Implementación de modelo multimodal para recomendar negocios de comida

Gabriel Astudillo Laroze
gastudillo@uc.cl

Miguel Fernández Pizarro
mafernandez17@uc.cl

Javier Ramos Di Consoli
javier.ramos@uc.cl

Ariel Reyes Pardo
adreyes@uc.cl

December 13, 2023

Abstract

En este informe se presentan los resultados de implementar un enfoque multimodal para recomendar locales de comida. Los experimentos realizados permiten concluir que el modelo CLIP, que considera imágenes y texto, tiene un rendimiento superior a BERT y ResNet-50. Asimismo, se puede concluir que los features de texto creados por CLIP explican en gran parte el rendimiento del modelo multimodal al trabajar con el dataset Yelp.

Keywords: Multimodal recommendation, YELP.

1 Introducción

La problemática planteada ya ha sido abordada mediante distintos enfoques como recomendación no personalizada, implícita y basada en contenido. Respecto a esta última, en la tarea 1 del curso se presentaron evaluaciones sobre texto e imagen de manera independiente, construyendo los embeddings con técnicas que no provienen del mismo modelo. En el caso de texto, se trabajó con TF-IDF y para imágenes se trabajó con ResNet-50.

Dado lo anterior, la propuesta de proyecto se centra en utilizar un enfoque multimodal para lograr construir embeddings desde un mismo modelo, estableciendo diferentes estrategias de agregación que permitan consolidar ambos formatos. Para ello, se trabaja con el modelo CLIP (Contrastive Language-Image Pretraining), comparando su rendimiento con ResNet-50 y BERT.

2 Métodos

En esta sección se describen los elementos centrales del proyecto: Dataset, modelos y propuesta de solución.

2.1 Dataset

El dataset considerado para la experimentación es Yelp [1]. Este conjunto de datos fue creado en el contexto de "Yelp Dataset Challenge", instancia que busca reconocer las mejores soluciones provistas por estudiantes e investigadores en el área de sistemas recomendadores.

Sobre el tipo de datos que incluye este dataset, se identifica información sobre los locales de comida, imágenes y reviews de clientes. En particular, dados los objetivos del proyecto desarrollado, este conjunto de datos es interesante de evaluar debido a la inclusión de imágenes y textos que pueden ser utilizados por un modelo multimodal como CLIP.

2.2 Propuesta de solución

La solución propuesta es implementar un modelo de recomendación basada en contenido que combine imágenes y texto. Para ello se compara enfoques unimodales y multimodales. Para procesar imágenes, se considera ResNet-50 [2]. Esta arquitectura residual se caracteriza por su profundidad de 50 capas, donde introduce el concepto de bloques residuales. Estos bloques permiten que la red neuronal aprenda las diferencias entre la entrada y la salida de cada capa, facilitando así el entrenamiento de redes muy profundas sin afectar el rendimiento. La clave de ResNet-50 radica en su capacidad para superar el problema de degradación del rendimiento al aumentar la

profundidad de la red, lo que la hace eficaz para tareas de visión por computadora como clasificación de imágenes, detección de objetos y otras aplicaciones de reconocimiento visual. La dimensión de los embeddings construidos es de 2048.

Por otro lado, para construir representaciones vectoriales de texto, se trabaja con el framework Sentence-Transformers de Hugging Face [3], usando como modelo base BERT [4]. Este modelo transformer permite obtener embeddings de dimensión 768. Estos embeddings capturan información contextual y semántica más rica en comparación con los enfoques tradicionales de representación de palabras. Gracias a su entrenamiento bidireccional en grandes cantidades de texto, BERT puede generar representaciones vectoriales de palabras y frases que reflejan su significado y contexto en un espacio vectorial de alta dimensionalidad. Estos embeddings son fundamentales para diversas tareas de procesamiento del lenguaje natural, ya que ayudan a mejorar el rendimiento en tareas como clasificación de texto, análisis de sentimientos, extracción de información y otras aplicaciones relacionadas con el entendimiento del lenguaje humano.

Como ya fue mencionado, la propuesta multimodal considera el uso del modelo CLIP. A grandes rasgos, esta arquitectura crea representaciones latentes de imágenes y texto, utilizando un enfoque Contrastive, en el que la representación numérica de un concepto en formato de texto está próximo a una imagen que refleje la idea central de dicho concepto. El modelo CLIP incluye un encoder tanto para imagen, como para texto. Ambos bloques construyen representaciones latentes de cada entrada. Es decir, si ingresan 10 imágenes que tienen asociadas 10 frases, cada encoder construye un vector latente para cada caso. Luego, el modelo se encarga de encontrar relaciones entre las representaciones vectoriales asociadas a las imágenes y a los textos. Para cuantificar esta relación, se considera como medida de similaridad el producto punto entre los vectores.

La primera etapa del proyecto consistió en la construcción de embeddings para texto e imágenes, siendo esta última la que fue más profundizada. En cuanto a texto, uno de los desafíos enfrentados fue la limitación de CLIP respecto al input que puede procesar. Este valor no puede superar 77 tokens, lo que significa una complicación relevante dado que la mayoría de los reviews supera ese número. Según la documentación del modelo [5], para aumentar el número de tokens de entrada es necesario realizar fine-tuning, proceso que no es factible de realizar por la limitación de recursos computacionales.

Para enfrentar la dificultad anterior, se decidió construir resúmenes usando el modelo BART-Large [6], ajustado mediante un proceso de fine-tuning con el dataset CNN Daily Mail.

Por otro lado, dado que el dataset Yelp también incluye información textual sobre los locales de comida, se tomó la decisión de construir otro embedding con estos datos. En este caso, a diferencia de los reviews, no existe problemas con el límite de tokens. De esta manera, los embeddings creados para texto consideran los siguientes criterios:

- Información del local (Nombre, categoría y ciudad)
- Review de la experiencia por parte de un usuario.
- Resumen de la concatenación de los dos casos anteriores.

En cuanto a las imágenes, se trabajó con ResNet-50 y CLIP, obteniendo embeddings para todas las fotos asociadas a los locales. En este aspecto, CLIP solo establece limitaciones para el tamaño de entrada pero el proceso de creación de representaciones vectoriales no tuvo complicaciones.

Para crear la representación multimodal, se concatenan los embeddings de texto e imágenes. En ese sentido, dado que no todos los locales tienen una imagen asociada, y con el fin de respetar la dimensionalidad de los embeddings, se optó por replicar la representación numérica de texto en caso de no tener fotografías.

En la tabla 1, se presenta un resumen con los embeddings construidos:

Entrada	SBERT	ResNet-50	CLIP
Información local	X		X
Review	X		X
Resumen	X		X
Imagen		X	X

Table 1: Embeddings creados.

Respecto a la forma de generar las recomendaciones, se considera un enfoque basado en el cálculo de similaridad de coseno. Para ello, se considera como función de agregación a la media, obteniendo finalmente una representación vectorial para cada negocio. Luego, se calcula la similaridad con todos los negocios con los cuales el usuario no ha tenido una interacción, generando un grupo de 10, 20 y 30 con los valores más altos en la medida antes mencionada.

Finalmente, respecto a las métricas de evaluación, se considera NDCG@k y MAP@k con k=10, 20 y 30.

3 Resultados

3.1 Evaluación texto

Como se puede observar en la Tabla 2 el mejor rendimiento fue alcanzado por CLIP, considerando los embeddings asociados a la información de los locales. Este hecho resulta interesante dado que permite afirmar que la representación creada por CLIP, que tiene una dimensión de 512, es mejor que la representación de BERT que tiene dimensión 768.

Modelo	MAP@30	NDCG@30
Random	0.0006	0.0033
Most popular	0.0025	0.0178
SBERT Reviews	0.0028	0.0160
SBERT Info-local	0.0028	0.0178
SBERT Resumen	0.0026	0.0163
CLIP-Reviews	0.0016	0.0080
CLIP-Info-local	0.0033	0.0212
CLIP-Resumen	0.0026	0.0154

Table 2: Resultados - texto.

Un segundo elemento que se puede notar es que, considerando solo texto, el modelo CLIP tiene mejores resultados que SentenceBERT al trabajar con información de local. Sin embargo, al considerar el resumen creado, SentenceBERT tiene un rendimiento ligeramente superior a CLIP. Sobre la representación creada sobre reviews, la comparación no es justa ya que CLIP solo logra procesar una porción menor del texto. De todas formas, se consideró relevante su inclusión para mostrar la mejora en el desempeño al corregir esta limitación usando un resumen.

3.2 Evaluación imagen

Al trabajar con imágenes, se puede observar que CLIP tiene un rendimiento superior a ResNet-50 al considerar la métrica MAP. Sin embargo, respecto a NDCG, ResNet tiene mejores resultados. En la Tabla 3 se presenta el detalle de las métricas.

Modelo	MAP@30	NDCG@30
Random	0.0006	0.0033
Most popular	0.0025	0.0178
ResNet-50	0.0020	0.0123
CLIP	0.0025	0.0114

Table 3: Resultados - imágenes.

En comparación con el baseline establecido, Most Popular, CLIP obtiene una métrica MAP superior pero un NDCG inferior.

3.3 Evaluación multimodal

En la Tabla 4 se presentan las métricas asociadas a la solución multimodal entre imágenes y los distintos embeddings de texto. Como se puede notar, el modelo multimodal que considera los embeddings de la información de cada local es el que obtiene el mejor resultado. En este caso, se supera el valor alcanzado por el baseline Most Popular.

Modelo	MAP@30	NDCG@30
Random	0.0006	0.0033
Most popular	0.0025	0.0178
CLIP_{multi} Review	0.0020	0.0122
CLIP_{multi} Info local	0.0035	0.0205
CLIP_{multi} Resumen	0.0029	0.0161

Table 4: Resultados - multimodal

Por otro lado, al comparar las métricas de los modelos que incluyen reviews y resúmenes, se puede observar un impacto positivo. Este hecho es esperable dado que la información más relevante está contenida en la información del local.

3.4 Mejores modelos

En la Tabla 5 se puede notar que los tres mejores resultados son alcanzados por el modelo CLIP, considerando diferentes embeddings.

Modelo	MAP@30	NDCG@30
CLIP_{mult} Img-info local	0.0035	0.0205
CLIP_{text} Info local	0.0033	0.0212
CLIP_{mult} Img-Resumen	0.0029	0.0161

Table 5: Modelos con mejores resultados

Un elemento destacable que se puede observar en la tabla anterior es que el modelo basado solo en texto, información de los locales, tiene un

rendimiento similar al modelo que incorpora texto e imágenes. Este resultado es consistente con lo observado en otras tareas multimodales que incluyen texto e imágen, donde se reporta que el modelo solo con texto logra un resultado que se aproxima al enfoque que agrupa texto e imagen [7].

Cabe señalar que en la sección de anexos se presenta el detalle de las métricas para MAP@k y NDCG@k con k=10, 20 y 30.

4 Conclusiones

La propuesta desarrollada logra mostrar los beneficios y limitaciones de trabajar con un enfoque multimodal.

Por un lado, al trabajar con este tipo de modelos multimodales, es importante considerar las restricciones asociadas a los inputs. Como se pudo observar en la sección de resultados, aplicar un modelo transformer como BART es una alternativa razonable para reducir la cantidad de tokens, sin perder rendimiento de forma significativa. Este enfoque se muestra como una mejor opción frente a la alternativa de truncar el texto.

Al comparar el rendimiento de CLIP con modelos unimodales como ResNet y SentenceBERT, se puede observar que las representaciones de CLIP son competitivas. En el caso de imágenes, los resultados obtenidos en las métricas evaluación dan cuenta de un rendimiento similar a lo alcanzado por la arquitectura convolucional. En cuanto a texto, CLIP tiene mejores resultados que SentenceBERT.

Profundizando en el rendimiento de CLIP se pudo comprobar que, para el datase Yelp, los embeddings con la información de los locales logra un rendimiento similar a la agregación entre la mirada de texto e imagen. Este hallazgo se considera relevante ya que da cuenta de la dificultad para establecer recomendaciones en función de imágenes del local. En ese sentido, sería interesante explorar un enfoque donde se establezcan estándares similares para las fotografías de cada negocio, resaltando diferencias relevantes.

5 Trabajo futuro

Como trabajo futuro, se proponen dos proyectos. En primer lugar, se podría replicar esta investigación considerando dataset distintos como, por ejemplo, crear un conjunto de datos a gran escala a partir de un sitio web de moda compartida, [8]. Los usuarios de este sitio web subieron conjuntos que contienen varios artículos de moda como componentes, según su inspiración en la combinación de prendas. Los conjuntos constan de prendas superiores, prendas inferiores, vestidos, zapatos, bolsos y otros accesorios. Algo similar es construido en [9].

Por otro lado, sería interesante evaluar el rendimiento de otras arquitecturas multimodales como BLIP-2 [10] y FLAVA [11]. En ese sentido, se podrían incorporar métricas que estén asociadas a criterios de eficiencia computacional.

References

[1] Yelp. Yelp Dataset. https://www.yelp.com/dataset, 2023.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016.

[3] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[7] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection, 2023.

[8] Polyvere. Polyvere Dataset. https://polyvore.ch/, 2023.

[9] Linlin Liu, Haijun Zhang, and Dongliang Zhou. Clothing generation by multi-modal embedding: A compatibility matrix-regularized gan model. *Image and Vision Computing*, 107:104097, 2021.

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[11] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model, 2022.

6 Anexos

Modelo-Embedding	MAP@10	MAP@20	MAP@30	NDCG@10	NDCG@20	NDCG@30
Random	0.0005	0.0006	0.0006	0.0019	0.0029	0.0033
Most popular	0.0019	0.0023	0.0025	0.0068	0.0126	0.0178
SBERT Reviews	0.0023	0.0027	0.0028	0.0069	0.0121	0.0160
SBERT Info-local	0.0023	0.0026	0.0028	0.0079	0.0128	0.0178
SBERT Resumen	0.0021	0.0024	0.0026	0.0063	0.0110	0.0163
ResNet-50 Img	0.0016	0.0018	0.0020	0.0052	0.0088	0.0123
CLIP_{text} Reviews	0.0014	0.0016	0.0016	0.0037	0.0059	0.0080
CLIP_{image} Img	0.0024	0.0025	0.0026	0.0058	0.0084	0.0114
CLIP_{multi} Img-Reviews	0.0016	0.0019	0.0020	0.0041	0.0081	0.0122
CLIP_{text} Info local	0.0027	0.0031	0.0033	0.0082	0.0143	0.0212
CLIP_{multi} Img-infolocal	0.0029	0.0033	0.0035	0.0081	0.0144	0.0205
CLIP_{text} Resumen	0.0022	0.0025	0.0026	0.0065	0.0106	0.0154
CLIP_{multi} Img-Resumen	0.0024	0.0027	0.0029	0.0073	0.0109	0.0161

Table 6: Resultados generales

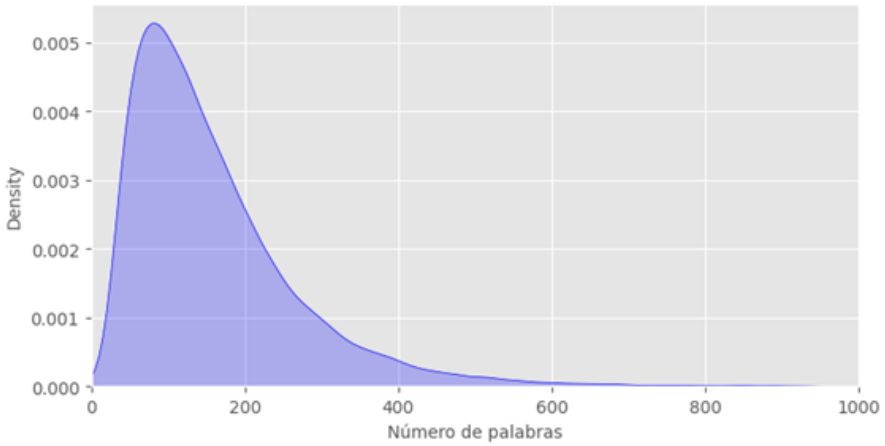

Figure 1: Distrinbución del largo texto original

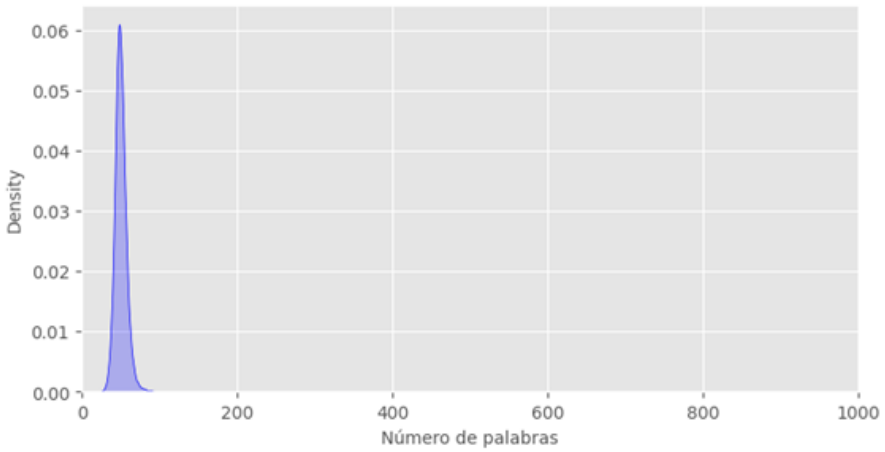

Figure 2: Distrinbución del largo texto resumen