Modelado escalable de usuarios basado en conjuntos para explicabilidad visual en sistemas de recomendación

Jorge García-Mateo[1], Jorge Paz-Ruza[2], y Carlos Eiras-Franco[2]

[1] Facultade de Informática, Universidade da Coruña, A Coruña, Spain
jorge.garcia.mateo@udc.es
[2] Universidade da Coruña, Grupo LIDIA - CITIC, A Coruña, Spain
{j.ruza, carlos.eiras.franco}@udc.es

Resumen La explicabilidad visual en sistemas de recomendación busca justificar las recomendaciones mediante imágenes relevantes para cada usuario. Sin embargo, modelos como ELVis, MF-ELVis y BRIE representan a cada usuario mediante embeddings aprendidos indexados por identificador, lo que hace que el número de parámetros crezca con la base de usuarios dificultando su escalabilidad a entornos reales, e impide el tratamiento directo de usuarios no vistos en entrenamiento. Proponemos una alternativa basada en Deep Sets que representa a cada usuario a partir del conjunto de imágenes de su historial visual. Esta formulación preserva la invariancia a permutaciones, elimina la tabla de embeddings de usuario y mantiene el objetivo de ranking por pares de BRIE mediante Bayesian Personalized Ranking. Evaluando la propuesta en seis conjuntos de datos de explicación de recomendaciones de restaurantes frente a modelos existentes, nuestro modelo propuesto alcanza un rendimiento comparable al de los modelos Estado del Arte en métricas de ranking, y reduce el número de parámetros hasta en un 97,6 % frente al BRIE original, y en una proporción aún mayor respecto a ELVis y MF-ELVis, a costa de un aumento del coste computacional como consecuencia de agregar dinámicamente el historial visual del usuario.

Keywords: Sistemas de recomendación · Explicabilidad visual · Deep Sets · Representación de usuarios · Escalabilidad

1. Introducción

Los sistemas de recomendación son un componente central en muchas plataformas digitales [14,16], donde ayudan a los usuarios a navegar grandes cantidades de información. A medida que estos sistemas influyen en la toma de decisiones, la explicabilidad cobra un papel relevante: no solo interesa recibir recomendaciones precisas, sino también comprender por qué se producen. En dominios como la restauración, el turismo o el comercio electrónico, la explicabilidad visual resulta especialmente atractiva; en esta línea, Díez et al. [3] proponen utilizar imágenes existentes asociadas a los ítems recomendados para construir

explicaciones personalizadas, seleccionando aquellas que mejor muestran por qué un ítem puede ser relevante para un usuario concreto.

Los enfoques existentes suelen representar a cada usuario mediante un embedding asociado a su identificador, que captura sus preferencias latentes. Aunque esta estrategia es eficaz en escenarios cerrados, introduce dos limitaciones prácticas: el número de parámetros crece con la base de usuarios y los usuarios no vistos no pueden representarse de forma natural. Este problema es especialmente relevante en plataformas dinámicas, donde la base de usuarios cambia continuamente.

Esta limitación también aparece en BRIE [9], un modelo reciente que aborda la selección de imágenes explicativas mediante ranking por pares, pero que mantiene una representación de usuario basada en identificadores. Para abordarla, proponemos sustituir sus embeddings de usuario por una representación basada en Deep Sets [15]. El historial visual de cada usuario se modela como un conjunto de imágenes y se agrega mediante una función invariante a permutaciones. De este modo, el número de parámetros deja de depender del número de usuarios, mientras se conserva el objetivo de ranking por pares del modelo base.

Las contribuciones principales son: i) una representación de usuario basada en conjuntos e independiente de $|U|$; ii) su integración en BRIE manteniendo BPR; y iii) una evaluación en seis ciudades de TripAdvisor considerando ranking, parámetros y coste computacional.

2. Fundamentos y trabajo relacionado

2.1. Formulación del problema

Sea U el conjunto de usuarios, I el conjunto de ítems recomendables y P el conjunto de imágenes asociadas a dichos ítems. Sea $P_i \subseteq P$ el conjunto de imágenes asociadas al ítem i. En esta tarea se asume que un sistema de recomendación previo ha seleccionado un par usuario–ítem (u, i), por lo que el objetivo no es decidir qué ítem recomendar, sino generar una explicación visual para esa recomendación.

Para ello, el problema se formula como una tarea de ranking personalizado de imágenes. Para cada usuario u e ítem recomendado i, el sistema calcula una puntuación R_{up} para cada imagen candidata $p \in P_i$, que estima su adecuación como explicación visual de la recomendación. De este modo, se busca aprender una función de ranking

$$f : U \times I \times P_i \to \mathbb{R}$$

que ordene las imágenes según su relevancia explicativa estimada para el usuario. Esta relevancia se refiere a la adecuación de la imagen para explicar un ítem ya recomendado, no a la relevancia del propio ítem.

2.2. Modelos previos de explicabilidad visual

Diversos trabajos han abordado la explicabilidad visual asignando puntuaciones personalizadas a imágenes candidatas a partir del contenido visual asociado

a usuarios e ítems. Esta línea de trabajo se relaciona con modelos de recomendación visual como VBPR [5] y con propuestas de explicaciones visuales [2]. Aquí nos centramos en modelos que ordenan imágenes explicativas mediante representaciones de usuario basadas en identificadores.

Entre ellos, ELVis combina un embedding de usuario con una representación visual proyectada y una función de similitud parametrizada mediante una red neuronal [3], mientras que MF-ELVis sustituye dicha función por un producto escalar en un espacio latente compartido [10]. Ambos ordenan imágenes a partir de sus puntuaciones, pero se entrenan como clasificación binaria. Por su parte, BRIE reformula la tarea como ranking por pares mediante *Bayesian Personalized Ranking (BPR)*, alineando mejor el entrenamiento con el objetivo de ordenar explicaciones visuales [9]. La Figura 1 resume estas diferencias.

En dicha figura, u denota el usuario y p una imagen candidata del ítem recomendado. En ELVis y MF-ELVis, el identificador de usuario se transforma en un embedding entrenable $\mathbf{U}_u \in \mathbb{R}^d$, mientras que el embedding visual de la imagen, extraído con una red Inception ResNet-v2 preentrenada, se proyecta al mismo espacio latente como $\mathbf{V}'_p \in \mathbb{R}^d$. ELVis concatena ambas representaciones y aprende la función de similitud mediante un MLP, mientras que MF-ELVis emplea directamente su producto escalar. En ambos casos, la salida se proyecta con una sigmoide para obtener $\hat{R}_{up} \in (0, 1)$ y el entrenamiento se realiza con entropía cruzada binaria. BRIE, en cambio, compara durante el entrenamiento una imagen positiva p con una imagen negativa p_{neg} y optimiza una pérdida BPR, por lo que sus puntuaciones no se restringen al intervalo $(0, 1)$.

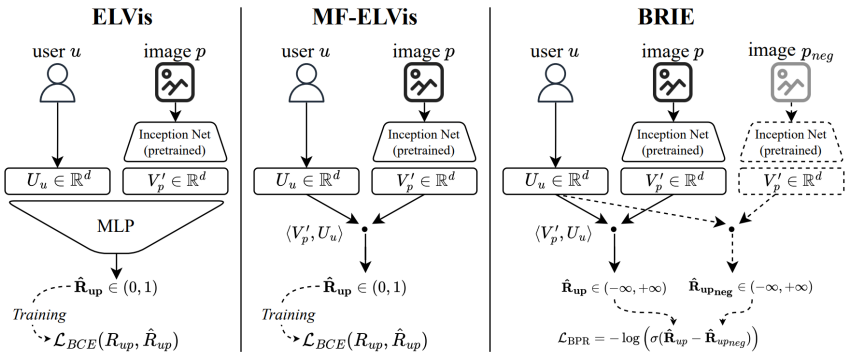


Figura 1. Comparación arquitectónica de ELVis, MF-ELVis y BRIE. Reproducido de [11].

No obstante, ELVis, MF-ELVis y BRIE representan a cada usuario mediante un embedding entrenable asociado a un identificador fijo. Este diseño es eficaz en escenarios cerrados, pero limita la escalabilidad: los parámetros crecen linealmente con $|U|$ y los usuarios nuevos no pueden manejarse de forma nativa sin reentrenamiento.

2.3. Deep Sets y modelado basado en conjuntos

Muchos problemas de aprendizaje toman como entrada un conjunto de elementos y no una secuencia ordenada. En estos casos, el modelo debe ser invariante a permutaciones, es decir, su salida no debe depender del orden en el que se presenten los elementos. Formalmente, una función f definida sobre conjuntos cumple esta propiedad si

$$f(\{x_1, \ldots, x_n\}) = f(\{x_{\pi(1)}, \ldots, x_{\pi(n)}\})$$

para cualquier permutación π de los índices.

Deep Sets proporciona una formulación general para construir funciones con esta propiedad. Bajo condiciones adecuadas, Zaheer et al. [15] muestran que una función invariante a permutaciones puede expresarse mediante una transformación individual de cada elemento, seguida de una agregación simétrica y una transformación final:

$$f(X) = \rho\left(\sum_{x \in X} \phi(x)\right), \tag{1}$$

donde ϕ codifica cada elemento del conjunto y ρ transforma la representación agregada. Esta estructura permite procesar conjuntos de cardinalidad variable preservando la invariancia respecto del orden.

Esta formulación se ha aplicado a entradas sin orden natural, como nubes de puntos [12], parches histológicos [6], colecciones de objetos [4] y atención sobre conjuntos [8]. De forma análoga, el historial visual de un usuario puede interpretarse como un conjunto de imágenes sin orden intrínseco.

3. Método propuesto

3.1. Visión general de la arquitectura propuesta

El modelo propuesto toma BRIE [9] como base y conserva su esquema de aprendizaje por ranking. La modificación principal consiste en reemplazar la representación de usuario basada en identificadores por un bloque Deep Sets. Así, se mantiene el procedimiento general de BRIE: las imágenes candidatas se proyectan a un espacio latente, se calculan puntuaciones usuario–imagen y el entrenamiento se optimiza mediante Bayesian Personalized Ranking.

En la práctica, las imágenes se representan mediante embeddings visuales precalculados de 1536 dimensiones extraídos con Inception ResNet-v2, siguiendo el mismo contexto experimental empleado en trabajos previos sobre estos conjuntos de datos [3,9]. Estos embeddings se transforman a una dimensión latente d mediante una capa totalmente conectada.

Sobre estas representaciones visuales se construye el embedding de usuario mediante un bloque Deep Sets. El historial visual se procesa mediante una codificación individual de sus imágenes, una agregación invariante a permutaciones y una transformación final, correspondientes a las funciones ϕ y ρ. En la configuración de referencia, ambas se implementan como perceptrones multicapa (MLPs)

con activaciones ReLU, permitiendo obtener embeddings de tamaño fijo a partir de historiales de longitud variable.

La Figura 2 ilustra la arquitectura general del modelo propuesto.

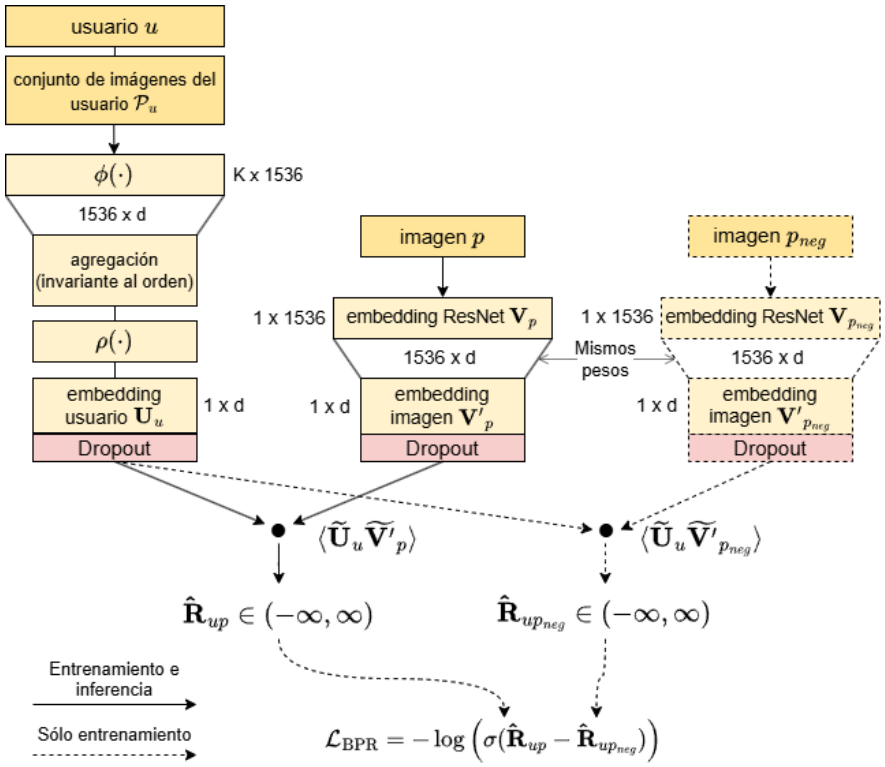


Figura 2. Arquitectura propuesta basada en BRIE con modelado de usuarios mediante Deep Sets. Basado en [9].

En un modelo basado en identificadores, el número de parámetros entrenables puede expresarse como

$$|\Theta_{\text{ID}}| \approx |U| \cdot d + |\Theta_{\text{rest}}|, \tag{2}$$

donde $|U|$ es el número de usuarios, d es la dimensión latente y $|\Theta_{\text{rest}}|$ agrupa el resto de parámetros compartidos del modelo. En cambio, en la formulación propuesta, la tabla de embeddings específica de usuario se sustituye por funciones compartidas de codificación de conjuntos, de modo que

$$|\Theta_{\text{DS}}| = |\Theta_\phi| + |\Theta_\rho| + |\Theta_{\text{rest}}|, \tag{3}$$

por lo que el número de parámetros deja de depender de $|U|$. Además, al generar la representación del usuario a partir de su historial visual, el modelo puede

representar usuarios no vistos durante el entrenamiento sin necesidad de añadir nuevos embeddings ni reentrenar la arquitectura, siempre que dispongan de imágenes asociadas. De este modo, se mejora la escalabilidad del sistema y se mitiga el problema de arranque en frío (*cold-start*) asociado a los enfoques basados en identificadores.

3.2. Representación de usuario con Deep Sets

Sea $X_u = \{x_{u,1}, \ldots, x_{u,n_u}\}$ el conjunto de imágenes positivas asociadas al usuario u. A diferencia de la formulación canónica con suma, empleamos agregación por media para evitar que la magnitud de la representación dependa del número de imágenes del historial.

A nivel conceptual, la representación del usuario se obtiene aplicando ϕ a las imágenes disponibles en su historial visual y agregando los embeddings resultantes mediante una media. Esta operación permite construir una representación de tamaño fijo a partir de historiales de distinta longitud y preserva la invariancia respecto al orden de las imágenes.

3.3. Integración en BRIE

Una vez obtenido $\mathbf{u}$, la compatibilidad con la imagen candidata positiva y la negativa se calcula mediante producto escalar, como en BRIE. Sean $\mathbf{v}_p^+$ y $\mathbf{v}_p^-$ sus representaciones latentes, respectivamente. El modelo calcula las puntuaciones

$$\hat{r}_{up^+} = \langle \mathbf{u}, \mathbf{v}_p^+ \rangle, \qquad \hat{r}_{up^-} = \langle \mathbf{u}, \mathbf{v}_p^- \rangle. \tag{4}$$

Durante el entrenamiento, para cada usuario u se muestrea una imagen positiva p^+ de su historial y una negativa p^- entre imágenes no asociadas a dicho usuario.

El entrenamiento se lleva a cabo con la función de pérdida Bayesian Personalized Ranking (BPR) [13]:

$$\mathcal{L}_{\text{BPR}} = -\log \sigma \left(\hat{r}_{up^+} - \hat{r}_{up^-} \right), \tag{5}$$

donde σ denota la función sigmoide. Así, se mantiene el objetivo de ranking de BRIE sustituyendo únicamente la representación de usuario.

4. Configuración experimental

4.1. Conjuntos de datos

Evaluamos el modelo sobre seis conjuntos reales de TripAdvisor utilizados previamente en explicabilidad visual [3,9]: Gijón, Barcelona, Madrid, Nueva York, París y Londres. Cada conjunto contiene imágenes de restaurantes subidas por usuarios, interacciones usuario–ítem, particiones predefinidas de entrenamiento, validación y test, y embeddings visuales fijos. Seguimos el protocolo

de BRIE, donde cada caso de test se asocia a un conjunto controlado de imágenes candidatas para el ítem recomendado.

La Tabla 1 resume las estadísticas básicas de los seis conjuntos de datos evaluados por ciudad.

Tabla 1. Estadísticas de los conjuntos de datos.

Ciudad	Usuarios	Restaurantes	Imágenes	Imágenes/usuario
Gijón	5.139	598	18.679	3,64
Barcelona	33.537	5.881	150.416	4,49
Madrid	43.628	6.810	203.905	4,67
Nueva York	61.019	7.588	231.141	3,79
París	61.391	11.982	251.636	4,10
Londres	134.816	13.888	479.798	3,56

4.2. Modelos de referencia y métricas de evaluación

Comparamos la propuesta con RND, CNT, ELVis, MF-ELVis y BRIE. RND ordena aleatoriamente, mientras que CNT construye para cada usuario un centroide visual como la media de los embeddings de sus imágenes de entrenamiento y puntúa cada imagen candidata según su cercanía a dicho centroide. Esta comparación permite evaluar el impacto de reemplazar embeddings de usuario basados en identificadores por una representación basada en Deep Sets manteniendo el objetivo de ranking.

El rendimiento se evalúa con las mismas métricas de ranking promediadas empleadas en trabajos previos: MRecall@10, MNDCG@10 y MAUC. Siguiendo esos trabajos, MRecall@10 y MNDCG@10 se calculan sobre los casos de test en los que el ítem recomendado proporciona al menos diez imágenes candidatas y el usuario tiene al menos diez imágenes en entrenamiento, mientras que MAUC se calcula sobre todos los casos de test.

Además de la calidad de ranking, analizamos el número de parámetros entrenables y el coste computacional asociado al modelado dinámico de usuarios.

4.3. Protocolo de entrenamiento y detalles de implementación

Los ejemplos negativos se generan dinámicamente durante el entrenamiento seleccionando aleatoriamente imágenes con las que el usuario no ha interactuado, siguiendo el mismo protocolo para todos los conjuntos de datos.

La selección del modelo se realiza sobre Barcelona mediante random search [1] sobre 25 configuraciones, variando la dimensión latente $d \in \{16, 64, 256, 1024\}$, la tasa de aprendizaje (lr) en $\{10^{-3}, 10^{-4}, 10^{-5}\}$ y la probabilidad de dropout en $\{0.2, 0.4, 0.6, 0.8\}$. La configuración final emplea $d = 64$, learning rate de 10^{-3}, dropout de 0.4 y un tamaño máximo de historial de $K = 20$.

En validación y test, la imagen candidata que se está puntuando se excluye del historial del usuario para evitar fugas de información. Una vez seleccionados los hiperparámetros en Barcelona, la misma configuración se reutiliza en el resto de

ciudades sin reajustes. Todos los resultados corresponden a una única ejecución por modelo y ciudad con semilla global fija. El modelo está implementado en PyTorch y se optimiza con Adam [7], batch size 2^{14}, un máximo de 100 épocas y early stopping basado en MAUC de validación. Los experimentos se realizaron sobre una GPU NVIDIA GeForce RTX 4050 Laptop.

5. Resultados y discusión

5.1. Rendimiento en ranking

Para el cálculo del número de parámetros se respetan las dimensiones latentes empleadas en los trabajos originales: $d = 256$ para ELVis, $d = 1024$ para MF-ELVis y $d = 64$ para BRIE.

La Tabla 2 presenta el rendimiento en ranking y el número de parámetros en las seis ciudades evaluadas.

Tabla 2. Resumen del rendimiento y tamaño de modelo (número de parámetros entrenables) en las seis ciudades evaluadas. Los mejores resultados se muestran en negrita y los segundos mejores aparecen subrayados.

	Gijón				Barcelona			
Modelo	Params	MRecall@10	MNDCG@10	MAUC	Params	MRecall@10	MNDCG@10	MAUC
RND	–	0,373	0,185	0,487	–	0,409	0,186	0,502
CNT	–	0,464	0,218	0,546	–	0,443	0,219	0,554
ELVis	1.840.641	0,521	0,262	0,596	9.110.529	0,597	0,327	0,631
MF-ELVis	6.836.224	0,538	0,285	0,592	35.915.776	0,557	0,293	0,596
BRIE	427.264	**0,607**	**0,333**	**0,643**	2.244.736	**0,630**	**0,368**	**0,663**
Propuesta	**209.472**	0,571	0,303	0,635	**209.472**	0,610	0,343	0,658
	Madrid				Nueva York			
Modelo	Params	MRecall@10	MNDCG@10	MAUC	Params	MRecall@10	MNDCG@10	MAUC
RND	–	0,374	0,171	0,499	–	0,374	0,168	0,502
CNT	–	0,420	0,203	0,557	–	0,431	0,217	0,563
ELVis	11.892.737	0,572	0,314	0,638	13.608.961	0,553	0,304	0,637
MF-ELVis	47.044.608	0,528	0,279	0,601	53.909.504	0,516	0,276	0,602
BRIE	2.940.288	**0,612**	**0,348**	**0,673**	3.369.344	**0,598**	**0,341**	**0,677**
Propuesta	**209.472**	0,597	0,338	0,668	**209.472**	0,577	0,328	0,672
	París				Londres			
Modelo	Params	MRecall@10	MNDCG@10	MAUC	Params	MRecall@10	MNDCG@10	MAUC
RND	–	0,459	0,209	0,502	–	0,342	0,155	0,500
CNT	–	0,499	0,245	0,557	–	0,400	0,200	0,562
ELVis	12.355.329	0,643	0,352	0,630	35.037.953	0,530	0,293	0,629
MF-ELVis	48.894.976	0,606	0,323	0,596	139.625.472	0,531	0,267	0,597
BRIE	3.055.936	**0,669**	**0,391**	**0,666**	8.726.592	**0,563**	**0,318**	**0,665**
Propuesta	**209.472**	0,661	0,375	0,661	**209.472**	0,549	0,312	0,663

La propuesta se sitúa como segundo mejor método en las seis ciudades evaluadas. Aunque BRIE obtiene los mejores valores en todas las métricas, las diferencias son exiguas, especialmente en MAUC; los resultados corresponden a una única ejecución con semilla global fija. Además, la configuración de la propuesta se seleccionó mediante un random search limitado en Barcelona y se reutilizó en el resto de ciudades sin reajustes específicos, por lo que los resultados no deben interpretarse como una optimización exhaustiva para cada conjunto de datos.

Las diferencias respecto a BRIE son algo mayores en MRecall@10 y MNDCG@10 que en MAUC. Esto indica que la representación basada en conjuntos mantiene un comportamiento competitivo a nivel global, aunque puede perder cierta precisión en las primeras posiciones del ranking frente a un modelo con embeddings específicos de usuario. En Londres, por ejemplo, BRIE alcanza un MAUC de 0,665 frente a 0,663 de la propuesta, mientras que los parámetros se reducen de 8.726.592 a 209.472, es decir, un 97,6 %. La reducción es todavía mayor respecto a ELVis y MF-ELVis, debido a las mayores dimensiones latentes empleadas por estos modelos en sus trabajos originales. En conjunto, estos resultados muestran que Deep Sets ofrece una alternativa competitiva cuando se prioriza la escalabilidad del modelo.

5.2. Eficiencia y escalabilidad

Más allá de la calidad del ranking, la principal ventaja de la propuesta es estructural: mientras ELVis, MF-ELVis y BRIE incluyen una tabla de embeddings proporcional a $|U|$, la arquitectura propuesta mantiene 209.472 parámetros en todas las ciudades. Esta diferencia se amplifica al aumentar la base de usuarios, ya que cada nuevo usuario requiere un nuevo vector entrenable en los modelos basados en identificadores. La Figura 3 ilustra esta diferencia: los modelos basados en identificadores crecen con la base de usuarios, mientras que la propuesta permanece constante.

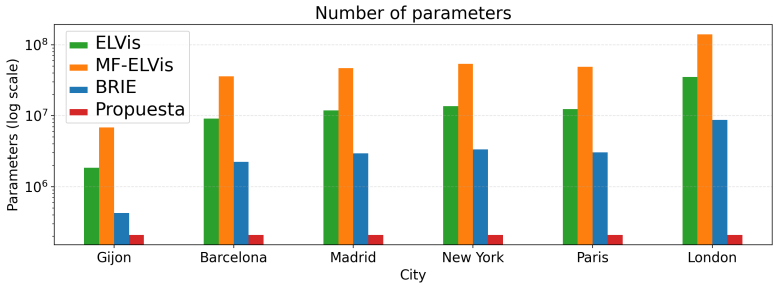


Figura 3. Número de parámetros entrenables por ciudad y modelo (escala logarítmica).

Esta mejora estructural tiene un coste computacional en inferencia: frente a una búsqueda directa en una tabla de embeddings, la propuesta debe codificar hasta K imágenes del historial mediante ϕ, agregarlas y aplicar ρ para obtener el embedding de usuario.

La Figura 4 muestra este efecto en términos de emisiones de CO_2 durante la inferencia, y se observa que la propuesta presenta mayores emisiones que BRIE debido al cálculo dinámico de la representación de usuario. Esta diferencia está correlacionada con $|\mathcal{P}_i|$, i.e. el número de posibles explicaciones por ítem; en los conjuntos de TripAdvisor, este efecto es visible en los restaurantes de Nueva York y París (que presentan muchas y pocas imágenes únicas por restaurante, respectivamente).

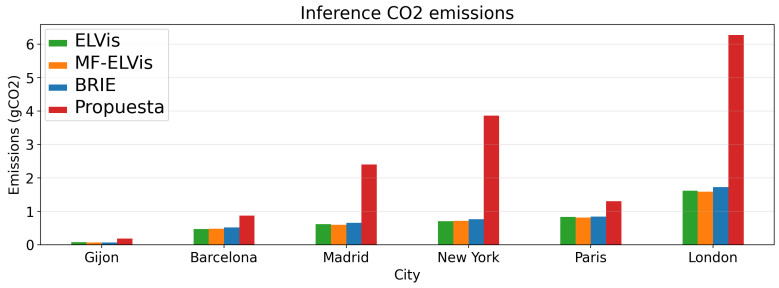


Figura 4. Emisiones de CO_2 en inferencia por ciudad.

El compromiso no debe interpretarse únicamente como un aumento del coste de inferencia. La propuesta sustituye la consulta directa de un embedding de usuario por la construcción dinámica de una representación agregada a partir de hasta K imágenes de su historial visual. Dado que ϕ se aplica a cada imagen, la complejidad de esta etapa respecto a K es de orden $O(K)$. Como K permanece fijo en todos los experimentos, el tiempo total de inferencia crece con el tamaño del conjunto evaluado, es decir, con el número de usuarios, casos de test e imágenes candidatas que deben puntuarse. En escenarios dinámicos con incorporación frecuente de nuevos usuarios, este coste podrá verse compensado por el ahorro asociado a no añadir embeddings específicos ni reentrenar el modelo.

5.3. Estudio de ablación de ϕ y ρ

Para analizar el impacto de las funciones ϕ y ρ, se realizó un estudio de ablación en el conjunto de Barcelona, evaluando distintas implementaciones de ambos componentes con distinto grado de complejidad. El resto del modelo se mantuvo fijo. La Tabla 3 recoge los resultados.

Tabla 3. Estudio de ablación del bloque Deep Sets en Barcelona. Todas las variantes usan $d = 64$, $lr = 10^{-3}$ y dropout 0.4, salvo Identity–Identity, que mantiene dimensión.

ϕ	ρ	MRecall@10	MNDCG@10	MAUC
Identity	Identity	0,544	0,287	0,612
MLP	MLP	0,610	0,343	0,658
Linear	MLP	0,616	0,350	0,657
MLP	Linear	**0,618**	**0,354**	**0,662**
Linear	Linear	0,613	0,350	0,659
MLP	Identity	0,616	0,351	0,660
Identity	MLP ($1536 \to d$)	0,614	0,348	0,659
Linear	Identity	0,613	0,347	0,655
Identity	Linear ($1536 \to d$)	0,607	0,345	0,655

Nota: Identity–Identity no proyecta a d y opera en dimensión 1536. Cuando ϕ = Identity, la proyección $1536 \to d$ se realiza en ρ. En las variantes MLP, la arquitectura empleada es $\text{Linear}(1536, d) \to \text{ReLU} \to \text{Linear}(d, d) \to \text{LayerNorm}(d)$.

El rendimiento es estable entre variantes, con diferencias pequeñas respecto a la configuración MLP–MLP. Aunque la configuración MLP–Linear obtiene los mejores resultados, la mejora es marginal, lo que indica que el principal beneficio procede de la formulación basada en conjuntos más que de la elección concreta de ϕ y ρ. La variante Identity–Identity obtiene el peor rendimiento, lo que sugiere que alguna transformación parametrizada entrenable resulta necesaria para adaptar los embeddings visuales al espacio de ranking. Sin embargo, las configuraciones con capas lineales o MLPs presentan resultados similares, lo que refuerza que el elemento más relevante es la representación del historial de usuario como un conjunto agregado, más que una mayor complejidad en ϕ o ρ.

6. Conclusiones

En este trabajo se ha presentado una alternativa escalable al modelado de usuarios basado en identificadores para explicabilidad visual. La propuesta reemplaza los embeddings fijos de usuario por una representación basada en Deep Sets, calculada dinámicamente a partir del historial visual, y mantiene el objetivo de ranking por pares de BRIE.

Los experimentos sobre seis conjuntos de datos de TripAdvisor muestran que el modelo alcanza un rendimiento cercano al de BRIE y se sitúa consistentemente como segundo mejor método en MRecall@10, MNDCG@10 y MAUC. Además, reduce sustancialmente los parámetros entrenables: en Londres, el conjunto más grande, pasa de 8.726.592 a 209.472, una reducción del 97,6 %.

Estos resultados muestran que las representaciones de usuario basadas en identificadores pueden sustituirse por representaciones derivadas del historial visual sin degradar sustancialmente la calidad del ranking. La principal ventaja es que el tamaño del modelo deja de depender del número de usuarios, facilitando su aplicación en bases de usuarios que cambian con frecuencia, incorporan usuarios no vistos o crecen continuamente.

El resultado principal refleja un compromiso entre escalabilidad y coste de inferencia: el modelo elimina la dependencia de embeddings específicos de usuario, pero debe calcular dinámicamente sus representaciones. Como trabajo futuro, se evaluará explícitamente su comportamiento con usuarios no vistos durante el entrenamiento a partir de su historial visual disponible, así como su generalización a otros dominios de recomendación visual, como moda, comercio electrónico y turismo. También se estudiarán mecanismos de reutilización o almacenamiento temporal de las representaciones de usuario y estrategias de actualización eficiente de los historiales.

Declaración de intereses Los autores declaran no tener intereses en conflicto.

Agradecimientos Este trabajo está financiado por MICIU/AEI y FSE+ (FPU21/05783), FEDER Una forma de hacer Europa (PID2023-147404OB-I00), FEDER/UE (PID2021-1280), Ministerio de Transformación Digital y Función Pública y 'NextGenerationEU'/PRTR (TSI-100925-2023-1), y Xunta de Galicia (ED431C 2022/44). CITIC está financiado por la "Consellería de Cultura, Educación e Universidade" a través de FEDER y la "Secretaría Xeral de Universidades" (ED431G 2023/01).

Referencias

1. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of Machine Learning Research **13**, 281–305 (2012)
2. Chen, X., Chen, H., Xu, H., Zhang, Y., Cao, Y., Qin, Z., Zha, H.: Personalized fashion recommendation with visual explanations based on multimodal attention network. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 765–774 (2019)
3. Díez, J., Pérez-Núñez, P., Luaces, O., Remeseiro, B., Bahamonde, A.: Towards explainable personalized recommendations by learning from users' photos. Information Sciences **520**, 416–430 (2020). https://doi.org/10.1016/j.ins.2020.02.018
4. Hartford, J., Graham, D., Leyton-Brown, K., Ravanbakhsh, S.: Deep models of interactions across sets. In: ICML. Proceedings of Machine Learning Research (2018)
5. He, R., McAuley, J.: Vbpr: Visual bayesian personalized ranking from implicit feedback. In: AAAI 2016. vol. 30, pp. 144–150 (2016). https://doi.org/10.1609/aaai.v30i1.9973
6. Hemati, S., Kalra, S., Meaney, C., Babaie, M., Ghodsi, A., Tizhoosh, H.: Cnn and deep sets for end-to-end whole slide image representation learning. In: Proceedings of the Fourth Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 143, pp. 301–311 (2021)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: ICML. Proceedings of Machine Learning Research, vol. 97, pp. 3744–3753 (2019)
9. Paz-Ruza, J., Alonso-Betanzos, A., Guijarro-Berdiñas, B., Cancela, B., Eiras-Franco, C.: Sustainable transparency on recommender systems: Bayesian ranking of images for explainability. Information Fusion (2024). https://doi.org/10.1016/j.inffus.2024.102497
10. Paz-Ruza, J., Eiras-Franco, C., Guijarro-Berdiñas, B., Alonso-Betanzos, A.: Sustainable personalisation and explainability in dyadic data systems. Procedia Computer Science **207**, 1017–1026 (2022). https://doi.org/10.1016/j.procs.2022.09.157
11. Paz-Ruza, J., Esteban-Martínez, D., Alonso-Betanzos, A., Guijarro-Berdiñas, B.: Sustainable techniques to improve data quality for training image-based explanatory models for recommender systems (2025). https://doi.org/10.48550/arXiv.2407.06740
12. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR. pp. 652–660 (2017)
13. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: Bpr: Bayesian personalized ranking from implicit feedback. In: (UAI '09). pp. 452–461 (2009)
14. Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.): Recommender Systems Handbook. Springer, New York, NY (2011). https://doi.org/10.1007/978-0-387-85820-3
15. Zaheer, M., Kottur, S., Ravanbakhsh, S., Póczos, B., Salakhutdinov, R., Smola, A.J.: Deep sets. In: Advances in Neural Information Processing Systems 30. pp. 3391–3401 (2017)
16. Zhang, Y., Chen, X.: Explainable recommendation: A survey and new perspectives. Foundations and Trends in Information Retrieval (2020). https://doi.org/10.1561/1500000066