When Attention Turn To Be Explanation. A Case Study in Recommender Systems*

Ricardo Anibal Matamoros Aragon[1,*,†], Italo Zoppis[1,*,†] and Sara Manzoni[1,*,†]

[1]Department of Informatics, Systems and Communication, University of Milano-Bicocca, 20126 Milano, Italy

Abstract

Recent studies in deep learning aim to understand how intermediate representations learned from *attention mechanisms* motivates the decisions of a predictive model and, consequently, provides information on the model's decision-making process. In fact, while the effectiveness of *attention* is a well-established topic, the power of attention coefficients to express explanations remains a somewhat controversial issue in the literature. In this work, we empirically evaluate the possibility of using *attention coefficients* to obtain faithful explanations for recommender systems. In particular, after showing how to use *attention* for explaining recommendations, we examine the robustness of our proposal.

Keywords

Attention Mechanism, Recommender System, Explainability

1. Introduction

Since its introduction, *attention mechanism* [1, 2, 3] has motivated researchers to evaluate its applications in designing faithfully explainable deep learning models [4, 5, 6]. However, while *attention* has become a mature tool to improve model performance, the ability of *attention coefficients* to express explanations remains somewhat controversial. The question found a fertile field of discussion after Jain and Wallace [7], not having found significant correlations with other *explainable* methods, concluded that attention alone could not help explain the predictions returned. An immediate response from Wiegreffe and Pinter intensified this research [8]. While acknowledging the importance of the discussion, they only supported few arguments raised by Jain and Wallace. Indeed, they reinforced the founding concept of their research: to deprive attention of a faithful exclusive explanatory role, however relegating it to the value of being able to coexist with several plausible explanations for a similar degree of faithfulness. The debate continues today: many insights on these topics include e.g., theoretical analyses of attention, the necessity to bring users in the loop, forcing attention to technically understand neural networks decisions [9, 10, 11]. In this paper we first propose a solution to use attentional coefficients in

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author.

†These authors contributed equally.

r.matamorosaragon@campus.unimib.it (R. A. M. Aragon)

https://www.unimib.it/ricardo-anibal-matamoros-aragon (R. A. M. Aragon)

0000-0002-1957-2530 (R. A. M. Aragon); 0000-0001-7312-7123 (I. Zoppis); 0000-0002-6406-536X (S. Manzoni)

Trace A: SC computation	Eq.n	Trace B: RC computation	Eq.n
$H_{sim} = W_A \cdot R^{(UU)^T}$	(1)	$\overrightarrow{h}^{(u),0} = W_B \cdot \overrightarrow{y}^{(u)}$	(1)
$\overrightarrow{e}_{sim} = \overrightarrow{w}_{sim} LeakyReLU(H_{sim})$	(2)	$e_{ui}^{(k)} = LeakyReLU(w_a(\overrightarrow{h}^{(u),k} \,\|\|\, \overrightarrow{h}^{(i),k}))$	(2)
$\overrightarrow{\alpha}_{sim} = Softmax(\overrightarrow{e}_{sim})$	(3)	$\alpha_{ui}^{(k)} = \frac{exp(e_{ui}^{(k)})}{\sum_{j \in N(u)} exp(e_{uj}^{(k)})}$	(3)
$\hat{H_{sim}} = \overrightarrow{\alpha}_{sim} \circ H_{sim}$	(4)		
$\overrightarrow{h}_{simcx} = \sum_j [\hat{H_{sim}}]_{\cdot,j}$	(5)	$\overrightarrow{h}^{(u),k+1} = \sigma\left(\sum_{j \in N(u)} \alpha_{uj}^{(k)} \overrightarrow{h}^{(j),k}\right)$	(4)

Table 1
Similarity and Recommendation coefficients (SCs & RCs): The RC, $\alpha_{ui}^{(k)}$, at the kth layer, is computed recursively for each u (and resource $i \in N(u)$). Equivalently, for each neighborood of u, we obtain an RC vector $\overrightarrow{\alpha}_{rec}^{(u),k}$ such that $\left[\overrightarrow{\alpha}_{rec}^{(u),k}\right]_i = \alpha_{ui}^{(k)}$.

order to understand the suggestion provided by recommender systems (section 2). We then use *counterfactual distribution* to test the robustness of our approach.

2. Methods

Recommendation systems (RS) represent important challenges for research and market [12, 13]. They act as information filtering suggesting the resources that are most relevant to certain users[14, 15, 16, 17]. Most of RSs operate now by using interactions and data collected from users on the assumption that people who have agreed in the assessment of certain items are likely to agree again in the future (collaborative filtering) [18, 19, 20]. We apply our investigation to neural-based collaborative filtering where both user-user and user-items interactions are typical neural network input [21, 22, 23]. The following definitions will be useful for better clarifying our analysis.

Heterogeneous graphs
A directed graph $G = (V, E, T^{(V)}, T^{(E)})$ is heterogeneous if nodes $v \in V$ and/or edges $e \in E$ are associated with mapping $\phi(v) : V \to T^{(V)}$ and $\varphi(e) : E \to T^{(E)}$, i.e., $\phi(v)$ and $\varphi(e)$ associate node types (Labels) and edge types (labels), respectively.

Bipartite Graph
A heterogeneous graph $G(V)$ is bipartite if its verteces V can be partitioned into two subsets I_1 and I_2 such that each edge $\forall e \in E$ has one of its two ends in I_1 and the other in I_2.

2.1. Attention based Explaination for Recommendation

User-item interactive behavior is formulated through a heterogeneous bipartite graph $G(V)$ with $V = U \cup I$, where V is partitioned by users U and Items I, respectively. Recommendations are then supplied by applying neural-based convolutional filtering with attention over $G(V)$ [24, 25]. In particular, we will make user-user similarities and past user rating participate to the (attention-based) recommendations through two sets of (attention) coefficients that we call here, *Recommendation* and *Similarity* coefficients.

Similarity coefficients (SCs) focus on similarities between user rating. Let $\overrightarrow{y}^{(u)} = [s_1, s_2, \dots s_n]$ be the profile of user u, reporting the scores u assigned to n resources. Moreover, let $R^{(UU)}$ be an $m \times m$ matrix (*User-User* matrix) whose row $\overrightarrow{s}^{(u)} = [R^{(UU)}]_{u,\cdot}$ gives the (normalized dot product) similarities between $\overrightarrow{y}^{(u)}$ and other m score profiles $\overrightarrow{y}^{(j)}$, i.e. for each pair of users (u, j) we have $[R^{(UU)}]_{u,j} = (\overrightarrow{y}^{(u)} \cdot \overrightarrow{y}^{(j)})/(||\overrightarrow{y}^{(u)}|| \cdot ||\overrightarrow{y}^{(j)}||)$ [1]. The SCs (annotated as $\overrightarrow{\alpha}^{(u)}_{sim}$) are computed in Tab. 1 from the embedded (matrix) $H^{(UU)}$ of $R^{(UU)^T}$ (Trace A, Eq. 1) by normalizing the attentional values $\overrightarrow{e}_{sim}$ with the Softmax function (Trace A, Eqs. 1-4). Output of this mechanism is an attention-based context vector h_{simcx} that is obtained summing all the scaled features (coloumns) in $H^{(UU)}$ (Trace A, Eq. 5).

Recommendation Coefficients (RCs) focus on relevant user-item evaluation (Tab. 1, Trace B). Intuition is to assign high coefficients to user-item pairs which are indeed good recommendations, low scores otherwise. In this case, RCs are obtained from the latent representation $\overrightarrow{h}^{(u),k}$ of the user' s score $\overrightarrow{y}^{(u)}$ (Trace B, Eq. 1) after k embedding layers (recursive computation reported in Trace B). Output of this mechanism is an attention-based context vector $\overrightarrow{h}^{(u),k+1}$ (Trace B, Eq. 4) obtained by properly scaling the latent representation of each resource j in the neighborhood of u [24, 26, 27].

A dense fully connected layer F_Θ parameterized by (learnable) Θ returns the estimated rating $\overrightarrow{\hat{y}}^{(u)} \in \mathcal{R}^n$ of n resources as a function of the aggregation of the 2 previous mechanism output, i.e., $\overrightarrow{\hat{y}}^{(u)} = F_\Theta(\overrightarrow{h}_{simcx} || \overrightarrow{h}^{(u),k+1})$. In this way, by interpreting $[\overrightarrow{\hat{y}}_u]_j$ as the probability of j being relevant to u, then we can suggest item $j^* = argmax_{j \in I}\{[\overrightarrow{\hat{y}}^u]_j \mid \exists v \in U : [\overrightarrow{\hat{y}}^u]_j \geq \delta, [\overrightarrow{\alpha}^{(u)}_{sim}]_v \geq \delta, [\overrightarrow{y}^{(v)}]_j \geq \delta\}$. Thus, we recommend relevant resources i to user u with high probability (w.h.p.) i.e., $[\overrightarrow{\hat{y}}^u]_j \geq \delta$ (for large $\delta \in [0, 1]$) among all those resources recommendable in I for which there is at least one user v similar to u w.h.p. ($[\overrightarrow{\alpha}^{(u)}_{sim}]_v \geq \delta$) who in the past has given i an high relevance ($[\overrightarrow{y}^{(v)}]_j \geq \delta$) [2]. A faithful decision should be motivated, for example, if we could make the attention coefficients as influential as possible for the selection of the elements; Since item selection depends on both score distribution (model output) and attention coefficients (here we will use RCs), we should reduce the divergence between them as much as possible. Equivalentely, this can be expressed by the loss $\mathcal{L} = \sum_j [\overrightarrow{y}^u]_j \log([\overrightarrow{\hat{y}}^u]_j) + \sum_j [\overrightarrow{y}^{(u)}]_j \log([\overrightarrow{\alpha}^{(u)}_{rec}]_j)$ where both the divergence between observed and predicted distribution (i.e., cross-entropy in the first term) and the divergence between observed and RC distribution (i.e., cross-entropy in the second term) is taken into account [3]. Because of the definition of both j^* (argmax of the score distribution) and the loss we get at the minimum (of the gradient optimization) an estimation of the ground probability accomplished with the requirement for the attention coefficient to reflect the prediction as best as possible. We then literally provide an explanation as follow: Resource j

[1] Here we use the following notation: $[A]_{i,j}$ the component in the ith row, jth coloumn of matrix A; $[A]_{i,\cdot}$ the ith row of A; $[\overrightarrow{y}]_i$ the ith component of vector $\overrightarrow{y}$.

[2] Note that, during a test, it is sufficient to limit the recommendation to those items that have not already been rated by the users being recommended, we call this user here as Target user.

[3] Please refer to table 1, Trace B, for the definition of $\overrightarrow{\alpha}^{(u)}_{rec}$.

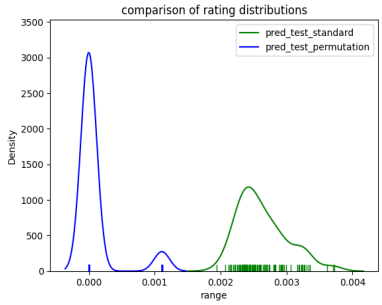

Figure 1: Scores before and after permutation.

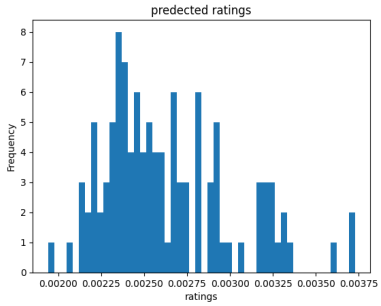

Figure 2: Predicted vs observed scores

has been suggested as relevant for u due to a relevant similarity of u with a user that, in the past, assigned a relevant score to the item suggested.

The second objective of our investigation was to check whether the score predicted could be different, if the attentional mechanism had emphasized different focus on the attending information or, in other way, if a different distribution of attentional coefficients (i.e. here we use recommendation coefficients) can provide similar suggestion. In this case nullifying the usefulness of attention as explanation. We considered a "counterfactual distribution" of attentional coefficient as [7]. Practically, we simply scramble the original attention weights $\hat{\alpha}$, re-assigning each value to an arbitrary, randomly sampled index (input feature).

3. Results, Discussion and Conclusion

The debate inherent to the attention explanation relationship has provided diverse literature among researchers. Despite an evidence that attention is not explanation in general, by properly adapting attention mechanism in task-oriented prediction (E.g.,[28, 29] for a review), we showed how attention coefficients can be correlated to user-item pairs in RS, thus motivating the choices of the neural decision-making process [30]. Our results (on the Movielens dataset [31] restricted to 20 users and 250 items) provide the following qualitative observations (Fig. 1). [4]

- When permuting the RC values, a change in the suggested scores is observed (Fig. 1).
- System obtains appreciable performances distributing the predicted score over the observed one (Fig. 2).

In conclusion, if the attention mechanism emphasized a different focus, the prediction would differ. This argumentation will naturally need to be quantified in our future research.

[4]We report preliminary qualitative results visualizing comparisons between distribution of observed scores, predicted scores, and RC attention coefficients.

References

[1] S. Chaudhari, V. Mithal, G. Polatkan, R. Ramanath, An attentive survey of attention models, ACM Transactions on Intelligent Systems and Technology (TIST) 12 (2021) 1–32.

[2] Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, Neurocomputing 452 (2021) 48–62.

[3] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[4] Y. Zhang, Q. Ma, Dual attention model for citation recommendation with analyses on explainability of attention mechanisms and qualitative experiments, Computational Linguistics 48 (2022) 403–470.

[5] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, S. Zhang, Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation, IEEE transactions on medical imaging 40 (2020) 699–711.

[6] H. Chefer, S. Gur, L. Wolf, Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 397–406.

[7] J. Burstein, C. Doran, T. Solorio, Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers), in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019.

[8] S. Wiegreffe, Y. Pinter, Attention is not not explanation, arXiv preprint arXiv:1908.04626 (2019).

[9] C. Chen, A. Ross, An explainable attention-guided iris presentation attack detector, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 97–106.

[10] Y. Zhang, X. Chen, et al., Explainable recommendation: A survey and new perspectives, Foundations and Trends® in Information Retrieval 14 (2020) 1–101.

[11] K. Fiok, F. V. Farahani, W. Karwowski, T. Ahram, Explainable artificial intelligence for education and training, The Journal of Defense Modeling and Simulation 19 (2022) 133–144.

[12] J. Lu, D. Wu, M. Mao, W. Wang, G. Zhang, Recommender system application developments: a survey, Decision support systems 74 (2015) 12–32.

[13] L. Marconi, R. Matamoros Aragon, I. Zoppis, S. Manzoni, G. Mauri, F. Epifania, et al., Approaching explainable recommendations for personalized social learning the current stage of the educational platform" whoteach", in: CEUR WORKSHOP PROCEEDINGS, volume 2742, CEUR-WS, 2020, pp. 104–111.

[14] A. Calero Valdez, M. Ziefle, K. Verbert, Hci for recommender systems: the past, the present and the future, in: Proceedings of the 10th ACM conference on recommender systems, 2016, pp. 123–126.

[15] F. Ricci, L. Rokach, B. Shapira, Introduction to recommender systems handbook, in: Recommender systems handbook, Springer, 2010, pp. 1–35.

[16] C. Zhou, M. Leng, Z. Liu, X. Cui, J. Yu, The impact of recommender systems and pricing strategies on brand competition and consumer search, Electronic Commerce Research

and Applications 53 (2022) 101144.
[17] P. Lops, M. De Gemmis, G. Semeraro, Content-based recommender systems: State of the art and trends, Recommender systems handbook (2011) 73–105.
[18] X. Su, T. M. Khoshgoftaar, A survey of collaborative filtering techniques, Advances in artificial intelligence 2009 (2009).
[19] J. L. Herlocker, J. A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM conference on Computer supported cooperative work, 2000, pp. 241–250.
[20] J. B. Schafer, D. Frankowski, J. Herlocker, S. Sen, Collaborative filtering recommender systems, in: The adaptive web: methods and strategies of web personalization, Springer, 2007, pp. 291–324.
[21] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: Proceedings of the 26th international conference on world wide web, 2017, pp. 173–182.
[22] M. Qu, J. Tang, J. Shang, X. Ren, M. Zhang, J. Han, An attention-based collaboration framework for multi-view network representation learning, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 1767–1776.
[23] C. Vassiliou, D. Stamoulis, D. Martakos, S. Athanassopoulos, A recommender system framework combining neural networks & collaborative filtering, in: Proceedings of the 5th WSEAS international conference on Instrumentation, measurement, circuits and systems, 2006, pp. 285–290.
[24] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al., Graph attention networks, stat 1050 (2017) 10–48550.
[25] M. Zhang, S. Wu, M. Gao, X. Jiang, K. Xu, L. Wang, Personalized graph neural networks with attention mechanism for session-aware recommendation, IEEE Transactions on Knowledge and Data Engineering 34 (2020) 3946–3957.
[26] J. B. Lee, R. A. Rossi, S. Kim, N. K. Ahmed, E. Koh, Attention models in graphs: A survey, ACM Transactions on Knowledge Discovery from Data (TKDD) 13 (2019) 1–25.
[27] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, Convolutional networks on graphs for learning molecular fingerprints, Advances in neural information processing systems 28 (2015).
[28] A. Bibal, R. Cardon, D. Alfter, R. Wilkens, X. Wang, T. François, P. Watrin, Is attention explanation? an introduction to the debate, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 3889–3900.
[29] X. Wang, X. He, Y. Cao, M. Liu, T.-S. Chua, Kgat: Knowledge graph attention network for recommendation, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 950–958.
[30] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, H. Zha, Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 765–774.
[31] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, Acm transactions on interactive intelligent systems (tiis) 5 (2015) 1–19.