Redes Neuronales para Lenguaje Natural

2024

Grupo de Procesamiento de Lenguaje Natural
Instituto de Computación


Multimodalidad

Introducción

Entendemos por “modalidad” a conjuntos diversos de tipos de datos

Por ejemplo: texto, audio, imágenes, video…

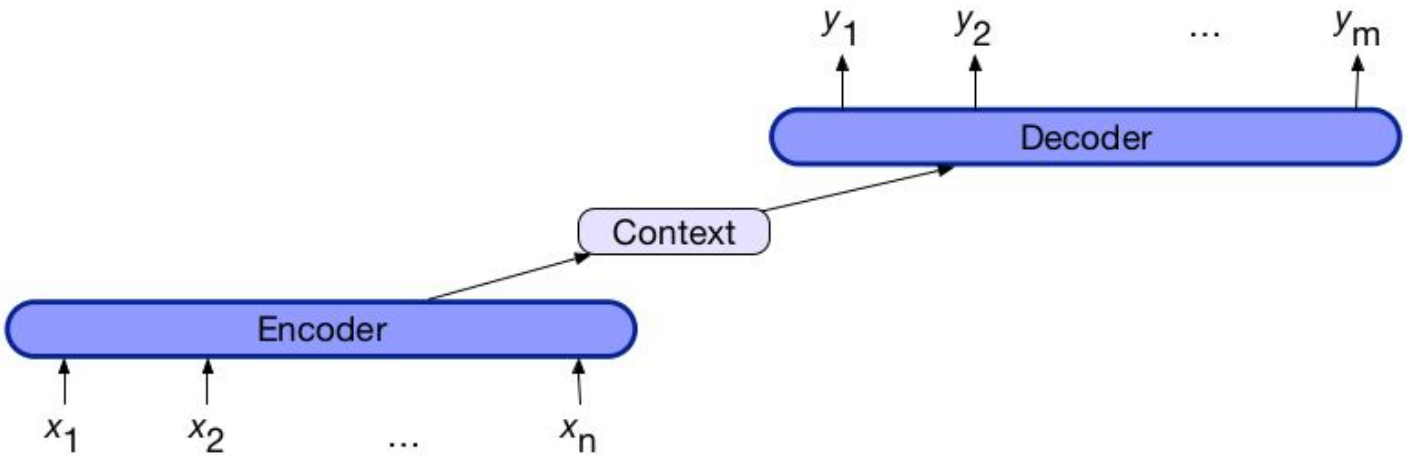
Los modelos multimodales son modelos que intentan utilizar datos de más de una modalidad, para:

- Traducir: speech2text, captioning de imágenes, generación de imágenes
- Alinear: segmentación de imágenes, diarización
- Clasificar: clasificación de imágenes, análisis de sentimiento en audio
- …

Arquitectura Encoder-Decoder (repaso)

Una red compuesta por dos subredes:

- **Encoder**: red que codifica la entrada
- **Decoder**: red que decodifica (y construye) la salida
- **Context vector**: “embedding” de toda la secuencia de entrada



Encoder-Decoder

Las redes neuronales son capaces de adaptarse a muchos tipos de entrada

Eso las hace particularmente apropiadas para datos multimodales

El enfoque clásico es pensar estos sistemas como encoder-decoder:

- Cada encoder y cada decoder puede estar trabajando en una modalidad diferente

Captioning de Imágenes

Show and Tell

Dada una imagen *I*

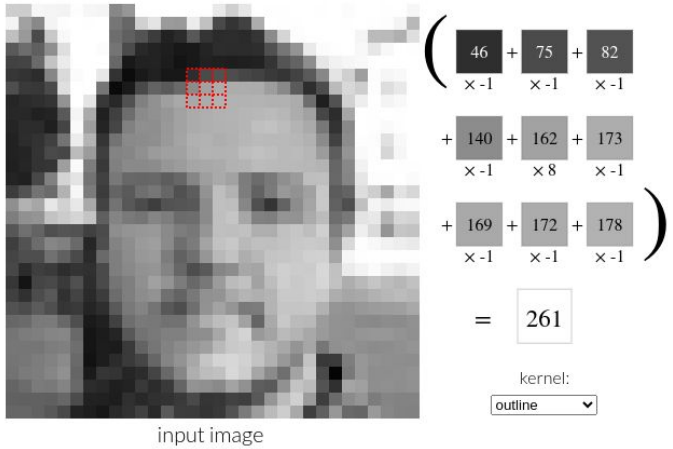
Buscamos una secuencia $S = \{S_1, S_2, \ldots S_n\}$ de palabras de un vocabulario tales que se maximice: $P(S|I)$

La solución está inspirada en el modelo encoder-decoder usado en traducción automática

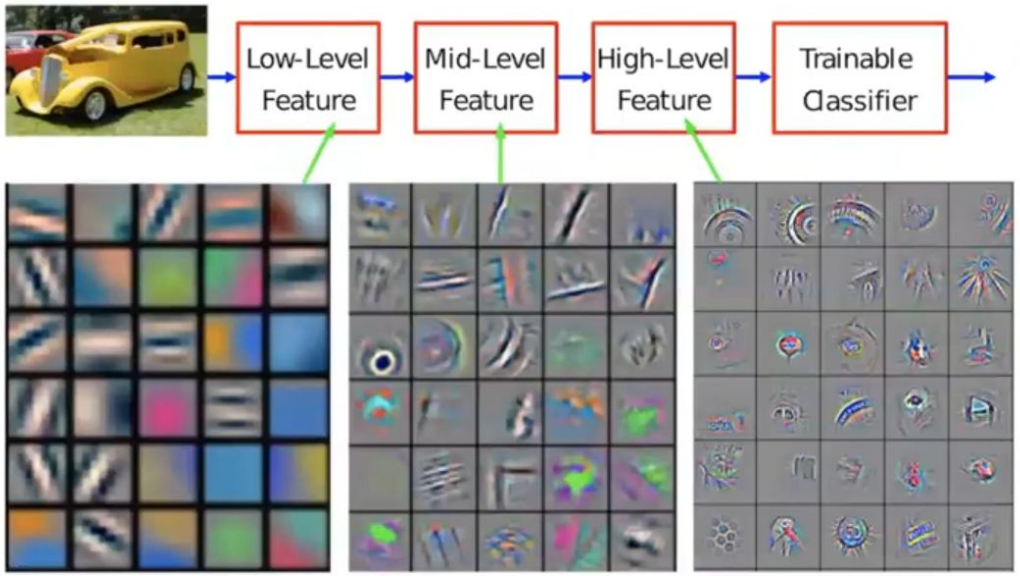
Pero en vez de usar LSTM para el encoder y el decoder, usan CNN (de imágenes) para el encoder y LSTM para el decoder

Show and Tell: A Neural Image Caption Generator
Vinyals, Toshev, Bengio, Erhan (2014)

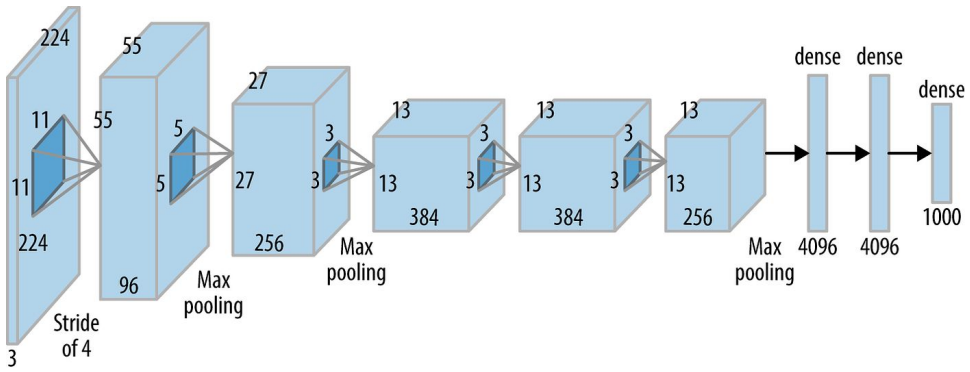
CNN (repaso)



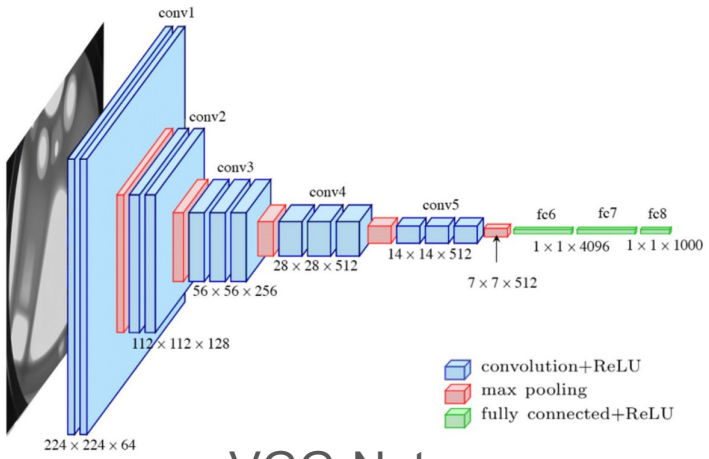
Convolución



Extracción de features y clasificación

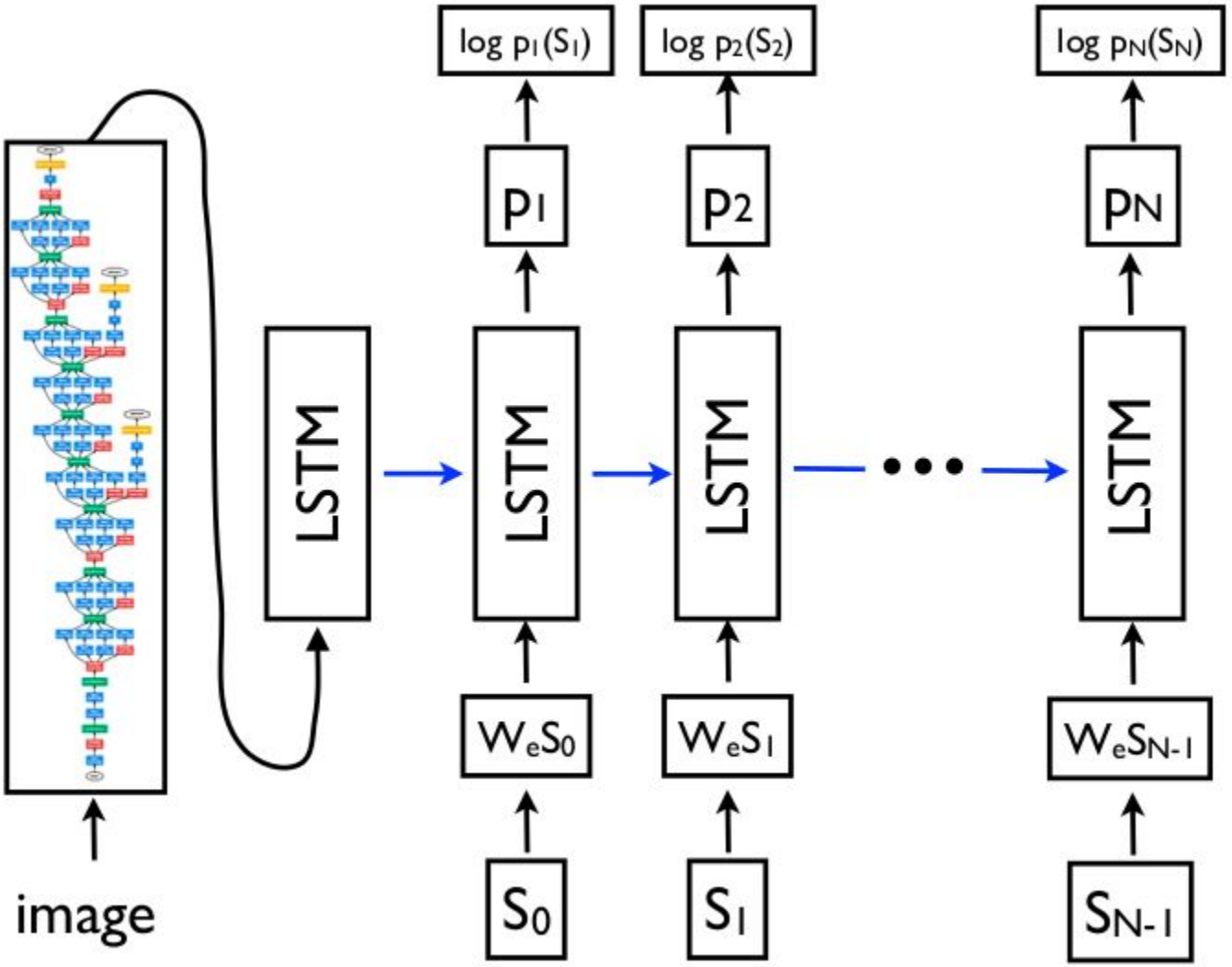


AlexNet



VGG-Net

Show and Tell



Show and Tell

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A skateboarder does a trick on a ramp.

A dog is jumping to catch a frisbee.

A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

A little girl in a pink hat is blowing bubbles.

A refrigerator filled with lots of food and drinks.

A herd of elephants walking across a dry grass field.

A close up of a cat laying on a couch.

A red motorcycle parked on the side of the road.

A yellow school bus parked in a parking lot.

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image

Show, Attend and Tell

El trabajo anterior estaba basado en el encoder-decoder simple usado para traducción

En este trabajo se adopta la idea del encoder-decoder con atención

En vez de usar la salida de la capa final de la CNN, obtienen varios (L) vectores de capas más bajas, les llaman *“anotaciones”*

Se usan dos mecanismos de atención entre las L anotaciones (embeddings de imagen) para producir los tokens:

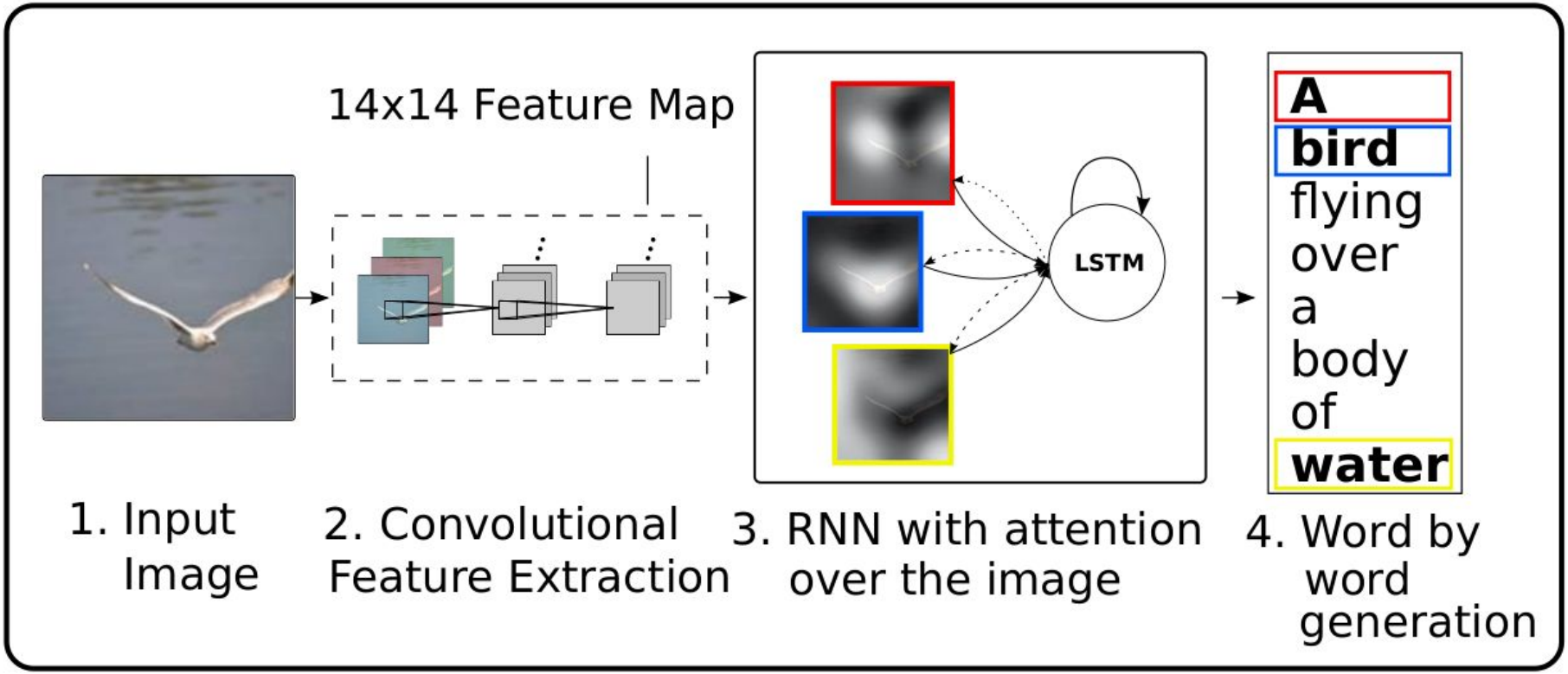
Soft attention y *Hard attention*

Show, Attend and Tell: Neural Image Caption, Generation with Visual Attention
Xu et al. (2015)

Show, Attend, and Tell

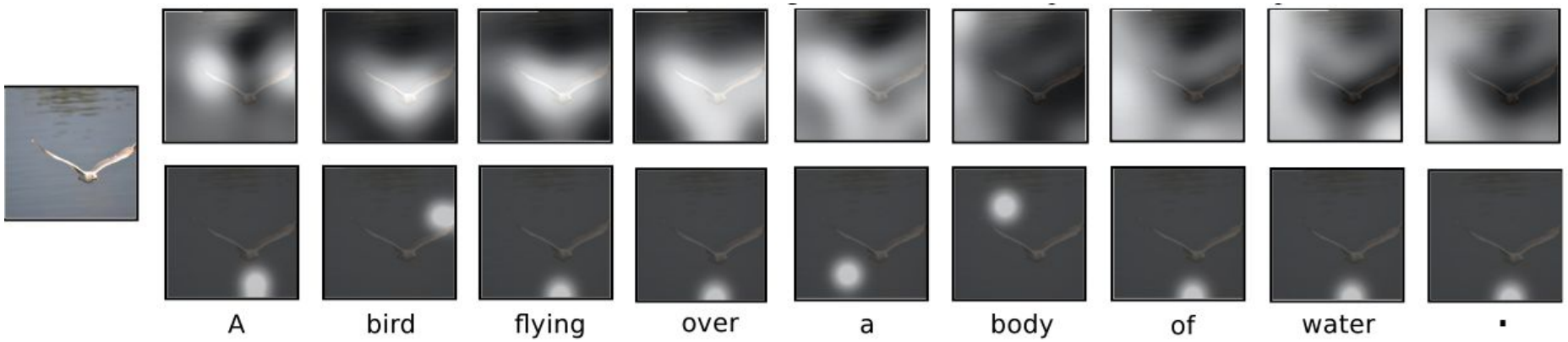
CNN como encoder de L vectores de la imágen (anotaciones)

LSTM para decodificar el caption, atención sobre los L vectores



Show, Attend, and Tell

Soft attention: el vector de contexto es una suma ponderada de las anotaciones, cubriendo una zona más grande pero difusa



Hard attention: el vector de contexto es principalmente una sola de las anotaciones, resultando en una atención más fija en ciertos puntos

Show, Attend, and Tell

Ejemplos que andan bien

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Show, Attend, and Tell

Ejemplos de errores

A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and
a hat on a skateboard.

A person is standing on a beach
with a surfboard.

A woman is sitting at a table
with a large pizza.

A man is talking on his cell phone
while another man watches.

El mecanismo de atención nos puede dar una idea de lo que salió mal

Show, Attend, and Tell

Resultados

		BLEU				
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04


Embeddings Imagen-Texto

CLIP

Es un modelo que permite codificar imágenes y texto de forma de que estén en el mismo espacio de embeddings

Entrenado con un corpus muchísimo más grande que los datasets existentes creados a mano

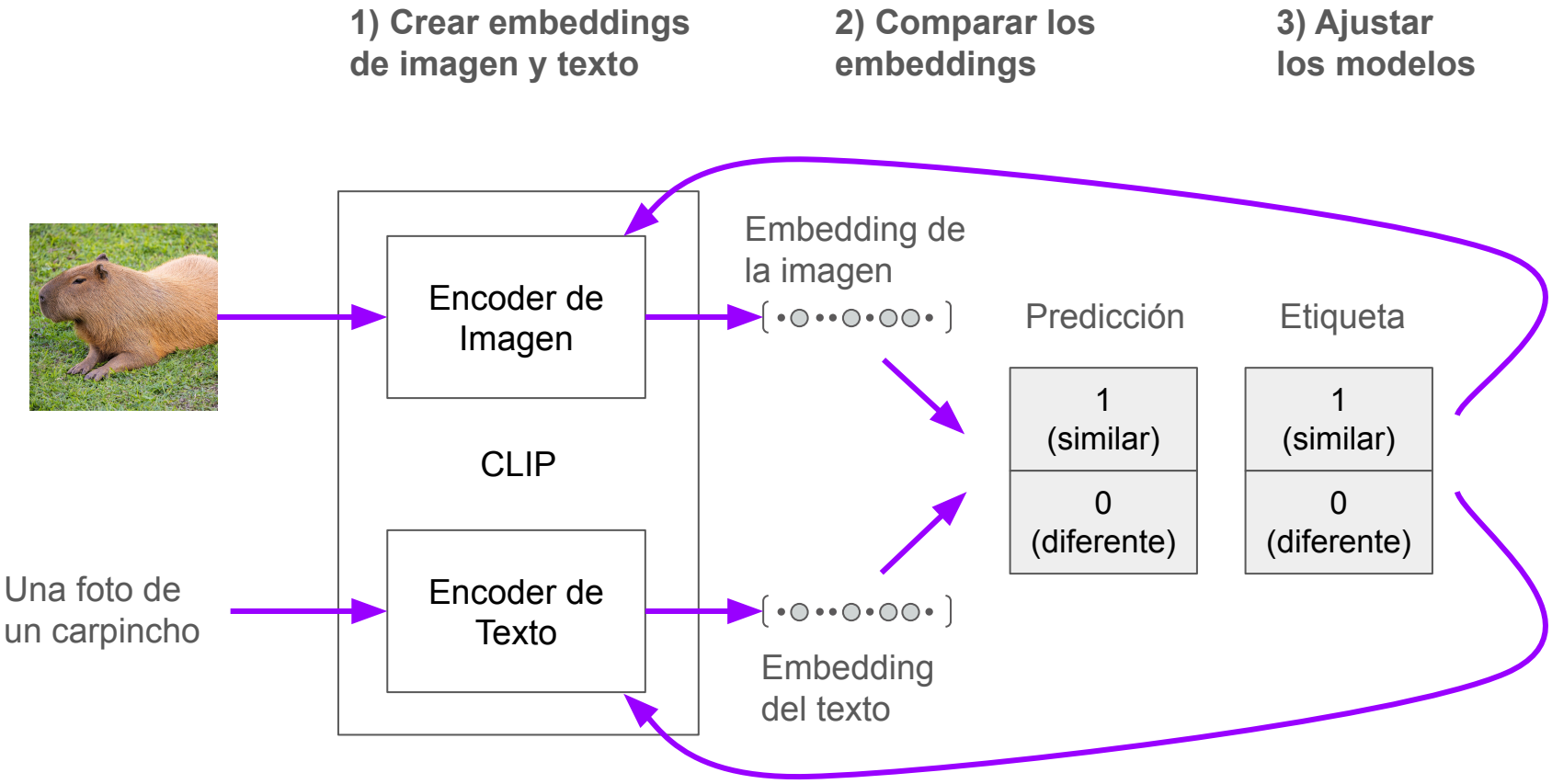
Se entrena de forma autosupervisada

Modelo con capacidades de reconocimiento zero-shot:

- Aunque no haya visto ejemplos de ciertas clases en el conjunto de entrenamiento, igual puede llegar a predecirlos

Learning Transferable Visual Models From Natural Language Supervision
Radford et al. (2021)

CLIP


1) Crear embeddings de imagen y texto
2) Comparar los embeddings
3) Ajustar los modelos
Encoder de Imagen
CLIP
Encoder de Texto
Una foto de un carpincho
Embedding de la imagen
Embedding del texto
Predicción
Etiqueta
1 (similar)
0 (diferente)
1 (similar)
0 (diferente)

CLIP

400 millones de ejemplos de entrenamiento!

- Una parte obtenidos de otros datasets
- La mayoría construido con búsquedas web
- Obtienen imágenes y su alt-text

Entrenamiento:

- Contrastive-loss: función de pérdida que “acerca” ejemplos positivos y “aleja” ejemplos negativos
- Los ejemplos positivos son pares del corpus
- Los ejemplos negativos son una imagen del corpus con una caption incorrecta sorteada

CLIP

Procesamiento del texto:

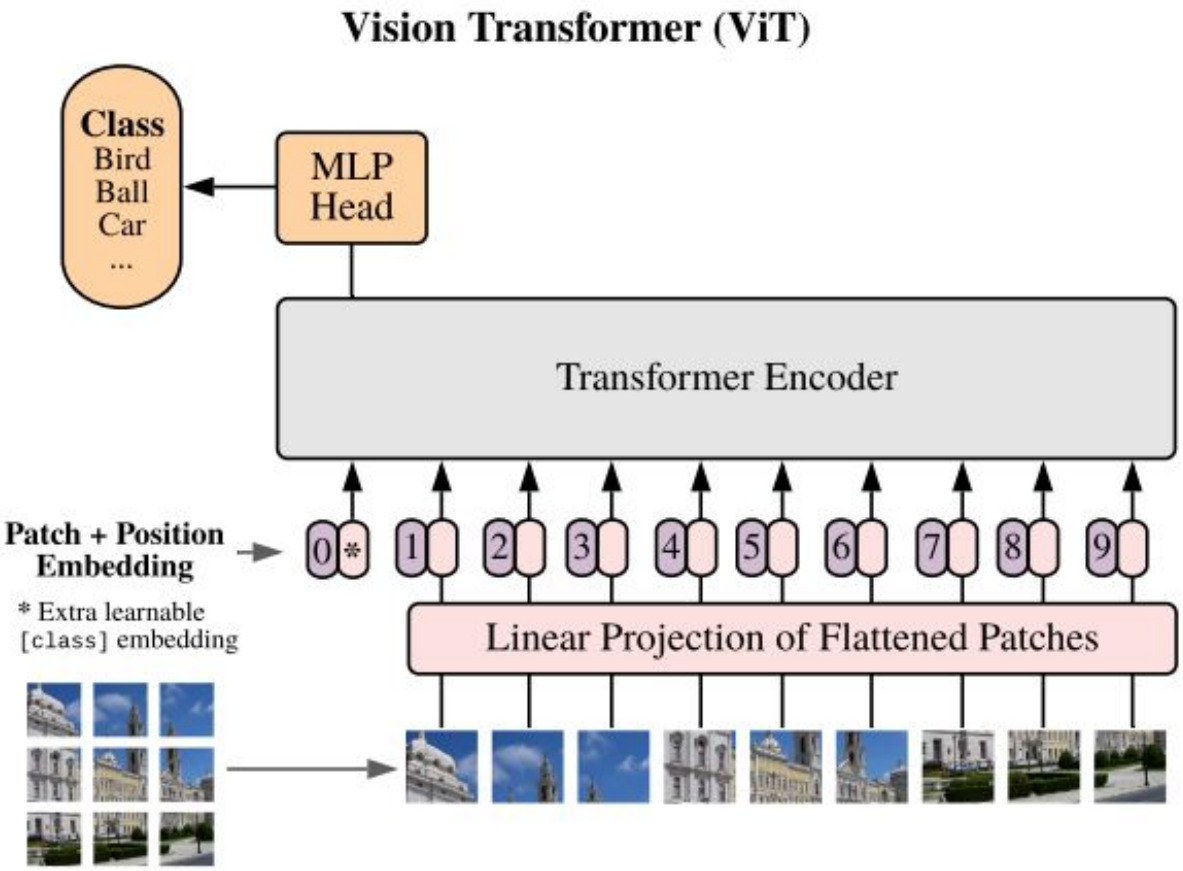
- Transformer

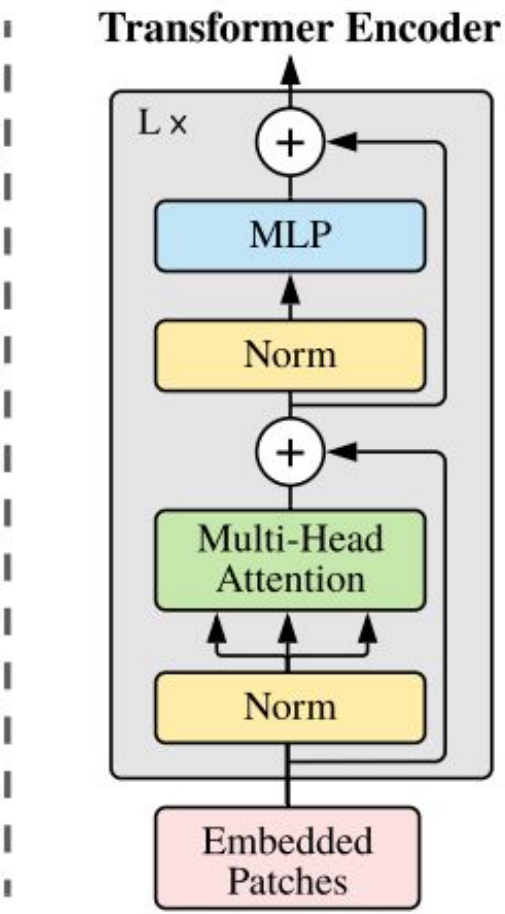
Procesamiento de la imagen:

- ResNet (CNN)
- Vision Transformer

Recordar que el embedding resultante en ambos casos debe pertenecer al mismo espacio (mismo tamaño)

Vision Transformer (ViT)





An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
Dosovitskiy et al. (2020)

Vision Transformer (ViT)

La misma idea del transformer de texto, pero con *patches* (fragmentos contiguos de imagen) en vez de tokens (palabras)

Los patches tienen un tamaño fijo (16x16) y no se solapan

Entrenamiento

- Supervisado: Tratamos de predecir la clase de una imagen
- Como BERT: Enmascaramos un patch y tratamos de predecir qué características debería tener al reconstruirlo

CLIP

¿Cómo se usa?

Supongamos que tenemos una imagen nunca vista en el dataset

Y un conjunto de posibles clases C

Durante el entrenamiento seguro se vieron ejemplos de muchos objetos (con nombres) distintos

…y con la tokenización sub-palabra se puede acceder a más nombres

Nos inventamos captions "*una foto de un X*" para cada X de C

Obtenemos embeddings de la imagen y las posibles captions, y comparamos con similitud coseno


Generación de Imágenes

DALL-E

Zero-Shot Text-to-Image Generation

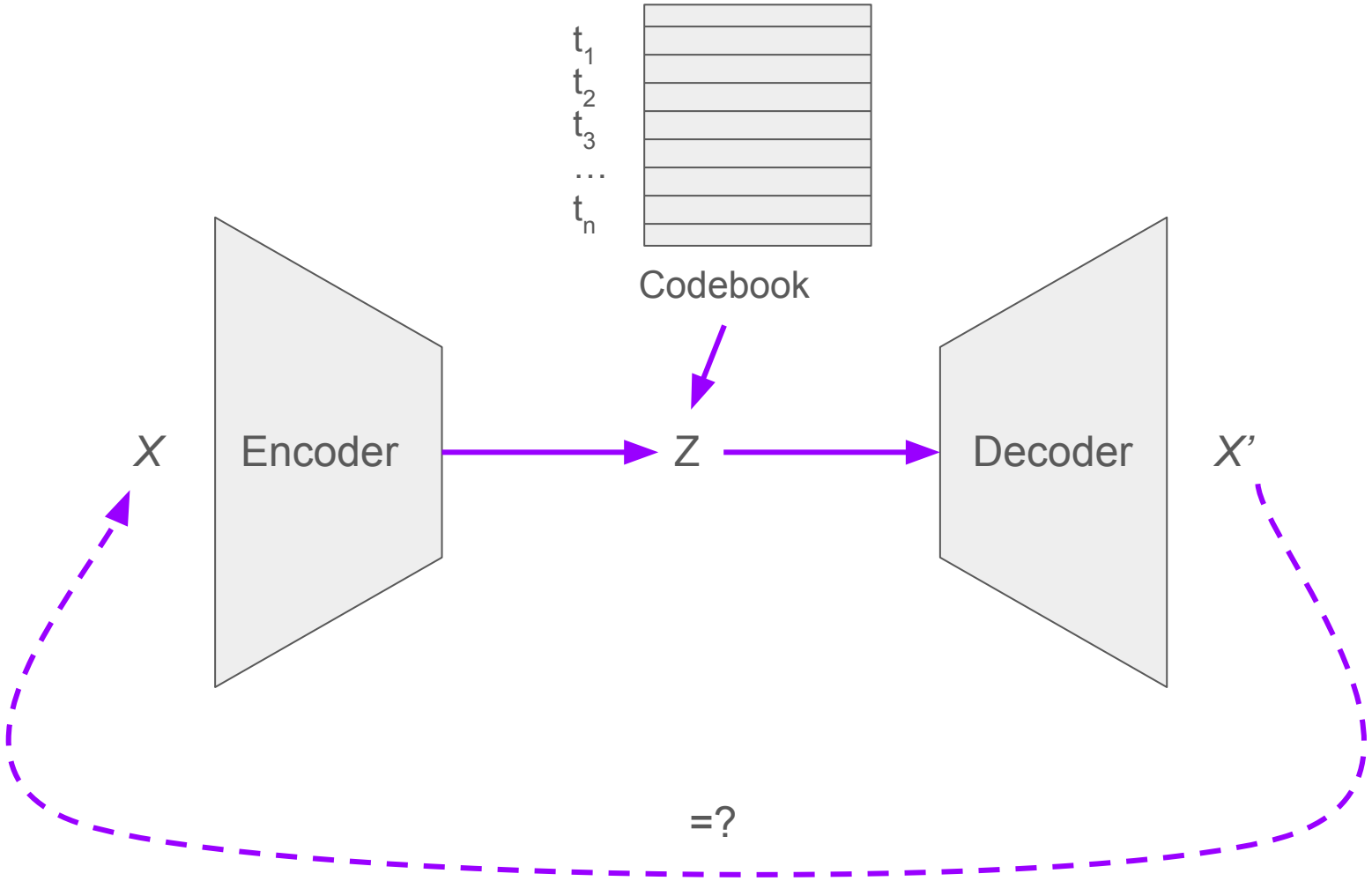
Modelo de generación de imágenes a partir de texto que se volvió muy popular

Separa el método de generación en dos pasos

- Autoencoder (dVAE) que aprende a codificar/decodificar una imagen
- Cálculo de la probabilidad de la imagen a partir de la probabilidad conjunta del texto y la codificación

Zero-Shot Text-to-Image Generation
Ramesh et al. (2021)

(discrete) Variational Autoencoder



DALL-E

Tenemos un autoencoder que puede codificar una imagen en un set de tokens del codebook o decodificar desde un set de tokens

Para obtener la imagen, se separa en el cálculo de dos probabilidades

$$p_{\theta,\psi}(x, y, z) = p_{\theta}(x \mid y, z)\, p_{\psi}(y, z)$$

imagen

caption

tokens (de la imagen)

Con un transformer autorregresivo se aprende a transformar la secuencia de texto en su secuencia de tokens de imagen asociada

Zero-Shot Text-to-Image Generation
Ramesh et al. (2021)

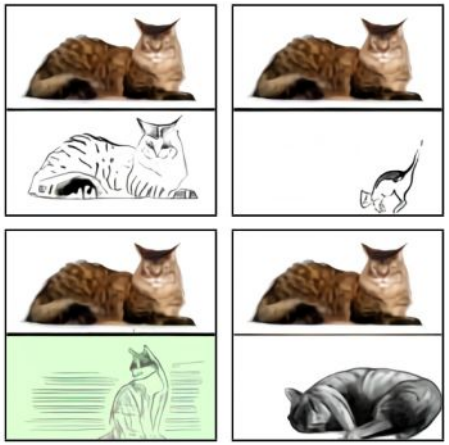
DALL-E

Ejemplos de generación a partir de caption

(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

(c) a neon sign that reads "backprop". a neon sign that reads "backprop". backprop neon sign



(d) the exact same cat on the top as a sketch on the bottom

Stable Diffusion

Modelo de difusión (denoising)

- La tarea de generar imágenes se ve como un problema de generación/eliminación de ruido
- Logra un nivel de detalle mucho mayor en las imágenes

Se comienza con una imagen aleatoria (ruido)

En cada paso se obtiene un perfil de ruido condicionado al encoding del texto (BERT o CLIP)

Se elimina ese ruido, obteniendo una imagen mejor a cada paso

High-Resolution Image Synthesis with Latent Diffusion Models
Rombach et al. (2022)

Stable Diffusion

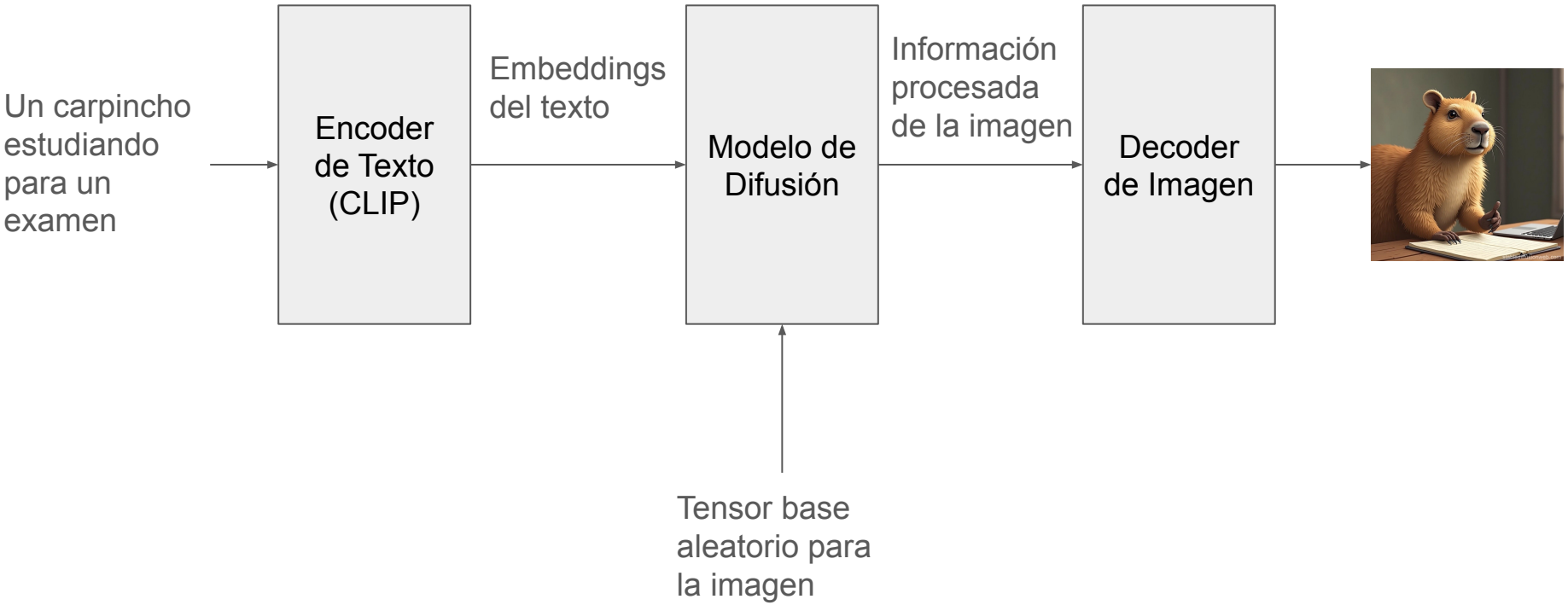
Un gaucho tomando mate en la rambla de Montevideo

Un carpincho estudiando para un examen
A capybara studying for an exam

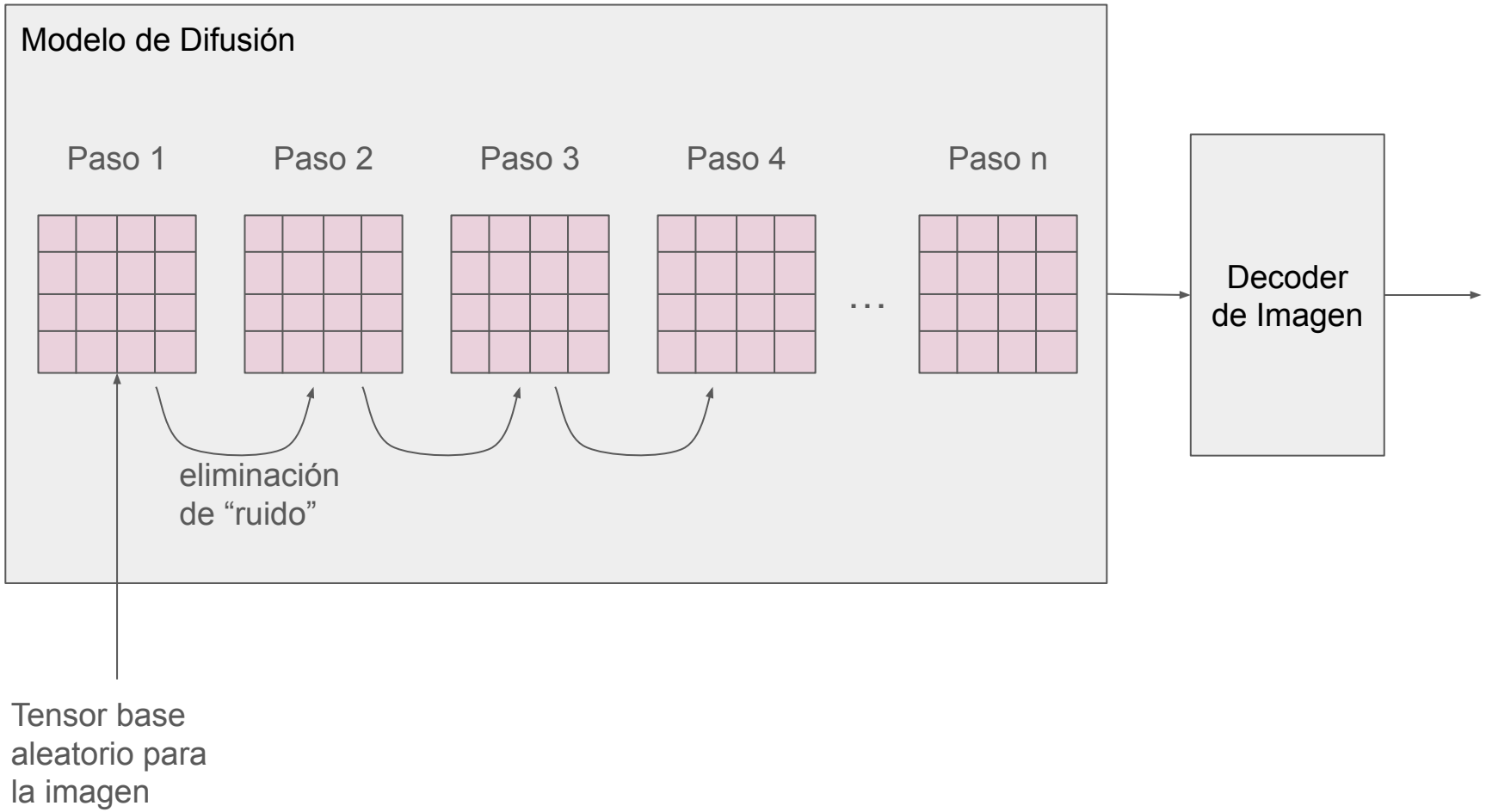
Godzilla aplastando edificios en Montevideo
Godzilla stomping on buildings in Montevideo

Generados con https://stablediffusionweb.com/

Stable Diffusion



Stable Diffusion



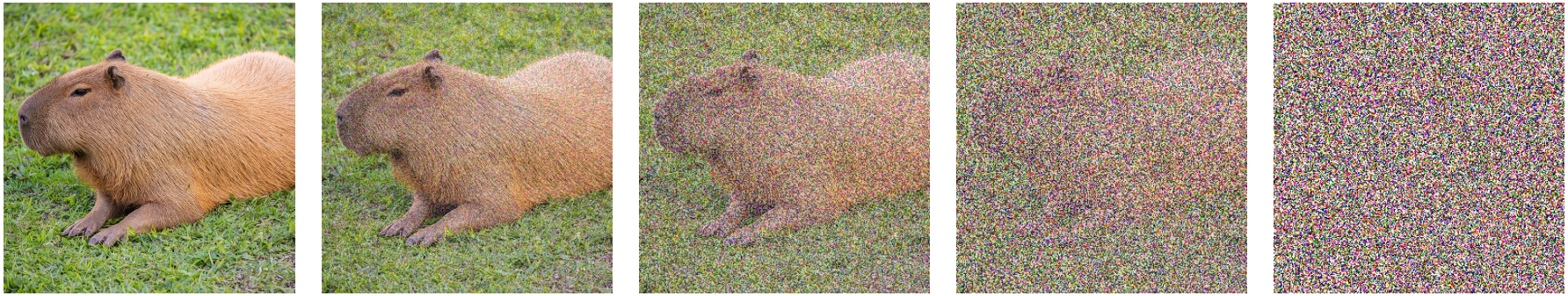
Stable Diffusion

Dada una imagen del conjunto de entrenamiento

Se le va agregando diferente cantidad de ruido

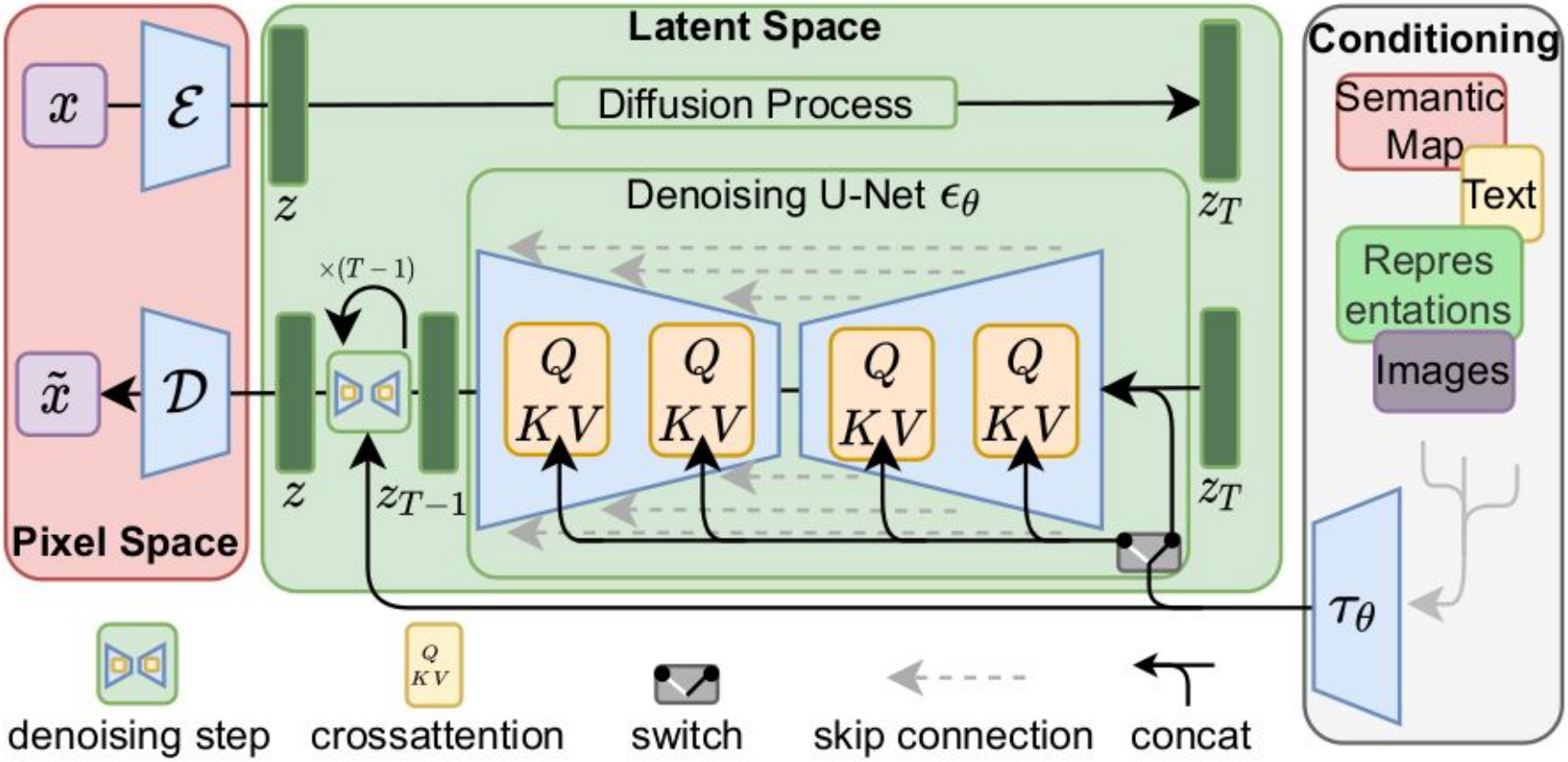
Con eso se construyen ejemplos de entrenamiento

La red aprende a eliminar el ruido



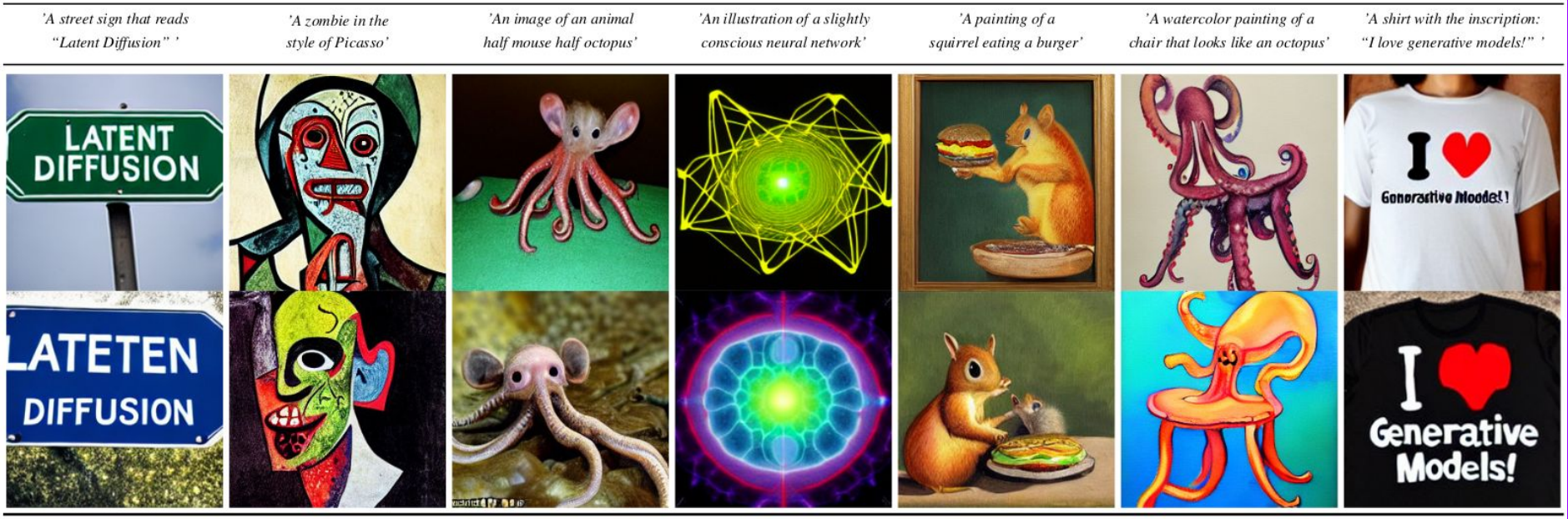
Lo que hace es predecir un *perfil de ruido* dado el texto de entrada

Stable Diffusion



Stable Diffusion

Ejemplos de generación a partir de caption



Stable Diffusion

Ejemplos de inpainting:

Completar una región de una imagen con otro contenido


LLMs Multimodales

LLMs Multimodales

Los LLMs que están saliendo en la actualidad en general incorporan características multimodales

Les podemos pedir que describan una imagen, que generen una imagen nueva, que transcriban un audio (?)

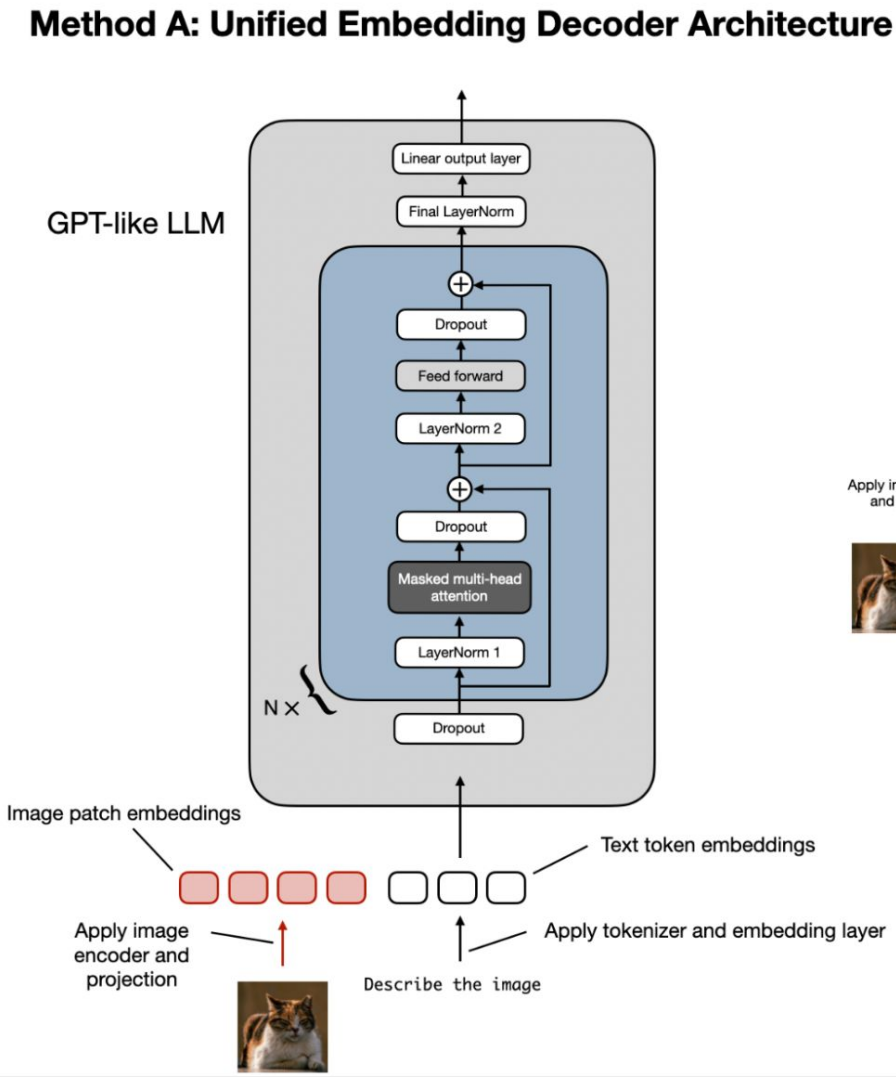
Dos tipos principales de arquitectura (según Raschka, 2024):

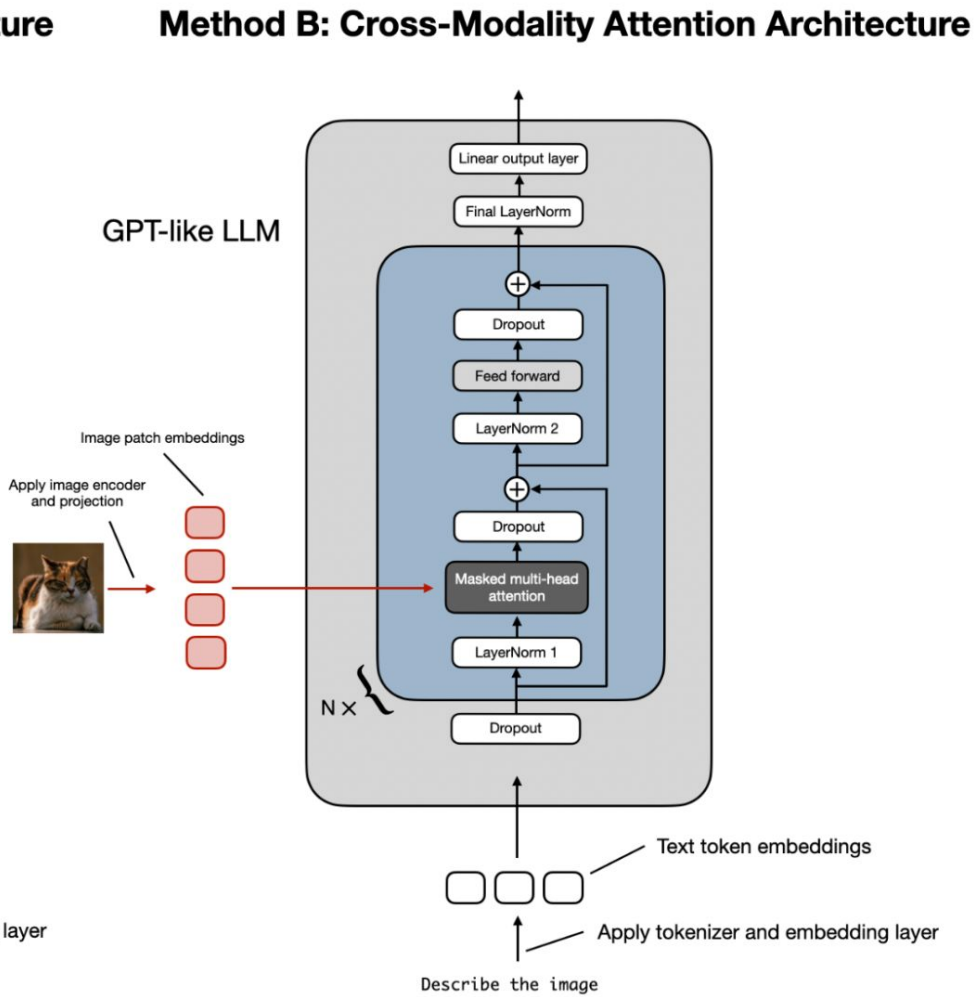
- Arquitectura de Decoder con Embedding Unificado

  (autorregresiva o decoder-only)
- Arquitectura de Atención Cross-Modality

  (cross-attention)

Dos arquitecturas





Representación de datos

En ambas arquitecturas, el núcleo del problema es representar los datos de diferentes modalidades de forma que sean compatibles

Los textos se separan en tokens

Las imágenes se separan en *patches*, como en ViT

Cada fragmento (token o patch) se pasa por su propia subred de encoding, para transformarlo en un embedding

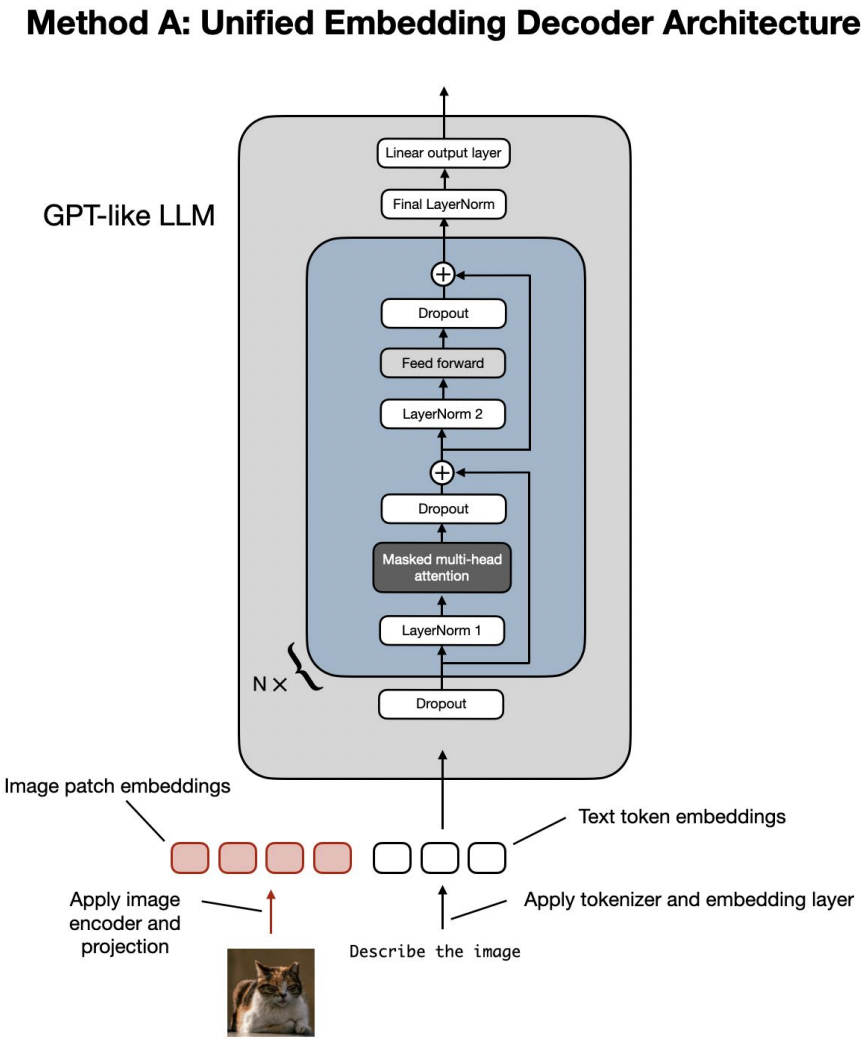
Los embeddings son los que se combinan con los mecanismos de atención

Decoder con Embedding Unificado (decoder-only)

Obtenemos los embeddings de imagen (patches) y de texto (tokens) a partir de su propia subred, pero existen en un espacio común (como en CLIP)

Se entrena de forma autorregresiva: cierta secuencia de patches se continúa con cierta secuencia de tokens

Ejemplo: modelo Pixtral 12B (Mistral multimodal)

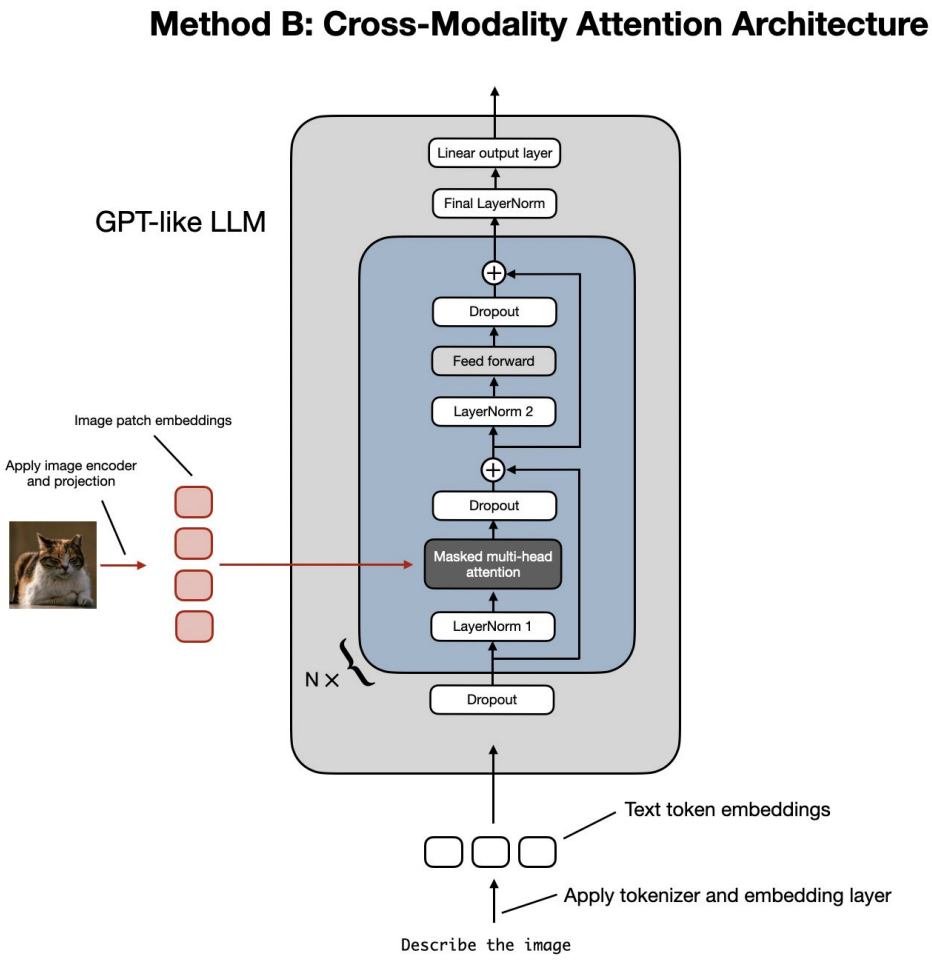


Atención Cross-Modality (cross-attention)

Se inspira en el transformer original (Vaswani, 2017), donde encoder y decoder se combinan con un cross-attention

Cross-attention combina la proyección Q de una modalidad (encoder) con las proyecciones K y V de la otra (decoder)

Ejemplo: familia de modelos Llama 3 multimodales



Entrenamiento

Pueden empezar entrenando cada modelo por separado

- Por ejemplo usar CLIP para la base de los embeddings
- O un ViT y un LLM

Luego se hace un instruction tuning con el modelo completo usando un dataset adaptado a tareas multimodales


Multimodalidad

Referencias

- The Illustrated Stable Diffusion - Jay Alammar - https://jalammar.github.io/illustrated-stable-diffusion/
- How OpenAI's DALL-E works? - Zain ul Abideen - https://medium.com/@zaiinn440/how-openais-dall-e-works-da24ac6c12fa
- Understanding Multimodal LLMs - Sebastian Raschka - https://sebastianraschka.com/blog/2024/understanding-multimodal-llms.html
- Papers…