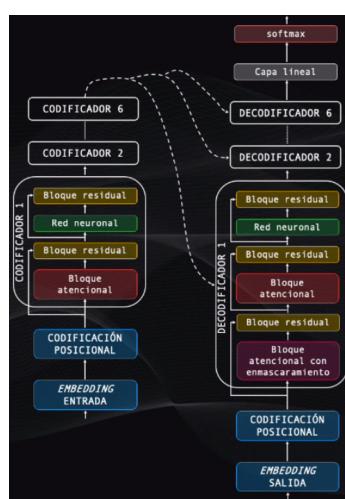


Figura 2.8: Red neuronal Transformer

Estos suponen sólo algunos ejemplos de los tipos de redes neuronales más comunes.

nes y ampliamente utilizadas. Cada tipo tiene sus propias características, fortalezas y debilidades, y es importante elegir el tipo adecuado según el problema específico que se esté abordando y el tipo de datos disponibles.

2.5. Stable Diffusion

Como hemos visto, hay diferentes redes neuronales destinadas a la generación tanto de texto como de imágenes. En el caso de Stable Diffusion, el modelo de IA utilizado en nuestro proyecto, funciona con un modelo de difusión latente basado en CNNs y en Transformers. Ya visto el comportamiento y funcionamiento de ambas redes neuronales en apartados anteriores, analizaremos más en profundidad de qué manera se integran y complementan en el modelo que utilizaremos más adelante.

Stable Diffusion es un modelo de Inteligencia Artificial generativa cuya principal función es transformar el texto a imagen, aunque también presenta otras funciones asombrosas como la transformación de imagen a imagen o incluso lo más novedoso hasta el momento, transformaciones de texto a vídeo. Las empresas desarrolladoras hicieron una colaboración conjunta entre CompVis LMU, Runway y Stability AI y el lanzamiento finalmente se produjo a mediados de 2022, es decir, esta poderosa herramienta relativamente nueva y los avances tecnológicos que ha alcanzado hasta la fecha son impresionantes.

2.5.1. Funcionamiento interno de Stable Diffusion

En principio, Stable Diffusion cuenta con un entrenamiento de más de 5 millones de imágenes proporcionado por el dataset Laion-5B que permite contar con una gran variedad de opciones de creación de imágenes, desde objetos, animales, paisajes y lugares, personas e incluso celebridades mundialmente conocidas con una calidad notable.

Comencemos entendiendo el funcionamiento del modelo de difusión latente, y para facilitar su explicación y comprensión, lo disecaremos en dos partes: difusión y el espacio latente. El proceso de difusión que sigue la generación de una imagen es, en primer lugar y partiendo de las millones de imágenes del dataset mencionadas, se procede a añadir ruido gaussiano gradualmente a través de una serie de pasos hasta que las fotografías pierden todo valor y terminan siendo irreconocibles. Se podría decir que se comporta como una cadena de Markov, al ser el ruido gaussiano una variable aleatoria y al depender cada paso exclusivamente de su anterior. Este primer método se conoce como difusión directa hacia delante o forward diffusion.

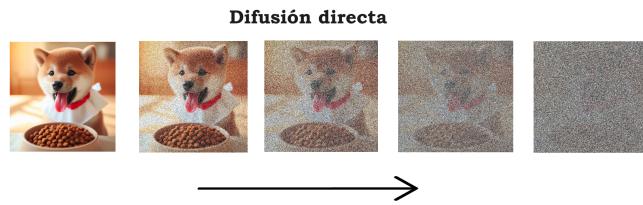


Figura 2.9: difusión directa de una imagen de un cachorro de Shiba Inu comiendo

Así mismo, cuando se procesa una petición dado un prompt, se parte de una imagen únicamente hecha de ruido aleatorio, es decir, un imagen sin nada relevante ni identificable en ella. Partiendo de esta imagen llena de ruido, se intenta revertir lo hecho anteriormente para volver a la imagen original quitando el ruido gradualmente. El ruido es aleatorio, y su aleatoriedad depende de un parámetro llamado seed o semilla que asocia el ruido generado con un número aleatorio. Por lo tanto, si se repite la semilla se volverá a generar exactamente el mismo ruido y tendríamos como resultado una réplica de una imagen ya generada con esa misma semilla. Este proceso se conoce como difusión inversa y su objetivo principal es que el modelo aprenda a eliminar el ruido completamente.

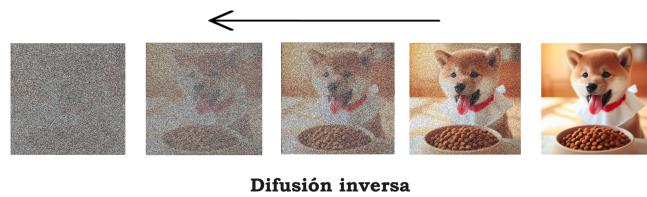


Figura 2.10: difusión inversa de una imagen de un cachorro de Shiba Inu comiendo

Una vez tenemos la imagen llena de ruido y la forma de calcular el ruido que hay en cada imagen para poder eliminarlo, se empieza el proceso al que llamamos sampler o muestreo. Aquí es cuando entra un componente importante llamado noise scheduler, cuya función es determinar la cantidad de ruido que se debe suprimir en cada paso para alcanzar la forma óptima y que se puedan evitar cambios bruscos entre paso y paso, que sea de forma gradual y que los detalles se vayan puliendo conforme la imagen vaya cobrando más forma. El proceso de muestreo se repite la cantidad de steps o pasos especificada por el usuario. Y de esta manera, concluimos con el primer componente del modelo de difusión latente: la difusión.

Pasemos a la segunda pieza del puzzle: el espacio latente. El proceso de difusión, por lo que hemos podido ver, es un proceso algo lento y costoso al tener que trabajar con los píxeles de una imagen. Si tenemos una imagen en color con la escala RGB de 512x512 píxeles, estamos hablando de una multiplicación de 3x512x512, es decir, un espacio de casi 80 mil dimensiones. Para solucionar este problema, se trabaja con el espacio latente que trata de reducir la imagen a una escala de 64x64 píxeles, asegurando mayor velocidad y una menor carga de trabajo al trabajar, de esta forma,

con unas dimensiones más pequeñas de apenas 12 mil.

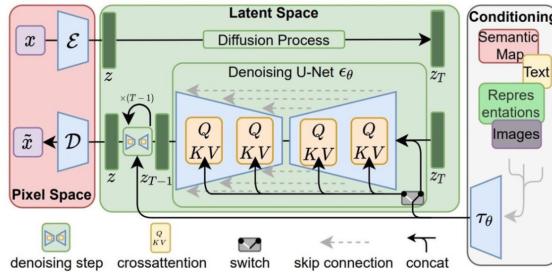


Figura 2.11: Representación gráfica del proceso de conversión al espacio latente

Ahora que entendemos el propósito y el funcionamiento de las dos variables del modelo de difusión latente, profundicemos más en cómo se lleva a cabo cada uno de estos dos procesos y qué mecanismos se utilizan para ello. El modelo se compone principalmente de 3 componentes:

Un codificador de texto basado en redes transformer, una red U-Net compuesta de dos redes ResNet y un autocodificador variacional (VAE).

Codificador de texto basado en Transformers: para producir la imagen que deseamos, previamente es necesario escribir una descripción detallada de la imagen y es por tanto, el primer paso. Este texto que le introducimos al modelo se conoce como prompt y es de lo que se encarga de procesar el codificador de texto. Lo primero que realiza el text-encoder de Stable Diffusion es a través de un modelo llamado CLIP (Contrastive Language-Image Pre-Training) que se encarga de ofrecer una descripción detallada de las imágenes a través de su propio tokenizador. El siguiente paso y como ya hemos visto en apartados anteriores, el transformer se encarga de realizar la fase de embedding en la que se transforman las palabras de texto en tokens que la red neuronal pueda entender y manejar, para que después, a través del método de self-attention, se decida qué palabras son las que más relevancia tienen.

Además, Stable Diffusion ha añadido a esto una pequeña variación y mejora que añade a esta última técnica, otra llamada cross-attention (o atención cruzada) con la que se permite crear relaciones entre los diferentes embedding y mejorar la precisión del resultado. Por ejemplo, si el prompt pedido es “unas flores pequeñas sobre una bicicleta azul”, solamente con la técnica de self-attention podría procesarse una imagen que fuera “una flor azul sobre una bicicleta pequeña”, lo cual es válido para esa arquitectura pero no es lo que el usuario ha pedido. En cambio, con la ayuda de la técnica de atención cruzada se crean relaciones que tienen más a menos distancia, en la que en este ejemplo la relación flores con azul tiene más distancia que la relación de flores con pequeña, y por tanto, es esta última la que se utilizaría al tener más relación y menos distancia.

Red neuronal U-Net: es una red neuronal convolucional entrenada para identifi-

car e intentar predecir la cantidad de ruido contenida en una imagen. Consta de un codificador y un decodificador en los que cada uno de estos componentes son, a su vez, bloques ResNet, que son redes convolucionales profundas compuestas por una gran cantidad de capas. La función del codificador se basa en reducir la calidad de la resolución de la imagen mientras que la función del decodificador es la contraria, generar la imagen en la máxima resolución posible. Entre ambos componentes se añaden conexiones de acceso directo para evitar la pérdida de información importante. Al final, lo que se pretende conseguir es que la red U-Net consiga determinar el ruido para posteriormente, poder conseguir una representación libre de ruido en la imagen final.

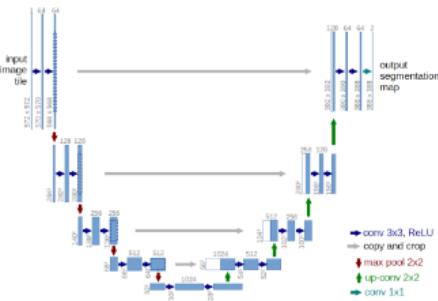


Figura 2.12: Arquitectura de una red U-Net

Autocodificador variacional VAE: es un tipo de red neuronal que al igual que las redes U-Net consta de dos componentes: un codificador y un decodificador, sin embargo sus funciones distan mucho las unas de las otras. El primer componente del autocodificador variacional se encarga de uno de los primeros pasos que realiza Stable Diffusion, que es convertir el espacio de píxeles de la imagen en un tensor dentro del espacio latente de menores dimensiones sustrayendo las características más relevantes de la imagen original y comprimiéndolas en el tensor latente. Y al final del proceso de difusión, se parte del tensor del espacio latente con el que se ha trabajado, para transformarlo en la imagen final generada en una escala de 512x512 píxeles.

2.5.2. Parámetros adicionales

2.5.2.1. Escala CFG

La escala CFG, o Classifier Free Guidance, es un parámetro de Stable Diffusion que controla la medida en la que la generación de imágenes se ciñe a la descripción introducida por el usuario. Esto significa que, si se establece un CFG más bajo, la inteligencia artificial tiene una mayor libertad para ser más creativa, mientras que un nivel más alto obliga a seguir en mayor medida las indicaciones.

A nivel de calidad de imagen puede tener diversos efectos. Este nivel debe ajustarse en función del prompt indicado. Para indicaciones breves es preferible elegir

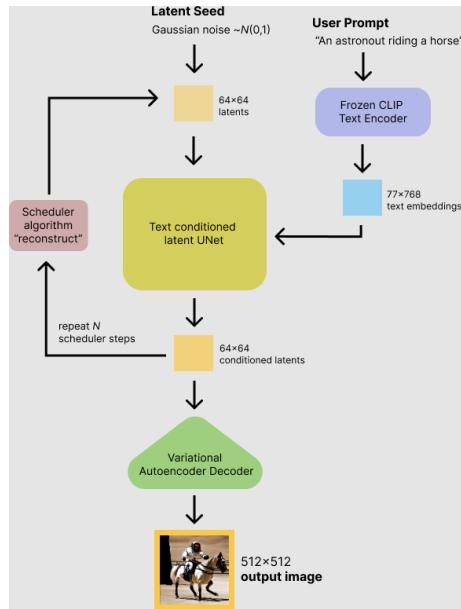


Figura 2.13: Representación gráfica del funcionamiento de Stable Diffusion

un CFG más bajo, de entre 2 y 6, mientras que en el caso en el que se pretenda una descripción muy precisa, se necesitará un nivel algo más alto, de entre 10 y 15. No obstante, lo más recomendable es ajustar este nivel entre 7 y 10, puesto que se encontrará un equilibrio entre creatividad y precisión.

Capítulo 3

Descripción del Trabajo

3.1. Elección del modelo

Como hemos podido ver, Stable Diffusion es una herramienta potente y eficaz con una buena estructura que cuenta con una gran base de datos en la que podemos encontrar una inmensa variedad de imágenes. Estas características la convirtieron en nuestra elección final, pero veamos más en profundidad el proceso de elección de esta Inteligencia Artificial generativa en comparación con la amplia variedad de las que hay presentes actualmente en el mercado.

En primer lugar, para entender nuestra elección es necesario poner en contexto el gran abanico de posibilidades de IAs generativas que hay hoy en día, sus prestaciones, características, y sobre todo, su accesibilidad.

Las IAs generadoras de imágenes más potentes del mercado y las que consideramos para desarrollar la base de nuestro trabajo son Midjourney, DALL-E, Leonardo AI y Stable Diffusion, todas ellas producen unos resultados bastante satisfactorios. Sin embargo, descartamos rápidamente las dos primeras: Midjourney producida por un laboratorio independiente y, DALL-E producida por la famosa empresa creadora de ChatGPT, OpenAI. El motivo fue que al ser ambas de pago, no podíamos tener acceso a su modelo de forma tan amplia como las demás, siendo prácticamente imposible acceder a ellas y mucho menos a poder entrenarlas.

Para acceder a los modelos utilizamos Hugging Face, la plataforma que cuenta con una amplia gama de bases de datos de todo tipo y de modelos de IA generativa de texto a imagen, imagen a imagen, imagen a texto y un largo etcétera. Intentamos buscar modelos de ambas en la plataforma sin éxito ya que lo máximo que encontramos eran imitaciones o pequeñas demos que no alcanzaban el nivel de calidad requerido.

Leonardo AI sí que es una opción algo más accesible ya que la plataforma sí que cuenta con un plan gratuito que te deja probar el modelo con una limitación

de 150 imágenes a generar al día, lo cual está muy bien. Además, la calidad es bastante buena y permite ajustar gran variedad de parámetros, como por ejemplo, el número de imágenes que deseamos que se generen al mismo tiempo, el estilo, la paleta de colores que deseamos, tamaño e incluso la resolución. El inconveniente era la accesibilidad al modelo y la información sobre este mismo, ya que era escasa y en Hugging Face no había acceso. En comparación con Stable Diffusion, la incertidumbre era mucho más alta y había muchísimas limitaciones a la hora de utilizarlo.

Todo ello nos llevó a elegir Stable Diffusion rápidamente, una herramienta open source, es decir, de código abierto que nos facilita mucho el entrenamiento de personas para obtener imágenes personalizadas de forma rápida y con alta calidad. Además, sus diferentes versiones nos permitían explorar aún más a fondo el modelo y saber cuál era el que encajaba con las características que presentaban las prestaciones de nuestros equipos. Stable Diffusion resultó ser la candidata ideal para que la creación de imágenes destinadas a los libros de vida fuera lo más sencilla, familiar y creativa posible.

Ahora expuestos todos los motivos de elección de Stable Diffusion, tanto en comparación con otras redes neuronales vistas en el Capítulo 2 como en comparación con otros modelos de IA Generativa, expondremos cómo implementamos la herramienta a lo largo de todo el trabajo y las dificultades a las que nos enfrentamos durante el proceso.

3.2. Entrenamiento con Stable Diffusion

Stable Diffusion cuenta, hasta el momento, con 3 grandes versiones que, paulatinamente, han ido mejorando la calidad en las imágenes generadas. Todas ellas son totalmente gratuitas y de libre acceso, la primera versión que se presentó fue Stable Diffusion 1 (en sus variaciones 1.4 y 1.5), seguida por Stable Diffusion 2 (con sus respectivas variaciones 2.0 y 2.1) y por último, Stable Diffusion XL (que cuenta con su variación XL Turbo).

La diferencia principal entre las dos primeras versiones es el tamaño de resolución de las imágenes ya que las primeras versiones trabajaban en un espacio de 512x512 píxeles y en la versión 2 dicho tamaño aumentó a 768x758. Además, se introdujeron algunas correcciones y mejoras como la técnica de inpainting, que se trata de la restauración de algunas partes de la imagen mejorando la calidad y los detalles de la misma o incluso, reemplazando ese área por lo especificado por el usuario en el prompt.

Con la última versión Stable Diffusion XL, se generan imágenes con una calidad excepcional, lo que supuso una gran mejora en el modelo al contar con un dataset mucho más extenso. El inconveniente con la versión XL, en nuestro caso, era la

limitación de que requiere una tarjeta gráfica demasiado potente, con la que, por desgracia, no contamos en nuestro equipo.

Respecto a la versión 2, si bien es verdad que no requiere tanta GPU como la versión XL, sí que requiere más que en la primera versión, al ser las imágenes con mayor resolución y, viendo la comparación en la calidad que presentaban los resultados de ambas, optamos por utilizar la versión 1.5 ya que era la que mejor se adecuaba a nosotros en términos de calidad y tiempo.

Ahora bien, existen varios métodos a través de los cuales se puede utilizar esta herramienta y hemos probado sus funcionalidades de diferentes maneras. En primer lugar, se puede utilizar mediante código escrito en Python a través de Google Colab, ya que la plataforma ofrece cuadernos en los que se trabaja de manera online y que además, proporciona una GPU en la nube a la que Google te da acceso. En concreto, esta GPU es la T4, que es la única opción que nos deja Google entre las que hay (A100 GPU, L4 GPU, V100 GPU) ya que se conoce que las demás son de pago. Otra alternativa es mediante la propia página de Stable Diffusion, que ofrece una demo para utilizar esta avanzada versión. Por último, ejecutar el modelo en local, consiguiéndolo descargar en la página Hugging Face, que incluye multitud de modelos de todo tipo, bases de datos, librerías y licencias para descargar y utilizar, por lo que hemos podido comprobar, presenta muy buenos resultados.

El hecho de probar un modelo de inteligencia artificial en un servidor no es concordante con nuestros objetivos del proyecto, puesto que necesitamos entrenar un modelo e incluirlo en una aplicación, de manera que el usuario pueda interactuar y conseguir imágenes personalizadas en un tiempo aceptable, por ello descartamos la opción de utilizar la demo que se encuentra en la página de Stable Diffusion.

Una vez que tenemos el modelo de generación de imágenes elegido, se debe ejecutar en nuestro ordenador y ver cuál es el rendimiento real. Esto quiere decir que la imagen debe generarse de manera correcta y sin deformaciones, y debe incluir todos los elementos solicitados en la descripción introducida. Además, debe realizar esta generación en un tiempo adecuado.

Para ello, el proceso más óptimo y que finalmente elegimos llevar a cabo tras gran cantidad de pruebas es, en primer lugar, realizando el entrenamiento de imágenes personales a través de la plataforma de Google Colab en internet y en segundo lugar, para la generación de imágenes desde nuestro ordenador optamos por la instalación de una interfaz, llamada NMKD Stable Diffusion GUI. Esta herramienta nos permite ejecutar localmente cualquier modelo de generación de imágenes a partir de texto, e incluso permite aceptar imágenes como input, es decir, generaciones de tipo imagen a imagen.

El principal de los objetivos que establecimos en la realización del proyecto era generar imágenes personalizadas del paciente en cuestión, y para ello es estrictamente necesario entrenar el modelo elegido.

Como se ha dicho anteriormente, el método elegido fue un cuaderno en Google Colab mediante Dreambooth, un modelo de generación de aprendizaje profundo, y que fue desarrollado en 2022 por un grupo de investigadores de Google Research y la Universidad de Boston. Este modelo nos permite añadir capas de entrenamiento a la inteligencia artificial para que reconozca objetos concretos. Esto es muy importante, porque es el mecanismo que consigue mejores resultados y con una velocidad aceptable, que era la utilización que queríamos otorgarle. Por consiguiente, podemos decir que la misión de esta tecnología es la de poder entrenar a modelos de inteligencia artificial para personalizarlo según tus necesidades.

Para realizar el entrenamiento de una forma correcta, lo primero que tenemos que tener claro es el elemento o token al que queremos dar una identidad. Por ejemplo, si seleccionamos una persona, debemos elegir unas imágenes en las que aparezca, de tal manera que, tras el entrenamiento, la IA pueda identificarla.

Lo ideal es que se elija un número considerable de fotografías, a partir de 10, las cuales tienen que cumplir ciertas características. Deben ser fotografías de buena calidad, bajo diferentes ángulos, escenarios y luces, se recomienda que como mínimo hayan 1 o 2 fotografías en las que la persona aparezca de perfil, mostrando 3 cuartos de la cara, de frente y si es posible que en alguna esté sonriendo (para que la IA pueda reconocer la expresión), de cuerpo entero, cintura para arriba y del rostro de cerca. Además, es importante que la ropa no sea siempre la misma, sino al ser así, el modelo podría interpretar como que la ropa forma parte de la persona y siempre se la generaría con la misma, lo cual no queremos que ocurra. Idealmente, las fotografías deberían alternar la luz y estar hechas tanto en interiores como en exteriores. Por último y esto es fundamental, estas deberán tener un tamaño igual o mayor a 512 x 512 píxeles, y deberán llamarse exactamente de la misma manera, con el identificador del token al que hagamos referencia. Además, es preferible que todas las imágenes tengan la misma extensión, ya sea .jpg o .png.

En el ejemplo de la figura 3.1 elegimos como persona de entrenamiento una que no fuese reconocida por nuestro modelo (al contar con un dataset de 5 mil millones de imágenes, ya de por sí reconoce a varias personas famosas sin necesidad de entrenarlas). En este caso, se trata de una actriz coreana llamada Jung Hoyeon, y el token que le otorgamos, como se puede ver, respondía bajo el nombre de "sqgkhoju". Lo ideal, es que la etiqueta no sea nombrada bajo una palabra que el modelo pueda reconocer, es decir, que el token no tenga significado. Si por un casual llamásemos al token "mujer", seguramente la IA no logre ni identificar ni asociar a la persona que hemos entrenado, y probablemente acabe generando la imagen de una mujer que no existe, que es justamente lo que queremos evitar.

Una vez que las imágenes cumplan con todos los requisitos, debemos utilizar el código abierto en la plataforma Google Colab para realizar el entrenamiento. El proceso que sigue el cuaderno es muy sencillo, en él sólo hace falta seguir una serie de pasos para completarlo. El primero es conectar el cuaderno a una cuenta de Google, para que se pueda guardar en el Drive asociado a esa cuenta una carpeta llamada "Fast-Dreambooth", en la que se guardarán todos los archivos que se generen durante

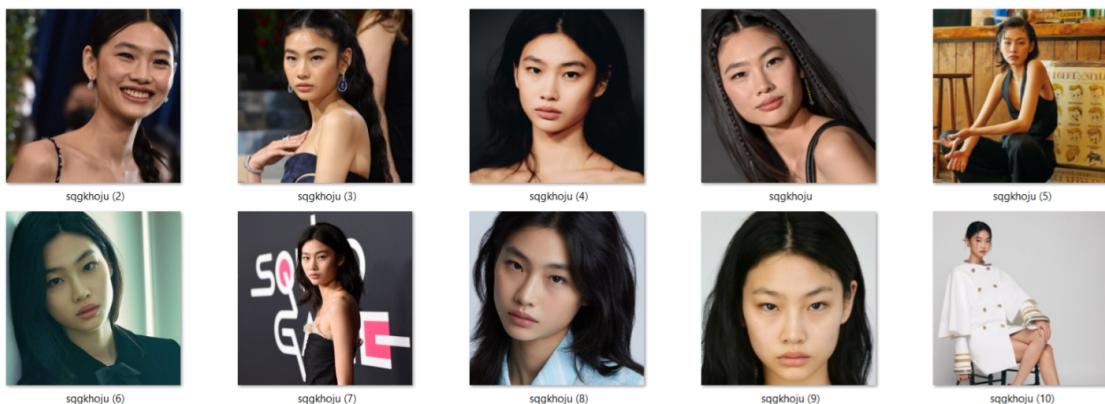


Figura 3.1: Dataset seleccionado para el entrenamiento de personas con Lora

el proceso. El siguiente paso es instalar las dependencias necesarias para ejecutar el código en python, seguido de establecer un nombre a la sesión en la que estamos trabajando, para que en un futuro cuando se hagan entrenamientos diferentes, se puedan distinguir unos de otros. En este paso se crea una carpeta llamada Sessions dentro de la carpeta mencionada recientemente. A su vez, esta carpeta contendrá otra bajo el nombre de la sesión que hayamos especificado en el cuaderno, y es ahí donde se guardarán todos los archivos que se creen en la ejecución.

A continuación, es turno de subir las imágenes previa y cuidadosamente seleccionadas. Ya sea seleccionándolas directamente desde la carpeta en la que las tengamos guardadas en local, o habiéndolas subido previamente a una carpeta de la cuenta de Google Drive, y proporcionar la ruta en la que están en la celda del cuaderno habilitada para ello.

El último paso, y uno de los más importantes, consiste en establecer algunos parámetros con los que se va a entrenar al modelo. El más importante y el único que nosotros hemos modificado, entre todos los que hay, es el número de steps. De modo que, en cuanto mayor sean, más tiempo tardará en generarse el archivo. Normalmente, se tardaba unos 20 o 25 minutos en terminar de entrenarse, lo cual hemos considerado que es bastante rápido. En el ejemplo de la figura 3.2 se puede ver la ejecución del progreso de entrenamiento en el que para una cantidad de 2000 pasos lleva 21 minutos y 36 segundos.

Tras la finalización, se creará un archivo de alrededor de 2 giga bytes, que contendrá el modelo de Stable Diffusion 1.5, con una capa de entrenamiento más, puesto que incorporará el elemento deseado. Con esto ya tendríamos un elemento de inteligencia artificial personalizado.

Este archivo, en formato ckpt, se podrá utilizar en la aplicación SKMD GUI más adelante para generar imágenes, y contendrá el elemento entrenado bajo el token seleccionado. Si posteriormente se pretende incluir elementos al modelo ya entrenado, también se puede realizar empezando de nuevo el proceso de entrenamiento y utilizando de base el archivo en extensión ckpt anterior. Cuando se realice este

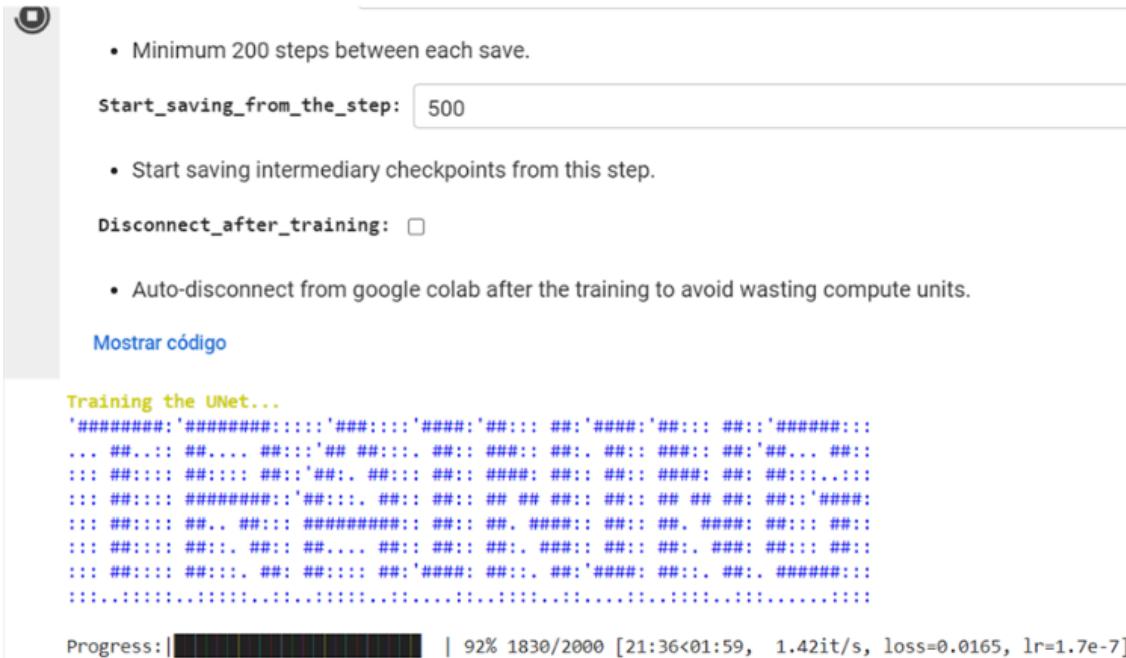


Figura 3.2: Procedimiento del entrenamiento mediante Dreambooth

Intento	Número de imágenes	Número de pasos	Veredicto
1	20	4000	
2	20	3000	
3	15	3000	
4	15	2000	
5	10	3000	
6	10	2400	

Tabla 3.1: Tabla de resultados obtenidos de entrenamiento

segundo entrenamiento, se podrán generar imágenes acerca de ambos elementos, lo cuál es muy útil para nuestros objetivos, ya que en un mismo modelo enfocado a un paciente, debe haber múltiples elementos. Sin embargo, más adelante veremos que este último aspecto ha supuesto uno de los grandes fallos que experimenta el modelo en cuanto a múltiples capas de entrenamiento.

Un aspecto muy importante a tener en cuenta, es que tras la realización de múltiples pruebas, los resultados óptimos que hemos obtenido ha sido seleccionando un conjunto de datos formado por 10 imágenes, y con 2400 pasos de entrenamiento. En la tabla 3.1

Hemos querido poner a prueba, no sólo las capacidades del modelo de entrenar a personas, ya que hemos podido comprobar sus puntos fuertes y débiles en la generación de seres humanos (los cuales podremos ver más adelante), sino de elementos que consideramos que también son de suma importancia a la hora de representar

recuerdos: animales y lugares.

El objetivo de entrenar al modelo con lugares es que estos no se encuentren en la base de datos de imágenes de Stable Diffusion, puesto que de ser así, no tendría sentido realizar el entrenamiento. De este modo, podremos lograr que el paciente pueda rememorar sitios emblemáticos para él o ella y que las imágenes generadas que emulan recuerdos consigan ser aún más personales. Por ejemplo, el parque de su vecindario, la casa de sus padres o incluso, su propio salón.

Las primeras pruebas que realizamos sobre un lugar se trataba de un edificio característico, que por supuesto no estaba incluido previamente. Hablamos de la basílica de Colmenar Viejo, Madrid. Las imágenes seleccionadas estaban hechas desde diferentes ángulos, alturas, luces y lejanías. En la figura 3.3 se pueden apreciar las características que reúnen las fotografías en cuestión y el token otorgado.

Además, quisimos comprobar si el número de imágenes y steps establecidos para las personas, producía resultados igual de satisfactorios para lugares. Y efectivamente, corroboramos la hipótesis de manera airosa. Ejemplos de ello, lo podemos ver en el apartado siguiente.



Figura 3.3: Dataset seleccionado para el entrenamiento con lugares

Para el entrenamiento de animales, seleccionamos 10 fotografías de un perro de la raza Shiba Inu en diferentes perspectivas, escenarios y mostrando distintas emociones para comprobar si la inteligencia artificial permitía entrenar con animales. Sin embargo, a la luz de los resultados vistos en personas y lugares quisimos comprobar la técnica de LORA que presenta Stable Diffusion. Las siglas LORA hacen referencia a *Low-Rank Adaptation of Large Language Models*, del inglés. Esta técnica favorece un equilibrio entre el tamaño del archivo y la eficiencia del propio entrenamiento y como vimos en el Capítulo 2 del Estado de la Cuestión, ha presentado imágenes de gran calidad en un tiempo excepcional.

El hecho de cambiar a LORA no supuso grandes cambios a la hora de entrenar el elemento deseado debido a que la selección de imágenes es exactamente igual. Donde realmente cambia el entrenamiento es en el cuaderno utilizado en Google Colab, este es diferente y por lo tanto, la forma de ejecución y los pasos a seguir también lo son.

Es de vital importancia saber cuáles son los parámetros que se deben ajustar para poder desarrollar el modelo de manera correcta y entender cada uno de ellos, dado que pueden resultar un poco más complejos en comparación con la técnica del cuaderno de Fast-Dreambooth.

Se debe indicar el nombre del proyecto. Es muy importante, porque aquí es donde obtendremos el resultado final, y es donde se debe incluir el dataset que mencionamos en el párrafo anterior. A continuación, el modelo a entrenar, como se ha explicado, hemos elegido el Stable Diffusion 1.5, por lo que es el que sirve de base para todos los entrenamientos escogidos. No obstante, esta técnica de entrenamiento tiene la peculiar característica de que se puede utilizar de base cualquier checkpoint desarrollado previamente, por lo que en caso de realizar un entrenamiento sobre personas, existe la posibilidad de elegir un modelo de base especializado en retratos. Esto garantiza que, seleccionando unas fotografías adecuadas y ajustando de manera correcta cada parámetro, los resultados sean bastante buenos.

Animales:

Para realizar una prueba del modelo de entrenamiento con Lora, lo primero fue elegir el animal y crear nuestro propio conjunto de datos, de manera que había que escoger 10 imágenes que tuviesen un tamaño igual o superior a 512 x 512 píxeles, y que la relación entre el ancho y el alto fuese 1 a 1. Además, al elegir el animal, se hizo la comprobación de que Stable Diffusion no lo incorporara ya en el modelo.



Figura 3.4: Dataset seleccionado para el entrenamiento con animales

Como se puede apreciar en la figura anterior, se hizo hincapié en la diversidad de las imágenes, para aportar un mayor valor al entrenamiento, de modo que se tuviese una visión completa del elemento a entrenar.

Ahora podemos decir que hemos logrado entrenar un modelo de generación de

imágenes incluyendo fotografías propias, y eso es algo que puede ser realmente útil para nuestros siguientes propósitos. Esto es porque podemos lograr que cualquier persona pueda incorporar las imágenes que considere oportunas para servir de apoyo al paciente. Lo cual consideramos un éxito en el desarrollo de nuestro trabajo.

3.3. Generación de imágenes con la interfaz de Stable Diffusion

3.3.1. Requisitos de instalación

La instalación se realizó desde el siguiente link <https://nmkd.itch.io/t2i-gui>
OS: Windows 10/11 64-bit

Minimum: GPU: Nvidia GPU with 4 GB VRAM, Maxwell Architecture (2014) or newer

Alternatively, with limited feature support: Any DirectML-capable GPU with 8 GB of VRAM RAM: 8 GB RAM (Note: Pagefile must be enabled as swapping will occur with only 8 GB!)

Disk: 10 GB (another free 5 GB for temporary files recommended)

Recommended: GPU: Nvidia GPU with 8 GB VRAM, Pascal Architecture (2016) or newer RAM: 16 GB RAM Disk: 12 GB on SSD (another free 5 GB for temporary files recommended)

****explicar extensamente los requisitos que se deben cumplir para instalar la app (GPU, cuda, memoria y etc)**

3.3.2. Funcionamiento y detalles de la interfaz

para que sirven los botones y ajustes etc

****explicar como funciona para integrar diferentes modelos, loras, establecer los parametros a la hora de generar la imagen**

3.3.3. Resultados de generación de imágenes

Para la ejecución, debemos seleccionar un modelo , introducir una descripción y ajustar el número de pasos, que determinarán la duración de la generación y la calidad de la imagen, en función del modelo escogido y las limitaciones que presenta la tarjeta gráfica, que como hemos explicado es fundamental para esta tarea. Uno de los puntos a favor de esta interfaz, y que resulta de gran interés para nuestro

proyecto, es que permite la utilización de modelos personalizados.

Para la personalización del modelo de Stable Diffusion hemos probado distintas vías y se ha llegado a distintos resultados y conclusiones:

Una vez entrenado el modelo, obtuvimos un total de 10 archivos en formato safetensor. Esto es debido a que, como hemos explicado con anterioridad, Lora tiene la ventaja de devolver múltiples resultados con un número de pasos distinto, con el objetivo de tener la posibilidad de probar cuáles son los steps ideales, y con cuáles habría falta o exceso de entrenamiento. En este caso, el safetensor adecuado fue el primero, puesto que a partir del segundo se podía apreciar la deformidad típica del sobreentrenamiento.



Figura 3.5: Mismo prompt con 1000, 2000 y 10.000 pasos de entrenamiento

En el momento en el que se descartan los demás archivos Safetensor, la siguiente labor es probar cuál es el número de pasos ideal para la generación de imágenes, y descubrir si con diferentes prompts, el modelo entrenado podría obtener buenos resultados.

**explicar mejor

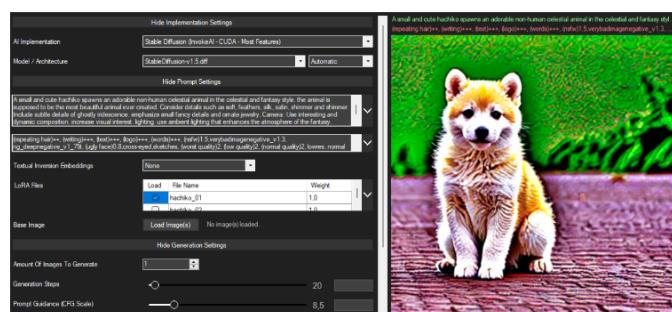


Figura 3.6: Imagen generada con el modelo Stable Diffusion 1.5, Lora Hachiko

Con diferentes estilos:

Añadiendo otros elementos:

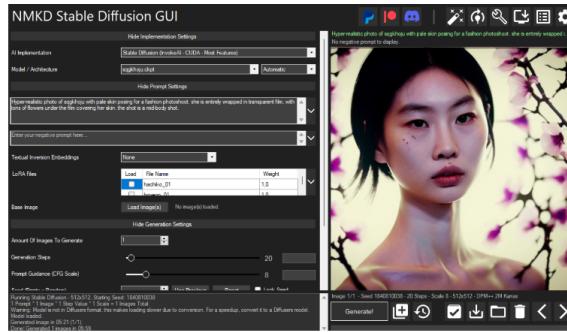


Figura 3.7: Resultados de entrenamiento de una persona con Dreambooth



Figura 3.8: Resultados de entrenamiento de una persona con diferentes estilos



Figura 3.9: Resultados de imágenes del lugar entrenado al estilo Van Gogh y viñeta



Figura 3.10: Resultados de imágenes del lugar en invierno y con una persona

Peligros con el sobreentrenamiento: Se produce cuando el modelo no se puede generalizar y se ajusta demasiado al conjunto de datos entrenados. Se debe principalmente a que el tamaño de los datos es demasiado pequeño y no contiene suficientes

muestras de datos para poder representar con precisión la totalidad de datos de entrada posibles. Otra razón es que el modelo se entrena durante demasiado tiempo en un solo conjunto de datos de muestra. En nuestro modelo, encontramos sobreentrenamiento cuando elegimos un número de steps muy elevado para un número de fotografías que no es lo suficientemente alto. Una manera de detectar que nuestro modelo está sobreentrenado es cuando no genera bien la cara de la persona, y se aprecian fallos en determinadas facciones, como en los ojos y la boca, en los cuales se aprecia deformidad.

Falta de pasos en el entrenamiento: En este caso, se produce cuando el modelo de datos no tiene la capacidad de capturar de forma precisa la relación entre las variables de entrada y de salida, de manera que existe un elevado índice de errores en el conjunto de datos de entrenamiento y en los datos no vistos. Se debe a que el modelo es demasiado simple, porque el tamaño de los datos es demasiado pequeño, o bien porque se necesita más tiempo de entrenamiento. En este caso, cuando generamos las imágenes, se evidencia que el modelo aún no ha aprendido lo suficiente acerca del elemento o token del que se ha realizado el entrenamiento, pues el resultado de la generación refleja una persona que no muestra ningún parecido con la realidad.

En cuanto a mezcla de personas: Para ampliar los horizontes de nuestro entrenamiento, hemos entrenado a una persona, asociando un token a ella para su identificación, sobre un modelo que previamente ya había sido entrenado con una persona y su token asociado, para comprobar si efectivamente se podía generar en una sola imagen una representación de las dos personas entrenadas, una junto a la otra. Es aquí donde se ha detectado una peculiaridad, ya que, a pesar de que un modelo genere buenos resultados de cada una de las dos personas, en el mismo momento que se solicita en un determinado prompt o descripción que se vean reflejados una o varias personas entrenadas en la misma fotografía, ningún modelo de los que se haya probado, ha generado buenos resultados de esta manera. Lo que finalmente se aprecia en la imagen generada, es que aparecen dos personas pero sus rostros son una mezcla de las características faciales y fisiológicas de ambas, produciéndose una deformidad en muchos de los casos y que los rostros aparezcan prácticamente duplicados.

**sñadir capítulo 4 para el funcionamiento de la aplicación

4.1. backend

4.2. frontend

Avances respecto al modelo.

Aplicación de Stable Diffusion genera imágenes aceptables en un tiempo normal (1-2 minutos como máximo). Existe la posibilidad de utilizar imágenes para poder



Figura 3.11: Duplicidad de elementos

incorporar al modelo de Stable Diffusion como Token. La aplicación contiene un apartado de training, pero tras probar incorporar escasas imágenes al modelo, el proceso es demasiado largo (varios días) y el resultado es una incógnita.

El modelo de .csv parece estar obsoleto, y al intentar el entrenamiento con las imágenes de personas sale un error en el paso 6, lo cual limita y habrá que identificar y trabajar con otro modelo de entrenamiento. LoRA puede ser una opción para ello. También hemos probado más modelos de Hugging Face a ejecutar por consola mediante Anaconda, pero los resultados no son los adecuados. Generan imágenes de calidad no muy buena en un tiempo muy elevado, entre 10 y 20 minutos. Por tanto, también decidimos descartar esta vía. La más razonable hasta ahora es la GUI de Stable Diffusion, ya que es la única que genera imágenes en un tiempo adecuado, y que permite la posibilidad de incorporar imágenes propias al modelo

Capítulo 4

Conclusiones y Trabajo Futuro

En primer lugar, podemos afirmar que, mediante la ayuda de la inteligencia artificial, se puede brindar un material fotográfico personalizado que sirva para colaborar con terapias ocupacionales, de manera que un paciente tenga la posibilidad de evocar momentos significativos, que de otra manera no tendría la capacidad de visualizar.

Con respecto a la generación de imágenes, se ha conseguido el objetivo de entrenar un modelo de inteligencia artificial de convertir texto a imagen, incluyendo a la persona que se considere. Este hecho significa que, en una terapia de reminiscencia, si se nos otorga un número considerable de imágenes en las que aparezca una persona concreta (a partir de diez), se puede trabajar en base a un modelo entrenado que reconozca a esa persona.

Adicionalmente, el modelo que se ha obtenido, puede servir de base para realizar un nuevo entrenamiento, lo cuál encaja a la perfección con nuestros objetivos, puesto que para un determinado paciente, podemos desarrollar un gran modelo que incluya el número de personas que se deseé.

Otro aspecto importante que debemos tener en cuenta y al que hemos llegado a la conclusión después de toda la investigación, es que para hacer funcionar nuestro modelo, al igual que cualquier otro de inteligencia artificial generativa de imágenes, se necesita una gran capacidad de memoria gráfica. Con la GPU de nuestro equipo, la generación siempre va a abarcar unos minutos, y cuanta más calidad se desee, más tiempo se incrementará. Esto es un hecho que todas las personas que se presten al servicio de generar las fotografías personalizadas deben conocer.

Respecto al entrenamiento de los modelos, podemos concluir que el número de pasos pasos con el que se realiza correctamente es 2400 para un entrenamiento de 10 imágenes, puesto que se han realizado múltiples pruebas con más pasos, en los

que se detecta sobreentrenamiento, y con menos pasos, en los que se detecta falta de entrenamiento.

En base a los resultados obtenidos, podemos afirmar que el entrenamiento que maximiza la calidad de las imágenes es el de Dreambooth.

En el caso de realizarse la aplicación, se incluyen más conclusiones

Propuestas de mejora

Tras la realización de este proyecto, nos gustaría que personas encargadas de asistir a pacientes con pérdida de memoria, pudiesen realizar una investigación partiendo de nuestro modelo, con el objetivo de estudiar psicológicamente en qué medida se está favoreciendo la reducción del estrés, se está potenciando la capacidad cognitiva del paciente y se está logrando la satisfacción de las personas. Este es el mayor objetivo que tenemos con esta inteligencia artificial, lograr el bienestar personal, y el alcance de este trabajo no nos permite comprobar si realmente hemos obtenido grandes resultados en el aspecto social. Adicionalmente, sería conveniente que si este proyecto fuese utilizado por los terapeutas, deberían tener un equipo con una gran tarjeta gráfica, para que se invirtiera así el menor tiempo posible en la generación de las imágenes. Tras haber utilizado este mismo modelo en la nube, donde utilizan servidores con GPUs de gran potencia, la obtención de resultados era instantánea, por lo que destacamos un gran margen de mejora que en caso de reducirse, agilizará en gran medida la terapia. Además, cabe resaltar que en nuestro modelo, hemos empleado la versión 1.5 de Stable Diffusion porque es la única que podía funcionar en local con la tarjeta gráfica de nuestro equipo. Con una mejora en este aspecto, se podría trabajar y entrenar la versión XL, que aporta una gran calidad a las imágenes, y se lograrían unos resultados aún mejores minimizando el tiempo de espera.

Introduction

Introduction to the subject area. This chapter contains the translation of Chapter 1.

Conclusions and Future Work

Conclusions and future lines of work. This chapter contains the translation of Chapter 4.

Contribuciones Personales

Estudiante 1: Sergio Llorente Hernando

Lo primero de todo fue reflexionar sobre los objetivos del proyecto, de manera que realicé una investigación sobre la terapia ocupacional, los libros de vida y las distintas inteligencias artificiales, con el objetivo de saber qué tecnología debemos utilizar para satisfacer las necesidades de los usuarios. Desde un primer momento hice una labor de búsqueda y prueba de diferentes modelos de inteligencia artificial generativa de imágenes. En primer lugar, la prueba era con servidores especializados y más adelante, hice una descarga de los modelos que generaban imágenes de calidad, con el objetivo de comprobar cuáles de los modelos podrían funcionar mejor y otorgar unos resultados satisfactorios.

El siguiente paso en el que me centré fue en conseguir hacer funcionar la inteligencia artificial generativa en local. Para ello, en primer lugar opté por probar los modelos en anaconda, lo cual no era una interfaz cómoda para el usuario y la generación de imágenes era lenta. A continuación empecé a trabajar con SDGUI, una aplicación con la que podía testar cualquier tipo de modelo, ya sea entrenado o no, y además de eso, ajustar todos los parámetros necesarios. Con esta aplicación sí que se generaban bien las imágenes. Esto sirvió para diseñar nuestro propio programa que incluyera el modelo generativo.

Para crear la aplicación, primero opté por crear un script de Python. Esto dio buenos resultados, porque mostraba la imagen al igual que en el programa de referencia, pero la interfaz no era buena para el usuario, y además haciendo pruebas en cualquier otro ordenador no obtuve resultados. Este hecho hizo que creara un entorno virtual para saber las dependencias que son necesarias para que funcione el modelo, y por otro lado, dividir el back y el front, con el objetivo de hacer más atractiva la aplicación mediante html, que es un lenguaje con el que ya estaba más familiarizado. Esto dio un mejor resultado, a pesar de no conseguir que el modelo de inteligencia artificial fuese uno entrenado por nosotros mismos.

La siguiente fase del trabajo, fue la del entrenamiento de los modelos, con el objetivo de tener la capacidad de personalizar la inteligencia artificial a petición de cualquier usuario. Fue un trabajo muy complejo en el que hice muchísimas pruebas con diferentes vías, con el objetivo de obtener los mejores resultados posibles. Fueron múltiples pruebas debido a que, para ofrecer al usuario la experiencia óptima, necesitamos conocer cuántas imágenes y cuántos pasos de aprendizaje se requieren. Tras la realización de estas pruebas se concluyó que la estrategia más óptima era utilizar un conjunto de 10 imágenes y 2400 pasos de entrenamiento. No obstante, hicieron falta múltiples entrenamientos erróneos o parcialmente correctos para conocer este hecho.

La siguiente fase fue realizar más entrenamientos con diferentes elementos y analizar los resultados. Además, aquí opté por examinar si un resultado de un entrenamiento podía servir de base para un próximo. La respuesta fue afirmativa y eso me llevó a añadir múltiples capas a un mismo modelo, lo cual podría ayudar a un usuario a tener un solo modelo personalizado que incluya todos los elementos que desee.

Estudiante 2: Isabella Romano Ramos

Lo primero que realicé fue la introducción del tema de estudio sobre el trabajo después de la primera sesión de reunión con el tutor en la que quedaron claros los aspectos generales que se iban a abordar, los objetivos principales y las tecnologías que se iban a utilizar.

Una vez definido esto, empezó la fase de investigación sobre los temas principales que abordaban el proyecto: el Alzheimer y la Inteligencia Artificial. Me embarqué en un proceso de recopilación de información para entender lo que ocurre en el cerebro que hace que se deteriore la memoria y que cause esta enfermedad que afecta a millones de personas, y así poder reflejarlo en la memoria. En cuanto a la IA, era un concepto totalmente nuevo para mí ya que no había visto nada sobre este campo en ninguna asignatura de la carrera. Por lo que opté empezar a entenderla por el principio, desde un concepto tan simple como el perceptrón, hasta examinar las redes neuronales, pasando por todos los tipos de modelos fundamentales que existen y saber diferenciar la función de cada uno. Por ende, me encargué de buscar la bibliografía que hablaba de la IA, más en concreto sobre el Deep Learning y las redes neuronales. De esta manera, lo pude redactar de manera que se entendiera lo mejor posible en el Capítulo 2 Estado de la cuestión.

Como se ha podido ver a lo largo de todo el proyecto, nos hemos centrado en el campo concreto de la Inteligencia Artificial generativa de imágenes y lo que hice fue investigar las diferentes alternativas de modelos ya implementados que podríamos utilizar para el entrenamiento de personas y la integración en nuestra futura aplicación. Tras examinar profundamente las prestaciones y disponibilidad de cada

uno, entre los tres candidatos finales, que se trataban de Midjourney, Dall-e y Stable Diffusion, se eligió usar el último por las limitaciones que nos mostraban los otros dos modelos. Una vez elegido, hice un análisis profundo de cómo está implementado el modelo de difusión estable que utiliza la herramienta y lo redacté extensamente en el Capítulo 2 Estado de la cuestión. Además, durante el proceso surgieron conceptos interesantes en relación con Stable Diffusion, como LORA y Dreambooth e incluso, consideré importante añadir la importancia de la ingeniería del prompt para la creación de imágenes a partir de texto.

Cuando pude comprender los conceptos teóricos que explicaban el funcionamiento de Stable Diffusion, se pasó a la parte práctica, es decir, a pensar en la estructura y usos de la aplicación que se iba a desarrollar como apoyo al modelo de IA generativa seleccionado. Finalmente, caímos en la cuenta de que las prestaciones que ofrecen mis dos equipos, tanto mi portátil como el ordenador de sobremesa, no cumplen los requisitos mínimos para poder ejecutar la aplicación en local ni realizar las pruebas necesarias sobre la generación de imágenes, al no contar con GPU en el portátil y no ser lo suficientemente potente la GPU en el ordenador de sobremesa. Es decir, era inviable que se pudieran generar imágenes en mis equipos.

Por ello, solo me podía limitar a entrenar el modelo que seleccionamos a través de la herramienta de Google Colab. Fue entonces cuando busqué en Hugging Face el modelo de Stable Diffusion más adecuado, que resultó ser el 1.5, y fui probando el número de pasos y de imágenes que producían los mejores resultados. Además, me encargué de organizar y estructurar el contenido de la memoria en la herramienta de Latex, utilizada para generar este mismo documento.

Para la aplicación, al no poder hacer pruebas en la generación de imágenes, me encargué de buscar e investigar sobre las mejores opciones para realizar el front en HTML. En el camino, pude dar con un software llamado Bootstrap que es muy eficiente y útil para facilitar el proceso del diseño del front, de forma fácil y sencilla.

Bibliografía

*Y así, del mucho leer y del poco dormir, se
le secó el celebro de manera que vino a
perder el juicio.
(modificar en Cascaras\bibliografia.tex)*

Miguel de Cervantes Saavedra

BAUTISTA, T., OETIKER, T., PARTL, H., HYNA, I. y SCHLEGL, E. *Una Descripción de LATEX 2_E*. Versión electrónica, 1998.

KNUTH, D. E. *The TeX book*. Addison-Wesley Professional., 1986.

KRISHNAN, E., editor. *LATEX Tutorials. A primer*. Indian TeX Users Group, 2003.

LAMPORT, L. *LATEX: A Document Preparation System, 2nd Edition*. Addison-Wesley Professional, 1994.

MITTELBACH, F., GOOSSENS, M., BRAAMS, J., CARLISLE, D. y ROWLEY, C. *The LATEX Companion*. Addison-Wesley Professional, segunda edición, 2004.

OETIKER, T., PARTL, H., HYNA, I. y SCHLEGL, E. *The Not So Short Introduction to LATEX 2_E*. Versión electrónica, 1996.

Este texto se puede encontrar en el fichero Cascaras/fin.tex. Si deseas eliminarlo, basta con comentar la línea correspondiente al final del fichero TFGTeXiS.tex.

*-¿Qué te parece desto, Sancho? – Dijo Don Quijote –
Bien podrán los encantadores quitarme la ventura,
pero el esfuerzo y el ánimo, será imposible.*

*Segunda parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes*

*-Buena está – dijo Sancho –; fírmela vuestra merced.
–No es menester firmarla – dijo Don Quijote–,
sino solamente poner mi rúbrica.*

*Primera parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes*

