

# Sistemas Recomendadores

## IIC-3633

Deep Learning en Sistemas Recomendadores  
Parte 2

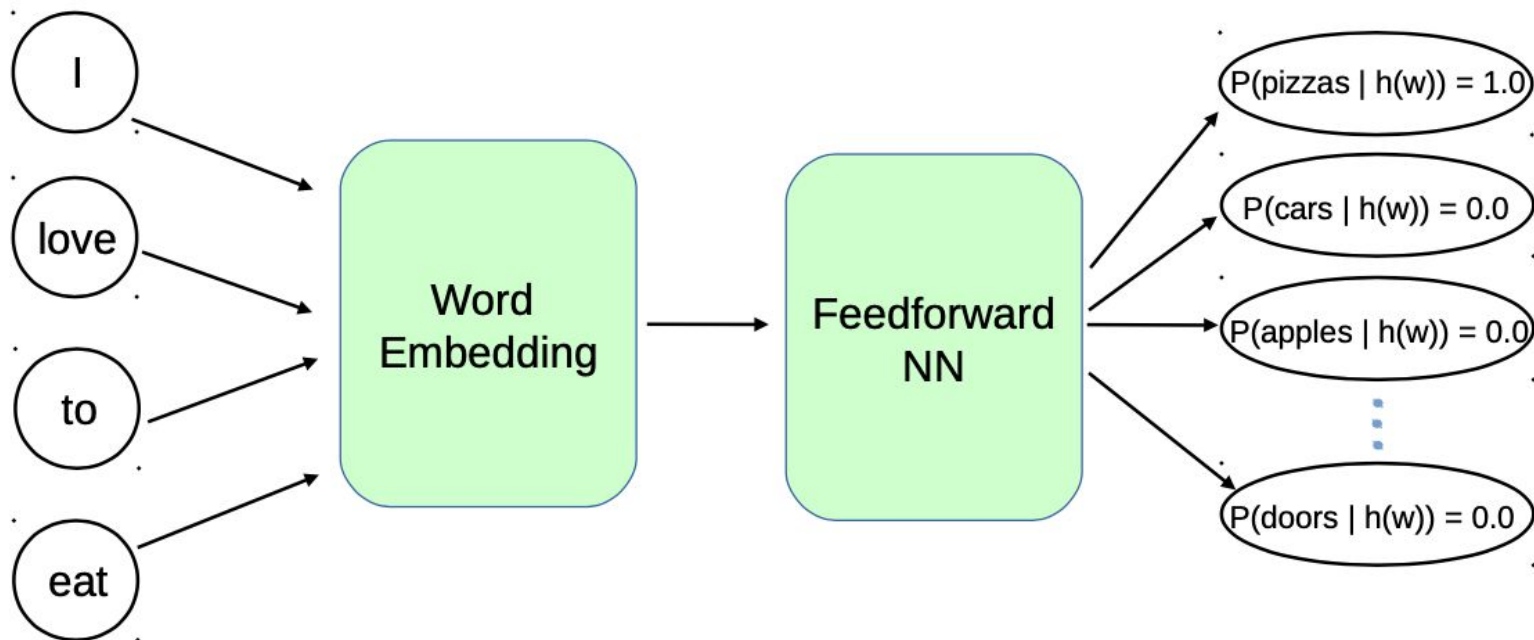
# Esta clase

1. Repaso modelos de lenguaje
2. Modelos multimodales
3. Deep Learning para recomendación (Multimodal)

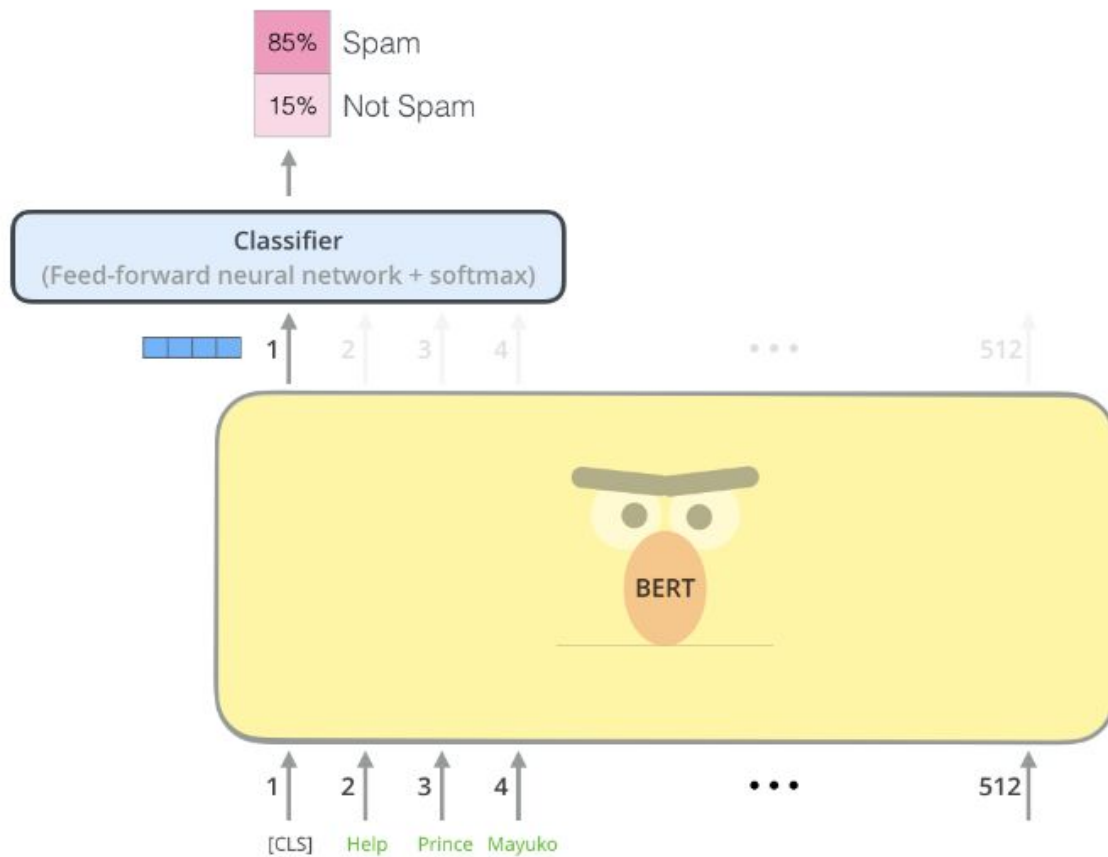
# Word2vec

$x = \text{"I love to eat"} \rightarrow \text{CONTEXTO } h(w)$

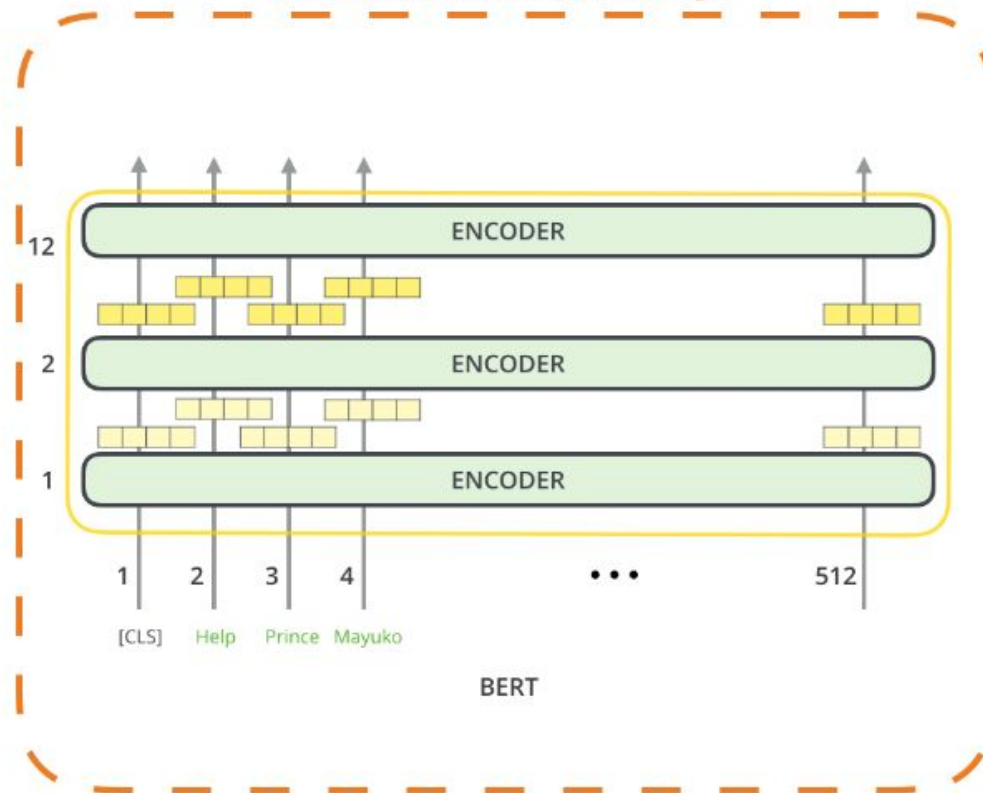
$y = ? \rightarrow \text{SIGUIENTE PALABRA}$



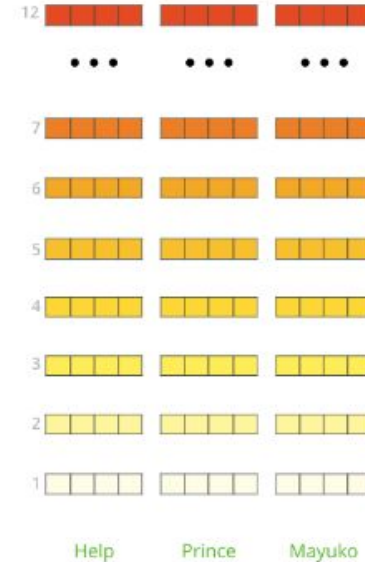
# BERT



### Generate Contextualized Embeddings

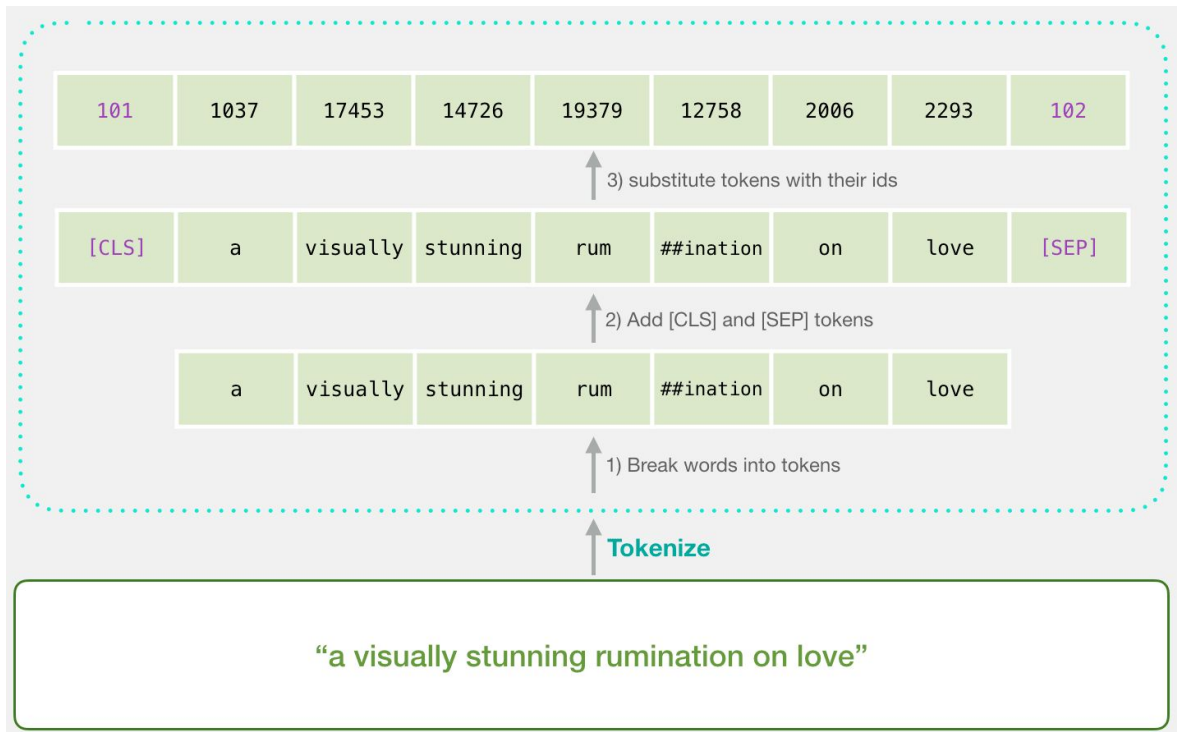


The output of each encoder layer along each token's path can be used as a feature representing that token.



But which one should we use?

# BERT tokenizer



BERT trae incorporado un tokenizer que:

1. Divide palabras en dos partes de manera que la segunda se pueda utilizar de nuevo.  
(ej. ama + **#dos**  
da +
1. Agregar tokens especiales [CLS] que representa un texto completo y [SEP] para denotar separación si tiene más de una oración.
2. Convertir cada token al índice en el vocabulario.

# Práctico

<https://colab.research.google.com/drive/1Loy1RyUORTo2CsRWuolnMtkXWm3SN7eo?usp=sharing>

# Uso de deep learning para representación multimodal



---

## Learning Transferable Visual Models From Natural Language Supervision

---

Alec Radford<sup>\*1</sup> Jong Wook Kim<sup>\*1</sup> Chris Hallacy<sup>1</sup> Aditya Ramesh<sup>1</sup> Gabriel Goh<sup>1</sup> Sandhini Agarwal<sup>1</sup>  
Girish Sastry<sup>1</sup> Amanda Aspell<sup>1</sup> Pamela Mishkin<sup>1</sup> Jack Clark<sup>1</sup> Gretchen Krueger<sup>1</sup> Ilya Sutskever<sup>1</sup>



# Motivación

Sacar provecho de información de imágenes y de texto en conjunto, capturando relaciones o patrones comunes entre ellos.

El objetivo es si se entrena un modelo con ambos tipos de dato juntos se obtengan mejores embeddings que cada uno por separado.

# Claves de CLIP: Contrastive Language Image Pre-training

**Multi-Modal:** CLIP entiende tanto imágenes como texto simultáneamente.

**Zero-Shot:** Predice descripciones de texto para imágenes no vistas sin fine-tuning.

**Open-source:** Accesible y extensible para diversos usos y desarrollos en IA.

**Contrastive learning:** aprender a acercar textos e imágenes que se refieren a lo mismo y alejar las que no tienen relación.

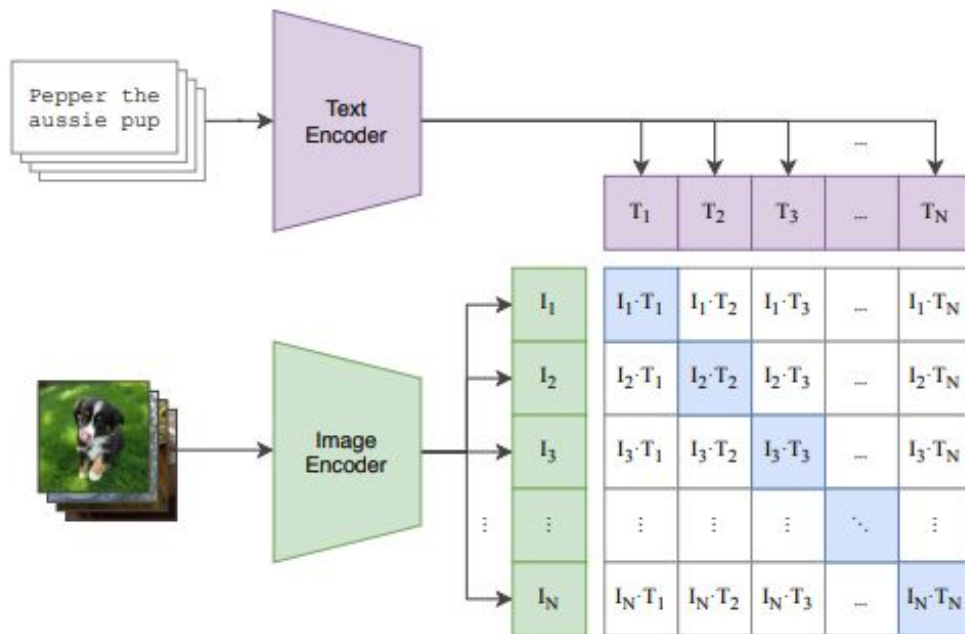
# Arquitectura y aprendizaje

El Text Encoder y el Image Encoder se entrenan juntos para acercar imágenes y textos que corresponden entre sí.

(1) Contrastive pre-training

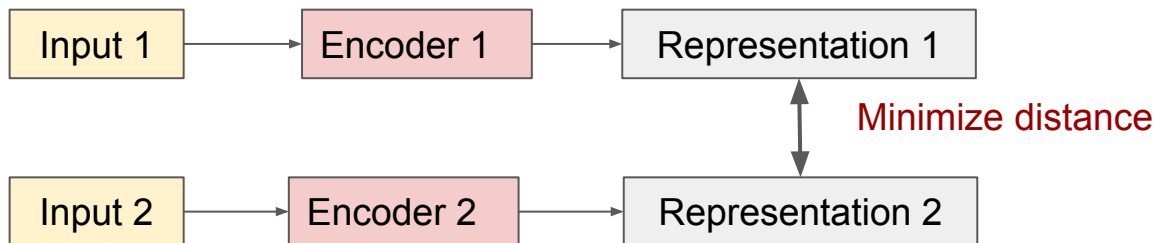
El **contrastive pre-training** busca maximizar la similaridad coseno de la diagonal de la matriz  $N \times N$  de los embeddings de imágenes y de textos.

Se actualizan los pesos de ambos encoders.

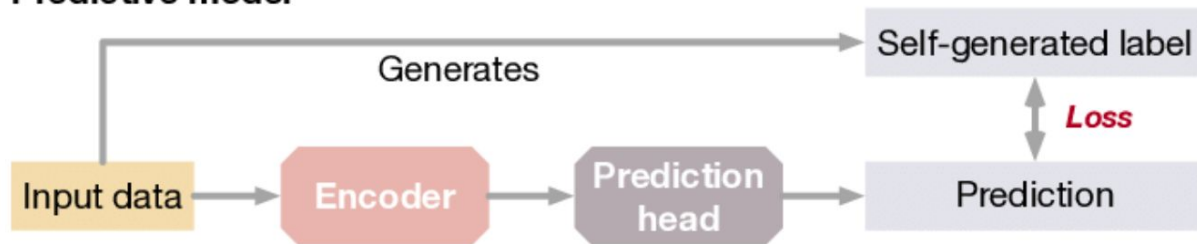


# Contrastive vs Predictive learning

## Contrastive Model



## Predictive model



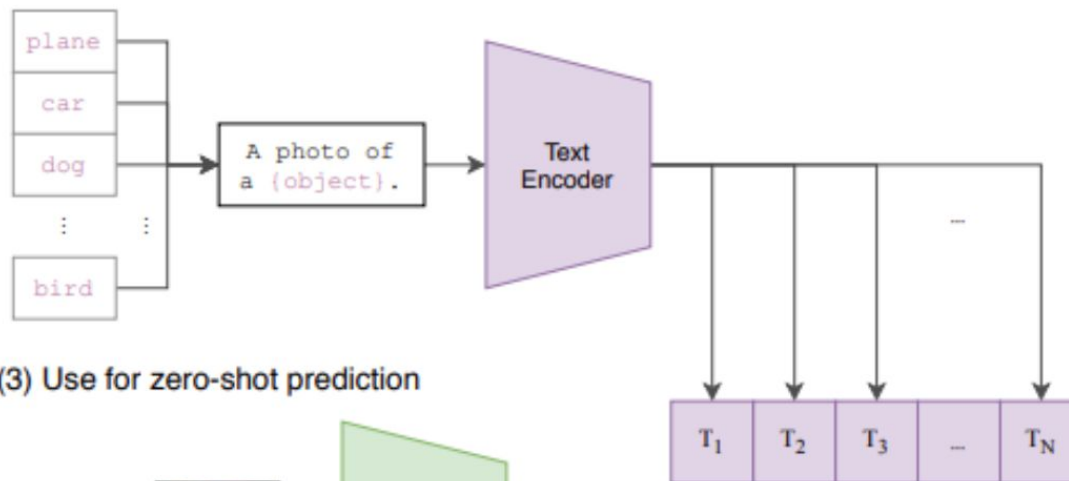
## Trucos para data augmentation

prompt engineering  
para data  
augmentation:

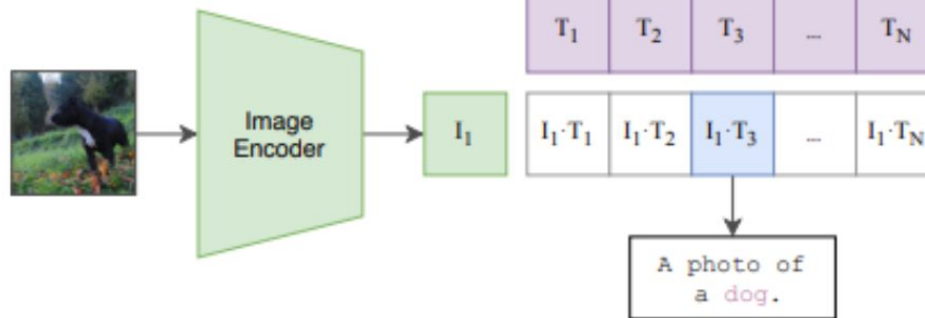
“a photo of a  
<object>”

usando un dataset  
de clasificación de  
imágenes.

(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



# Predicción zero-shot

El concepto de zero-shot es utilizar el conocimiento aprendido durante el entrenamiento para abordar tareas que no se vieron durante este proceso.

Dado que CLIP se entrena para entender tanto texto como imágenes, puede relacionar cualquier imagen con cualquier texto.

Incluso si nunca ha visto esa combinación particular durante el entrenamiento.



# ¿Por qué es útil?

## **Flexibilidad:**

- Permite a los usuarios aplicar CLIP a una variedad de tareas sin necesidad de entrenamiento adicional.
- Hace el match entre un texto de largo variable con una imagen.
- Ej. para clasificación necesitamos explícitamente el nombre de la clase.

## **Eficiencia:**

- Facilita el uso del modelo en escenarios prácticos y diversos, ahorrando tiempo y recursos sin la necesidad de fine-tuning.

# Resultados

# Baseline: Visual n-grams

## Learning Visual N-Grams from Web Data

Ang Li\*  
University of Maryland  
College Park, MD 20742, USA  
angli@umiacs.umd.edu

Allan Jabri Armand Joulin Laurens van der Maaten  
Facebook AI Research  
770 Broadway, New York, NY 10025, USA  
{ajabri, ajoulin, lvdmaaten}@fb.com

Paper estado del arte en 2021 para predecir  
n-gramas de texto dado un embedding de una  
imagen usando CNNs.

Aceptado en ICCV 2017.



**Predicted  $n$ -grams**  
lights  
Burning Man  
Mardi Gras  
parade in progress

**Predicted  $n$ -grams**  
GP  
Silverstone Classic  
Formula 1  
race for the

**Predicted  $n$ -grams**  
navy yard  
construction on the  
Port of San Diego  
cargo

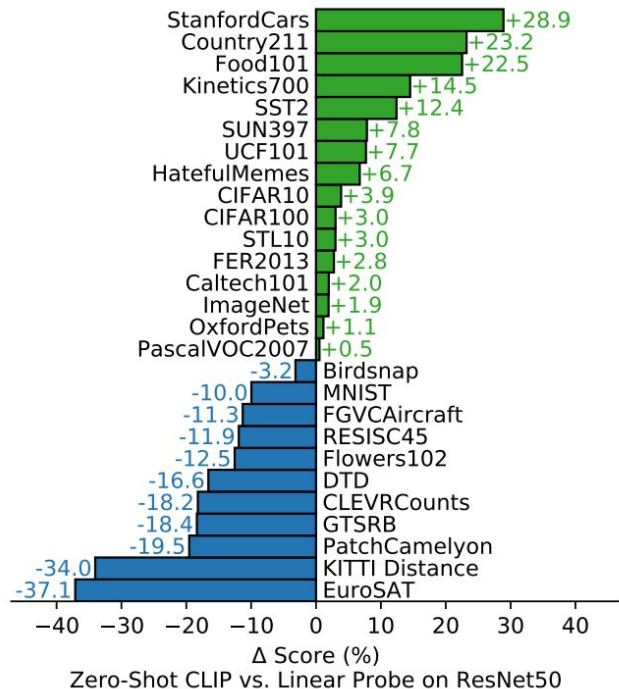
Figure 1. Four high-scoring visual  $n$ -grams for three images in our test set according to our visual  $n$ -gram model, which was trained *solely* on *unsupervised* web data. We selected the  $n$ -grams that are displayed in the figure from the five highest scoring  $n$ -grams according to our model, in such a way as to minimize word overlap between the  $n$ -grams. For all figures in the paper, we refer the reader to the supplementary material for license information.

## Resultados accuracy: CLIP zero-shot vs Visual N-Grams

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	<b>98.4</b>	<b>76.2</b>	<b>58.5</b>

*Table 1.* Comparing CLIP to prior zero-shot transfer image classification results. CLIP improves performance on all three datasets by a large amount. This improvement reflects many differences in the 4 years since the development of Visual N-Grams (Li et al., 2017).

# Resultados en tareas y datasets.



Tareas de Computer Vision donde CLIP Zero-Shot supera (o empeora) con respecto al mejor modelo para cada tarea.

*Figure 5. Zero-shot CLIP is competitive with a fully supervised baseline.* Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

# Generalización de clip a ejemplos fuera de distribución de ImageNet original.

		Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	$\Delta$ Score
ORIGINAL	ImageNet		76.2	76.2	0%
	ImageNetV2		64.3	70.1	+5.8%
	ImageNet-R		37.7	88.9	+51.2%
	ObjectNet		32.6	72.3	+39.7%
	ImageNet Sketch		25.2	60.2	+35.0%
	ImageNet-A		2.7	77.1	+74.4%

Otras  
versiones

# Limitaciones de CLIP

Le va mal en algunos dominios por ejemplo:

- diferenciar modelos de autos
- diferenciar especies de flores
- diferenciar modelos de aviones
- contar número de objetos en una imagen
- no le va bien con texto escrito a mano
- no sabe generar captions, solo sabe decir si un texto corresponde a una imagen.
- no le va bien para textos largos.

# Aplicaciones de CLIP



# Modelos generativos de imágenes dado un prompt

Stable diffusion  
Dall-E

Usan el encoder de CLIP + U-NET, Image Generator y estrategias para mejorar calidad de la imagen de salida de manera iterativa agregando ruido aleatorio.

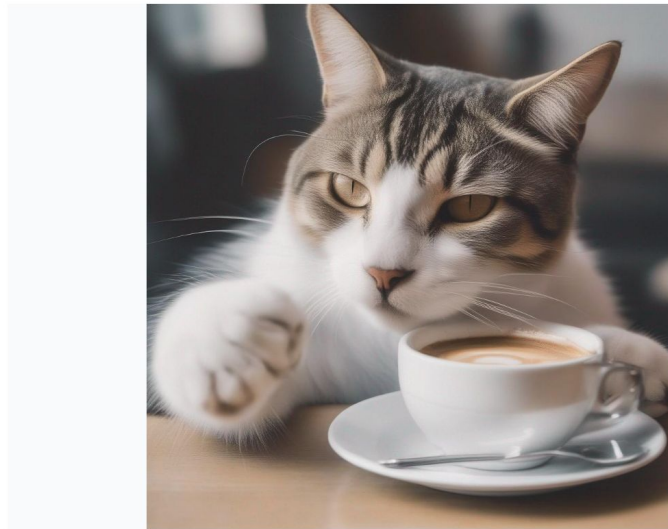
⚡ Hosted inference API ⓘ

📄 Text-to-Image

a cat drinking a coffee

Compute

Computation time on gpu: 6.920 s

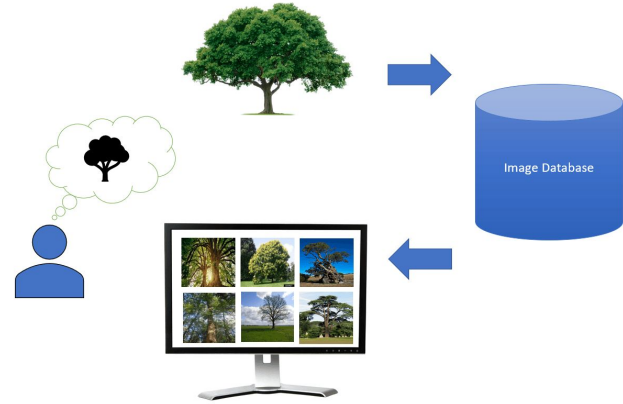


</> JSON Output

🖼 Maximize

# Otras aplicaciones

- Búsqueda de imágenes (visual search).
- Generación de descripciones de imágenes (image captioning).
- Verificación de contenido.
- Análisis de sentimiento visual.
- Apoyo en diagnósticos médicos utilizando imagen de entrada.



## FINDINGS

the patient was imaged in a lordotic position, which distorts the mediastinal contours. within that limitation, the lungs are clear without consolidation or edema. the mediastinum is otherwise unremarkable. the cardiac silhouette is within normal limits for size. no effusion or pneumothorax is noted. no displaced fractures are evident.

# práctico CLIP

[https://colab.research.google.com/drive/1HJaNd\\_2dPPy2Pz6ttzjVasDa9tsaUQ2c?usp=sharing](https://colab.research.google.com/drive/1HJaNd_2dPPy2Pz6ttzjVasDa9tsaUQ2c?usp=sharing)

# Recomendación multimodal

Product  
description

Product  
image

Related  
products

The screenshot displays an Amazon product page for a Polo Ralph Lauren polo shirt. The main product image is a maroon polo shirt with "HOUSTON" printed in gold across the chest. To the left of the main image are three smaller thumbnail images showing different views of the shirt. To the right of the main image is the product description and pricing information. Below the main image is a section titled "Customers who viewed this item also viewed" showing a row of 12 related polo shirts. At the bottom right, there is a "Recommended from our brands" section showing four more polo shirts.

**Product Description:**

Polo Ralph Lauren  
Polo Ralph Lauren Mens Custom Slim Fit Mesh City Polo Shirt  
★★★★☆ (295 ratings) | 64 answered questions  
Price: **\$54.75 - \$134.99**  
Fit: As expected (79%)  
Size: Select Size Chart  
Color: Burgundy Houston

**Features:**

- 100% Cotton
- Polo Ralph Lauren
- Custom Slim Fit
- Features embroidered pony, city, and crest logos
- Two riveted vent gromets under each arm
- Mesh knit with tennis tails

**Customers who viewed this item also viewed:**

Product	Price
Polo Ralph Lauren Mens Custom Slim Fit Big Pony Polo Shirt	\$67.00 - \$127.50
Polo Ralph Lauren Mens Big Pony Custom Slim Fit Three Button Crest Polo	\$55.75 - \$89.99
Polo Ralph Lauren Mens Custom Slim Fit Big Pony Large Polo Shirt	\$41.75 - \$89.00
Polo Ralph Lauren Mens Custom Slim Fit Big Pony Mesh Crest Polo	\$74.99 - \$95.00
Polo Ralph Lauren Mens Big Pony Custom Slim Fit Big Pony Crest Polo	\$54.99 - \$99.99
Polo Ralph Lauren Mens Classic Fit Big Crested Pony Polo Shirt	\$42.00 - \$165.00
Polo Ralph Lauren Mens Big Pony Country Custom Fit Mesh Polo Shirt	\$55.75 - \$99.00
Polo Ralph Lauren Mens Classic Fit Big Crest Big Pony Polo Shirt	\$64.99 - \$98.50
Polo Ralph Lauren Mens Big Pony Custom Slim Fit Mesh Polo Shirt	\$124.75
Polo Ralph Lauren Mens Custom Slim Fit Big Pony Crest Polo Shirt	\$64.99 - \$95.00
Polo Ralph Lauren Mens Big Pony Custom Slim Fit Crested Crest Polo	\$64.99 - \$99.50

**Recommended from our brands:**

Product	Price
Amazon Essentials	\$20.00
Amazon Essentials	\$19.00
Hawaiian Breese	\$29.99
Something for Everyone	\$12.99 - \$16.99

Page 1 of 5

# Extracción automática de features

También conocida como “representation learning”

Texto:

- TF-IDF
- Sentence embeddings

Imágenes:

- CNN

Grafos / Interacciones

- Matrix factorization
- Graph embeddings

# Filtrado colaborativo vs Basado en contenido

## Collaborative filtering

- Asume que el comportamiento determina la compra / interacción
- Tiene mayor capacidad de generar recomendaciones personalizadas en especial MF que aprende vector de usuario.

## Basado en contenido

- Asume que el contenido determina la compra / interacción
- Soluciona el problema de cold-start.

Podemos utilizar ambos juntos?

Si, ya que tanto MF como Deep Learning son extensibles

# Multimodalidad en recomendación

Idea principal: aprender de multiples “modalidades” al mismo tiempo.

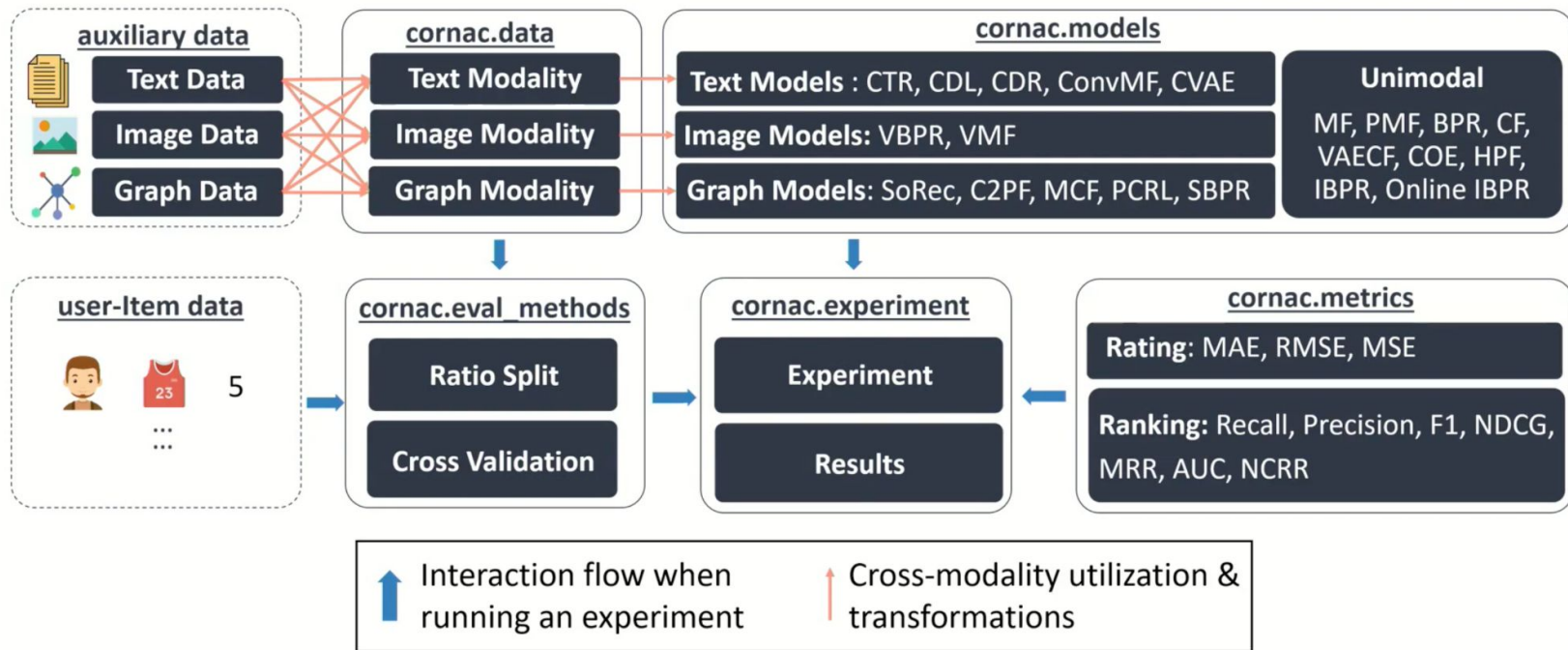
1. Preference feedback
2. Contenido de items (imagen, texto, grafos, etc..)
3. Contenido de usuarios (biografia , red social, etc..)

...

Otras?



# Herramienta para recomendación multimodal: CORNAC



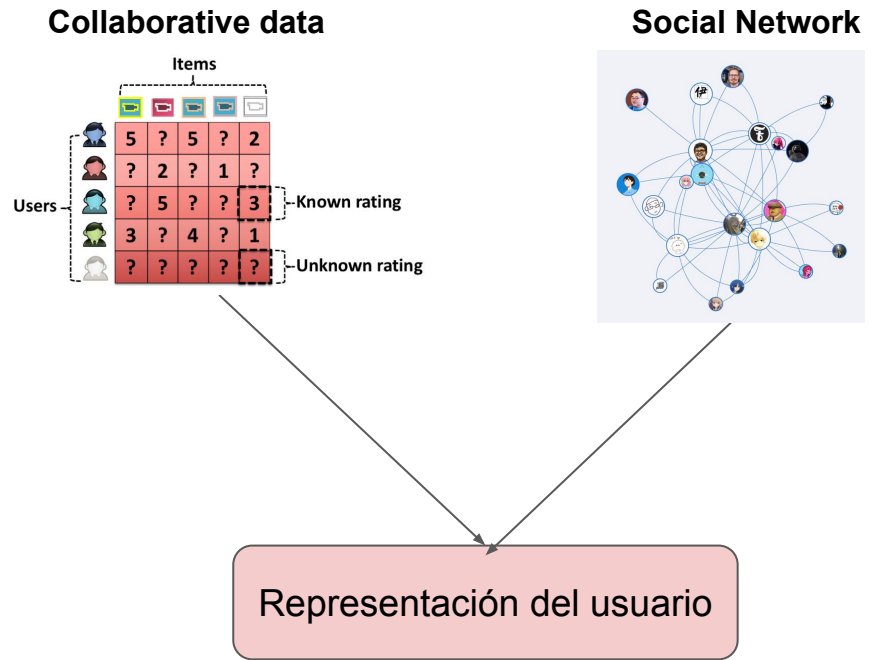
CORNAC: MULTIMODAL RECOMMENDATION LIBRARY

<https://cornac.readthedocs.io/en/latest/>

Multimodalidad para mejorar representación de usuarios

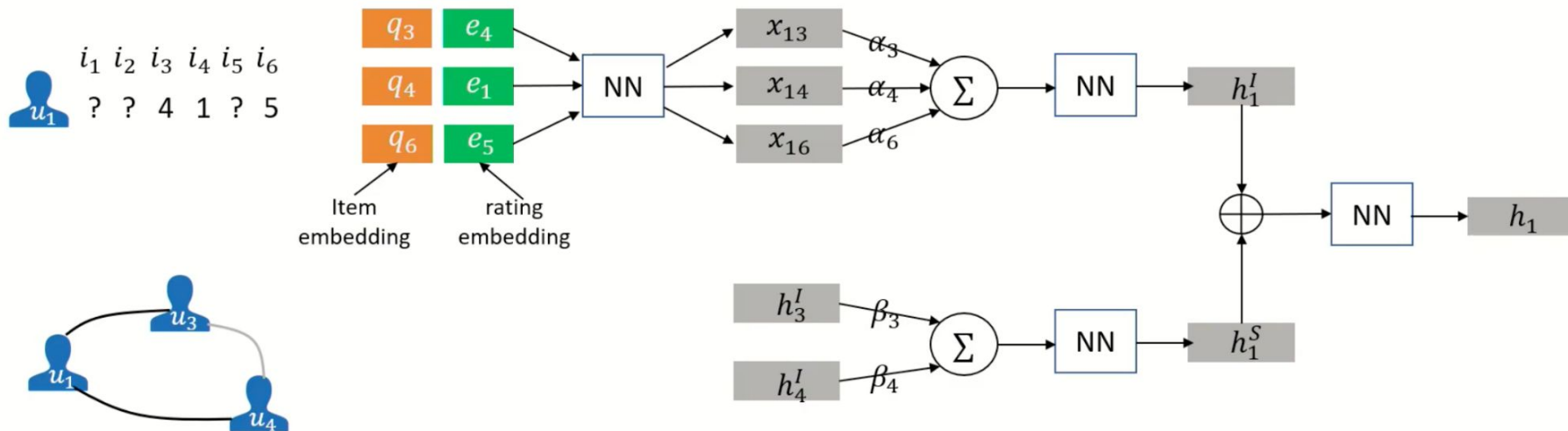
Se pueden utilizar modelos para extraer información de ambas modalidades:

- Filtrado colaborativo.
- Red social de usuarios.



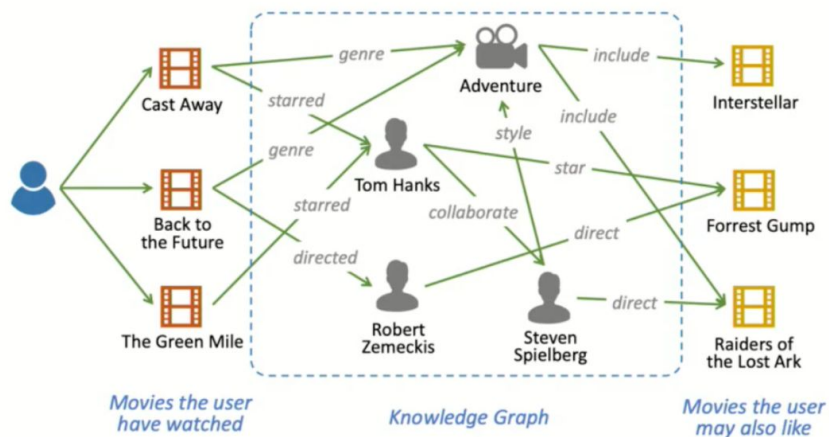
# GNN-based user representation

Fan, Wenqi, et al. "Graph neural networks for social recommendation." WWW. 2019.



Multimodalidad para mejorar representación de items

Nos interesa extraer información de relaciones entre ítems



**Knowledge graph**



Goodthreads Men's Slim-fit Long-Sleeve Solid Oxford Shirt

★★★★☆ · 404 customer reviews | 30 answered questions

Price: \$25.00

Fit: As expected (77%)

Color: Blue

- 100% Cotton
- Imported
- Machine Wash
- This classic, versatile shirt provides a clean, buttoned-up look with a special wash for a soft feel
- Model is 6'1" and wearing a size Medium
- Slim fit: closer-fitting in the chest, slightly tapered through the waist for a tailored look



Goodthreads Men's Slim-fit Long-Sleeve Solid Oxford Shirt  
★★★★☆ · 404 customer reviews | 30 answered questions  
Price: \$25.00



Goodthreads Men's Slim-fit Long-Sleeve Solid Oxford Shirt  
★★★★☆ · 404 customer reviews | 30 answered questions  
Price: \$25.00



Goodthreads Men's Slim-fit Long-Sleeve Solid Oxford Shirt  
★★★★☆ · 404 customer reviews | 30 answered questions  
Price: \$25.00



Goodthreads Men's Slim-fit Long-Sleeve Solid Oxford Shirt  
★★★★☆ · 404 customer reviews | 30 answered questions  
Price: \$25.00

**Ítems adquiridos juntos**

# Por qué es importante la información de items adquiridos juntos

Rara vez consumimos items muy parecidos a los que ya hemos adquirido. eg. lentes de sol.

Consumimos ítems que son:

- complementarios: polera y pantalón.
- alternativas: poleras con distinto estilo / color.

Son aspectos que no se capturan solo con el texto o la imagen del ítem.

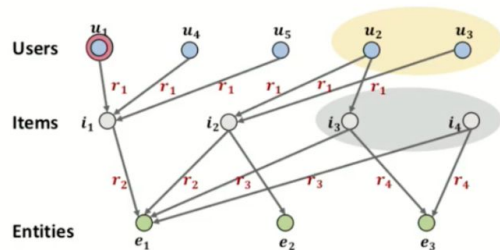


**items sustitutos y complementarios.**

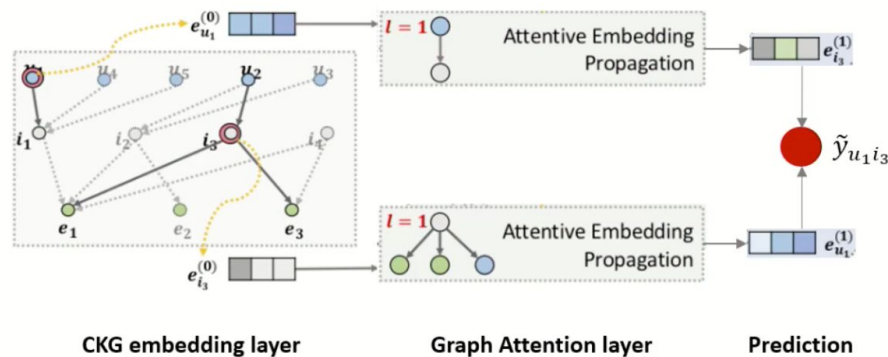


# Knowledge Graph Attention Network (KGAT)

Wang, X., He, X., Cao, Y., Liu, M., & Chua, T. S. Kgat: Knowledge graph attention network for recommendation. *SIGKDD*. 2019 (pp. 950-958).



**Collaborative Knowledge Graph (CKG)**  
User-Item graph  $\cup$  Knowledge Graph



**Illustration of KGAT model**

- Embed entities and relations
- Graph Attention based representation of entities
- Predict user-item interactions

Gracias!