Sistemas Recomendadores
IIC-3633

Deep Learning en Sistemas Recomendadores
Parte 2

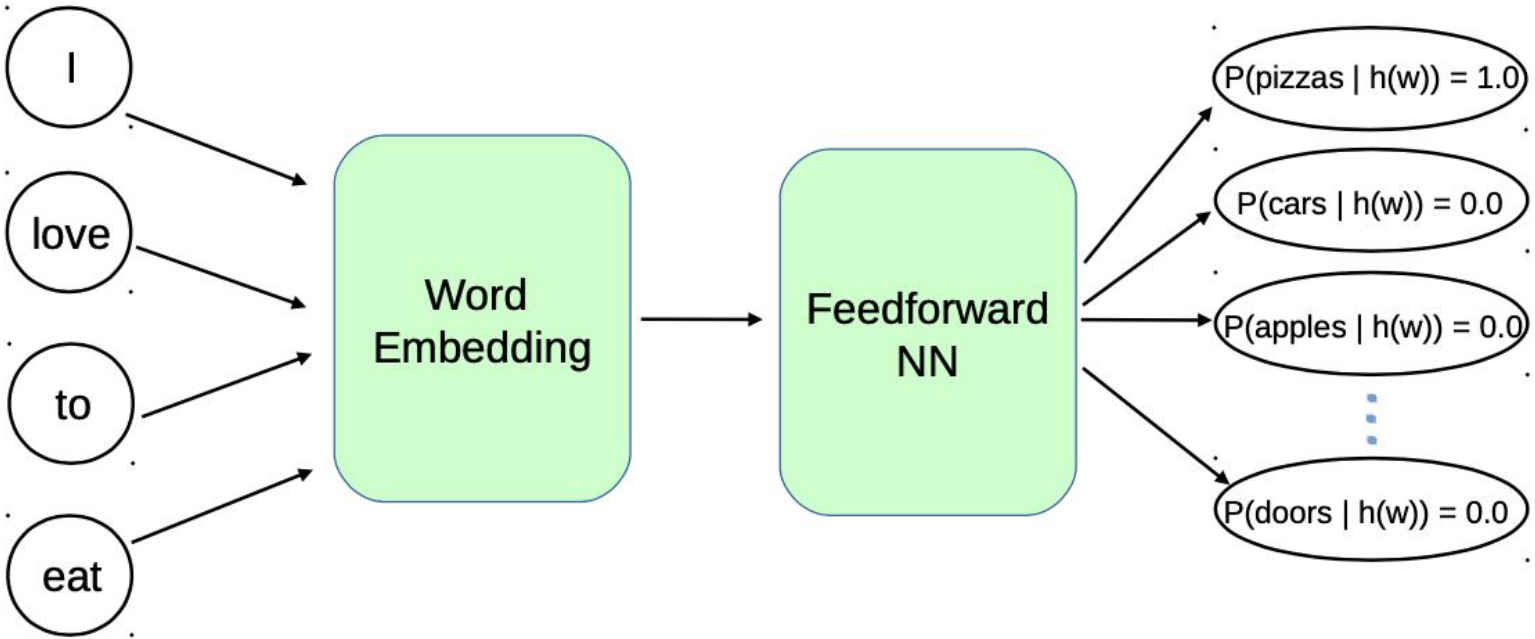
Esta clase

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

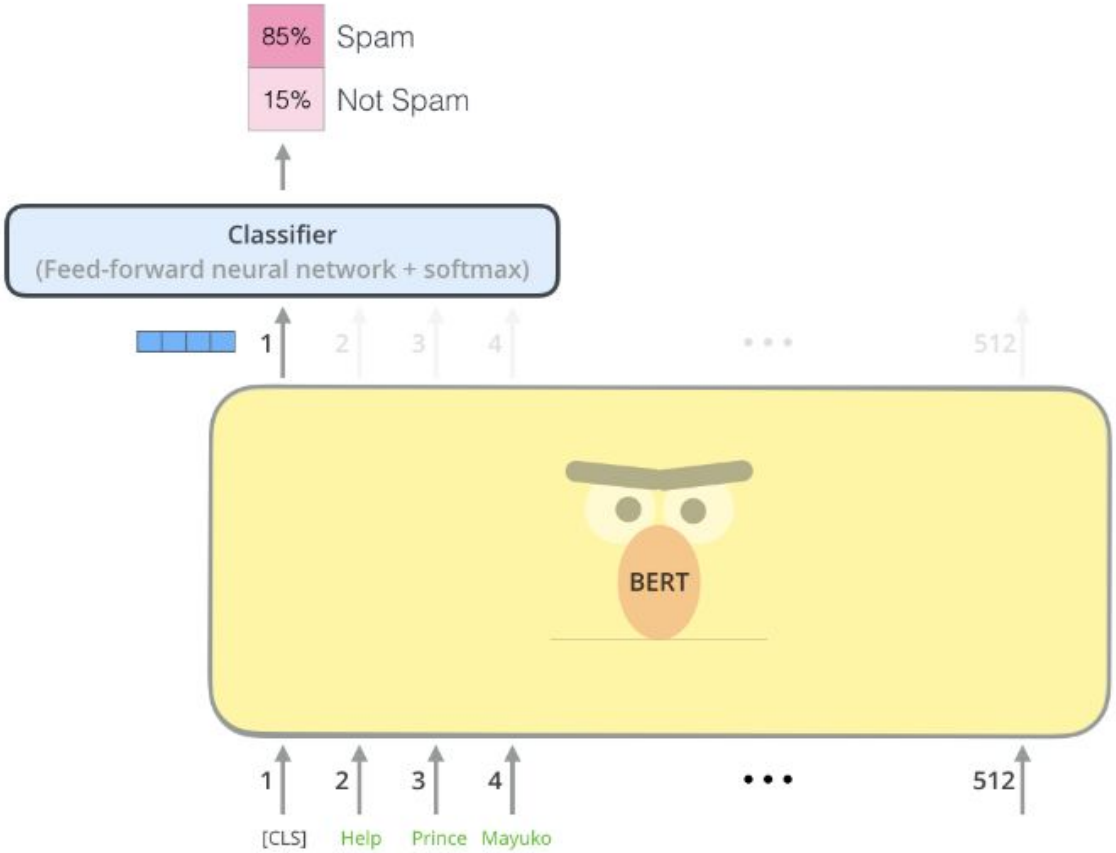
Word2vec

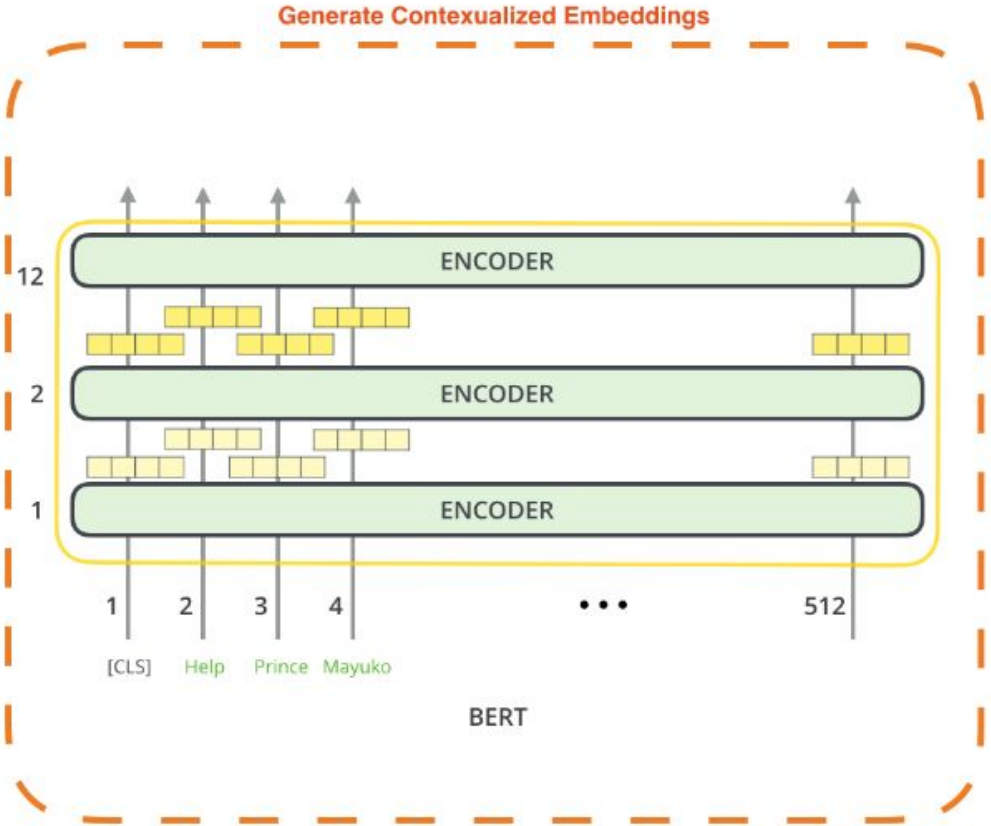
x = “I love to eat” → CONTEXTO h(w)

y = ? → SIGUIENTE PALABRA

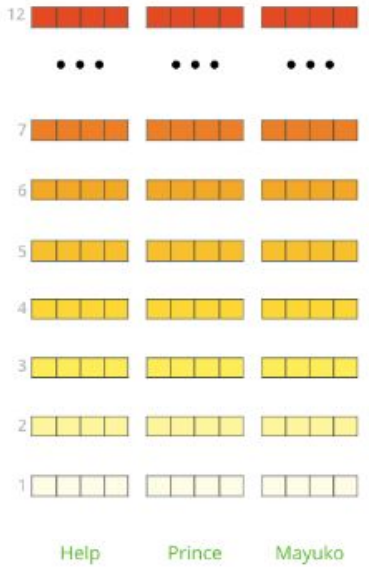


BERT


85%
Spam
15%
Not Spam
Classifier
(Feed-forward neural network + softmax)
1
2
3
4
512
BERT
1
2
3
4
512
[CLS]
Help
Prince
Mayuko

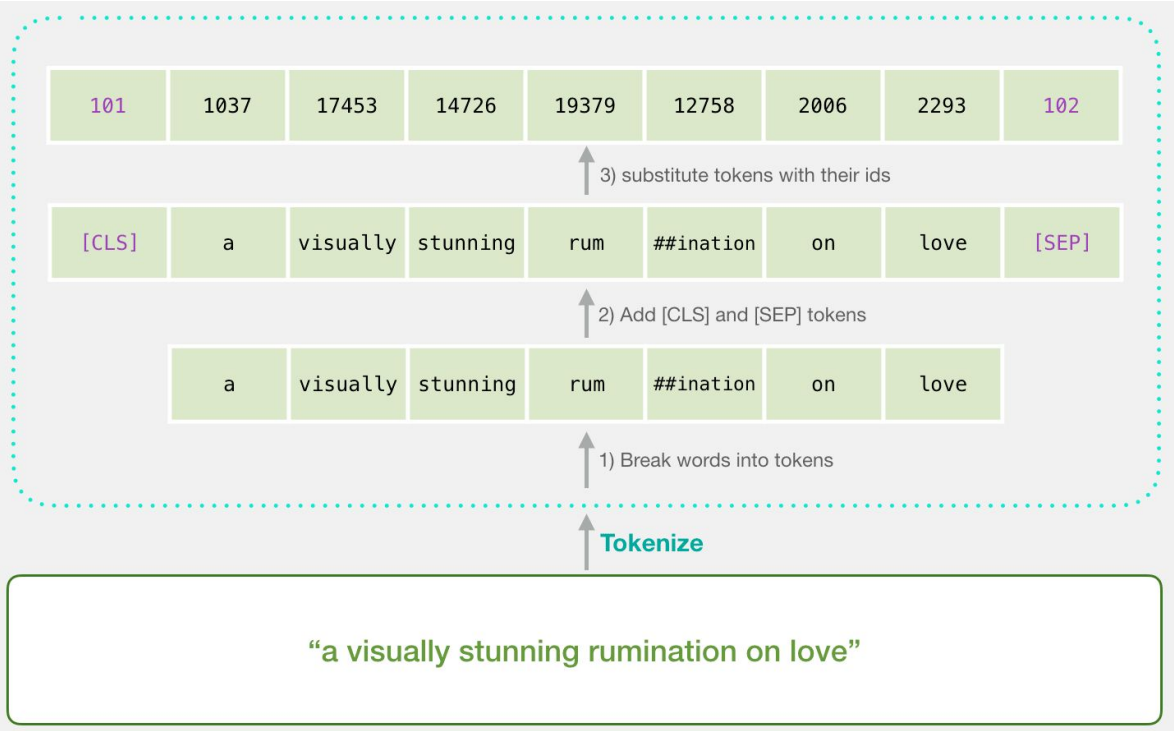


The output of each encoder layer along each token's path can be used as a feature representing that token.



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/1Loy1RyUORTo2CsRWuoInMtkXWm3SN7eo?usp=sharing

Uso de deep learning para representación multimodal

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

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 una 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.

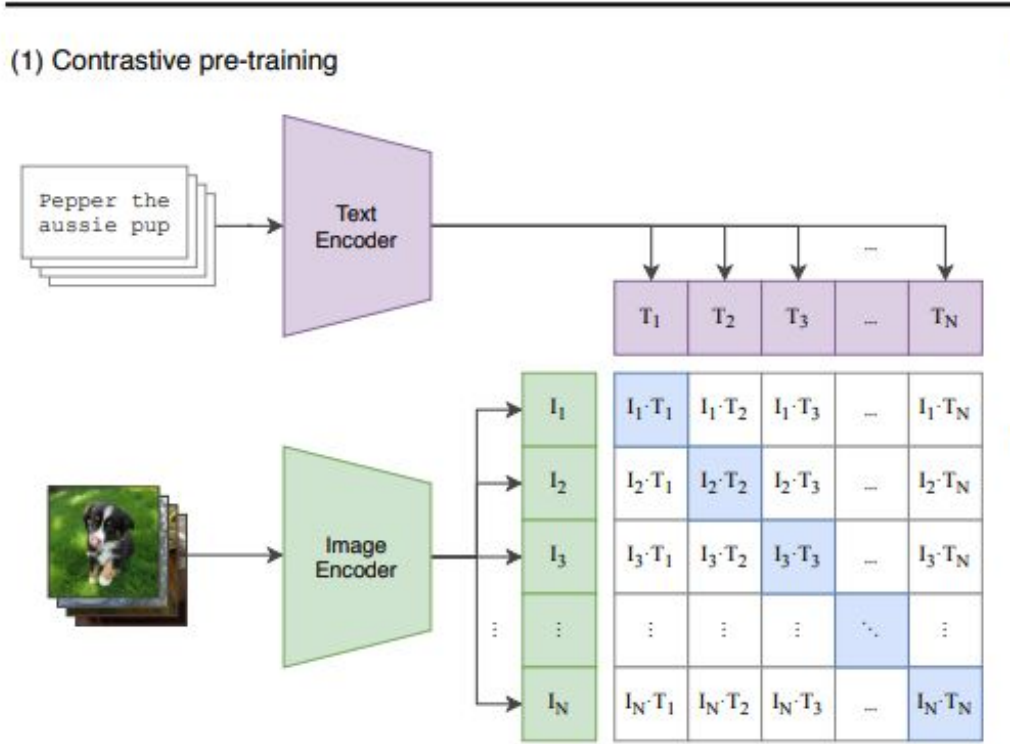
Contrastive learning: aprender a acercar textos e imágenes que se refieren a lo mismo y alejar las que no tienen relación.

Arquitectura y aprendizaje

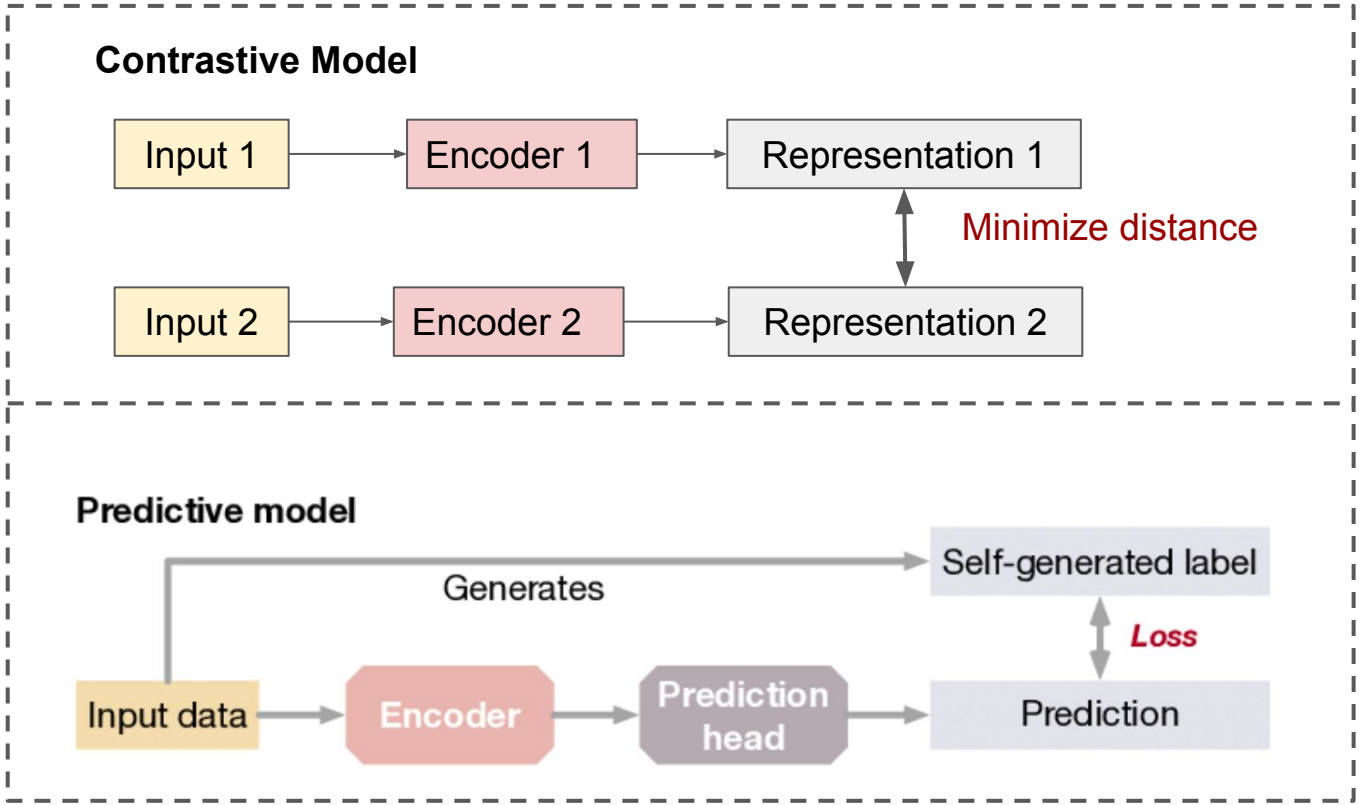
El Text Encoder y el Image Encoder se entrenan juntos para acercar imágenes y textos que corresponden entre sí.

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

Se actualizan los pesos de ambos encoders.



Contrastive vs Predictive learning

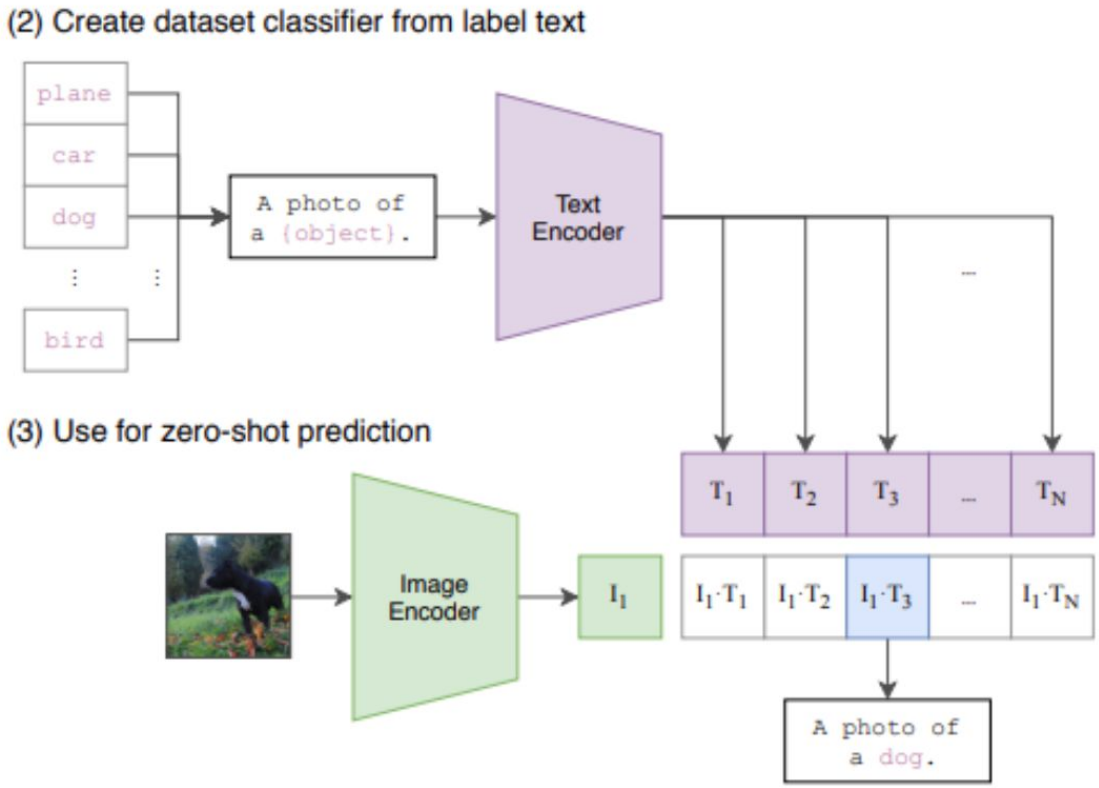


Trucos para data augmentation

prompt engineering para data augmentation:

“a photo of a <object>”

usando un dataset de clasificación de imágenes.



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 imágen.
- 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	**98.4**	**76.2**	**58.5**

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.

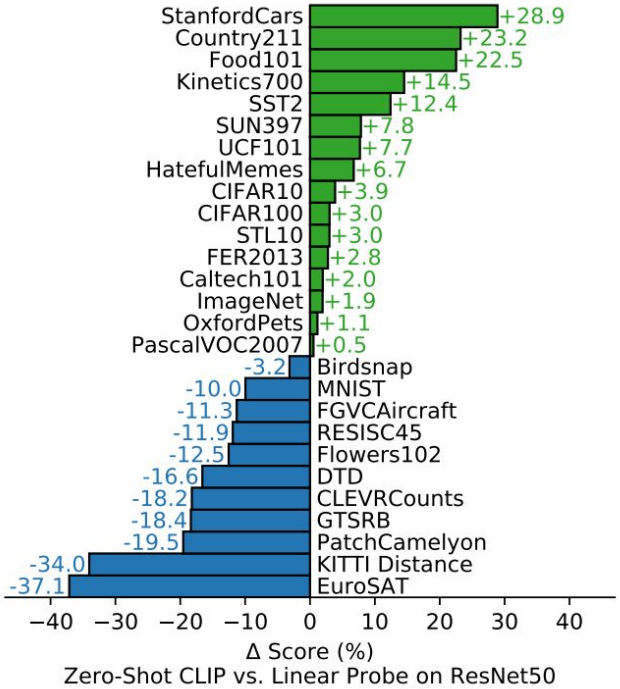
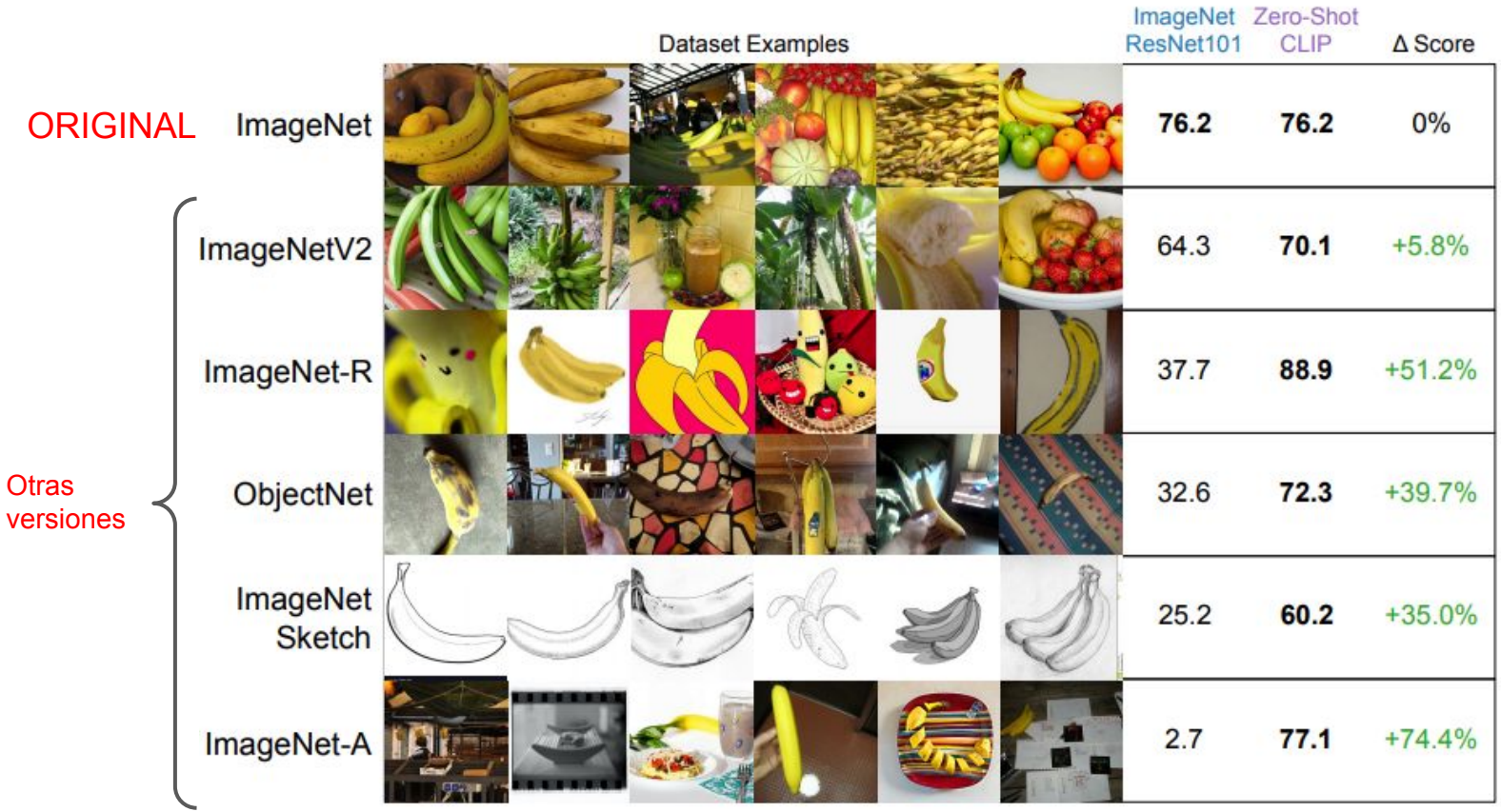


Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

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

Generalización de clip a ejemplos fuera de distribucion de ImageNet original.



Limitaciones de CLIP

Le va mal en algunos dominios por ejemplo:

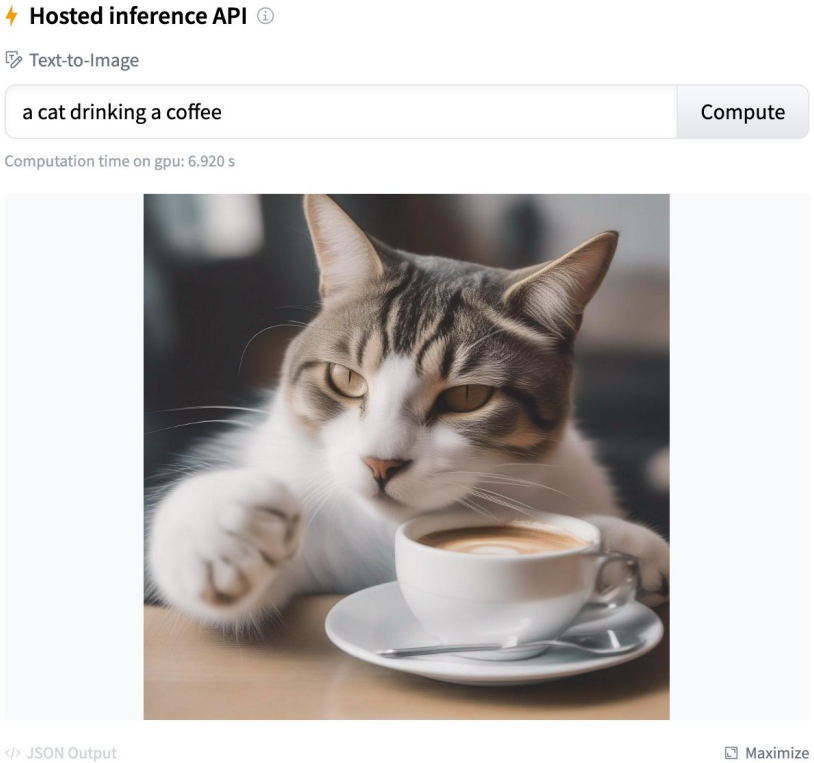
- diferenciar modelos de autos
- diferenciar especies de flores
- diferenciar modelos de aviones
- contar número de objetos en una imagen
- no le va bien con texto escrito a mano
- no sabe generar captions, solo sabe decir si un texto corresponde a una imagen.
- no le va bien para textos largos.

Aplicaciones de CLIP

Modelos generativos de imágenes dado un prompt

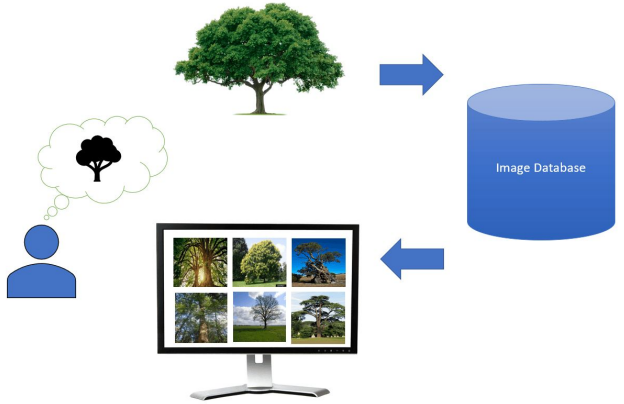
Stable diffusion
Dall-E

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.



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

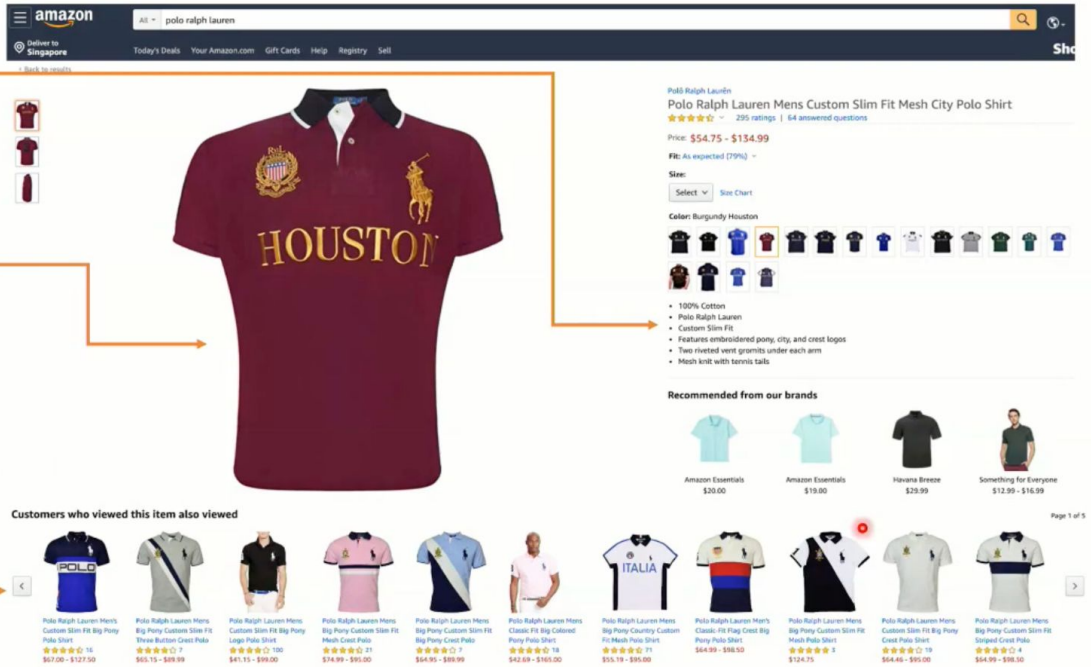
https://colab.research.google.com/drive/1HJaNd_2dPPy2Pz6ttzjVasDa9tsaUQ2c?usp=sharing

Recomendación multimodal

Product description

Product image

Related products



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

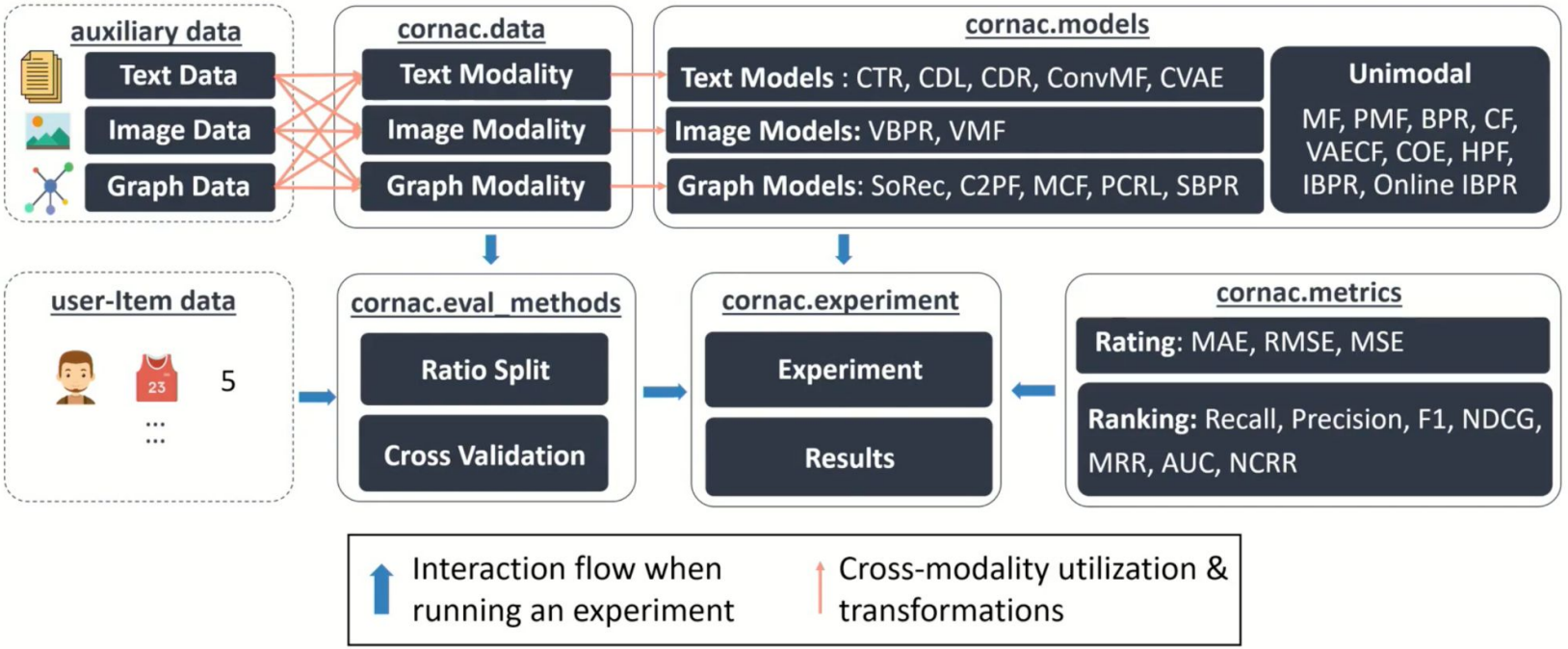
Idea principal: aprender de multiples “modalidades” al mismo tiempo.

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



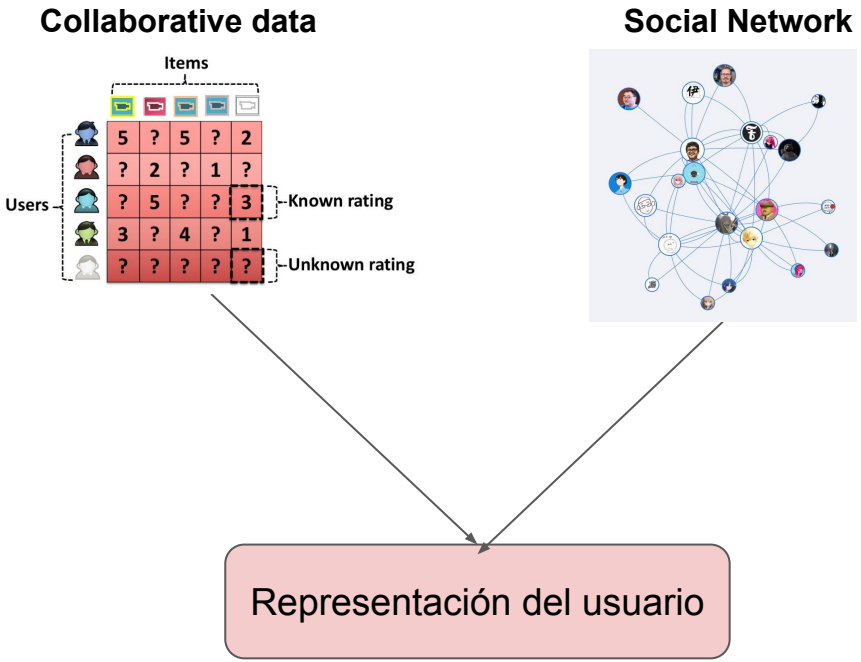
CORNAC: MULTIMODAL RECOMMENDATION LIBRARY

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

Multimodalidad para mejorar representación de usuarios

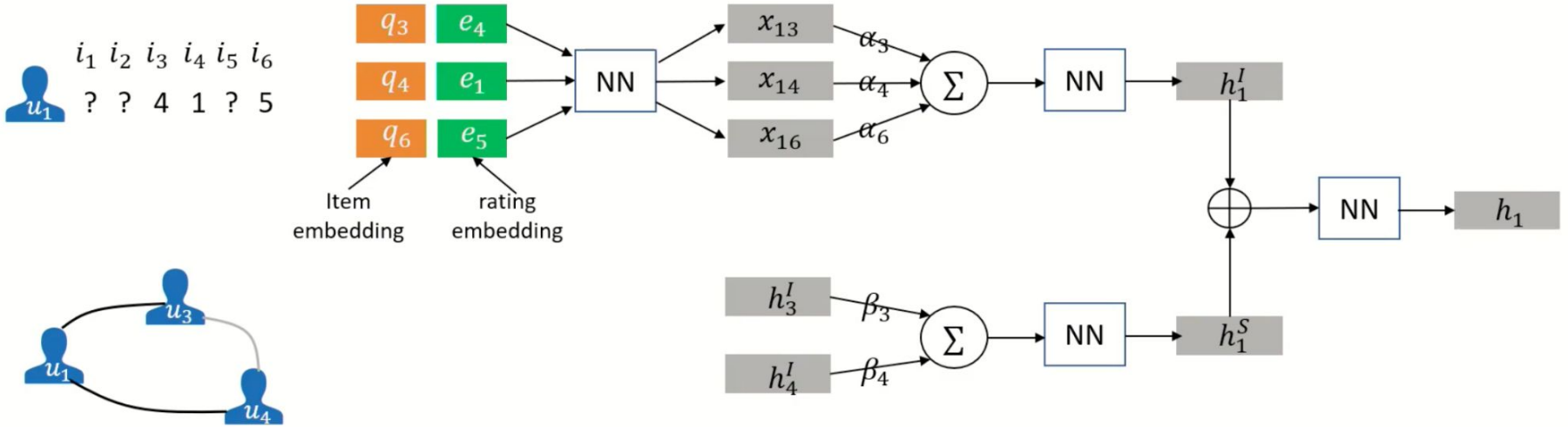
Se pueden utilizar modelos para extraer información de ambas modalidades:

- Filtrado colaborativo.
- Red social de usuarios.



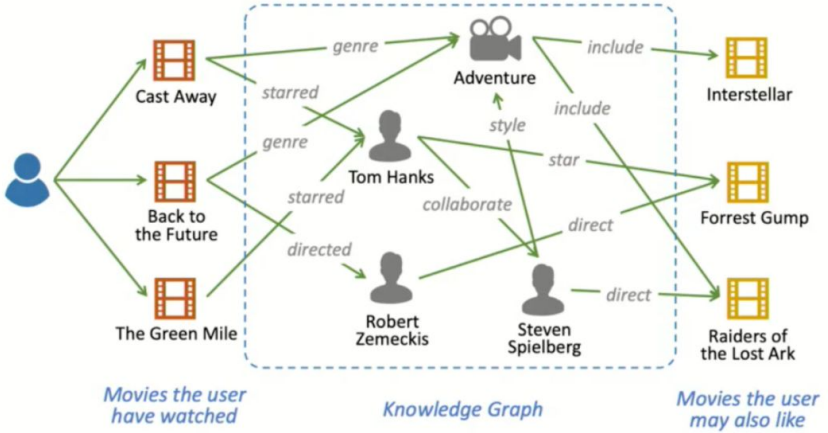
GNN-based user representation

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

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:

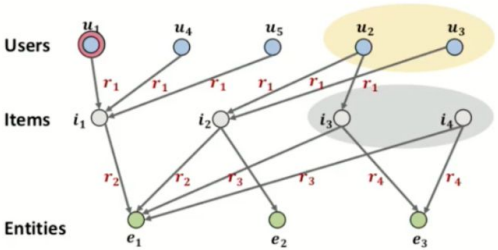
- complementarios: polera y pantalón.
- alternativas: poleras con distinto estilo / color.

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

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 ∪ Knowledge Graph

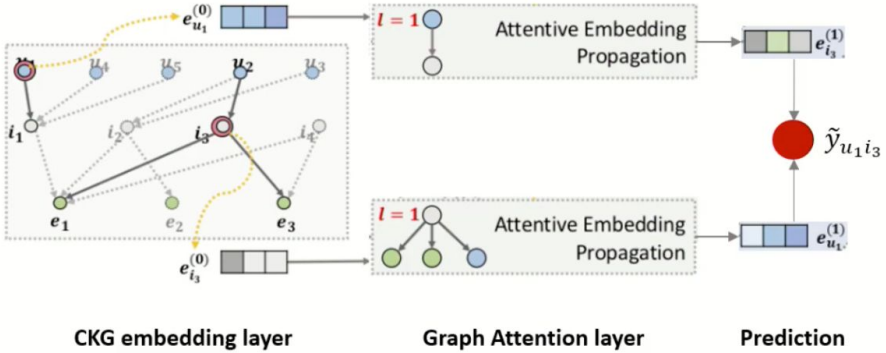


Illustration of KGAT model

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

Gracias!