CLIP para recomendación multimodal

Luis Arias, Benjamín Lillo, Darwin Sanhueza

Departamento de Ciencias de la Computación
Pontificia Universidad Católica de Chile
Santiago, Chile

Abstract

Dada una alta disponibilidad de datos en diferentes formatos, es necesaria una forma de utilizarlos de forma conjunta en sistemas recomendadores. En este trabajo se explora la recomendación multimodal de negocios de comida utilizando CLIP para aprovechar datos en formato de texto e imágenes de los negocios del *dataset* de Yelp. Se realizan recomendaciones tradicionales basadas en texto, imágenes y *feedback* implícito con el objetivo de compararlas con la recomendación con CLIP, para la cual se implementaron cinco técnicas que varían en cuanto a la representación de los ítems y usuarios. De los resultados, se observa que la quinta técnica de recomendación con CLIP presentó mejor desempeño que las otras técnicas, pero comparándose con las recomendaciones basadas en *feedback* implícito, estas últimas presentaron un mejor desempeño. Se concluye entonces que la recomendación multimodal logra demostrar una mejoría con respecto a la recomendación unimodal basada en texto o imágenes, pero no así con las recomendaciones basadas en *feedback* implícito, demostrando su efectividad.

Introducción

Actualmente, hay una muy alta disponibilidad de datos en distintos formatos, y la combinación de la información proveniente de texto e imágenes presenta una muy buena oportunidad para mejorar la calidad de los sistemas de recomendación. El desafío se encuentra en la comprensión de la información entregada por formatos de imagen y texto de manera conjunta, y este trabajo aborda esta problemática aplicando un modelo multimodal avanzado, CLIP (*Contrastive Language-Image Pre-training*), que ha demostrado su efectividad siendo capaz de entender de manera conjunta la información compartida entre texto e imágenes [1].

Tradicionalmente, los sistemas de recomendación se han basado en información unimodal, ya sea texto o imágenes, lo que limita su capacidad para comprender relaciones y contextos complejos. Por lo tanto, proponemos desarrollar un sistema de recomendación que aproveche las representaciones semánticas de CLIP, logrando así una comprensión más profunda de los elementos multimodales en cuestión, explorando su aplicación en un *dataset* de Yelp que contiene interacciones entre usuarios y negocios de comida.

Dataset

El set de datos a utilizar es un subconjunto del *dataset* de Yelp[1], que contiene 87730 reseñas, representando interacciones entre usuario e ítem correspondiente al negocio de comida, repartiendo un 80% como set de entrenamiento, 10% como set de validación y el 10% restante para el *set de testing*. Cada interacción contiene: id del usuario, id del negocio con el que interactuó, *rating*, id de la reseña, texto de la reseña, etiquetas adicionales (*useful*, *funny*, *cool*) y fecha.

Se utiliza también un set con los datos de cada negocio, conteniendo para cada uno: id del negocio, nombre, dirección, ciudad, estado, código postal, latitud, longitud, está abierto, atributos, categoría y horario de atención. Junto con esto, se tiene un set de fotos para cada negocio y *metadata* de cada foto.

Metodología

Para estudiar la recomendación multimodal, se establecen objetivos que buscan explorar las dimensiones de este enfoque. En primer lugar, se desarrollan recomendaciones como *random* y *most popular*, esto con el fin de ser *baselines*, además recomendaciones basadas en feedback implícito, también, por otro lado, se realizan recomendaciones basadas en texto e imágenes. En segundo lugar, se plantea comparar estas recomendaciones unimodales con las recomendaciones multimodales hechas con CLIP, con el objetivo de evaluar y contrastar la efectividad de recomendaciones. Para esto se especifica en los siguientes puntos como se llevan a cabo estas recomendaciones

- Para la recomendación basada en texto, se extrajeron los *embeddings* de los textos disponibles en las *reviews* para los negocios de comida con

[1] Disponible en https://www.yelp.com/dataset.

BERT Large[2]. Este es un modelo de procesado de lenguaje natural desarrollado por Google, que captura el contexto de las palabras de forma bidireccional utilizando *masked language modeling* (MLM) [2].

- Para la recomendación basada en imágenes, los *embeddings* de las imágenes se extrajeron utilizando tres redes neuronales convolucionales preentrenadas: ResNet101, un modelo de 101 capas de profundidad capaz de clasificar imágenes entre 1000 categorías diferentes[3]; VGG16, de 16 capas; y MobileNet, una red diseñada para ser eficiente en entornos de dispositivos móviles [4].
- Se realizan, además, recomendaciones basadas en *feedback* implícito con ALS y BPR, estos modelos con distintos factores latentes para determinar el mejor modelo, obteniendo que el modelo ALS con 100 factores latentes y el modelo BPR con 50 factores latentes.

En cambio, en el contexto de la recomendación multimodal se emplea el modelo CLIP[3]. Este enfoque implica que CLIP toma un par imagen-texto como entrada, para aprender un espacio multimodal de *embeddings*, eso se logra, ya que CLIP entrena conjuntamente un *encoder* de imágenes y un *encoder* de texto para maximizar la similitud coseno (ecuación (1))

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} = \frac{\sum_1^n A_i B_i}{\sqrt{\sum_1^n A_i^2}\sqrt{\sum_1^n B_i^2}} \qquad (1)$$

del *emebedding* de la imagen y texto del par correcto y a la vez minimizar la similitud coseno de los *embeddings* de imagen y texto de pares incorrectos. Esta arquitectura se observa en la Figura 1.

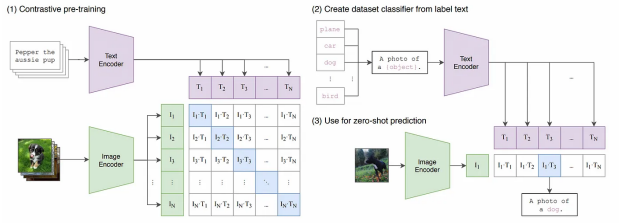

Figure 1: *Arquitectura CLIP [5].*

Entonces, para las recomendaciones multimodales se ocupa el modelo CLIP pre-entrenado de OpenAI con el fin de extraer los *embeddings* de las imágenes, los cuales se almacenan en un diccionario donde la clave es el ID de la foto y el valor es el *embedding*.

[2] Disponible en https://huggingface.co/bert-large-uncased.
[3] Disponible en https://huggingface.co/openai/clip-vit-base-patch32.

A continuación se realizan cinco técnicas para recomendación con CLIP, que varían en cuanto a la representación de los ítems y de los usuarios:

La primera técnica representa a cada ítem como el promedio de los *embeddings* de sus imágenes, en el caso de los usuarios se ocupan las imágenes que consumió y se calcula el *embedding* correspondiente. Se entrena un modelo de recomendación basado en contenido y se obtienen sus métricas.

La segunda técnica consiste en representar a los usuarios e ítems como el promedio de los *embeddings* provenientes de sus *reviews* que consumió el usuario y a los ítems se representa a partir de las *reviews* que recibió ese ítem. Se entrena un modelo de recomendación basado en contenido y se obtienen sus métricas.

La tercera técnica consiste en que el usuario está representado por el promedio de los *embeddings* de las *reviews* de los items, mientras que los ítems están representados por el promedio de los *embeddigs* de sus imágenes. Se entrena un modelo de recomendación los ítems que sean más similares al vector del usuario

La cuarta técnica es similar a la técnica anterior, ahora que el usuario está representado por el promedio de los *embeddings* de las imágenes de los ítems que consumió, mientras que los ítems están representados por el promedio de los *embeddings* de las review que recibió el ítem.

Por otra parte, la quinta técnica consiste en mezclar la técnica 1 y 2, es decir, los usuarios e ítems son representados por un *embedding* que es el promedio de los *embeddings* de *review* de los usuarios a ítems y además se agregan los *embeddings* las imágenes de los ítems.

Finalmente, para cada proceso de recomendación se calculan las métricas de desempeño MAP y nDCG para 10, 20 y 30 elementos.

En la siguiente sección se reportan los resultados obtenidos de las recomendaciones para distintos modelos.

Resultados

En la tabla 1 se hace un resumen de todos los resultados para distintos modelos mencionados en el apartado de la metodología.

Primero se observan que las métricas obtenidas de recomendaciones basadas en solo texto con BERT Large son más bajas que la recomendación basada en imágenes con MobileNet en nDCG, pero mejores en MAP, de igual manera el desempeño es bajo para ambos casos, esto indican que las recomendaciones basadas *embeddings* de estos datos no logran capturar de manera efectiva las preferencias de los usuarios.

	MAP@10	MAP@20	MAP@30	nDCG@10	nDCG@20	nDCG@30
Most popular	0.00184	0.00223	0.00245	0.00675	0.01251	0.01794
Random	0.00035	0.00043	0.00050	0.00121	0.00221	0.00387
Rec. con Texto (BERT Large)	0.00226	0.00247	0.00255	0.00498	0.00808	0.01007
Rec. con Imágenes (ResNet101)	0.00110	0.00140	0.00152	0.00343	0.00764	0.01063
Rec. con Imágenes (VGG16)	0.00117	0.00138	0.00147	0.00398	0.00697	0.00908
Rec. con Imágenes (MobileNet)	0.00162	0.00182	0.00197	0.00509	0.00797	0.01174
BPR	0.00369	0.00511	0.00494	0.01107	0.02004	0.03090
ALS	0.00709	0.00826	0.00866	0.02182	0.03743	0.05283
Tec.1: item=promedio embs img	0.00194	0.002158	0.00228	0.00642	0.00963	0.01273
Tec.2: item=promedio embs text	0.00226	0.002576	0.00272	0.00708	0.01151	0.01539
Tec.3: texto-imagen	0.00108	0.00128	0.00138	0.00365	0.00642	0.00908
Tec.4: imagen-texto	0.00017	0.00030	0.00036	0.00121	0.00310	0.00465
Tec.5: items=promedio embeddings reviews e imágenes	0.00246	0.00279	0.00295	0.00664	0.01129	0.01517

Table 1: *Métricas de los recomendadores.*

Estas métricas son un *baseline* para ver la efectiva de CLIP. Entre los distintos experimentos y técnicas mencionadas previamente, se observa lo siguiente, las técnicas 1 y 2 obtienen, obtienen resultados similares entre sí, lo que es esperable, ya que al ocupar CLIP los *embeddings* deben ser similares entre las imágenes y texto. Las técnicas 3 y 4 obtienen muy malos resultados, estos corresponden a representaciones distintas a los usuarios e ítems. Entonces, combinando la técnica 1 y 2, es decir, *embeddings* de imágenes y texto, se logra un rendimiento más equilibrado, superando las recomendaciones únicamente basadas en texto (BERT Large) o imágenes (MobileNet), pero no así con las recomendaciones basadas en *feedback* implícito. También sobran superar el *baseline* de una recomendación aleatoria, pero no así el *most popular*.

Conclusiones

En este trabajo se hicieron distintos tipos de recomendaciones como *random*, de *feedback* implícito, recomendaciones basadas en contendido con solo imágenes y texto, esto con el fin de compararlo con recomendaciones con el modelo multimodal CLIP, esto se implementó utilizando cinco técnicas que varían en cuanto a la representación de los ítems y de los usuarios. En los resultados se comparó con técnicas tradicionales como recomendación basada en texto, en imágenes, y en *feedback* implícito a través de las métricas MAP, nDCG para 10, 20 y 30 elementos. Se obtuvo que, si bien las recomendaciones con CLIP no fueron las mejores de todas las realizadas, siendo superadas por las recomendaciones basadas en *feedback* implícito, sí se logra superar a las recomendaciones basadas en texto e imagen, demostrando su efectividad al combinar los datos provenientes de texto e imagen.

En el siguiente apartado se comenta sobre el trabajo futuro y las oportunidades de mejora.

Trabajo futuro

En el futuro se pueden realizar acciones para complementar este trabajo y llegar a nuevas conclusiones, como por ejemplo profundizar en la elección y obtención de *embeddings* de las imágenes y texto, y así analizando su efectividad. Por ejemplo, respecto a las imágenes, detectar las imágenes que no son descriptivas de un lugar de comida, ya que filtrando estas imágenes se podrían obtener imágenes descriptivas de una comida y aprovechar de mejorar el modelo multimodal. Otro ejemplo, respecto al texto, es realizar un análisis de sentimientos respecto a las *reviews* otorgadas por los usuarios, ya que de esta manera se podría determinar el contexto/sentimiento de esa *review* y filtrar por ejemplo las *reviews* positivas que recibió el local. Por otro lado, incluir información de *feedback* implícito a la multimodalidad para ver su efecto en las recomendaciones, ya que las recomendaciones por *feedback* implícito son las que obtienen los mejores resultados y crean un ensamble entre multimodalidad e interacciones.

References

[1] Yangguang Li et al. *Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm*. 2022. arXiv: `2110.05208 [cs.CV]`.

[2] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: `1810.04805 [cs.CL]`.

[3] *Mathworks ResNet101 overview*. URL: https://www.mathworks.com/help/deeplearning/ref/resnet101.html#.

[4] Aurelia Michele, Vincent Colin, and Diaz D. Santika. "MobileNet Convolutional Neural Networks and Support Vector Machines for Palmprint Recognition". In: *Procedia Computer Science* 157 (2019). The 4th International Conference on Computer Science and Computational Intelligence (ICCSCI 2019) : Enabling Collaboration to Escalate Impact of Research Results for Society, pp. 110–117. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2019.08.147. URL: https://www.sciencedirect.com/science/article/pii/S1877050919310658.

[5] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].