A Lightweight Transformer for Next-Item Product Recommendation

M. Jeffrey Mei
jmei@wayfair.com
Wayfair LLC
Boston, MA, USA

Cole Zuber
czuber@wayfair.com
Wayfair LLC
Boston, MA, USA

Yasaman Khazaeni
ykhazaeni@wayfair.com
Wayfair LLC
Boston, MA, USA

ABSTRACT

We apply a transformer using sequential browse history to generate next-item product recommendations. Interpreting the learned item embeddings, we show that the model is able to implicitly learn price, popularity, style and functionality attributes without being explicitly passed these features during training. Our real-life test of this model on Wayfair's different international stores show mixed results (but overall win). Diagnosing the cause, we identify a useful metric (average number of customers browsing each product) to ensure good model convergence. We also find limitations of using standard metrics like recall and nDCG, which do not correctly account for the positional effects of showing items on the Wayfair website, and empirically determine a more accurate discount factor.

CCS CONCEPTS

• **Applied computing** → **Online shopping**; • **Information systems** → **Recommender systems**; • **Computing methodologies** → **Neural networks**; **Learning latent representations**; **Neural networks**.

KEYWORDS

transformers, style

ACM Reference Format:
M. Jeffrey Mei, Cole Zuber, and Yasaman Khazaeni. 2022. A Lightweight Transformer for Next-Item Product Recommendation. In *Sixteenth ACM Conference on Recommender Systems (RecSys '22), September 18–23, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3523227.3547491

1 INTRODUCTION

Identifying personalized recommendations to provide customers a better e-commerce experience has become increasingly important with the rise of online shopping and large catalog sizes. At Wayfair, we want to recommend the right products to customers at every step of their purchasing journey so that they can find their home needs more efficiently. These products may be very different at different stages of their purchasing journey. For example, customers may change their preferences over time (e.g. they may see a new kind of design/material/style on Wayfair that they like but were previously unaware of).

Earlier collaborative methods like matrix factorization use customers' overall set of product interactions (browsing history) and learn to connect items in some latent space, to identify and recommend products that are similar to a customer's overall browsing history. These methods do not distinguish how recently an item was viewed, and are thus slow to react to a change in customer preferences. To address this, models can add some sequential awareness. These have their own issues, e.g. difficulties in learning long-term dependencies (Markov Chains, Recurrent Neural Networks) and slow, non-parallelizable training (Long-Short Term Memory networks). Transformers attempted to address this by being easily parallelizable and being able to learn long-term dependencies using its attention mechanism [4]. SASRec incorporated this transformer architecture into a sequence-aware recommender system [2].

In this paper, we present Wayfair's Multi-headed Attention Recommender System (MARS), derived from SASRec, that uses a sequential input of viewed items to provide product recommendations that better match customers' latest preferences. Like SASRec, it uses a learned item embedding added to a learned positional embedding, with binary cross-entropy loss and a sigmoidal final activation function. To improve performance, we explored other additive embeddings for funnel stages (view, add to cart, order) and category (sofa/bed/rug/etc.). We analyze the learned item embeddings and find correlations to price, style, popularity and other attributes. This means that these features can be directly learned from the model and need not be passed in as inputs, greatly simplifying both training and inference.

For evaluation, we find that standard metrics like recall/nDCG do not accurately account for real-world positional effects, leading to a mismatch between expected and actual outcome (i.e. nDCG might suggest our model would win but a real-life A/B test might show a failure). We empirically determine the ideal discount for evaluating nDCG on next-item orders and find that is very closely approximated by the mean reciprocal rank for our use case. The contributions of our paper are as follows:

- We successfully deploy a transformer for next-item recommendation in the e-commerce domain, and show that the learned item embeddings are able to link item attributes like style, price, functionality and popularity.
- We identify a useful diagnostic metric for ensuring model convergence before training, which is the average number of distinct customers viewing each item during a specific time frame.
- We empirically show that mean reciprocal rank (MRR) is better for offline evaluation than recall/nDCG for estimating real-world impact; we optimize the discount rate for nDCG and show it closely approximates MRR.

RecSys '22, September 18–23, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9278-5/22/09.
https://doi.org/10.1145/3523227.3547491

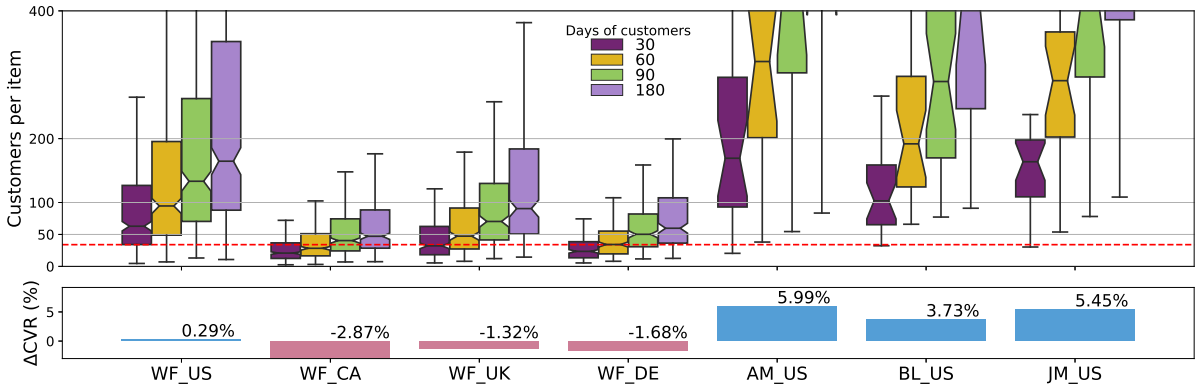

Figure 1: Top: distribution of average number of customers browsing each item in a product category. Bottom: observed change in order rate (ΔCVR) for each of Wayfair's stores for an A/B test using a model trained on 30 days of customers' interaction data for each store. It is clear that data thinness is causing poor A/B test results for WF_CA, WF_UK and WF_DE. The red line shows the 25th percentile for WF_US (34 customers per item), as a general target for other stores to match.

2 MODEL ARCHITECTURE

The architecture for MARS largely follows SASRec [2], with some changes to improve negative sampling and resurfacing of previously-viewed items. The sole input to the model is an ordered list of a customer's viewed items (including duplicates), and the set of items to score (and rank). We may not need to score/rank all items (e.g. if customers are browsing only sofas, we may want to recommend only sofas). For training, we used 30 days of customer interactions.

Our MARS A/B test was an overall win, although with considerable variation between different stores (Fig. 1). Wayfair has different stores for customers in US (WF_US), UK (WF_UK), Canada (WF_CA), and Germany (WF_DE). The US market, including specialty retailers Joss&Main (JM_US), AllModern (AM_US), BirchLane (BL_US), accounts for 80% of total revenue.

3 RESULTS

3.1 Training Data Quality

Fig. 1 shows the strong correlation between the average number of customers viewing each item and the overall online test result, measured as the change in order rate during an A/B test. Using the 25th percentile of WF_US as the baseline (~34 customers per item), the UK store needs 90 days of customers and the Canada/Germany stores need 180 days to match this threshold. A subsequent A/B test using these increased training set sizes showed positive ΔCVR for all stores.

3.2 Evaluation

For offline evaluation, we compared the mean reciprocal rank, recall and nDCG metrics. Because MARS was better at some (but not all) positions, this meant that the overall predicted outcome using these three metrics did not always agree. For example, for Wayfair Canada, our offline evaluation using recall and nDCG showed MARS underperforming control for the first 4 positions, and winning for all positions after that, whereas the mean reciprocal rank showed MARS slightly underperforming overall. The actual A/B test was not a win for this store (Fig. 1), although our A/B test showed similar results by position compared to our offline evaluation (i.e. underperforming for top 4 positions, then winning after that). This suggests that recall (which is not weighted by position) is inadequate for evaluation because it does not account for positional effects, and that the default discount of $\log_2$ for nDCG is too low for real-world use.

To empirically determine what the correct discount rate should be, we plot order rates by position (Fig. 2). We find that the decrease in order rates by position follows a monomial relationship; fitting the coefficient using a log-log plot (Fig. 2) gives a monomial of $y \propto x^{-1.02}$. Because our predictions have precisely one correct answer (relevancy = 1; all other predictions have relevancy = 0), calculating the (mean) nDCG simplifies to $\frac{1}{N}\sum_{x=1}^{N}\frac{1}{\log_2(x+1)}$. Substituting our empirical discount gives $\frac{1}{N}\sum_{x=1}^{N}\frac{1}{x^{1.02}} \approx \frac{1}{N}\sum_{x=1}^{N}\frac{1}{x}$; this last term is just MRR [5].

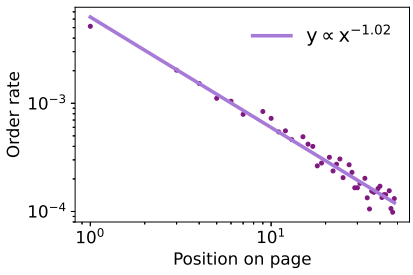

Figure 2: Order rates by position

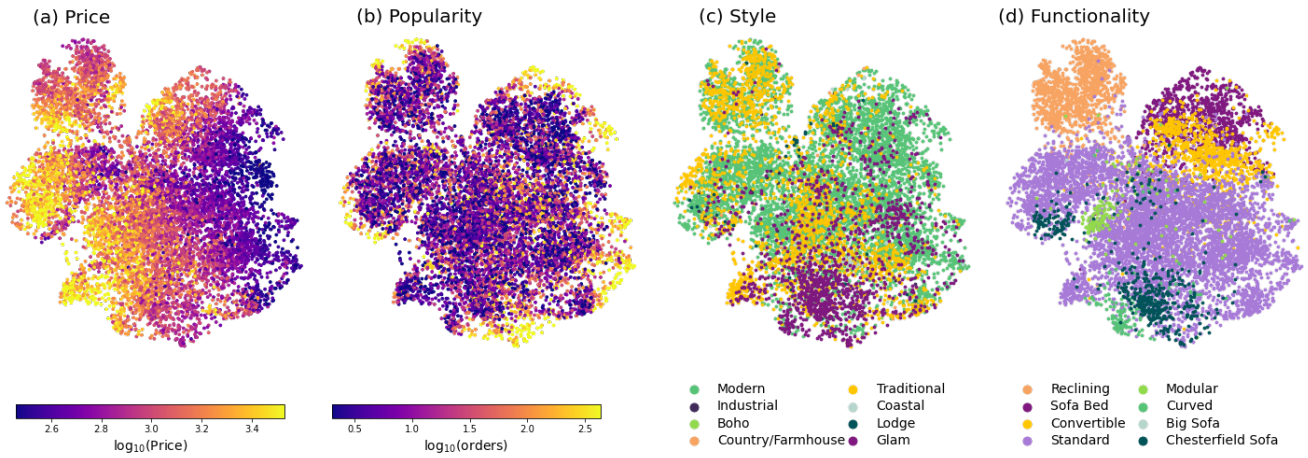

Figure 3: 2D projections (via UMAP [3]) of learned MARS item embeddings, correlated with various externally-provided attributes, for the top 10000 sofas in the WF_US catalog. MARS can learn these attributes despite not using any of this information during training.

4 DISCUSSION

4.1 Learned attributes

We expect that the learned item embeddings from MARS correspond strongly to inherent features that inform customer browsing. We compare the learned item embeddings for the top 10000 sofas in the WF_US catalog and compare to externally-provided attributes (not used in training) like price and style (Fig. 3). MARS has learned to cluster top-sellers (Fig. 3b), but has sub-grouped them by style/functionality (Fig. 3 c,d). The clustering for functionality is strongest; this suggests that customers care a lot about whether their sofa is 'reclining', 'convertible' and so on. 'Convertible' sofas are shown as an intermediate zone between 'standard' and 'sofa bed', suggesting that MARS does not treat these attributes categorically, but rather learns how to connect them on a continuum. For example, some customers who are looking for a 'sofa bed' may be satisfied with a 'convertible' sofa, but not a 'reclining' one.

4.2 Similarity of browsed items within a customer's browse history

We assume a customer's preferences are consistent across their browsed items. To identify this similarity in our model's learned embeddings, we compare the cosine similarity between different items within a customer's browse history against random items [1]. We see significantly more similarity between items that a customer browses as opposed to random items (Fig. 4), both for in-category and cross-category browse. We find the difference becomes more significant when removing the category signal which dominates the learned item embedding. For in-category browse, this high similarity may be because customers browse items that share similar physical attributes like functionality (Fig. 3d) or material, but for cross-category browse, items may share no physical attributes (e.g. sofas and coffee tables), and instead may share less tangible attributes like style or value. This similarity may be used to inform cross-category recommendations (e.g. to recommend a coffee table after a customer has purchased a sofa, before the customer has even viewed any coffee tables).

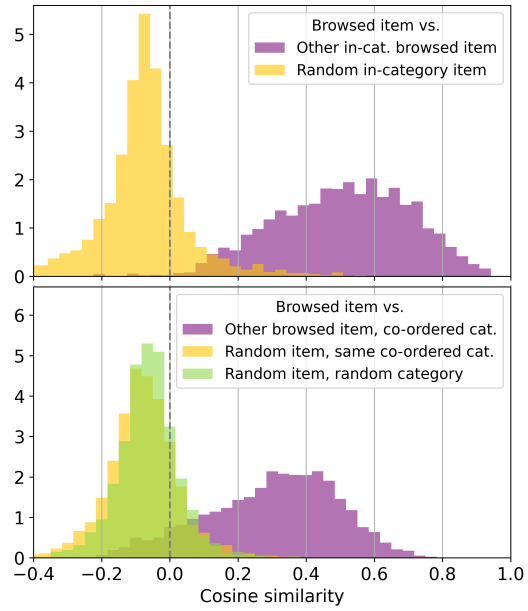

Figure 4: The cosine similarity of browsed items for both in-category (top) and cross-category (bottom) browsed items is higher than for random items. This suggests that MARS is learning both category-specific (e.g. material) and global (e.g. style) attributes that inform customer browse.

5 CONCLUSION

In this paper, we implement MARS, a transformer for next-item recommendation in the e-commerce domain. We demonstrate that attributes like item functionality, style, price and popularity can be learned implicitly by the model without being passed in for training, and find that these embeddings show a high degree of similarity within customers' in-category and cross-category browse. We also estimate a better discount factor for nDCG using real order rates by position and find this implies that mean reciprocal rank is a much better evaluation metric for our e-commerce use case.

6 SPEAKER BIO

Jeffrey Mei has been working as a data scientist at Wayfair for 1.5 years, where he makes models that better help customers find the right item from Wayfair's vast catalog. He completed his PhD in Oceanographic Engineering at the Massachusetts Institute of Technology in 2020, where he used computer vision and deep learning techniques to quantify and explain the relationship between snow surface textures and the underlying sea ice thickness. He is interested in interpreting learned embeddings and tweaking model architectures to better fit e-commerce needs.

REFERENCES

[1] Hao Jiang, Aakash Sabharwal, Adam Henderson, Diane Hu, and Liangjie Hong. 2019. Understanding the Role of Style in E-commerce Shopping. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3112–3120.

[2] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive Sequential Recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.

[3] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3, 29 (2018).

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in neural information processing systems* 30 (2017).

[5] Ellen M Voorhees et al. 1999. The TREC-8 Question Answering Track Report. In *TREC*, Vol. 99. 77–82.