Judge a Book by its Cover: Recomendación Multimodal de Libros

Nicolás Herrera, Vicente Steidle, Lucas Vidal
Departamento de Ciencia de la Computación, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Santiago, Chile
nicolash02@uc.cl, vsteidlef@uc.cl, lucas.vidal@uc.cl

Motivación

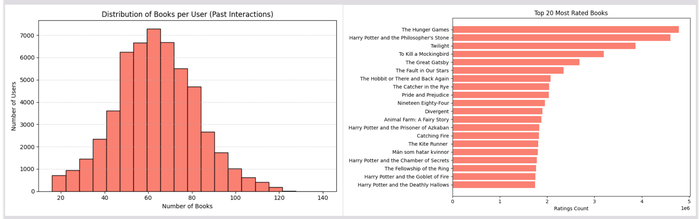
Goodreads ofrece decenas de millones de títulos disponibles, dificultando la tarea de encontrar nuevas lecturas que se alineen con los intereses particulares de cada lector. Si bien los enfoques clásicos de filtrado colaborativo han demostrado ser efectivos al capturar patrones de comportamiento colectivo, dichos modelos tienden a ignorar información valiosa contenida en el contenido textual y visual de los ítems.

Objetivos

- Explorar distintos esquemas de representación de contenido:
  - BERT y BERT-Large para texto.
  - VGG16 y ResNet50 para imágenes.
- Explorar sistemas de recomendación multimodal de libros que integren señales textuales y visuales mediante técnicas de aprendizaje profundo y métodos basados en contenido:
  - Neural Collaborative Filtering (**NCF**).
  - Visual Bayesian Personalized Ranking (**VBPR**).

Dataset

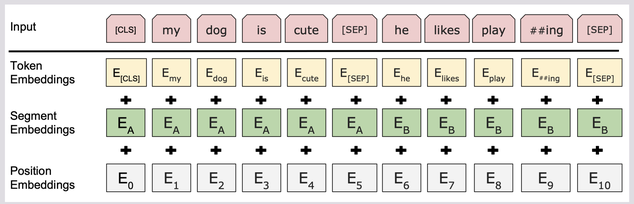
- Goodreads es una comunidad virtual de catalogación de lecturas, con más de 650k usuarios y 10M de libros.
- El dataset consiste en 3.3M de interacciones entre 53k usuarios 4k libros, sin ratings asociados a las interacciones.



- De los 53k usuarios, la mayoría cuenta con varias interacciones en el conjunto de entrenamiento. Como no hay ratings asociados a cada interacción, se pueden tomar las interacciones como información o feedback implícito.
- Cada libro tiene información adicional asociada: géneros, ratings, descripciones, imagen de portada, entre otros.

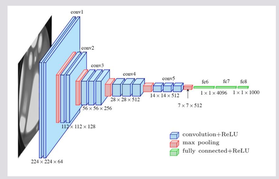
Features

- **Representación de texto**: utilizamos embeddings generados por BERT y BERT-Large a partir de las descripciones de cada libro. Esto nos entrega vectores de 768 y 1024 dimensiones, respectivamente, para cada libro.

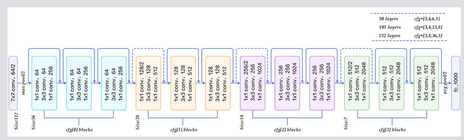


Generación de embeddings con BERT

- **Representación de imágenes**: para cada libro, se generaron features a partir de su portada utilizando VGG16 y ResNet50.



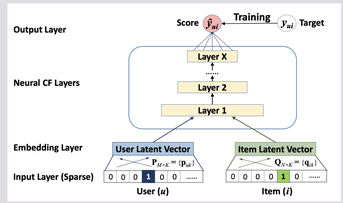
Arquitectura de VGG16



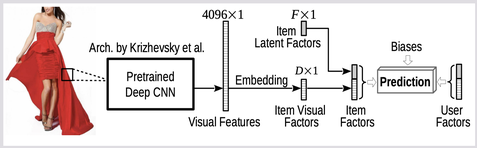
Arquitectura de ResNet50

Enfoques de Recomendación

- **Métodos colaborativos clásicos:**
  - Random
  - Item-Based Collaborative Filtering (ICBF)
  - Alternating Least Squares (ALS)
  - Bayesian Personalized Ranking (BRP)
  - Factorization Machines (FM)
- **Métodos basados en contenido:**
  - NCF
  - VBPR



Framework de NCF



Framework de VBPR

Referencias

[1] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018. http://arxiv.org/abs/1810.04805 .
[2] Iwana, B. K. and Uchida, S. Judging a book by its cover. CoRR, abs/1610.09204, 2016. http://arxiv.Org/abs/1610.09204 .
[3] Javaji, S. R. and Sarode, K. Multi-bert for embeddings for recommendation system, 2023. https://arxiv.org/abs/2308.13050 .
[4] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. CoRR, abs/2103.00020, 2021. https://arxiv.org/abs/2103.00020 .

Evaluación y Resultados

Método	P@5	NDCG@5	MAP@5	AUC	Div@5	Nov@5
ALS	**0.2660**	**0.2907**	0.1881	0.9198	0.2774	0.8243
BPR	0.1040	0.1108	0.0592	0.9183	0.2794	**0.9245**
IBCF	0.1840	0.1933	0.1229	0.8279	0.3239	0.5571
LightFM (warp)	0.0459	0.1523	**0.2783**	0.9409	0.3209	0.4974
LightFM (w. BERT)	0.0440	0.1109	0.2226	0.9393	0.3110	0.4633
LightFM (w. ResNet)	0.0559	0.1145	0.2276	0.9398	0.2988	0.5214
NCF (Vanilla)	0.1320	0.1531	0.0913	0.9579	0.3053	0.5888
NCF (w. BERT embeddings)	0.1340	0.1495	0.0840	0.9563	0.3062	0.6285
NCF (w. ResNet images)	0.1140	0.1290	0.0735	0.9532	0.3104	0.7012
NCF (w. BERT and ResNet)	0.1460	0.1597	0.0945	0.9592	0.3125	0.6395
VBPR (w. BERT embeddings)	0.1400,	0.1626	0.0936	0.9549	0.2958	0.6591
VBPR (w. VGG16 images)	0.1820	0.2069	0.1217	0.9607	0.6462	0.6875
VBPR (w. BERT and VGG16)	0.1920	0.2141	0.1275	**0.9635**	0.9515	0.6395

(a) $K = 5$

Método	P@10	NDCG@10	MAP@10	AUC	Div@10	Nov@10
ALS	**0.1930**	**0.2309**	0.1148	0.9198	0.2814	0.8311
BPR	0.0890	0.0984	0.0382	0.9183	0.2808	**0.9293**
IBCF	0.1600	0.1738	0.0815	0.8279	0.3212	0.5933
LightFM (warp)	0.0450	0.1472	**0.2793**	0.9409	0.3115	0.5543
LightFM (w. BERT)	0.0440	0.0936	0.2155	0.9393	0.3255	0.5175
LightFM (w. ResNet)	0.0490	0.1029	0.2451	0.9398	0.2975	0.5724
NCF (Vanilla)	0.1320	0.1531	0.0621	0.9579	0.3041	0.6243
NCF (w. BERT embeddings)	0.1340	0.1433	0.0585	0.9563	0.3097	0.6591
NCF (w. ResNet images)	0.1180	0.1254	0.0522	0.9532	0.3080	0.7228
NCF (w. BERT and ResNet)	0.1150	0.1344	0.0593	0.9592	0.3070	0.6785
VBPR (w. BERT embeddings)	0.1210,	0.1413	0.0608	0.9549	0.2913	0.6983
VBPR (w. VGG16 images)	0.1650	0.1810	0.0837	0.9607	0.6543	0.7109
VBPR (w. BERT and VGG16)	0.1550	0.1785	0.0813	**0.9635**	**0.9550**	0.7228

(b) $K = 10$

Método	P@20	NDCG@20	MAP@20	AUC	Div@20	Nov@20
ALS	**0.3020**	**0.2918**	0.1416	0.9198	0.2847	0.8438
BPR	0.1440	0.1294	0.0474	0.9183	0.2847	**0.9361**
IBCF	0.2470	0.2216	0.0997	0.8279	0.3134	0.6534
LightFM (warp)	0.0410	0.1472	0.2141	0.9409	0.3039	0.6359
LightFM (w. BERT)	0.0420	0.1268	0.2034	0.9393	0.3257	0.5794
LightFM (w. ResNet)	0.0430	0.1109	**0.2226**	0.9398	0.3110	0.4633
NCF (Vanilla)	0.1055	0.1890	0.0788	0.9579	0.3058	0.6716
NCF (w. BERT embeddings)	0.1110	0.1924	0.0751	0.9563	0.3109	0.6897
NCF (w. ResNet images)	0.0990	0.1715	0.0672	0.9532	0.3089	0.7407
NCF (w. BERT and ResNet)	0.0975	0.1788	0.0735	0.9592	0.3085	0.7216
VBPR (w. BERT embeddings)	0.1045,	0.1899	0.0765	0.9549	0.2963	0.7380
VBPR (w. VGG16 images)	0.1205	0.2356	0.0977	0.9607	0.6549	0.7513
VBPR (w. BERT and VGG16)	0.1275	0.2369	0.1023	**0.9635**	**0.9636**	0.7567

(c) $K = 20$

Conclusiones y Trabajo a Futuro

Conclusiones

- **Precisión vs. diversidad/novedad:** No existe un "mejor absoluto": ALS es imbatible en accuracy, mientras que VBPR multimodal equilibra bien diversidad y ranking global. BPR sigue siendo la opción para maximizar novedad.
- **Valor de la multimodalidad**: La incorporación conjunta de texto e imagen en NCF y VBPR mejora significativamente AUC y diversidad, lo que sugiere que las portadas aportan señales complementarias a las descripciones.
- **Recomendaciones prácticas:** Para escenarios donde la métrica clave sea la precisión pura, ALS es adecuado. Si se busca un trade-off entre calidad de ranking, diversidad y novedad, VBPR con embeddings de BERT y VGG16 es la mejor apuesta.

Trabajo futuro

- Explorar modelos basados en **CLIP o finetuning multimodal end-to-end** podría potenciar aún más la coherencia entre señales visuales y textuales.
- Explorar **modelos de ensamble dinámicos** (por ejemplo, uno basado en interacciones, otro en texto, otro en imágenes) **ajustando automáticamente la contribución** de cada modelo por usuario.