Generación de Imágenes mediante IA para Arca Continental

Alexia Elizabeth Naredo Betancourt - A00830440

Eduardo Martínez Martínez - A01023975

Fabián Trejo Díaz Barreiro - A01423983

Francia García Romero - A01769680

Miguel Ángel Bermea Rodríguez - A01411671

Samantha Daniela Guanipa Ugas - A01703936

Natural Language Processing

Monterrey, N.L.; a 24 de noviembre de 2023

Introducción

En el contexto contemporáneo del marketing empresarial, la evolución tecnológica ha desencadenado una revolución en la generación de contenido visual. Este informe se sumerge en el panorama de la generación de imágenes impulsada por la inteligencia artificial (IA), específicamente en el ámbito publicitario de Arca Continental. El propósito es examinar detalladamente cómo esta innovación tecnológica ha transformado la creación de anuncios publicitarios para esta empresa líder en la industria de bebidas.

La relación entre la inteligencia artificial y el proceso creativo ha permitido a Arca Continental trazar una nueva ruta en la forma en que concibe y proyecta su identidad visual en el mercado. Este reporte explorará minuciosamente las estrategias, algoritmos y tecnologías de punta empleadas en la generación de imágenes, desde la gestación conceptual hasta la materialización de anuncios publicitarios de gran impacto.

Se analizará con detenimiento cómo la implementación de algoritmos de IA ha redefinido la personalización en la publicidad, permitiendo una conexión más profunda con los consumidores y, por ende, generando un retorno de inversión significativo. Además, se examinarán las consideraciones éticas intrínsecas a esta revolucionaria práctica, incluyendo aspectos relacionados con la autoría, la transparencia y la equidad en la representación visual.

Investigación Preliminar de Modelos

Dall-E

¿Qué es?

“DALL-E es un modelo de inteligencia artificial desarrollado por OpenAI, la misma organización detrás de GPT-3.5, que tiene la capacidad de generar imágenes a partir de descripciones textuales.” (Emilio Romero, 2023)

¿Cómo funciona?

Según Emilio Romero, DALL-E utiliza una técnica de aprendizaje profundo conocida como Red Generativa Antagónica (GAN) que son sistemas de IA que constan de dos partes principales: un generador y un discriminador. El generador toma una descripción textual y la convierte en una imagen utilizando una red neuronal profunda que aprende a mapear las palabras de la descripción en una representación visual. Además de eso, esta red neuronal utiliza una técnica llamada codificación por atención, que le permite prestar atención a partes específicas de la descripción para generar detalles visuales más exactos y precisos. Por otro lado, el discriminador tiene la tarea de evaluar si las imágenes generadas por el generador son reales o falsas. A medida que el generador mejora, el discriminador se vuelve más exigente y preciso en la detección de imágenes generadas artificialmente.

Aplicaciones

Entre sus principales aplicaciones se encuentran el arte y diseño como guía o inspiración para algunos artistas o diseñadores. Además de la publicidad y el marketing ya que puede generar imágenes llamativas y atractivas para productos y servicios. Como complemento, se utiliza también en el ámbito de la investigación científica para visualizar conceptos abstractos. También se utiliza para la generación de contenido en medios digitales creando portadas llamativas para los sitios web, redes sociales y más.

Dalle-2

¿Cómo funciona?

DALL-E 2 representa la evolución de una arquitectura de vanguardia en el campo de la inteligencia artificial, desarrollada por OpenAI. Esta segunda iteración se basa en los principios fundamentales de la arquitectura GPT-3, un modelo de lenguaje

generativo preentrenado. Sin embargo, su distinción principal radica en su capacidad para comprender y generar imágenes a partir de descripciones de texto, lo que amplía significativamente su alcance y aplicación en comparación con su predecesor.

En términos de su arquitectura, DALL-E 2 construye sobre la base de GPT-3, pero se enfoca en la generación de imágenes. Esta extensión implica un aumento en la capacidad de producir imágenes de alta resolución con un nivel de detalle sorprendente. A diferencia de la versión anterior, DALL-E 2 ofrece versatilidad en la generación de imágenes, permitiendo la creación de contenido visual en diversas dimensiones. Además, se ha trabajado intensamente en mejorar la coherencia visual de las imágenes generadas, lo que se traduce en una mayor realismo y consistencia en la representación visual.

Uno de los aspectos destacados de DALL-E 2 es su capacidad para comprender y generar una amplia gama de conceptos y objetos, lo que se debe en parte a su entrenamiento con un conjunto de datos más extenso y diversificado. Esto amplía significativamente su utilidad en aplicaciones que van desde el diseño gráfico hasta la creación de arte, publicidad, diseño de productos y generación de contenido visual para diversas plataformas.

Aplicaciones

En cuanto a aplicaciones notables, DALL-E 2 se ha convertido en una herramienta valiosa en campos que requieren generación de contenido visual, como el diseño gráfico y la creación artística. También se utiliza en la generación de ilustraciones para publicaciones impresas y digitales, y en la producción de contenido visual para marketing y publicidad. Además, su capacidad para generar imágenes coherentes y detalladas lo hace relevante en la industria del entretenimiento, particularmente en la creación de prototipos de diseño de productos y en la generación de gráficos y elementos visuales para videojuegos.

Vertex AI

¿Qué es?

Es una plataforma de aprendizaje automático administrada por Google Cloud. Unifica todas las herramientas y servicios de Google Cloud relacionados con la inteligencia artificial y el aprendizaje automático. Cumple con el objetivo de facilitar la creación, implementación y mantenimiento de modelos de IA personalizados.

¿Cómo funciona?

Vertex AI combina flujos de trabajo de ingeniería de datos, ciencia de datos e ingeniería de aprendizaje automático. La plataforma incluye las herramientas AutoML y AI Platform en una interfaz de API unificada, una biblioteca cliente y una interfaz de usuario. AutoML se encarga de entrenar modelos en conjuntos de datos de imagen, video, texto, etc., sin necesidad de que el usuario escriba un código específico. Por otro lado, AI Platform ayuda a ejecutar un código de entrenamiento personalizado. Vertex AI permite la unión de ambas herramientas y sus funciones, así como elegir la opción de entrenamiento para guardar modelos, implementarlos e incluso solicitar predicciones.

Cabe mencionar que Imagen on Vertex AI [servicio Generative AI] genera costos cuando se usa y que ciertas funcionalidades de IA generativa de imágenes están disponibles en distintas etapas de lanzamiento; por ejemplo, la generación de imágenes, edición de imágenes, ajustes del modelo y entrenamiento tienen una disponibilidad general restringida por lo que tienes que unirte al programa de testers de Google Cloud y esperar ser aceptado en la waitlist.

Aplicaciones

Vertex AI ha sido utilizado para diversas aplicaciones, por ejemplo, Enterprise Search y Conversations en Vertex AI permite a las organizaciones crear aplicaciones de búsqueda y chat utilizando sus datos en pocos minutos. Vertex AI Conversation permite usar lenguaje natural para definir qué respuestas se desea que tenga el chatbot.

Además los clientes de esta plataforma tienen acceso a cientos de modelos base, incluyendo versiones de código abierto. En casos particulares se ofrecen modelos específicos para las industrias de ciberseguridad y empresas de ciencias biológicas y sanitarias como el modelo Med-PaLM 2 [chatbot a lo Bard o ChatGPT pero para fines médicos] en la reconocida Mayo Clinic en Rochester, Minnesota.

En cuanto a Imagen on Vertex AI [servicio Generative AI], la empresa ModiFace [pertenece a L'Oréal, empresa líder en el mercado de los cosméticos] utiliza esta plataforma para estar a la vanguardia en términos de IA en la industria de la belleza.

Midjourney

¿Qué es?

Midjourney es un programa y servicio de inteligencia artificial generativa. Midjourney genera imágenes a partir de descripciones en lenguaje natural (prompts). Está

alojado por el laboratorio de investigación independiente Midjourney, Inc., con sede en San Francisco.

¿Cómo funciona?

Midjourney se utiliza a través de la plataforma Discord, aunque existe una API que se puede utilizar con Python. Existe una lista de comandos que va desde crear imágenes hasta aumentar su calidad o juntar imágenes.

Aplicaciones

Su más notoria aplicación ese la de aumentar la calidad de imágenes. A diferencia de Dall-e o Stable Diffusion, Midjourney no es muy bueno al interpretar un prompt, pero es mejor al momento de reimaginar una imagen.

Stable Diffusion

¿Qué es?

Un motor de generación de imágenes creado por la universidad LMU Munich en colaboración con la startup Runway. Permite la generación texto-imagen, imagen-imagen e inpainting (restauración o rellenado de partes de la imagen). Es gratuito (con acceso a herramientas básicas, para mayor personalización existen planes de pago), de código abierto, y no incluye restricciones (a excepción de las éticas y legales) para la generación de imágenes, además su uso y distribución tanto personal como comercial está permitido bajo la licencia Creative ML OpenRAIL-M.

¿Cómo funciona?

Utiliza una variante del modelo de difusión llamada modelo de difusión latente (LDM), entrado para eliminar ruido gaussiano de las imágenes de entrenamiento, mediante añadir este ruido a las imágenes durante el entrenamiento de manera controlada para que el modelo aprenda a eliminarlo y mejorar la calidad de las imágenes. Stable Diffusion sigue un proceso de tres partes: un codificador variacional (VAE) que comprime la imagen a un espacio latente, la aplicación de ruido gaussiano iterativo a la representación latente, y un bloque U-Net que elimina el ruido de la representación, por último un decodificador VAE genera la imagen final.

Es accesible a través de una REST API que ofrece las herramientas anteriormente mencionadas, la capacidad de especificar la cantidad de imágenes a generar, configurar su alto y ancho, además soporta formatos PNG y JPG.

Aplicaciones

Algunas apps y/o sitios web que usan Stable Diffusion son: My Story Bot: Utilizado para generar libros de cuentos para niños de manera automática. Art Design: Da sugerencias de contenido atractivo y creativo para SEO. Iconik AI: Genera iconos profesionales que cumplen con las pautas de diseño de iOS, Android y sitios web. Senti NFT: Crea imágenes y las convierte en tokens no fungibles (NFT).

DeepAI Image Generator

¿Qué es?

Es una herramienta de generación de imágenes a partir de texto. El software es de código abierto y para poder utilizarlo solo se debe de crear una cuenta en la plataforma. Tiene distintas librerías disponibles que permiten obtener distintos acabados. Aunque algunos de los servicios y librerías relacionadas con la generación básica de imágenes se pueden utilizar de forma gratuita, servicios más especializados, dentro de los cuales se encuentra el uso del API, tiene un costo.

DeepAI API permite a desarrolladores tener a su alcance funciones como búsqueda semántica, traducción, parafraseo, generación de imágenes, colorización, identificación de objetos, etc... por lo cual es bastante útil para el uso personalizado que genere valor a una cierta actividad.

¿Cómo funciona?

Es un modelo de lenguaje basado en transformadores, que es un tipo de arquitectura de red neuronal que fue introducida en un artículo de investigación por Vaswani y otros en 2017. El modelo Transformer está compuesto por un codificador y un decodificador, cada uno de ellos formado por una pila de capas similares. El codificador procesa secuencias de entrada, como frases en lenguaje natural, y genera una codificación para cada token en la secuencia. Luego, el decodificador utiliza estas codificaciones para generar una secuencia de tokens de salida, que pueden ser palabras o frases correspondientes a la secuencia de entrada.

Aplicaciones

Una de las principales innovaciones del Transformer es el mecanismo de autoatención. Este mecanismo permite al modelo ponderar la importancia de diferentes tokens en la secuencia de entrada al generar la secuencia de salida. Específicamente, en cada capa del codificador o decodificador, el modelo calcula una matriz de atención que refleja la similitud entre cada par de tokens en la secuencia de entrada o salida. Luego, esta matriz de atención se aplica a las

codificaciones y decodificaciones para producir una suma ponderada que captura las partes más importantes de la secuencia de entrada o salida.

El Transformer también utiliza conexiones residuales y normalización de capa para ayudar a superar problemas con gradientes que desaparecen en redes más profundas. Estos mecanismos aseguran que los gradientes puedan fluir suavemente a través de la red durante el entrenamiento, incluso cuando se utilizan muchas capas. En general, el modelo de lenguaje basado en Transformer es una herramienta poderosa para el procesamiento de lenguaje natural. Su capacidad para capturar las relaciones entre palabras y frases, y para generar respuestas contextualmente informadas, lo ha convertido en un enfoque popular para una amplia gama de aplicaciones de inteligencia artificial, incluyendo chatbots, traducción automática y generación de texto.

Primera Implementación de Solución Propuesta

Esta implementación consta de dos partes: generación de texto y descripción del anuncio a crear y generación de la imagen como tal.

Generación de Texto

Para la generación automática de textos de promociones en nuestro proyecto, hemos integrado con éxito la API de GPT-4 de OpenAI. La configuración inicial implicó establecer una conexión segura y eficiente con la API, garantizando que nuestro sistema interactúe sin problemas con el avanzado modelo de lenguaje.

Con los datos de nuestro inventario, que incluyen detalles esenciales como nombres de productos, características y precios, GPT-4 se encarga de elaborar descripciones promocionales. Esta tarea se realiza mediante un proceso automatizado donde el modelo recibe la información del producto y, aprovechando su habilidad para entender y generar texto natural, crea promociones únicas y atractivas. Hemos prestado especial atención a que estas descripciones no solo sean informativas, sino que también capten la atención de los clientes potenciales, destacando las cualidades y beneficios de cada artículo.

Un aspecto crucial de este proceso ha sido la fase de optimización. A través de pruebas iterativas, hemos ajustado los parámetros de la generación de texto, como longitud, estilo y tono, para alinear las promociones con nuestras preferencias. Esta optimización es un proceso dinámico, que evoluciona constantemente en respuesta a las tendencias del mercado y los comentarios de los usuarios.

Generación de Imágenes

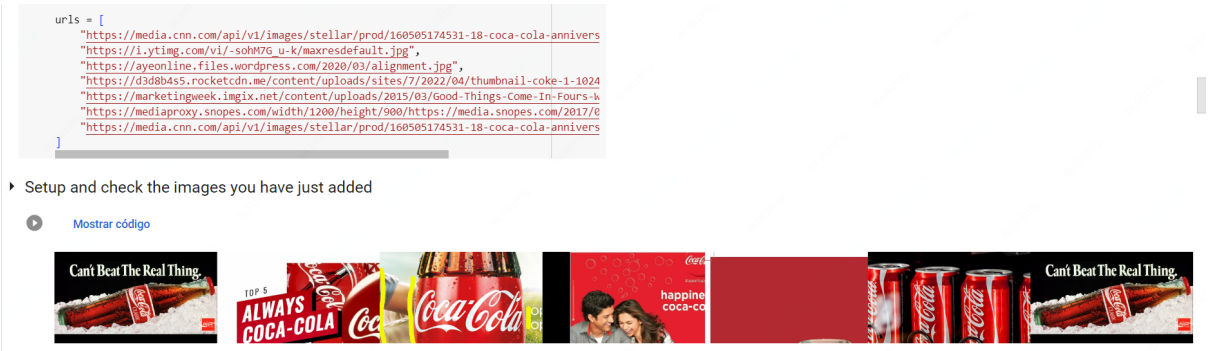
El modelo utilizado es de Stable Difussion (cuenta con licencia del tipo MIT*). En la primera sección del código se instalan e importan las librerías necesarias para el uso del modelo. Posteriormente, se encuentra la sección correspondiente a Fine-Tuning

Fine-Tuning

El fine tuning en lo relativo a la generación de imágenes por AI es un proceso de ajuste fino de un modelo pre-entrenado para adaptarlo a tareas específicas de generación de imágenes. Implica tomar un modelo de inteligencia artificial, como

una red neuronal convolucional (CNN) o una red generativa adversarial (GAN), que ya ha sido entrenado en una gran cantidad de datos, y luego afinarlo con datos adicionales o tareas de generación de imágenes específicas. El modelo que estamos utilizando ya ha sido entrenado con una amplia base de datos de imagenes. Sin embargo, no nos consta que ese dataset incluyera imagenes especificas al producto que buscamos anunciar. Por lo tanto, recurrimos al fine tuning para agregar nuevas imagenes asociadas con palabras clave para un ajuste de entrenamiento final del modelo.

Para poder realizar esto en la solución implementada, incluimos las urls a imagenes que nos interesa incluir. Despues, especificamos una etiqueta y descripción relacionada con estas imagenes, para que asi el modelo pueda relacionar un prompt dado con las imagenes. Una vez que se cargan las imagenes al modelo, en el código se definen las funciones de entrenamiento y se procede con el mismo.



Finalmente se incluye una sección interactiva en donde se puede especificar un prompt, y el modelo genera la imagen relacionada con el texto proporcionado como input.

- Esta licencia permite el uso comercial de las imágenes generadas por el modelo.

Segunda Implementación de Solución Propuesta

Entrega de Reto: Segundo MVP

El enfoque de esta segunda entrega busca mejorar la calidad del flyer final, a través de la generación de fondo y texto para luego agregar una imagen real del producto. El objetivo es garantizar que la representación del producto en el anuncio sea lo más fiel posible a la realidad.

Componentes

1. Text Generation:
    - Input: Prompt (texto descriptivo).
    - Output: Texto a incluir en el anuncio.
2. Background Image:
    - Input: Descripción de la imagen de fondo.
    - Output: Imagen de fondo seleccionada.
3. Choosing Product Image:
    - Input: Nombre del producto.
    - Output: Imagen del producto seleccionada aleatoriamente.
4. Ensemble of All Components:
    - Input:
        - Texto descriptivo.
        - Ubicación del texto en el anuncio.
        - Objeto (imagen del producto).
        - Ubicación y escala del objeto en el anuncio.
    - Output: Anuncio publicitario (flyer) generado.

Código

1. Configuración Inicial

1.1 Instalación de Bibliotecas

Se instalan las bibliotecas necesarias, incluyendo diffusers, accelerate, transformers, ftfy, gradio, y bitsandbytes. Estas bibliotecas proporcionan herramientas esenciales para el entrenamiento y la ejecución del modelo.

1.2 Importación de Módulos

Se importan los módulos a usar como torch, PIL, y otros desde diffusers, transformers, y bitsandbytes. Estos para la manipulación de datos, el procesamiento de imágenes y la configuración del modelo.

2. Configuración de Conceptos

2.1 Conceptos Iniciales

Se establecen los conceptos que el modelo aprenderá durante el proceso de entrenamiento. En este caso, se configura el prompt de instancia y se especifican las URL de las imágenes asociadas a este.

2.2 Descarga de Imágenes

Se descargan las imágenes asociadas con el nuevo concepto utilizando las URL proporcionadas. Estas imágenes se almacenan en una carpeta local para su posterior procesamiento y entrenamiento.

2.3 Configuración del Modelo

Se configuran parámetros esenciales para el entrenamiento, como el modelo preentrenado a utilizar, el prompt de instancia, y la clase previa.

3. Entrenamiento del Modelo

3.1 Configuración de Clases

Se establece una clase de DreamBoothDataset para gestionar el conjunto de datos de entrenamiento. Esto incluye la manipulación de imágenes de instancia y, opcionalmente, imágenes de clase para la preservación previa.

3.2 Entrenamiento

El modelo se entrena utilizando el conjunto de datos configurado. Durante el entrenamiento, se aplican técnicas específicas, como la adición de ruido a las representaciones latentes y la preservación de clase previa, para mejorar la fidelidad y generalización del modelo.

4. Ejecución del Modelo Entrenado

4.1 Configuración del Pipeline

Se configura un pipeline utilizando el modelo entrenado y un scheduler específico para la ejecución del modelo en Colab. Este pipeline se utiliza para generar anuncios de productos basados en el fondo y el texto proporcionados.

4.2 Generación de Anuncios

El modelo entrenado se utiliza para generar anuncios de productos. Se configuran parámetros como el prompt de fondo, el número de muestras y el diseño de la cuadrícula de imágenes resultante.

5. Selección de Imagen del Producto

5.1 Búsqueda de Imágenes en Google Drive

Se implementa una función que busca y selecciona aleatoriamente una imagen del producto desde una carpeta en guardada Google Drive.

5.2 Visualización de la Imagen Seleccionada

La imagen seleccionada se visualiza utilizando el visor de imágenes predeterminado en Colab.

-foto de la coca-

6. Ensamble de Todos los Componentes

6.1 Escala y Colocación de Imágenes

Se escalan las imágenes del producto según la configuración proporcionada y se colocan en el fondo generado. Esto se para crear una composición visual coherente al resto del flyer.

6.2 Adición de Texto

Se añade texto al fondo en una posición específica, proporcionando información adicional o promocional en el anuncio.

6.3 Generación del Flyer

El resultado final es un flyer que incluye el fondo, el texto y la imagen aleatoria del producto, guardado en formato JPEG.

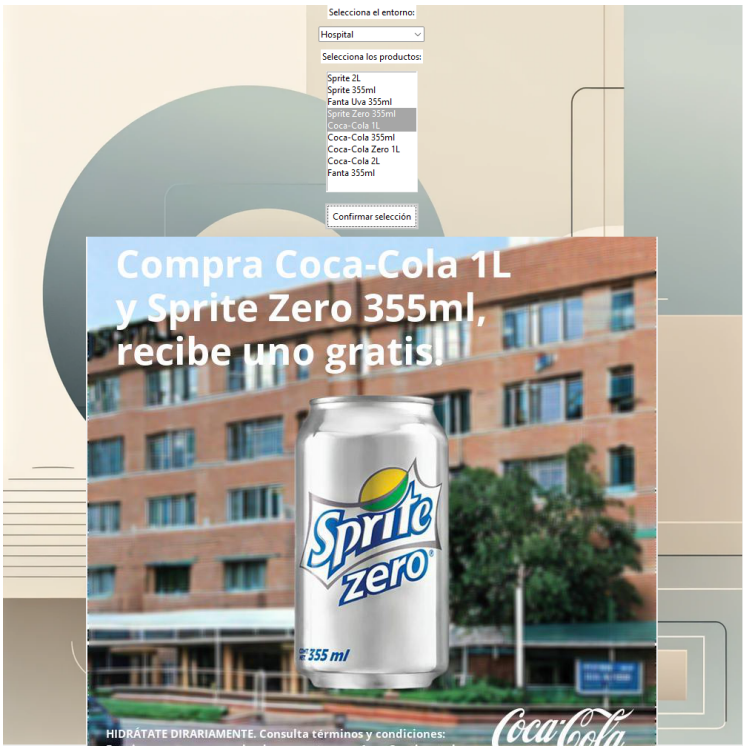
Resultados

Como primeros resultados se obtuvieron buenos colores pero una pobre descripción del producto como se puede observar en la siguiente imagen.

Progresivamente se fueron obteniendo mejores resultados, obteniendo un mejor despliegue del producto, pero careciendo de descripción.

Al decidir crear un collage con una imagen de fondo generada por inteligencia artificial, una imagen del producto obtenida en PNG, y un texto del producto creado con un modelo de Open AI, se comenzaron a obtener mejores resultados ya que se visualizaba como un comercial más que solo como una imagen generada por IA.

Por último, como resultado final, se creó una UI donde el usuario seleccionará el entorno que desea que se despliegue, además del producto seleccionado, y se imprimirá la imágen final con el producto seleccionado de un catálogo de imágenes, teniendo el resultado deseado desde el inicio del proyecto.



Conclusiones

En este proyecto, se aprendieron valiosas lecciones sobre la IA generativa y cómo puede aplicarse en el mundo real. Descubrimos que, aunque la inteligencia artificial ofrece soluciones avanzadas y sofisticadas, a veces la clave del éxito radica en la simplicidad y eficacia de combinar distintos outputs de manera creativa.

El desafío de fusionar las imágenes generadas por Stable Diffusion con los textos creados por GPT-4 fue uno de los aspectos más intrigantes del proyecto. Nos dimos cuenta de que, aunque cada tecnología es poderosa por sí sola, su verdadero potencial se desbloquea cuando se integran de manera eficiente. Aprendimos a valorar la importancia de la interoperabilidad entre diferentes sistemas de IA y cómo esto puede conducir a resultados innovadores.

Uno de los hallazgos más importantes fue que, a pesar de la complejidad inherente de la IA, soluciones relativamente simples, como combinar el output de ambas herramientas con código, pueden ser increíblemente efectivas. Esto nos enseñó que no siempre es necesario reinventar la rueda; a veces, la mejor solución es la que integra y maximiza lo que ya existe.

En términos de IA generativa, este proyecto nos ayudó a comprender mejor sus capacidades y limitaciones. Aprendimos que la IA puede generar contenido creativo y relevante, pero el toque humano sigue siendo crucial para guiar y perfeccionar el proceso. En conclusión, este proyecto no solo fue un ejercicio en el uso de tecnologías avanzadas sino también una lección en la importancia de la simplicidad, la integración y el equilibrio entre la innovación tecnológica y la intuición humana.

Referencias

Romero, E. (2023). Qué es DALL-E y cómo usar esta inteligencia artificial para crear imágenes. Recuperado de: https://www.inesem.es/revistadigital/diseno-y-artes-graficas/que-es-dall-e-y-como-usar-esta-inteligencia-artificial-para-crear-imagenes/

Geekflare. (2023). Google Cloud Vertex AI: Esto es lo que necesita saber. Recuperado el 9 Octubre 2023, desde https://geekflare.com/es/google-clouds-vertex-ai/

Google Cloud. (2023). Vertex AI | Google Cloud. Recuperado el 9 Octubre 2023, desde https://cloud.google.com/vertex-ai

AI Chat. (n.d.). DeepAI. Retrieved October 10, 2023, from https://deepai.org/chat

Stable diffusion and dreambooth API - generate and finetune dreambooth stable diffusion using API. (n.d.). Stable Diffusion And Dreambooth API - Generate and Finetune Dreambooth Stable Diffusion Using API. Retrieved October 10, 2023, from https://stablediffusionapi.com/

(N.d.). Stablediffusionweb.com. Retrieved October 10, 2023, from https://stablediffusionweb.com/

Offert, F., & Phan, T. (s/f). A Sign That Spells: DALL·E 2, Invisual Images and The Racial Politics of Feature Space. Retrieved from https://ar5iv.org/abs/2211.06323