

# Cursos Extraordinarios

verano 2025

**“Inteligencia Artificial y Grandes  
Modelos de Lenguaje: Funcionamiento,  
Componentes Clave y Aplicaciones”**

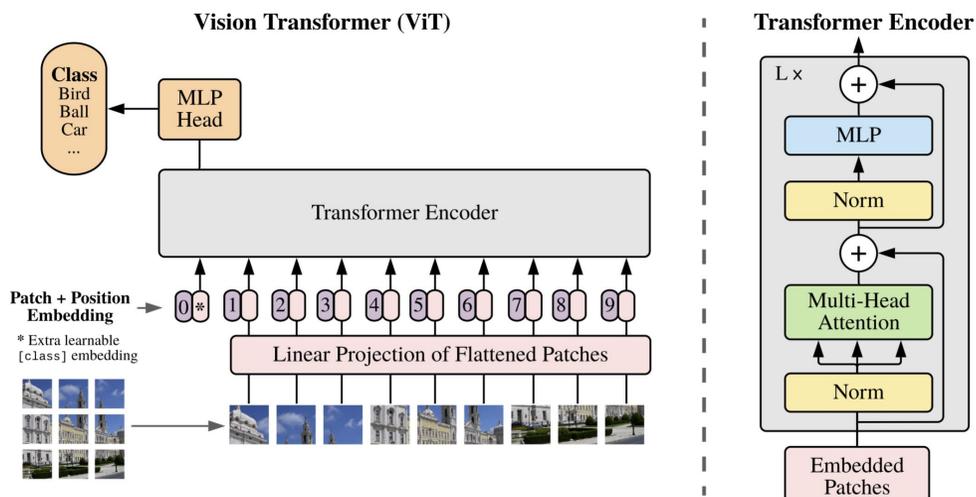
Zaragoza, del 30 de junio al 02 de julio de 2025

# Modelos multimodales

- **Visual Transformer(Dosovitskiy 2020)**

*Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Uszkoreit, J. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.*

- Se descompone la imagen en parches (16x16 patches) de izquierda a derecha
- Imagenet 88.55% (i.e. ResNExt 101: 85.4%) Top1 accuracy



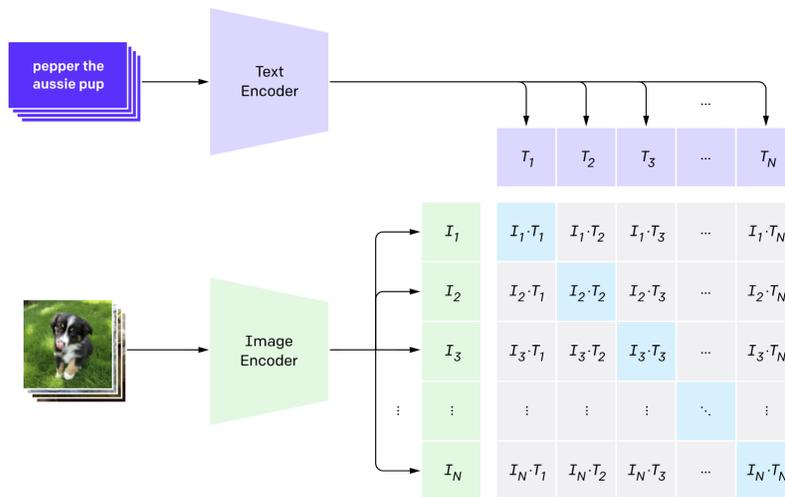
Model	Layers	Hidden size $D$	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Tamaño del modelo y número de parámetros

# Modelos multimodales

- **CLIP (Contrastive Language-Image Pre-Training)**

- Se entrenan dos redes una para representar texto y la otra imagen
- Contrastive loss
  - Compara todas las representaciones
    - Favorece que los valores de la diagonal sean altos
    - Penaliza los valores fuera de la diagonal



```
# image_encoder - ResNet or Vision Transformer
# text_encoder - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l] - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_i + loss_t)/2
```

<https://openai.com/blog/clip/>

# Modelos de difusión

## • Difusión

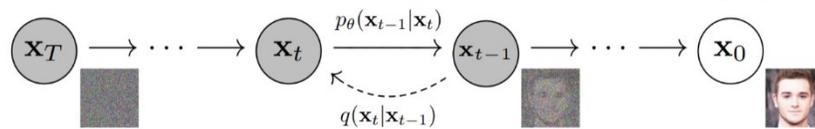
Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising diffusion probabilistic models*. *Advances in Neural Information Processing Systems*, 33, 6840-6851.

Dhariwal, P., & Nichol, A. (2021). *Diffusion models beat gans on image synthesis*. *arXiv preprint arXiv:2105.05233*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). *Learning transferable visual models from natural language supervision*. *arXiv preprint arXiv:2103.00020*.



- Proceso de difusión:
  - Proceso directo añadimos ruido a una entrada hasta que es irreconocible



$$p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, t), \boldsymbol{\Sigma}_{\theta}(\mathbf{x}_t, t))$$

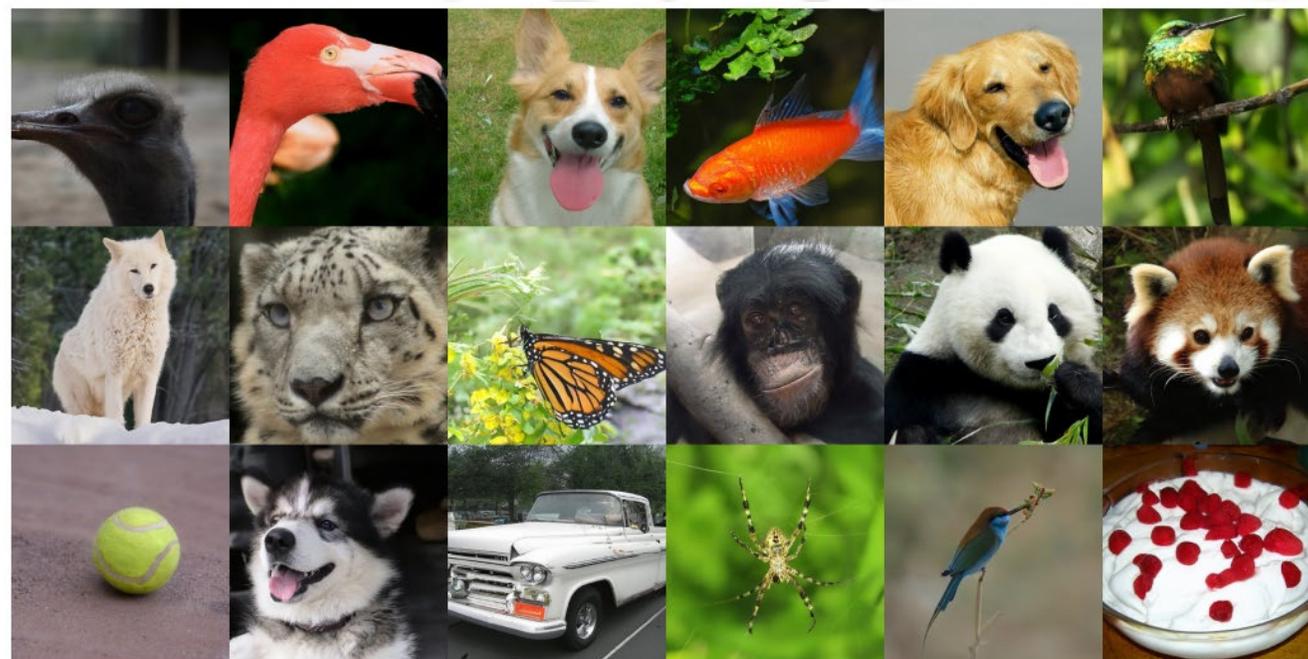
- Sentido “inverso” se limpia el ruido de la señal iterativamente (mediante una red)

# Modelos de difusión

- Modelos de difusión
  - Gran calidad de generación, superando a los modelos GAN



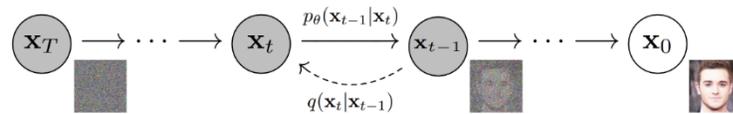
Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239



Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. arXiv preprint arXiv:2105.05233.

# Modelos multimodales

- Difusión dirigida por texto



Cómo incorporamos información del texto o de una clase?

- Generación no condicionada,
  - donde  $\mu, \Sigma$  se predicen de una red con entrada  $x_{t+1}$

$$p_{\theta}(x_t|x_{t+1}) = \mathcal{N}(\mu, \Sigma)$$

- Generación condicionada

$$p_{\theta, \phi}(x_t|x_{t+1}, y) = Z p_{\theta}(x_t|x_{t+1}) p_{\phi}(y|x_t)$$

$\uparrow$  Generación No condicionada  
 $\uparrow$  Dependiente del texto

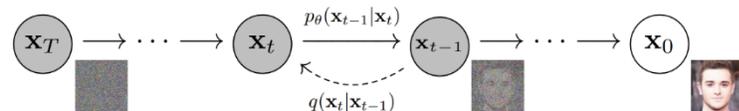
$$\mathcal{N}(\mu + s \Sigma \nabla_{x_t} \log p_{\phi}(y|x_t), \Sigma)$$

(Aproximación)

Se usa CLIP para este gradiente  
Dirige la generación hacia imágenes  
Con mayor similitud (se añade a la media)

# Modelos multimodales

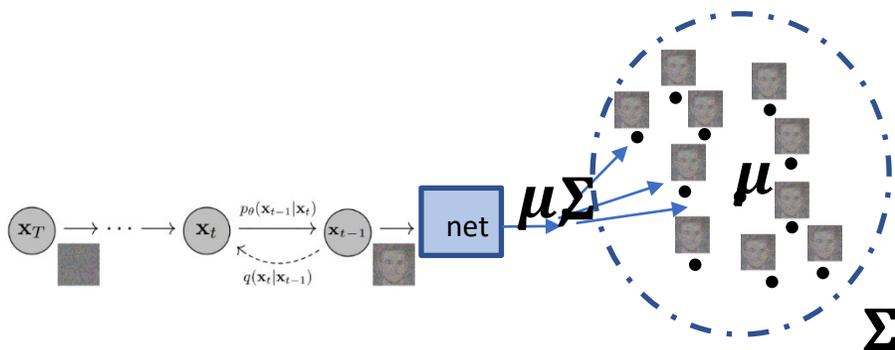
- Difusión dirigida por texto



← Cómo incorporamos información del texto o de una clase?

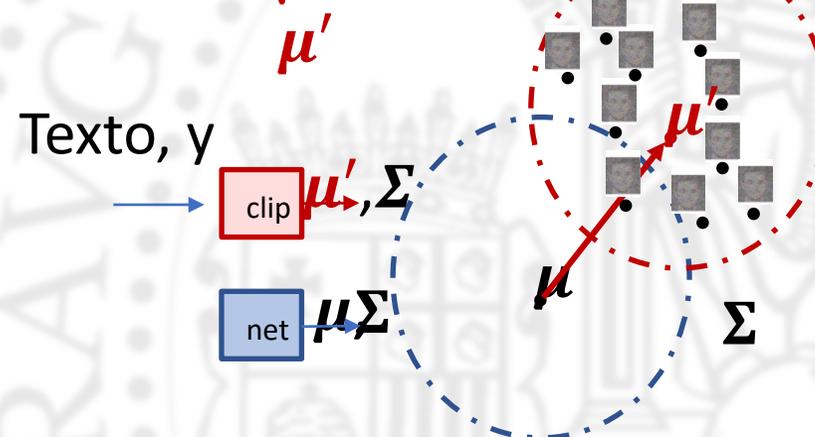
Generación no condicionada

$$p_\theta(x_t|x_{t+1}) = \mathcal{N}(\mu, \Sigma)$$



Generación condicionada

$$\mathcal{N}(\underbrace{\mu + s\Sigma \nabla_{x_t} \log p_\phi(y|x_t)}_{\mu'}, \Sigma)$$



# Diffusion models

- CLIP + diffusion -> Dalle, dalle2, midjourney, stable diffusion,...
- RAMESH, Aditya, et al. *Hierarchical text-conditional image generation with clip latents*. arXiv preprint arXiv:2204.06125, 2022, vol. 1, no 2, p. 3.
  - Los primeros modelos se realizaron con Gans+ CLIP y tenían muy baja calidad
  - Mediante la utilización de métodos de difusión unidos a los gradientes que proporciona CLIP la calidad empezó a ser cada vez mejor y cada modelo superaba a los anteriores en cuestión de meses



the angel of air. unreal engine  
[@arankomatsuzaki](https://twitter.com/arankomatsuzaki)



treehouse in the style of studio ghibli animation  
[@danielrusruss](https://twitter.com/danielrusruss)



unreal engine gaudi house  
 in a field of poppy

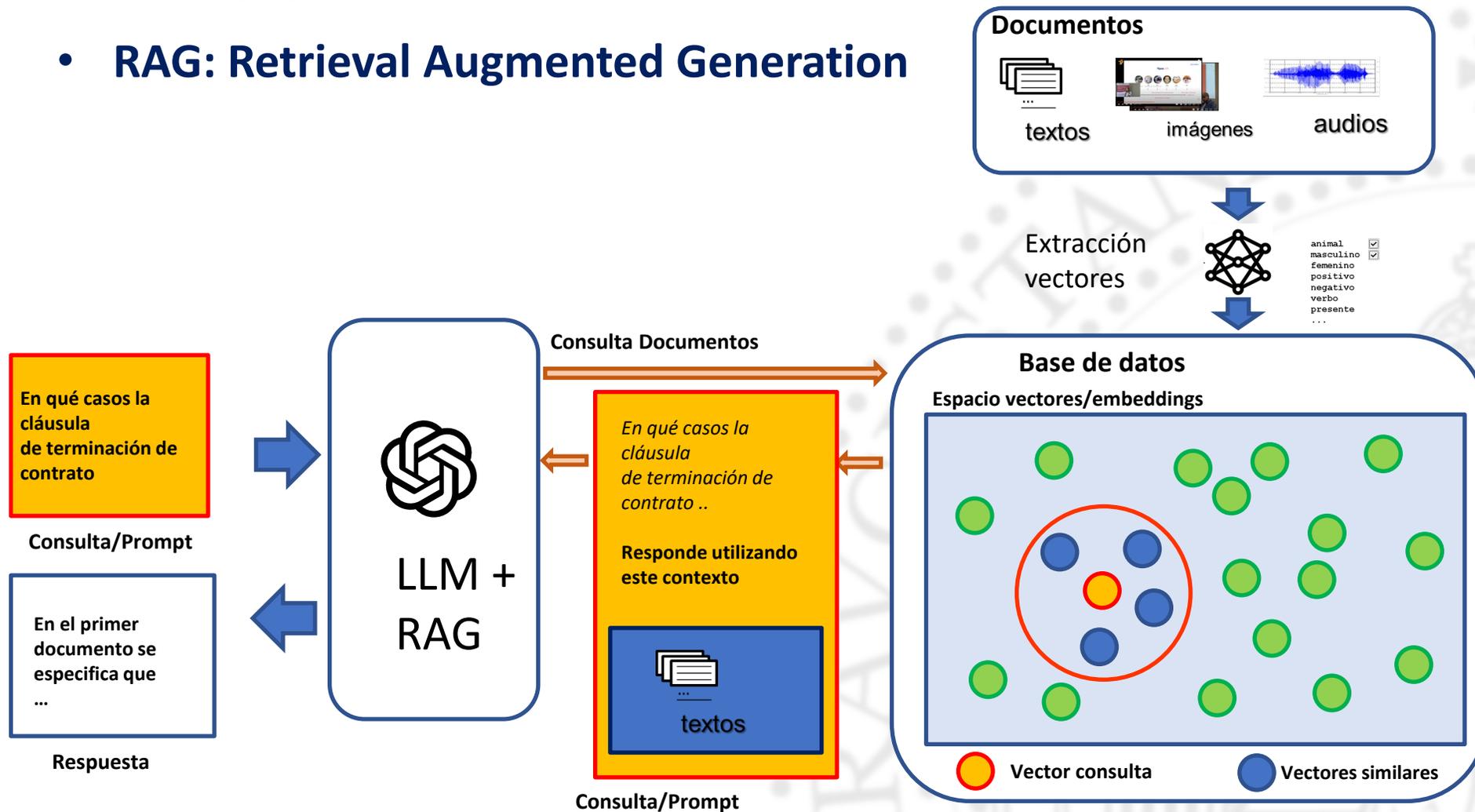


A wooden Spanish laptop of 1650  
 found the library of El Escorial

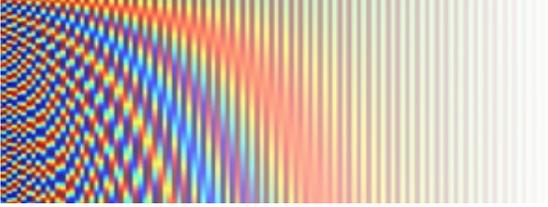
- Se observó por primera vez que algunas palabras modificaban la calidad de la imagen obtenida o la realizaban en un estilo particular: unreal engine, trending on artstation, in the style of \*\*\*

# RAG

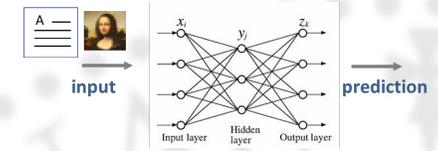
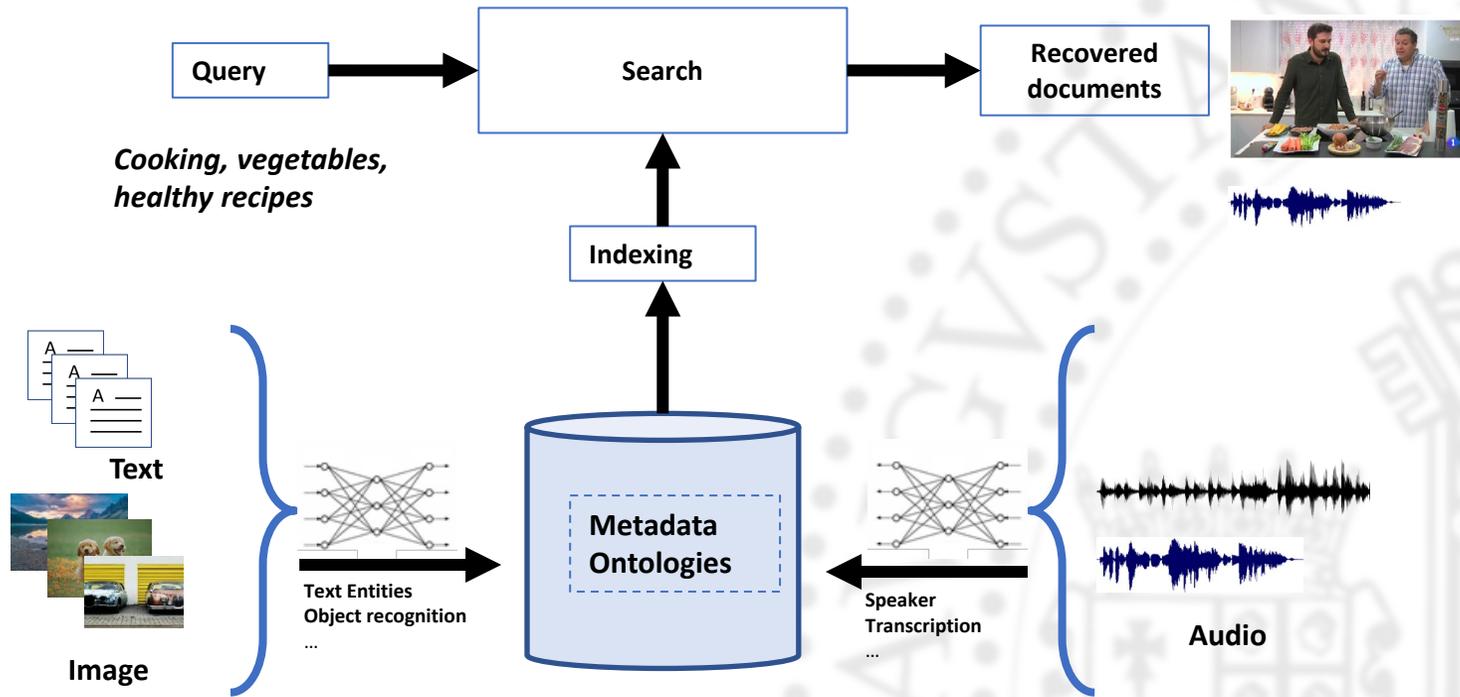
- RAG: Retrieval Augmented Generation



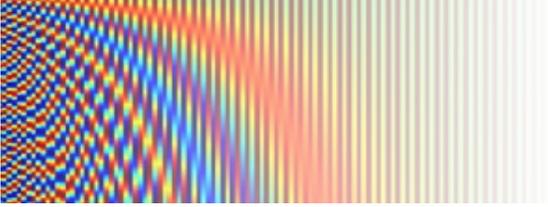
# Búsqueda



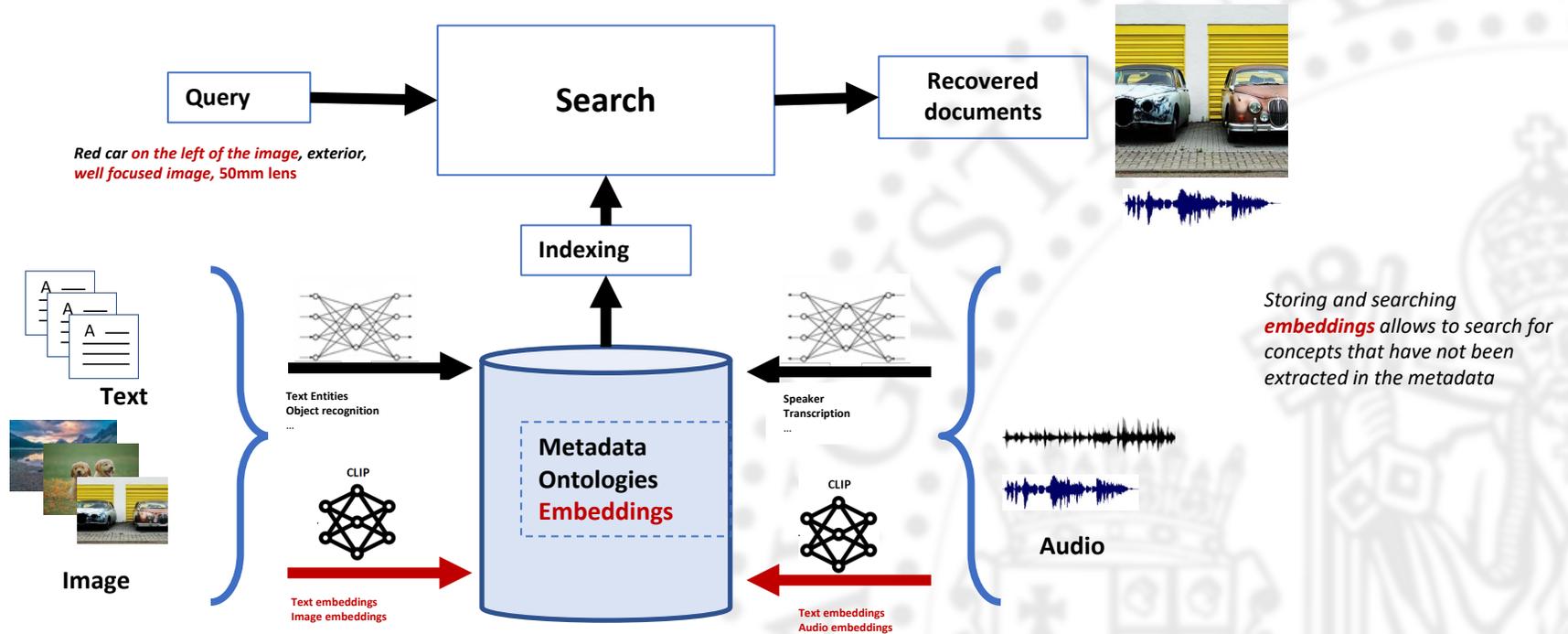
## Búsqueda y recuperación de información



# Búsqueda



## Búsqueda y recuperación de información

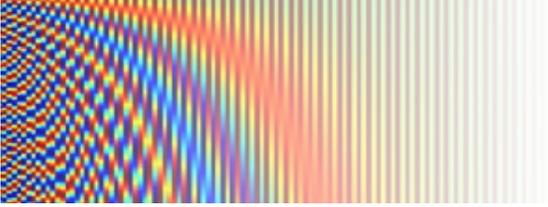


# RAG

## • Ejemplo de sistema RAG

1. Transcribimos los ateneos mediante un sistema estado del arte (basado en transformers)
2. Obtenemos representaciones vectoriales de los textos y su resumen (transformer)
3. Al crear resúmenes se corrigen muchos errores del reconocedor por el contexto





# RAG

- **Transformers: ejemplo asistente sobre Ateneos**

Qué oportunidades se han presentado relacionadas con el coche de hidrógeno.

-----

OPORTUNIDADES DE LA TECNOLOGÍA DEL HIDRÓGENO COMO VECTOR ENERGÉTICO. Fernando Palacín Arizón. 25/10/2017

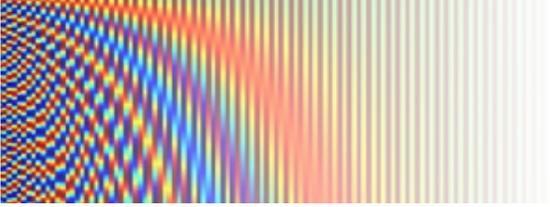
Las oportunidades que se presentan relacionadas con el coche de hidrógeno incluyen una mayor autonomía en comparación con los coches eléctricos, un tiempo de carga más rápido de 2-3 minutos al recargar el depósito de hidrógeno en lugar de la batería, y una reducción en la necesidad de mantenimiento. Además, se menciona que los fabricantes de automóviles están desarrollando coches de hidrógeno y se espera que mejoren en términos de prestaciones y costos en los próximos años.

BMW I, LA ECONOMÍA CIRCULAR APLICADA A LA AUTOMOCIÓN. Leonardo Santamaría Hernández. 16/05/2018

El texto menciona que la compañía tiene un acuerdo de intercambio con Toyota en el que les proporcionaron motores y ellos proporcionaron tecnología para desarrollar nuevos modelos con hidrógeno. Aunque no se proporcionan detalles específicos sobre las oportunidades relacionadas con el coche de hidrógeno, se menciona que en el futuro es posible que cambien a vehículos de hidrógeno.

LA ENERGÍA: UNA OPORTUNIDAD PARA EL DESARROLLO ECONÓMICO. Rafael Moliner. 29/04/2015

En la charla se menciona que el coche de hidrógeno es una alternativa a la crisis del petróleo y al cambio climático. Se habla sobre tecnologías existentes para implementar la economía del hidrógeno en el transporte. Se menciona que se puede producir hidrógeno a partir del carbono sin emitir CO2 y a precios más competitivos. También se menciona la posibilidad de almacenar electricidad en pilas en lugar de producir hidrógeno por electrólisis. Se afirma que el futuro de los coches es híbrido, con batería para una autonomía eléctrica de 200-300 km y una batería de combustible para mayor autonomía. Además, se habla del desarrollo actual de laboratorio que sugiere que se puede duplicar la capacidad de los supercondensadores.



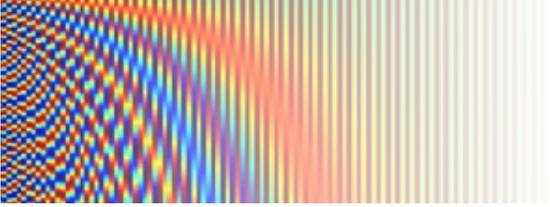
# RAG

- **Transformers: ejemplo asistente sobre Ateneos**

¿ Qué oportunidades se presentan para Teruel en el campo de la aeronáutica ?

-----

En el campo de la aeronáutica, Teruel tiene oportunidades en el negocio de reciclaje de aviones, ya que se espera que se reciclen más de 15,000 aviones en la zona, lo que representa el 10% de todas las aeronaves del mundo. Además, la creación de un 30% más de rutas aéreas en los próximos 20 años representa una ventaja para este negocio en Teruel. También se destaca que el aeropuerto de Teruel es considerado maravilloso y tiene tasas baratas, lo que atrae a la aviación ejecutiva o general desde varios lugares de España y de otros países europeos. Por lo tanto, hay oportunidades para el crecimiento y desarrollo en el campo de aeronáutica en Teruel.



# RAG

- **Transformers: ejemplo asistente sobre Ateneos**

¿Cuáles son los retos más importantes a los que se enfrenta el Aragonés como lengua?

-----

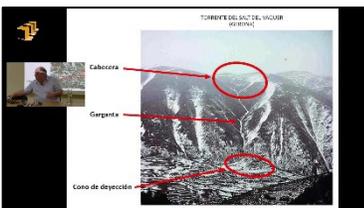
Los retos más importantes a los que se enfrenta el aragonés como lengua son la estigmatización y la falta de transmisión intergeneracional. El hablante menciona que los hablantes tienen derecho a abandonar su lengua, pero esto a menudo ocurre debido a la estigmatización que rodea al aragonés. Además, la falta de transmisión de padres a hijos es crucial para la supervivencia de la lengua. Por lo tanto, es importante que la lengua se viva y se transmita a las nuevas generaciones, y que se aprecie como parte del patrimonio lingüístico de Aragón. También se menciona la importancia de construir una estrategia de política lingüística a largo plazo para el aragonés, basada en el consenso y la participación de la comunidad. Esto implica promover el aragonés de una manera práctica y cotidiana, así como utilizarlo como una estrategia lingüística en lugar de centrarse solo en su estudio filológico.



Q5.txt

# Casos de uso

- Transformers: ejemplo asistente sobre Ateneos



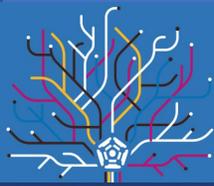
'slide with a landscape picture of mountain village'

- Ateneo 340, título/autor/año: DESAFÍOS DEL CAMBIO CLIMÁTICO: IMPACTOS, MITIGACIÓN Y ADAPTACIÓN. José M<sup>a</sup> Cudrat Prats. 26/10/22 .

- Ateneo 318, título/autor/año: LOS PINOS NO SON HIJOS DEL FRANQUISMO: SU "MALA PRENSA" ES UN MITO. Luis Alfonso Gil Sánchez. 17/02/21

- Ateneo 341, título/autor/año: EL AGUA NO SALE DE LOS RÍOS:HAGAMOS RESTAURACIÓN HIDROLÓGICO-FORESTAL DE CUENCAS. Ignacio Pérez-Soba Díez del Corral. 9/11/22

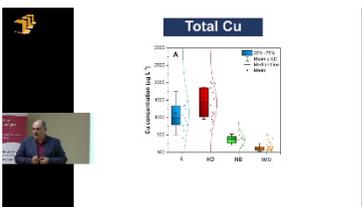
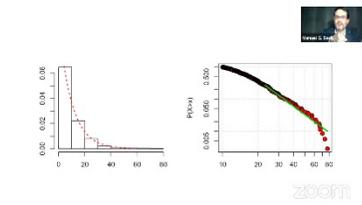
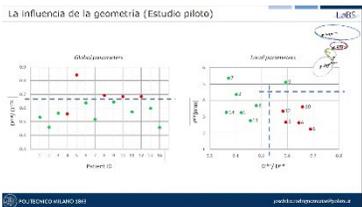
- Ateneo 337, título/autor/año: AL MARGEN DE PREJUICIOS URBANOS: CORTAR NO ES PERJUDICAR. Miguel Cabrera. 27/04/22 . data/es/ateneo/YC4m\_ss07h8/YC4m\_ss07h8.img/r200215.jpg





# RAG

- Ejemplo RAG, búsquedas visuales



'slide with a results graph plot'

- Ateneo 347, título/autor/año: EXPERIENCIAS DE MEDICINA in-silico. José Félix Rodríguez Matas. 1/03/23 .

- Ateneo 346, título/autor/año: MAiCRO: CÓMO DETECTAR DAÑO ACOLAR CON UNA FOTOGRAFÍA. Alejandro Consejo. 15/02/23 .

- Ateneo 316, título/autor/año: INVESTIGANDO SOBRE LA INVESTIGACIÓN: MÁS ALLÁ DEL TALENTO. Manuel G. Bedia. 20/01/2021 .

- Ateneo 350, título/autor/año: MICROMUESTREO, ESPECTROMETRÍA ATÓMICA E INTELIGENCIA ARTIFICIAL, NUEVAS POSIBILIDADES PARA EL DIAGNÓSTICO MÉDICO. Martín Resano. 19/04/23 .

# RAG

## • Ejemplo RAG, búsquedas visuales



'image with a person standing next to a slide with hands raised'

- Ateneo 275, título/autor/año: LA AERONÁUTICA DESDE UNA NUEVA PERSPECTIVA DE GÉNERO.  
Alejandro Ibrahim Perera. 01/02/2017. data/es/ateneo/90rMFE6Fi6M/90rMFE6Fi6M.img/r200154.jpg

- Ateneo 260, título/autor/año: ATAPUERCA Y EVOLUCIÓN HUMANA: ¿CÓMO SABEMOS QUIÉN ES QUIÉN Y CUÁL ES SU ANTIGÜEDAD? .  
Gloria Cuenca Bescós. 18/11/2015.

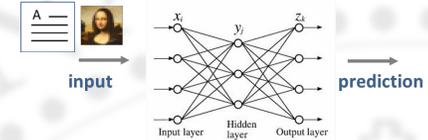
- Ateneo 306, título/autor/año: EMPRENDIMIENTO EN LA ERA DE LA TRANSFORMACIÓN DIGITAL.  
Ana Monreal Vidal. 4/12/2019

- Ateneo 329, título/autor/año: SERVICIOS E INFRAESTRUCTURA DE AWS Y LA PRÓXIMA REGIÓN EN ARAGÓN.  
Javier Ramírez. 15/12/21 .

# Search and description

## Descripción imágenes o vídeos

Los modelos de lenguaje recientes son capaces de analizar fragmentos de vídeo e imágenes y responder a preguntas



## GPT-4V(ision)



"Alright, HERE WE GO! The game is **TIED at 115!**  
 The crowd is on their FEET!  
 The clock's ticking down! Dribbling at the top...  
 He's waiting for the **FINAL SHOT!**  
 JUST UNDER 6 SECONDS LEFT!  
 He steps back...  
 TAKES THE THREE...  
 YES! HE NAILS THE THREE-POINTER AT THE BUZZER!  
 OH, BABY, WHAT A SHOT!  
 The crowd is going NUTS!  
 Teammates are ECSTATIC! Celebration on the court as CONFETTI flies!  
 The arena is ELECTRIC, FANS are hugging, the team is swarmed with joy! What a CLUTCH PERFORMANCE to win the GAME **and the SERIES!**  
 Unbelievable FINISH! Absolute MADNESS in the arena! And the hugs keep coming, WHAT A NIGHT!"

## Scoring panel

- Tied at 115
- Final shot (13.4s 4th quarter)
- Series is 3 - 1

# Search and description

## Descripción imágenes o vídeos

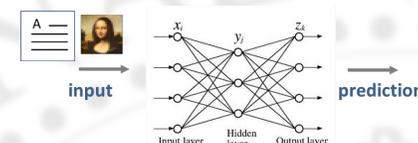
Los modelos de lenguaje recientes son capaces de analizar fragmentos de vídeo e imágenes y responder a preguntas

### GPT-4Vision



### Resuelve el problema

Para resolver este acertijo visual, debemos buscar una conexión o patrón entre las imágenes mostradas: un caballo, una hoja, un avión, una luna, un elefante, una casa y una oreja. Una forma de encontrar la solución es intentar asociar cada imagen con una palabra o letra que pueda formar el nombre de una prenda de vestir. En este caso, la estrategia podría ser pensar en las iniciales de cada objeto en un idioma específico (como el español): Caballo Hoja Avión Luna Elefante Casa Oreja. Ahora, si tomamos la primera letra de cada palabra enumerada, obtenemos: "C", "H", "A",



# Modelos multimodales

- Qwen 2.5 VL/Qwen2.5-Omni

Bai, Shuai et al. Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923

Jin Xu, et al. Qwen2.5-Omni Technical Report. arXiv preprint arXiv:arXiv:2503.20215

- Modelos multimodales open source creados por Alibaba

<https://github.com/QwenLM/Qwen2.5-VL>

