MEJORANDO LA RECOMENDACIÓN DE IMÁGENES MEDIANTE MODELOS MULTIMODALES CON ENCODERS VISUALES Y TEXTUALES

JOAQUÍN DE FERRARI, NICOLÁS SCHIAFFINO, GABRIEL VENEGAS

MOTIVACIÓN

Los sistemas de recomendación de imágenes actuales enfrentan dos desafíos clave que limitan su eficacia:

1. **La Ambigüedad Semántica**: una imagen por sí sola no revela el contexto completo (p. ej., un vestido para una "boda en la playa" vs. una "gala formal"), lo que confunde a los modelos basados únicamente en el análisis visual.
2. **El Problema del "Cold Start"**: los modelos tradicionales fallan al recomendar imágenes nuevas, ya que carecen de un historial de interacciones (clics o "me gusta") para estimar su relevancia.

Estos problemas son críticos en plataformas donde el contenido visual es el núcleo, como redes sociales o galerías de arte, pues la incapacidad de hacer recomendaciones precisas y novedosas afecta directamente la experiencia del usuario y la propuesta de valor del servicio.

PROPUESTA

La propuesta consiste de un framework de software modular y código abierto diseñado para la experimentación sistemática con modelos de recomendacion multimodales. Se implementó un pipeline completo, desde preprocesamiento hasta evaluación de modelos.

Componentes Clave del Framework:

- *Encoders Soportados:* Visuales (**ResNet, CLIP, DINO**) y Textuales (**Sentence-BERT, BERT**).
- *Fusión de Características:* Mecanismo de **Multi-Head Self-Attention** para ponderar y combinar la información de cada modalidad.
- *Predicción:* Red MLP final que procesa las características fusionadas.

A raíz de esto se realizan dos hipótesis:

1. El modelo multimodal será notablemente superior en métricas de recomendación comparado con métodos tradicionales como iKNN, random o uKNN.
2. Tanto la información textual como visual pueden aportar a hacer mejores recomendaciones, aunque la información visual, al provenir de un sólo frame de un video completo, puede no ser representativa para hacer recomendaciones efectivas.

Código y Recursos:
`https://github.com/Joacodef/PixelRec_Multimodal`

DATASET

PixelRec es un conjunto de datos multimodal a gran escala utilizado principalmente para comparar y desarrollar sistemas de recomendación que aprovechan los píxeles de imágenes en bruto, particularmente para contenido de video corto.

Estadístico	Cantidad
Número de Usuarios	50000
Número de ítems	82865
Total Interacciones	989494
Densidad (%)	0.0239

Table 1: Descripción dataset de interacciones en la versión **PixelRec50K**, utilizada para este trabajo.

Figure 1: Ejemplo de imágenes presentes en **PixelRec50K**

ARQUITECTURA DEL MODELO

El núcleo del framework es el `MultimodalRecommender`, una red que procesa y fusiona las distintas modalidades para predecir la probabilidad de interacción $\hat{y}_{ui}$.

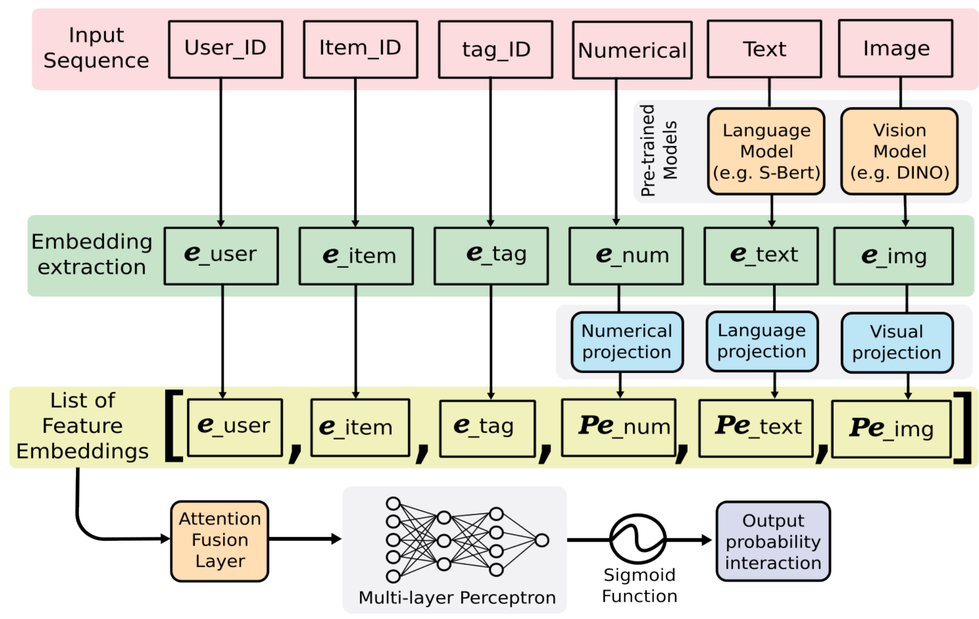

Figure 2: Flujo de datos a través de la arquitectura multimodal propuesta.

La fusión de modalidades $\{\mathbf{e}_{user}, \mathbf{e}_{item}, \mathbf{e}_{tag}, \mathbf{Pe}_{num}, \mathbf{Pe}_{text}, \mathbf{Pe}_{img}\}$ se realiza mediante un mecanismo de atención que permite al modelo ponderar su importancia.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

EJEMPLOS & RESULTADOS

Figure 3: Imagen **i213993** de Título "*When a drummer is placed in front of the drummer at a children's gala, the drummer gradually loses control.*", perteneciente a la categoría **performance**, cuenta con **6,460,723** vistas. Es el **mejor resultado** porque se recomienda como primera opción para el usuario correcto con una probabilidad del **91.16%**, la más alta lograda para una interacción que realmente ocurrió.

Figure 4: Imagen **i76860** de Título *"What is the experience of a giant centipede crawling on your body? What is the experience of having a centipede on your body?"*, perteneciente a la categoría **Daily Life**, cuenta con **3,311,261** vistas. Es el **peor resultado** porque se recomienda con probabilidades cercanas al **90%** para muchos usuarios de manera incorrecta; es un muy buen distractor.

Método	Recall@5	NDCG@5	Recall@10	NDCG@10	MRR	Novelty_P
Random	0.2360	0.1373	0.4728	0.2166	0.1409	**0.9523**
MostPopular	0.4617	0.2913	0.7212	0.3786	0.2745	0.9271
Item-kNN	0.2348	0.1387	0.4932	0.2406	0.1659	0.9501
User-kNN	0.2580	0.1637	0.4860	0.2357	0.1616	0.9509
Multimodal	0.5025	0.3260	0.7651	**0.4170**	**0.3111**	0.9232
Ablación						
Multi sin visual	**0.5088**	**0.3318**	**0.7715**	0.4167	0.3088	0.9222
Multi sin textual	0.3879	0.2363	0.6542	0.3219	0.2220	0.9339

Table 2: Resultados para *Top-k Retrieval* con k=10 calculados con el 30% de los usuarios de **PixelRec50K**, un sampling negativo de 20 items con generación random. El modelo multimodal consta de encoder visual *Dino* y textual *sBERT*.

CONCLUSIONES & TRABAJO FUTURO

Los resultados validan las hipótesis propuestas, demostrando la superioridad del modelo multimodal sobre los baselines tradicionales. La fusión de características mediante *Multi-Head Self-Attention* se mostró eficaz para ponderar y combinar las distintas modalidades.

Un hallazgo clave del estudio de ablación es que la información textual constituye una señal más potente que la visual en este contexto. De hecho, el modelo sin la modalidad de imagen ('Multi w/o visual') presenta un rendimiento comparable, e incluso ligeramente superior en métricas como Recall@10. Proponemos que esto se debe a que el texto (título y descripción) es una fuente de información ya sintetizada y curada por el autor, ofreciendo patrones más claros que los que se pueden extraer de imágenes visualmente muy heterogéneas y complejas.

Como trabajo futuro, se explorará el ajuste fino de los encoders pre-entrenados para especializar las representaciones, se investigarán mecanismos de fusión más avanzados, se implementarán estrategias de muestreo negativo más sofisticadas para robustecer el entrenamiento.

REFERENCIAS

[1] Y. Cheng, Y. Pan, J. Zhang, et al. An Image Dataset for Benchmarking Recommender Systems with Raw Pixels. *arXiv preprint arXiv:2309.06789v2*, 2023.
[2] Q. Liu, Y. Zhang, C. Chen, et al. Multimodal Recommender Systems: A Survey. *ACM Computing Surveys*, 56(9), 2024.
[3] H. Zhou, S. M. Faria, B. Wang, and O. Vechtomova. A Comprehensive Survey on Multimodal Recommender Systems. *arXiv preprint arXiv:2307.08272*, 2023.
[4] E. Jeung, J. Jung, and C. Park. A Multimodal Recommender System Using Deep Learning Techniques Combining Review Texts and Images. *Electronics*, 13(1), 2024.
[5] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.