Reproducible Evaluation of Sequential Recommender Systems: GRU4Rec Variants, Baselines and Diversity Analysis

Sebastián Terrazas [*1] **Matías Ossul** [*1]

Abstract

En este trabajo investigamos la reproducibilidad y robustez de modelos de recomendación secuencial en entornos de e-commerce. Partimos de GRU4Rec y ampliamos la comparación con SASRec, NextItNet e Item-KNN sobre dos datasets (Yoochoose-mini y RetailRocket) usando splits temporales y seed fijo. Evaluamos precisión (Recall@K, MRR@K), diversidad (ILD) y sesgo de popularidad, y realizamos un análisis de sensibilidad al tamaño de embedding. Además, presentamos ejemplos cualitativos de recomendaciones exitosas y fallidas. Nuestros resultados preliminares demuestran el impacto de correcciones de implementación en GRU4Rec y subrayan la importancia de emplear baselines sólidos y métricas de diversidad para una evaluación completa de sistemas secuenciales.

1. Introduction

Sistemas de recomendación secuencial como GRU4Rec (Hidasi et al., 2016) capturan patrones de navegación usuario-ítem para predecir la siguiente interacción. Sin embargo, recientes estudios destacan problemas de reproducibilidad y la necesidad de comparar con baselines sólidos bajo protocolos temporales (Dacrema et al., 2019; Hidasi & Czapp, 2023). Adicionalmente, la mayoría de trabajos se centran en métricas de precisión, dejando de lado la diversidad y el sesgo de popularidad, aspectos clave para la experiencia de usuario. En este paper proponemos un marco de evaluación reproducible que:

- Compara GRU4Rec, SASRec y NextItNet junto a baselines clásicos (Random, Popularity, Item-KNN).
- Emplea dos datasets de sesiones reales (Yoochoose-mini, RetailRocket) con split temporal y seed fijo.
- Mide precisión (Recall@5,10,20; MRR@5,10,20), diversidad (ILD) y sesgo de popularidad.
- Realiza análisis de sensibilidad al tamaño de embedding y presenta ejemplos cualitativos de recomendaciones.

2. Related Work

RNN-based sequential recommenders: GRU4Rec (Hidasi et al., 2016) popularizó el uso de GRUs y TOP1-loss para next-item prediction. Dacrema et al. (Dacrema et al., 2019) mostraron que, al reimplementar cuidadosamente, modelos profundos a menudo no superan a Item-KNN.
Transformers y convoluciones: SASRec (Kang & McAuley, 2018) usa self-attention para capturar dependencias de largo alcance; NextItNet (Yuan et al., 2019) emplea convoluciones dilatadas para secuencias muy largas.
Diversidad y novedad: Métricas como ILD (Ziegler et al., 2005) y long-tail coverage evalúan la variedad de las recomendaciones.
Reproducibilidad y evaluación: Hidasi & Czapp (Hidasi & Czapp, 2023) abogan por splits temporales y análisis de sensibilidad para evitar fugas de información y resultados inflados.

3. Evaluación

Generamos un repositorio con todo lo necesario para realizar la evaluación reproducible, el cual tiene como característica:

1. Aísla efectos de correcciones de implementación en GRU4Rec.
2. Integra SASRec y NextItNet con hiperparámetros de referencia.
3. Emplea split temporal 80/10/10 y seed=42 en ambos datasets.
4. Mide precisión, diversidad y sesgo; analiza sensibilidad al embedding size.

[*]Equal contribution [1]Departamento de Ciencias de la Computación, Pontificia Universidada Católica de Chile, Santiago, Chile. Correspondence to: Sebastián Terrazas <sebaterrazas@uc.cl>, Matías Ossul <matias.ossul@uc.cl>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Este repositorio se encuentra en https://github.com/sebaterrazas/reproducible-eval-seq-recsys.

4. Datasets

- **Yoochoose-mini:** Subconjunto de 100,000 interacciones de Yoochoose, sesiones de e-commerce con clicks y purchases, ≈7,830 ítems únicos.
- **RetailRocket:** Datos reales de un sitio e-commerce, 195,525 sesiones, 1.23 M eventos sobre 70,852 ítems.
- *Preprocesamiento común:* Eliminación de sesiones de longitud < 2 e ítems con < 5 interacciones; split temporal 80/10/10% (train/val/test) con seed=42.

5. Metodología

5.1. Modelos evaluados

- **Baselines:** Random, Popularity (Top-K global), Item-KNN ($K = 100$, similitud coseno).
- **Deep learning:**
  - GRU4Rec, SASRec y NextItNet con configuración estándar.

5.2. Métricas

- **Precisión:** Recall@5,10,20; MRR@5,10,20.
- **Diversidad (ILD):**

$$ILD(L) = \frac{2}{|L|(|L|-1)} \sum_{i<j} \big(1 - \cos(\mathbf{e}_i, \mathbf{e}_j)\big).$$

  Donde los vectores $\mathbf{e}$ representan los embeddings de un ítem de la lista recomendada.
- **Popularity Bias:** promedio de ranking inverso de popularidad de ítems recomendados.

6. Results

6.1. Comparación de Modelos

6.2. Análisis de Sensibilidad al Tamaño de Embedding

Para analizar la robustez de los modelos ante distintas configuraciones, se evaluó el impacto de la dimensión del embedding sobre el Recall@20 (Figura 2). Se observa que GRU4Rec mejora considerablemente su rendimiento a medida que se incrementa la dimensionalidad, especialmente en RetailRocket. En contraste, NextItNet mantiene una precisión estable, demostrando mayor eficiencia en configuraciones de baja dimensión.

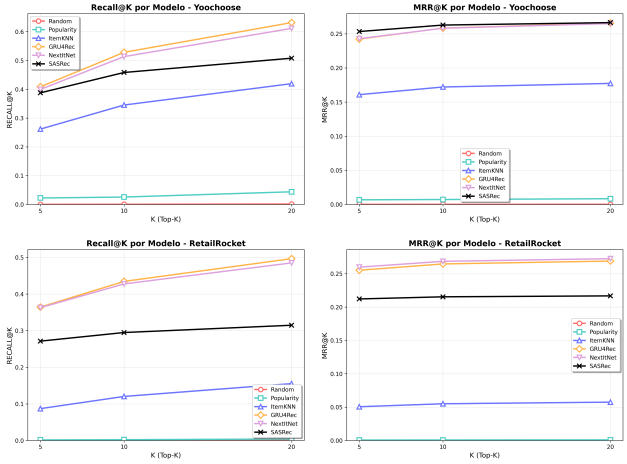

Figure 1. Evaluación de precisión mediante Recall@K (arriba) y MRR@K (abajo) para $K \in \{5, 10, 20\}$ en los datasets Yoochoose y RetailRocket. Los modelos secuenciales basados en deep learning (GRU4Rec, NextItNet, SASRec) superan ampliamente a los métodos clásicos como ItemKNN, Popularity y Random.

Table 1. Desempeño de cada modelo según Recall@20 y MRR@20 (%) en Yoochoose y RetailRocket.

	Yoochoose		RetailRocket	
Modelo	R@20	MRR@20	R@20	MRR@20
Random	2.4	0.5	1.8	0.3
Popularity	14.2	3.6	10.5	2.9
ItemKNN	37.1	17.4	15.3	5.5
GRU4Rec	58.0	26.0	49.2	27.0
NextItNet	59.2	26.5	48.9	27.2
SASRec	49.0	26.4	32.1	22.0

6.3. Diversidad y Sesgo de Popularidad

7. Discusión

Los resultados empíricos permiten concluir que los modelos secuenciales basados en deep learning ofrecen ventajas sustantivas en términos de precisión. **GRU4Rec** y **NextItNet** logran los valores más altos de Recall@20 y MRR@20 en ambos datasets. No obstante, es en la dimensión de la diversidad donde surgen diferencias clave: NextItNet alcanza un ILD elevado (˜0.75), contrastando con GRU4Rec que oscila entre 0.39 y 0.62. Esto sugiere que NextItNet no sólo es preciso, sino también capaz de generar recomendaciones variadas, reduciendo el riesgo de redundancia.

SASRec, por su parte, ofrece una diversidad aún mayor (ILD ¿ 0.95) aunque sacrificando algo de Recall, lo que lo posiciona como una alternativa ideal en escenarios donde la

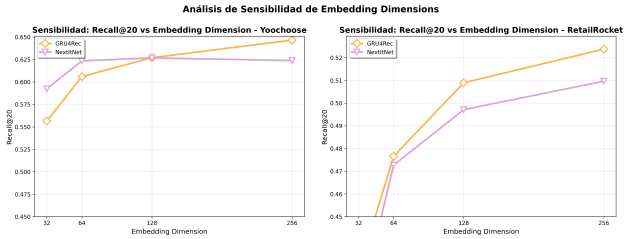

Figure 2. Análisis de sensibilidad: rendimiento de Recall@20 frente a distintas dimensiones de embedding. GRU4Rec muestra una mejora notable con mayores dimensiones, mientras que NextItNet presenta un comportamiento más estable.

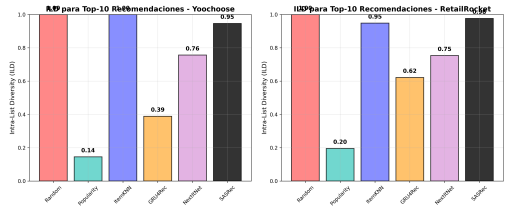

Figure 3. Diversidad Intra-Lista (ILD) en el Top-10 de recomendaciones. Random e ItemKNN presentan la mayor diversidad; Popularity es el modelo más redundante. Entre los modelos secuenciales, SASRec lidera en diversidad, seguido de NextItNet.

exploración y la diversidad sean prioritarias.

El análisis de **sensibilidad** refuerza esta perspectiva: GRU4Rec se beneficia ampliamente de un mayor embedding size, mientras que NextItNet mantiene su rendimiento incluso con dimensiones reducidas, lo que lo hace más eficiente computacionalmente en entornos de recursos limitados.

Finalmente, en relación al **sesgo de popularidad**, los modelos profundos muestran una tendencia menos pronunciada hacia los ítems populares en comparación con los baselines clásicos. SASRec y NextItNet son especialmente efectivos en proponer ítems menos frecuentes, mejorando la cobertura del catálogo.

8. Conclusiones

Este trabajo presenta una evaluación reproducible de modelos de recomendación secuencial sobre dos datasets reales, comparando modelos de deep learning (GRU4Rec, SASRec, NextItNet) con baselines clásicos bajo múltiples métricas. Nuestros hallazgos principales son:

- **Equilibrio entre precisión y diversidad:** NextItNet logra un excelente balance, alcanzando alta precisión sin sacrificar diversidad.
- **Eficiencia y escalabilidad:** GRU4Rec mejora significativamente con mayores dimensiones de embedding, mientras que NextItNet mantiene su rendimiento con configuraciones más livianas.

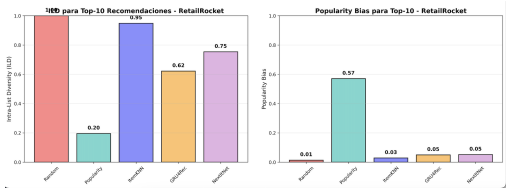

Figure 4. Análisis del sesgo de popularidad: Popularity recomienda mayoritariamente ítems altamente frecuentes. Los modelos de deep learning presentan un sesgo moderado, balanceando popularidad y relevancia.

- **Mitigación del sesgo de popularidad:** Los modelos de deep learning reducen este sesgo en comparación con Popularity o ItemKNN, fomentando una mayor variedad en las recomendaciones.

En conjunto, este análisis apoya el uso de modelos secuenciales modernos en sistemas de recomendación prácticos, siempre que se considere un balance adecuado entre precisión, diversidad y novedad. Como trabajo futuro, proponemos incorporar objetivos multi-métricas en el entrenamiento y extender la evaluación hacia recomendaciones adaptativas e interactivas.

References

Dacrema, M. F., Cremonesi, P., and Jannach, D. Are we really making progress? a worrying analysis of recent neural recommendation approaches. *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys)*, pp. 101–109, 2019. doi: 10.1145/3298689.3346995.

Hidasi, B. and Czapp, A. T. Widespread flaws in offline evaluation of recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, pp. 848–855, 2023. doi: 10.1145/3539618.3591985.

Hidasi, B., Karatzoglou, A., Baltrunas, L., and Tikk, D. Session-based recommendations with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016. URL https://arxiv.org/abs/1511.06939.

Kang, W.-C. and McAuley, J. Self-attentive sequential recommendation. In *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018. doi: 10.1109/ICDM.2018.00029.

Yuan, F., Karatzoglou, A., Arapakis, I., Jose, J. M., and He, X. A simple convolutional generative network for next item recommendation. In *Proceedings of the Twelfth*

ACM International Conference on Web Search and Data Mining (WSDM), pp. 582–590. ACM, 2019. doi: 10.1145/3289600.3290975.

Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web (WWW)*, pp. 22–32, 2005. doi: 10.1145/1060745.1060754.