

Inteligencia Artificial Aplicada para la Economía

Profesores

Profesor Magistral

Camilo Vega Barbosa

Asistente de Docencia

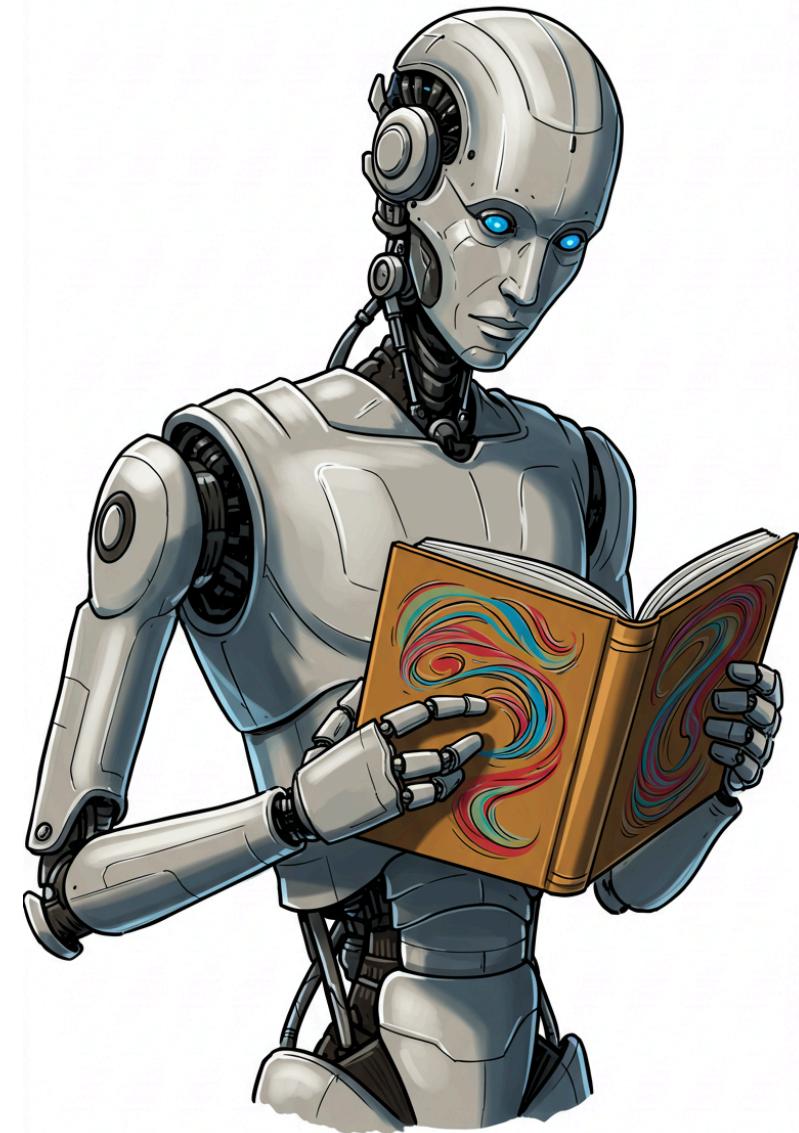
Sergio Julian Zona Moreno





Modelos de Visión, Imágenes y Video

Nuevos sentidos en las IA





La evolución de la IA en visión computacional

- 🔍 La inteligencia artificial ha revolucionado cómo las máquinas interpretan imágenes, pasando de simples clasificadores a sistemas que generan contenido visual fotorrealista a partir de texto.
- 💡 Tres arquitecturas clave impulsan este avance: las **Redes Neuronales Convolucionales (CNN)** para reconocimiento de patrones, los **modelos Diffusion** que generan imágenes mediante reducción de ruido, y los **Diffusion Transformers (DiT)** que combinan ambos enfoques para crear sistemas de generación visual de vanguardia.
- 🚀 Estas tecnologías han transformado sectores como medicina, seguridad y diseño, abriendo nuevas posibilidades creativas y difuminando la frontera entre contenido real y generado artificialmente.



CNN: Las redes que revolucionaron la visión artificial

- 🕒 Las Redes Neuronales Convolucionales nacieron inspiradas en el córtex visual biológico, donde Hubel y Wiesel descubrieron en 1959 que las neuronas visuales responden a regiones específicas del campo visual. Este descubrimiento inspiró a Yann LeCun quien, en 1989, desarrolló LeNet-5, la primera CNN moderna para reconocimiento de dígitos manuscritos.
- 🔍 El principio fundamental de las CNN es la convolución, una operación matemática que aplica filtros para detectar patrones locales como bordes, texturas y formas. Combinada con operaciones de pooling que reducen dimensionalidad y capas fully-connected para la clasificación final, las CNN lograron lo que parecía imposible: enseñar a las máquinas a "ver" de forma similar a los humanos.



AlexNet: El punto de inflexión en la visión computacional

En ImageNet 2012, AlexNet redujo el error de clasificación de 26% a 15%, marcando el inicio de la era moderna de visión computacional.

Fue desarrollado por Alex Krizhevsky, Ilya Sutskever (Que luego cofundaría OpenAI) y Geoffrey Hinton (**Nobel de Física en 2024**)

AlexNet introdujo innovaciones clave como:

- ReLU como función de activación
- Dropout para reducir sobreajuste
- Aumento de datos para mejorar generalización



Ilya Sutskever, co-autor de AlexNet y CoFundador de OpenAI

Ilya Sutskever dejó OpenAI en 2024 y fundó "Safe Superintelligence (SSI)", una nueva startup de inteligencia artificial enfocada en la seguridad.



Anatomía y flujo de una CNN

Componentes clave

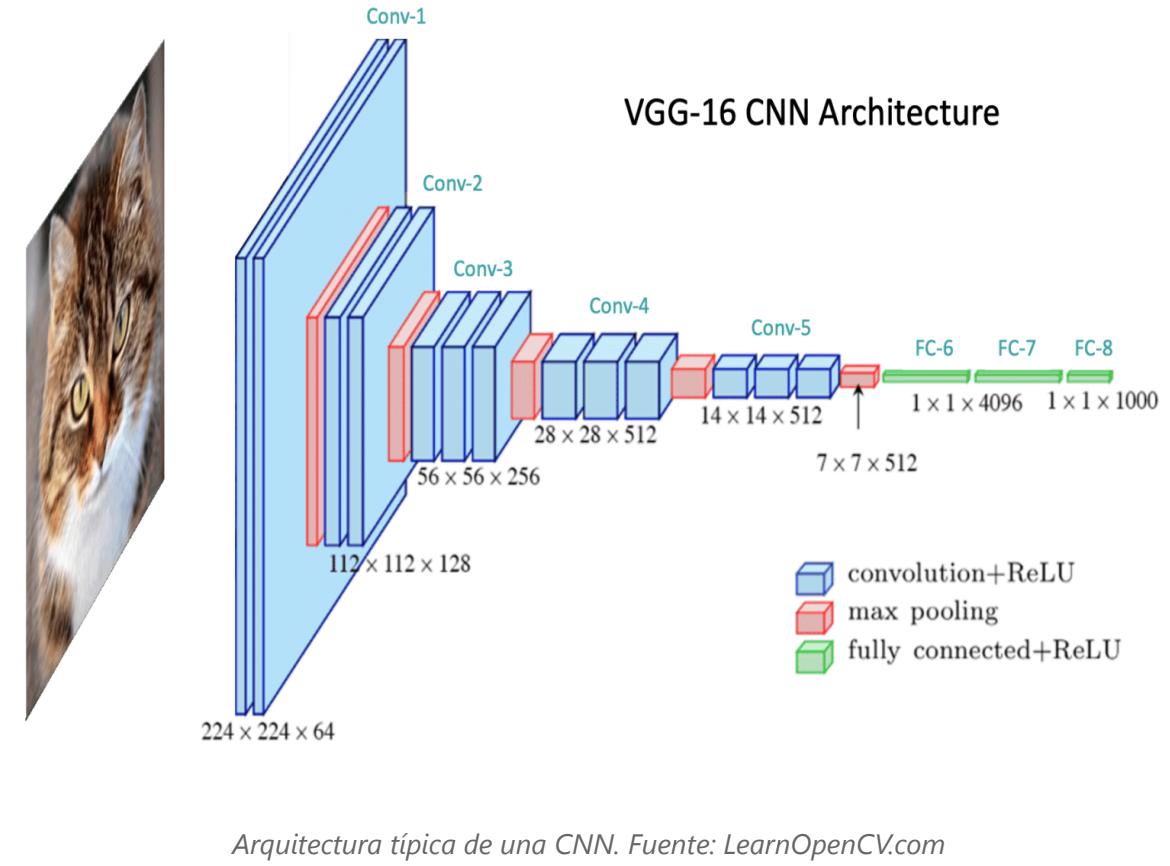
1. **Capa convolucional:** Extrae características mediante filtros deslizantes
2. **Activación (ReLU):** Introduce no-linealidad
3. **Pooling:** Reduce dimensiones preservando información relevante
4. **Capas fully-connected:** Combina todas las características para la decisión final



Anatomía y flujo de una CNN

Flujo de procesamiento

- La imagen de entrada atraviesa múltiples capas convolucionales
- Las primeras capas detectan características simples (bordes, texturas)
- Las capas profundas capturan patrones complejos (formas, objetos)
- La red aprende jerarquías de características cada vez más abstractas





Filtros Convolucionales: Los ojos de la CNN

Los filtros convolucionales son **matrices numéricas entrenables** que detectan patrones visuales específicos. Cada filtro se desliza sobre la imagen para generar un "mapa de características".

Imagen (parte)	Filtro (detector de borde vertical)	Resultado
$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$	\times	$\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$
	=	$\begin{vmatrix} 0 \\ 3 \\ 2 \end{vmatrix}$

Ejemplo: detección de borde vertical

Las primeras capas detectan elementos básicos (bordes, texturas), mientras que capas más profundas reconocen formas y objetos completos.

Funcionamiento de los filtros:

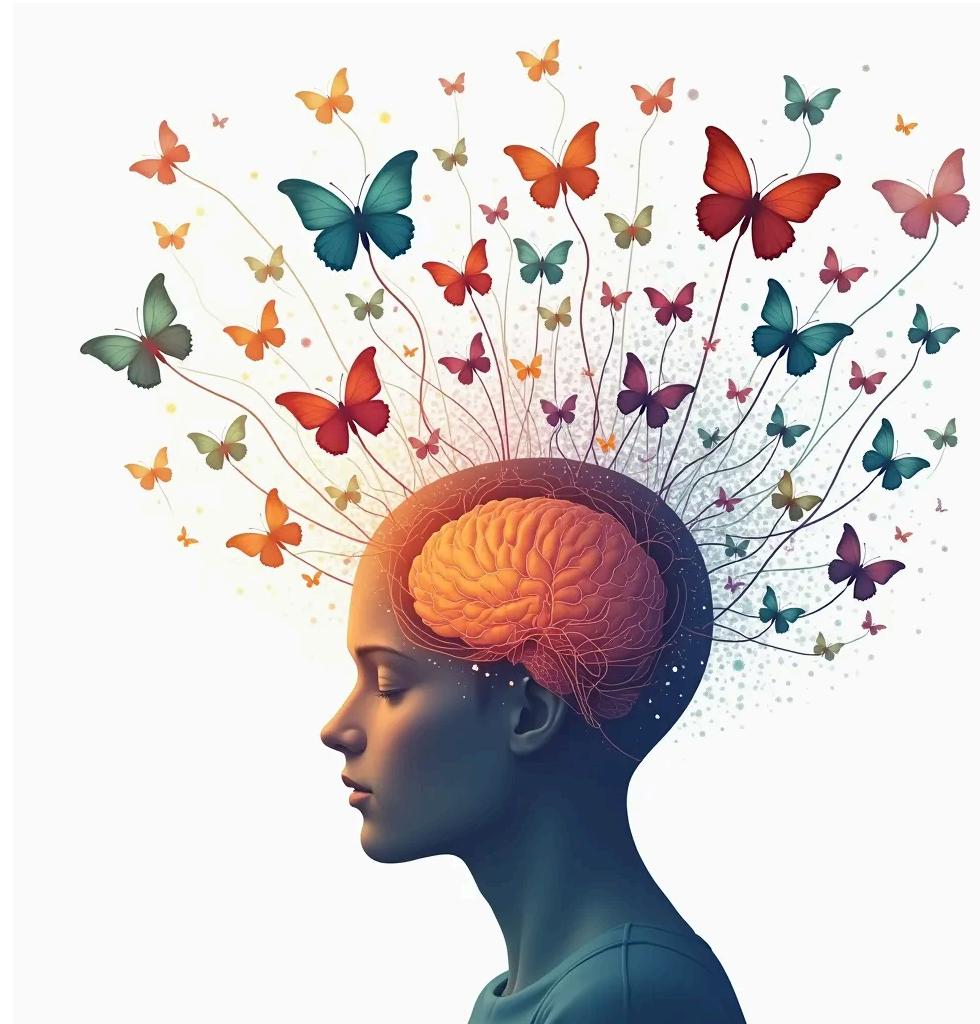
- Se posicionan sobre una región de la imagen
- Multiplican cada valor del filtro por el pixel correspondiente
- Suman todos los productos para obtener un único valor
- Se desplazan a la siguiente posición y repiten



Diffusion Models: El arte del ruido

Los modelos de difusión representan una revolución en la generación de imágenes, transformando ruido aleatorio en contenido visual detallado y realista mediante un proceso inspirado en la termodinámica.

El principio fundamental es sorprendentemente intuitivo: primero aprenden a destruir información gradualmente añadiendo ruido a las imágenes, para luego invertir este proceso y reconstruir imágenes realistas desde el caos.





Origen y evolución de los modelos de difusión

- 💡 El concepto tiene raíces en la física estadística, específicamente en los procesos de difusión donde las partículas se mueven de áreas de alta concentración a baja concentración hasta alcanzar equilibrio.
- 🔬 En 2015, Sohl-Dickstein y colaboradores propusieron el primer modelo de difusión para generación de imágenes, pero su adopción masiva llegó años después con el trabajo de Ho, Jain y Abbeel (2020) quienes introdujeron los "Diffusion Probabilistic Models" (DDPM).
- 🚀 La explosión de popularidad ocurrió en 2022 con el lanzamiento de Stable Diffusion, DALL-E 2 y Midjourney, llevando esta tecnología al público general y transformando industrias creativas enteras.

⚙️ ¿Cómo funcionan los modelos de difusión?

El proceso consta de dos fases fundamentales:

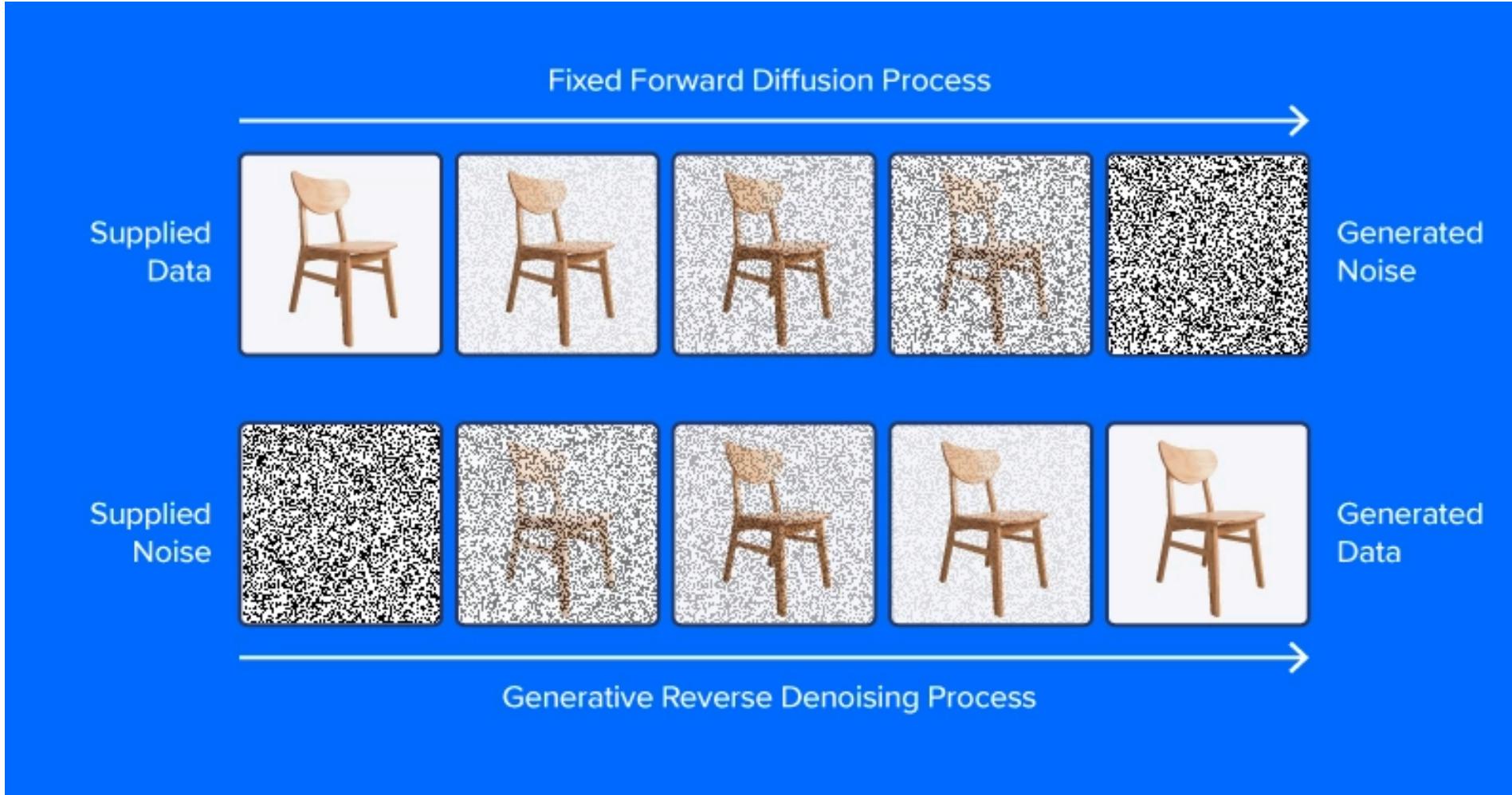
1. **Proceso de difusión directa (Forward):** Añade ruido gaussiano gradualmente a una imagen hasta convertirla en ruido puro
2. **Proceso de difusión inversa (Reverse):** Aprende a eliminar el ruido paso a paso para reconstruir la imagen

La magia está en el entrenamiento: el modelo aprende a predecir el ruido añadido en cada paso, lo que le permite posteriormente revertir el proceso con imágenes nunca vistas.

En esencia, un modelo de difusión no genera imágenes directamente, sino que aprende a "limpiar" ruido gradualmente hasta formar una imagen coherente.



El proceso de difusión en acción



Arriba: Proceso de difusión directa (Forward) que añade ruido progresivamente. Abajo: Proceso generativo inverso (Reverse) que elimina ruido para crear la imagen. Fuente: Contentstack.io



Matemática de la difusión: Entendiendo el proceso

Proceso de difusión directa:

- Comienza con una imagen real x_0
- Aplicamos ruido gaussiano en T pasos
- En cada paso t : $x_t = \sqrt{1-\beta_t} \cdot x_{t-1} + \sqrt{\beta_t} \cdot \epsilon$
- Donde β_t es la programación de ruido y ϵ es ruido gaussiano
- Al final, x_T es prácticamente ruido puro

Proceso de difusión inversa:

- Partimos de ruido puro x_T
- Aprendemos a predecir el ruido añadido en cada paso

La función objetivo que optimizamos durante el entrenamiento:

$$L = \mathbb{E}_{x_0, \epsilon, t} [\|\epsilon - \epsilon_\theta(x_t, t)\|^2]$$

Donde:

- x_0 es la imagen original
- ϵ es el ruido añadido
- ϵ_θ es nuestra red que predice el ruido
- t es el paso de difusión

Lo sorprendente es que esta simple pérdida cuadrática es suficiente para generar imágenes de alta calidad.

Text-to-Image: Guiando la difusión con palabras

Durante el entrenamiento, las imágenes con sus etiquetas crean una **base de datos de ruido etiquetado**.

Cuando escribimos un prompt como "**silla aguacate**", el modelo:

1. **Vectoriza** los conceptos "silla" y "aguacate"
2. **Combina** sus representaciones en el espacio latente
3. **Guía** el proceso de denoising para manifestar ambos conceptos en una imagen coherente



Ejemplo de "silla aguacate" generada por un modelo de difusión. La combinación de conceptos demuestra la capacidad del modelo para fusionar representaciones semánticas.

Componentes de un sistema Text-to-Image

Arquitectura de un sistema text-to-image basado en difusión:

1. **Codificador de texto:** Convierte prompts en representaciones vectoriales (CLIP, T5)
2. **U-Net condicionado:** Red que predice el ruido incorporando información textual
3. **Programación de ruido:** Controla la velocidad del proceso de difusión
4. **Muestreador:** Algoritmo que mejora la eficiencia de generación (DDIM, etc.)

La magia ocurre en cómo estos componentes trabajan juntos: el texto sirve como "brújula semántica" que guía la eliminación de ruido hacia conceptos específicos, permitiendo crear entidades visuales que combinan conceptos de formas creativas y nunca antes vistas.

🤝 Transformers y Difusión, la unión perfecta

- ⌚ Los Diffusion Transformers (DiT) representan la convergencia de dos paradigmas poderosos: combinan la capacidad de los transformers para modelar relaciones a larga distancia con la potencia generativa de los modelos de difusión.
- ✳️ A diferencia de las CNNs, los DiT pueden capturar patrones globales en imágenes gracias a su mecanismo de atención, permitiendo generar imágenes más coherentes estructuralmente y con mayor comprensión del contexto.

DiT fue propuesto en el paper "**Scalable Diffusion Models with Transformers**" (Peebles & Xie, 2022)  [arxiv.org/pdf/2212.09748](https://arxiv.org/pdf/2212.09748.pdf)





Anatomía de un Diffusion Transformer

Un DiT combina lo mejor de dos mundos: el modelo de **difusión** genera las imágenes de cada fotograma con alta calidad visual, mientras que el **transformer** proporciona coherencia secuencial entre fotogramas a través de su mecanismo de atención.

Componentes esenciales:

1. **Tokenización visual:** Divide cada fotograma en "tokens" o parches
2. **Embedding posicional:** Añade información temporal y espacial
3. **Atención cruzada entre fotogramas:** Permite que cada parche "observe" parches relevantes de otros fotogramas
4. **Denoiser condicional:** Elimina ruido guiado por el contexto de la secuencia completa

El mecanismo de atención: La clave del éxito

A diferencia de los transformers para texto donde las palabras se atienden entre sí, en DiT:

- La atención opera entre parches de **distintos fotogramas**
- Cada fotograma "consulta" a los anteriores y posteriores
- El modelo aprende qué partes deben mantener consistencia
- Esta coherencia permite transiciones fluidas en videos

En esencia, el DiT utiliza transformers para "recordar" cómo deben evolucionar las escenas a lo largo del tiempo, mientras que los modelos de difusión se encargan de generar los detalles visuales de cada fotograma individual.



Conclusiones: Un nuevo paradigma visual

- Las CNN revolucionaron la visión artificial enseñando a las máquinas a reconocer objetos y patrones visuales
- Los modelos de difusión transformaron la generación de imágenes, permitiendo crear contenido visual original de alta calidad
- Los DiT representan la nueva frontera, combinando lo mejor de los transformers y la difusión para crear sistemas visuales con mayor comprensión contextual y capacidades multimodales

La evolución de estas arquitecturas no solo ha expandido las capacidades técnicas de la IA, sino que ha democratizado la creación visual, redefiniendo nuestra relación con la tecnología y abriendo nuevas posibilidades creativas para todos.

Material Complementario: Recursos Interactivos



Para explorar en profundidad

- Visualización interactiva de CNN:
Explora cómo "ve" una red neuronal convolucional en cada capa
[LinkedIn - CNN Visualization](#)
- Diffusers en acción:
Experimenta el proceso de denoising en tiempo real
[FLUX.1-dev en Hugging Face](#)
- Generación de video con IA:
Crea videos con Veo 2 de Google
[Google AI Studio - Generate Video](#)



Herramientas recomendadas

- Para CNN:
 - [TensorFlow Playground](#)
 - [CNN Explainer](#)
- Para Diffusion Models:
 - [Hugging Face Diffusers](#)
 - [Stable Diffusion WebUI](#)
- Para DiT y Video:
 - [RunwayML Gen-2](#)
 - [Pika Labs](#)



Recursos del Curso

📌 Plataformas y Enlaces Principales

📁 GitHub del curso

🔗 github.com/CamiloVga/IA_Aplicada

🤖 Asistente IA para el curso

🔗 Google Notebook LLM