Cursos Extraordinarios

Verano 2024

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

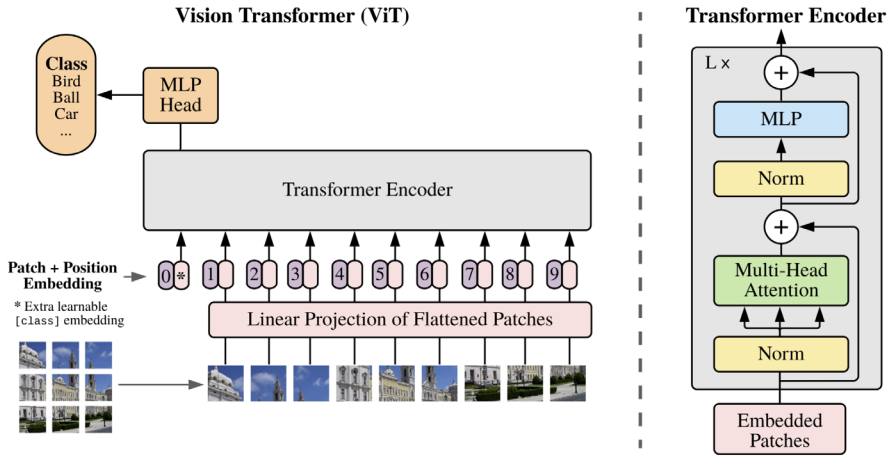
Zaragoza, del 3 al 5 de julio

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

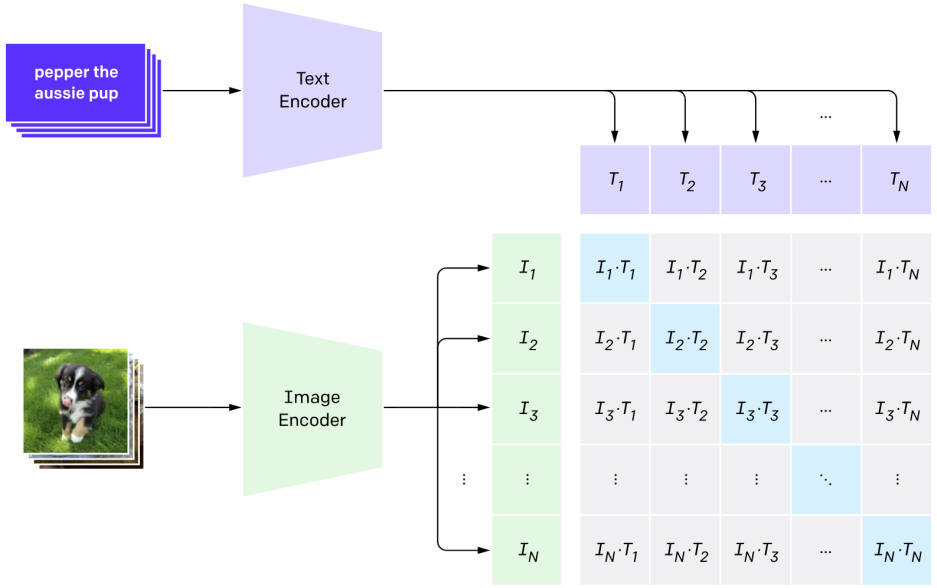


Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Model size and parameters

Modelos multimodales

- **CLIP (Contrastive Language-Image Pre-Training)**
  - Se entrenan dos redes una para representar texto y la otra imágen
  - Contrastive los
    - 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/

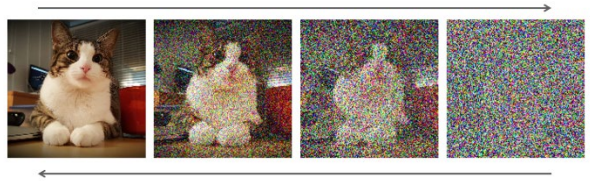
Modelos multimodales

• Difusión

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

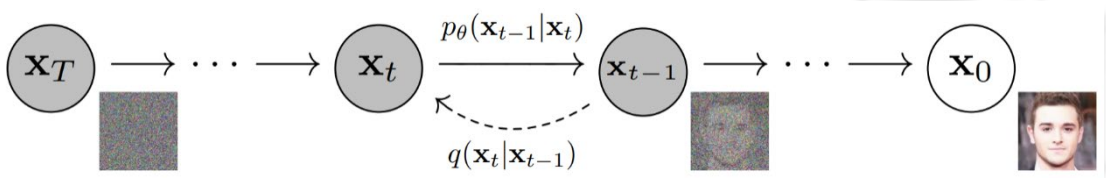
Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. arXiv preprint arXiv:2105.05233.

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

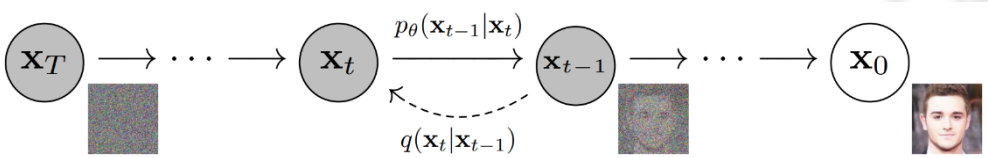


$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

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

Modelos multimodales

- **Difusión dirigida por texto**



Cómo incorporamos información del texto o de una clase?

  - Generación no condicionada,
    - donde μ, Σ se predicen de una red con entrada x_{t+1}

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

  - Generación condicionada

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

Constante

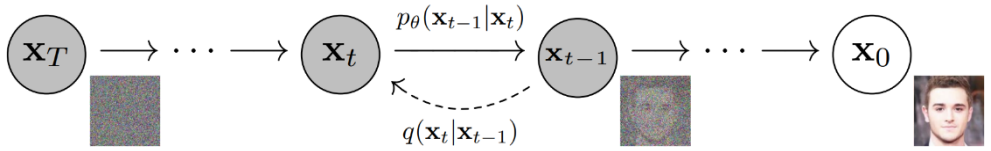
Aproximada por una serie de Taylor, el primer término **gradiente**

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

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

Modelos multimodales

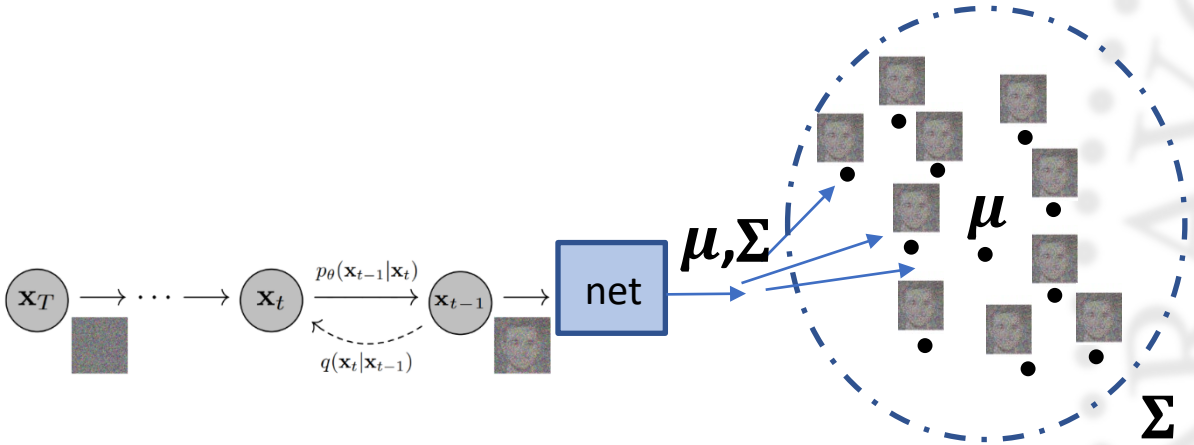
- **Difusión dirigida por texto**



Cómo incorporamos información del texto o de una clase?

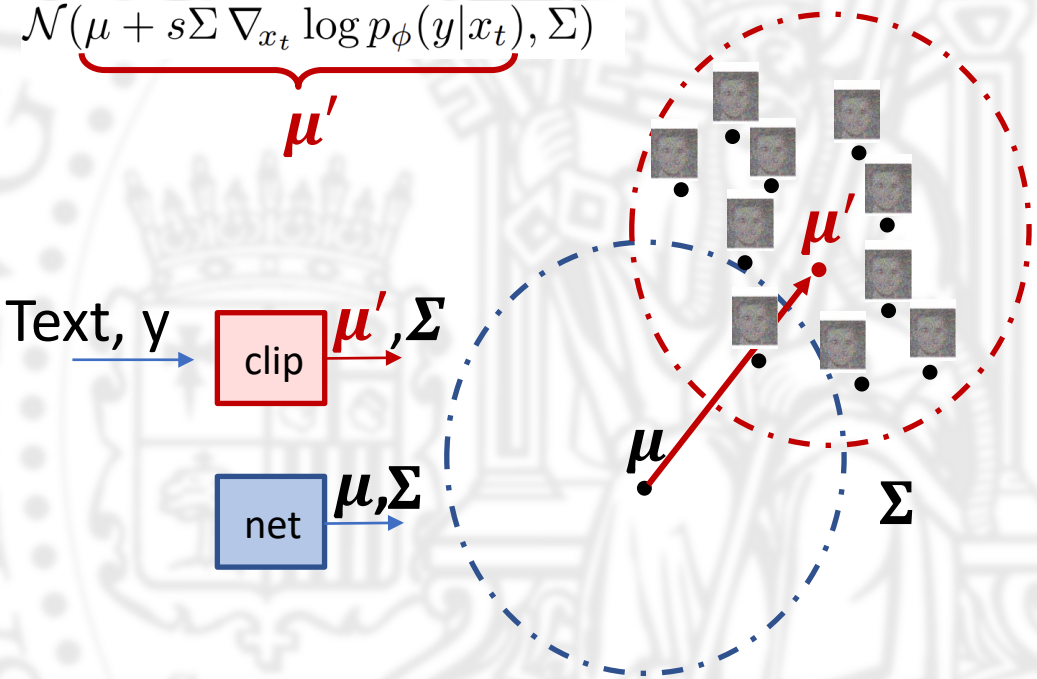
Generación no condicionada

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



Generación condicionada

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



Modelos multimodales

- **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

treehouse in the style of studio ghibli animation
@danielrussruss

unreal engine gaudi house in a field of poppy

VQGAN+CLIP 2021

A wooden Spanishlaptop of 1650 found the library of El Escorial

Difussion+CLIP 2022 (dalle2)

  - 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 ***

Modelos multimodales

- **Recuperación de información**
  - Video y audio

Modelos multimodales

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

Modelos multimodales

- **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.

Modelos multimodales

- **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.
```

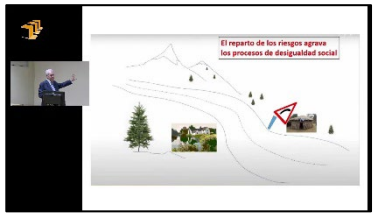
Modelos multimodales

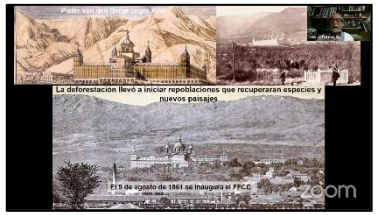
- **Transformers: ejemplo asistente sobre Ateneos**

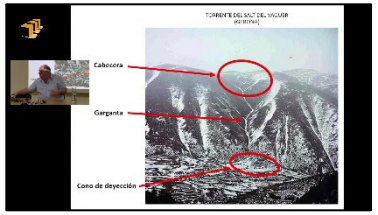
```
¿Cuáles son los retos más importantes a los que se enfrenta el Aragonés como
lengua?
------------------------------------------------------------
Los retos más importantes a los que se enfrenta el aragonés como lengua son la estigmatización y la falta de
transmisión intergeneracional. El hablante 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.
```

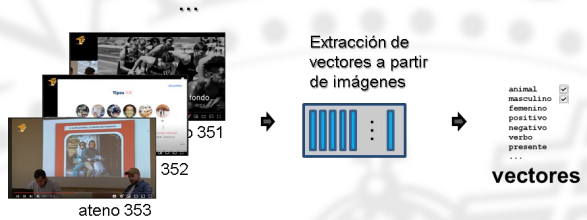
Modelos multimodales

- **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 Mª Cudrat Prats. 26/10/22 .

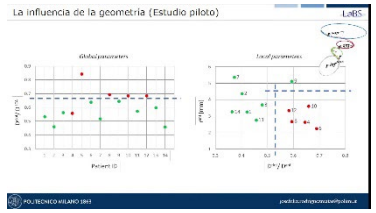
- Ateneo 318, título/autor/año: LOS PINOS NO SON HIJOS DEL FRANQUISMO: SU "MALA PRENSA" ES UN MITO. Luis Alfonso Gil Sánchez. 17/02/21

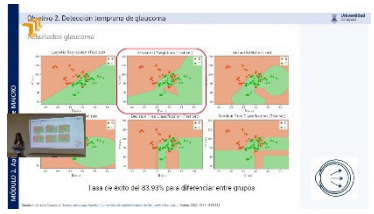
- Ateneo 341, título/autor/año: EL AGUA NO SALE DE LOS RÍOS:HAGAMOS RESTAURACIÓN HIDROLÓGICO-FORESTAL DE CUENCAS. Ignacio Pérez-Soba Díez del Corral. 9/11/22

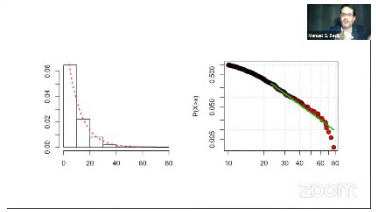
- Ateneo 337, título/autor/año: AL MARGEN DE PREJUICIOS URBANOS: CORTAR NO ES PERJUDICAR. Miguel Cabrera. 27/04/22 . data/es/ateneo/YC4m_ss07h8/YC4m_ss07h8.img/r200215.jpg

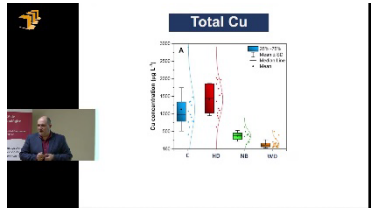
Modelos multimodales

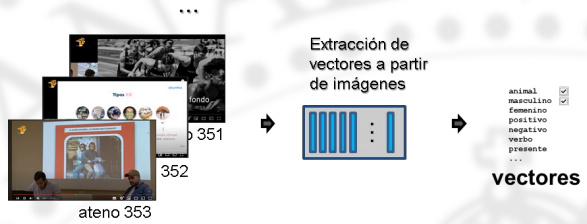
- **Ejemplo RAG, búsquedas visuales**











'slide with a results graph plot'

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

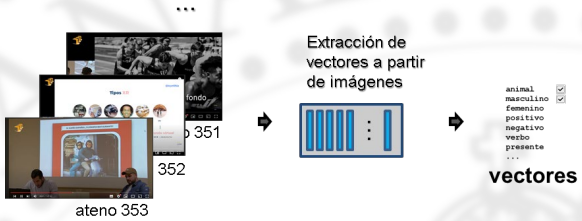
- Ateneo 346, título/autor/año: MAiCRO: CÓMO DETECTAR DAÑO ACOLAR CON UNA FOTOGRAFÍA. Alejandro Consejo. 15/02/23 .

- 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 .

Modelos multimodales

- **Ejemplo RAG, búsquedas visuales**



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

- **Ateneo 275, título/autor/año: LA AERONAÚTICA 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**

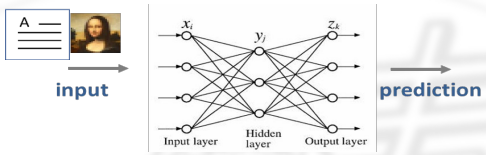
- **Ateneo 329, título/autor/año: SERVICIOS E INFRAESTRUCTURA DE AWS Y LA PRÓXIMA REGIÓN EN ARAGÓN. Javier Ramírez. 15/12/21 .**

Modelos multimodales

- **Descripción de segmentos de video**
  - Descripción de un segmento de video con el estilo de un comentarista deportivo

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

Modelos multimodales

GPT4 vision

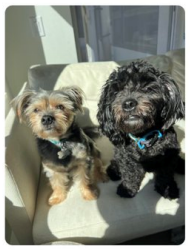
-

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", "L", "E", "C", "O". Observando estas letras, podemos armar la palabra "CHALECO", que es una prenda de vestir. Así que la respuesta es "chaleco".

Modelos multimodales

GPT4 vision

- 

What breed are my dogs?

CHATGPT

Based on the image you've provided:

The dog on the left appears to have characteristics of a Yorkshire Terrier, with its tan face and black body, as well as its silky fur texture.

The dog on the right seems to resemble a Poodle or Poodle mix, given its curly black coat.

However, it's essential to note that accurately determining a dog's breed based solely on an image can be challenging. For a definitive answer, you might consider genetic testing or consulting with a

Dime cómo se llama esta planta y qué cuidados requiere.

La planta en la imagen es conocida comúnmente como "Lirio de la Paz" o "Espatifilo" (Spathiphyllum). El Espatifilo es una planta muy popular en interiores gracias a su elegante follaje y sus flores distintivas en forma de espata blanca.

Aquí tienes algunos cuidados básicos para mantener un Espatifilo saludable:

Luz: Prefiere la luz indirecta o filtrada. La exposición directa al sol puede quemar sus hojas.

Riego: Le gusta el suelo constantemente húmedo, pero sin encharcamiento. Es importante dejar que el suelo se seque ligeramente entre riegos.

Humedad: Aprecia una buena humedad ambiental. Si el aire es demasiado seco, especialmente en invierno, se pueden pulverizar sus hojas con agua.

Temperatura: Prefiere temperaturas cálidas y consistentes, entre 18°C y 26°C.

Fertilización: Puede ser abonado con un fertilizante equilibrado para plantas de interior aproximadamente cada 6-8 semanas durante la temporada de crecimiento.

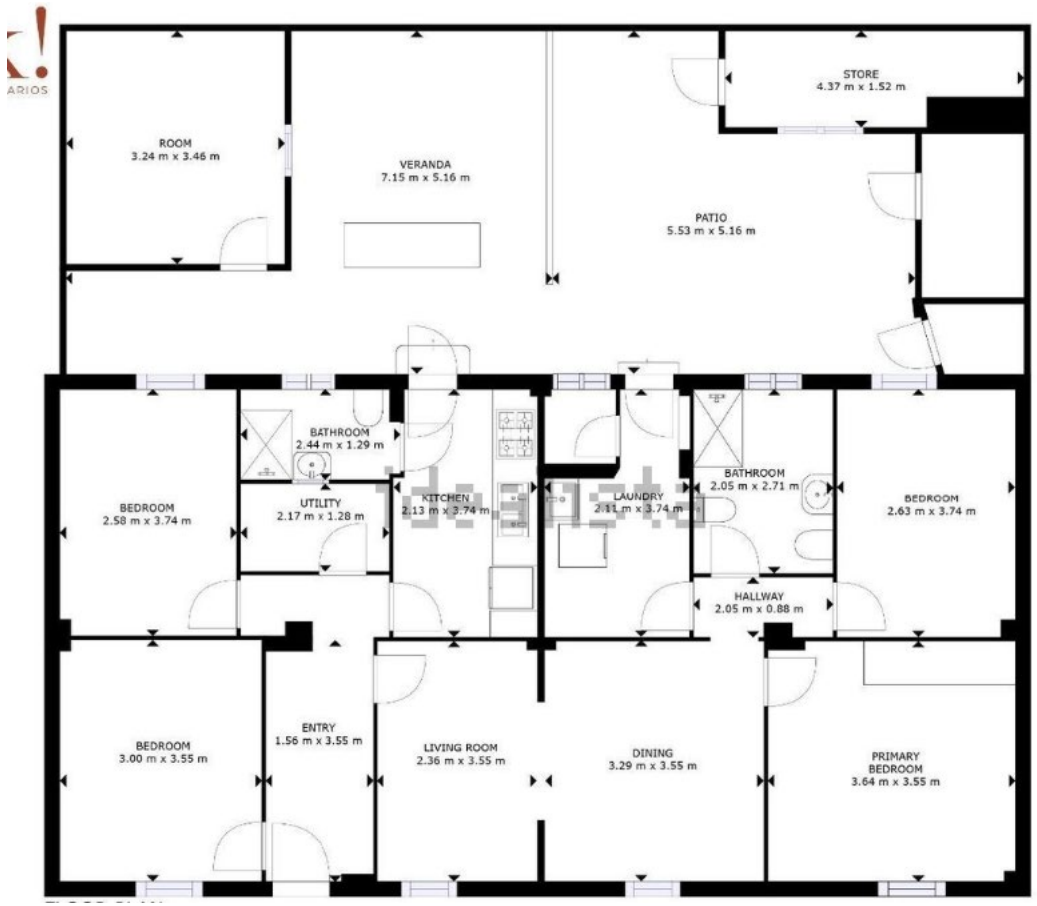
Poda: Eliminar las flores desvanecidas y las hojas muertas o amarillentas ayuda a promover un nuevo crecimiento.

Trasplante: Puede ser necesario trasplantar cada par de años si la planta crece demasiado para su maceta actual.

Como con todas las plantas, es importante monitorear su espécimen individual y ajustar los cuidados según sea necesario.

Modelos multimodales

GPT4 vision

- 

How many sq meters has every room in this house.

To calculate the area of each room in square meters, you multiply the width by the length for each room. Please note that the calculations are based on the measurements provided and might be subject to rounding errors. Here are the approximate calculations:

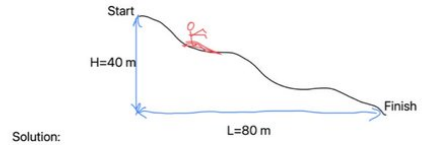
Room: 3.24 m x 3.46 m = 11.21 sq m
Veranda: 7.15 m x 5.16 m = 36.89 sq m
Store: 4.37 m x 1.52 m = 6.64 sq m
Patio: 5.53 m x 5.16 m = 28.54 sq m
Bedroom (Bottom Left): 2.58 m x 3.74 m = 9.65 sq m
Utility: 2.17 m x 1.28 m = 2.78 sq m
Kitchen: 2.13 m x 2.74 m = 5.84 sq m
Laundry: 2.11 m x 3.74 m = 7.89 sq m
Bathroom (Middle Left): 2.05 m x 2.71 m = 5.56 sq m
Hallway: 2.05 m x 0.88 m = 1.80 sq m
Bedroom (Middle Right): 2.63 m x 3.74 m = 9.84 sq m
Bathroom (Bottom Right): 2.44 m x 1.29 m = 3.15 sq m
Entry: 1.56 m x 3.55 m = 5.54 sq m
Living Room: 2.36 m x 3.55 m = 8.38 sq m
Dining: 3.29 m x 3.55 m = 11.68 sq m
Primary Bedroom: 3.64 m x 3.55 m = 12.92 sq m
Bedroom (Top Left): 3.00 m x 3.55 m = 10.65 sq m
These are the areas of each labeled room according to the floor plan provided.

Modelos multimodales

Prompt

Here's a solution to a Physics problem by a student.

23. A skier slides down the frictionless slope as shown. What is the skier's speed at the bottom?



Solution:

The conservation of energy means that the energy of the skier is same at the beginning and at the end.

Start: $E = mgL$

End: $E = \frac{1}{2}mv^2$

$\Rightarrow mgL = \frac{1}{2}mv^2$

$\Rightarrow v = \sqrt{2gL} = \sqrt{2 \times 9.81 \times 80} = 39.6\ m \cdot s^{-1}$

Try to reason about the question step by step.

Did the student get the correct answer? If the solution is wrong, please explain what is wrong and solve the problem.

Make sure to use LaTeX for math and round off the final answer to two decimal places.

Response

The student did not get the correct answer. The student made a mistake in the calculation of the potential energy at the start of the slope. The potential energy at the start of the slope is given by mgh, where m is the mass of the skier, g is the acceleration due to gravity, and h is the height of the slope. The student used the length of the slope, L, instead of the height, H, in the calculation of the potential energy.

The correct solution is:

The conservation of energy means that the total energy of the skier is the same at the beginning and at the end of the slope.

At the start of the slope, the skier has only potential energy, which is given by mgh.

At the end of the slope, the skier has only kinetic energy, which is given by $\frac{1}{2}mv^2$

Therefore, we can write:

$$mgh = \frac{1}{2}mv^2$$

Canceling out the mass m on both sides, we get:

$$gh = \frac{1}{2}v^2$$

Solving for v, we get:

$$v = \sqrt{2gh}$$

Substituting the given values, we get:

$$v = \sqrt{2 \times 9.81 \times 40} = 28.01 \text{ m/s}$$

Therefore, the skier's speed at the bottom of the slope is 28.01 m/s.