Counterfactual Explanations for Neural Recommenders

Khanh Hiep Tran, Azin Ghazimatin, Rishiraj Saha Roy

Max Planck Institute for Informatics, Germany

Motivation

- Understanding recommendations increases **trust** and **satisfaction** of users
- Neural recommenders are state-of-the-art but too **complex**
- Existing explanation methods for neural models are **limited**:
  - Rely on **attention**
  - **Not tangible or actionable** for end-users
  - Have **privacy** concerns
  - Need **external information**

↳ Need **counterfactual explanations**: A set of the user's own actions, when removed, produces a different recommendation.

Approach

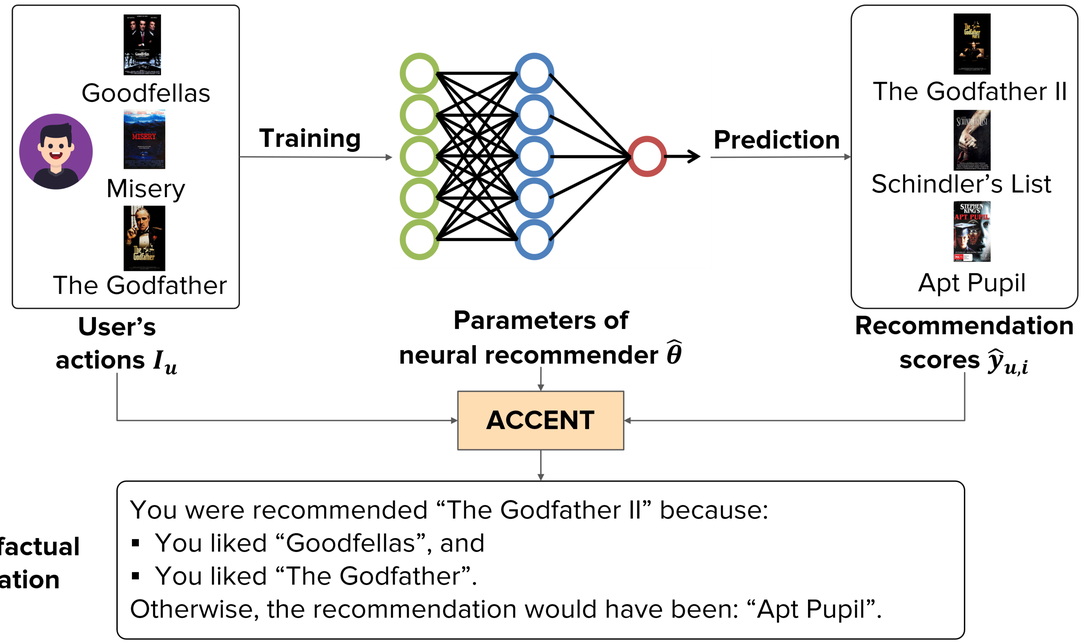
- Extend the basic idea in **PRINCE** [2] to neural models
- Estimate influence of actions on item scores using **Fast Influence Analysis (FIA)** [1]
- Extend influence on one item to **influence on pairs of items**
- **Iteratively closing the gap** between the recommendation and the replacement item with most influential actions

↳ **ACCENT**: **Ac**tion-based **C**ounterfactual **E**xplanations for **N**eural Recommenders for **T**angibility.

Method

For each replacement item **rec***:

- **For each** interaction **z** of user u:
  - **Calculate** the **influence** of **z** on scores of **rec** and **rec*** (using FIA)
- **Sort** interactions by **influence on the score gap**
- **Add** interactions to the result set **until the gap is filled**
- **Update** the smallest result set found so far



Candidate top-k set of replacement items		k = 5		k = 10		k = 20	
Recommender model	**Explanation model**	**CF percentage**	**CF set size**	**CF percentage**	**CF set size**	**CF percentage**	**CF set size**
NCF [3]	**Pure FIA [1]**	54.20	9.08	56.19	9.46	55.75	9.50
	FIA [1]	55.97	7.98	56.19	7.80	55.75	7.84
	ACCENT (proposed)	**57.30**	**4.73***	**57.74**	**4.69***	**57.08**	**4.62***
RCF [5]	**Pure Attention [5]**	73.01	9.36	73.45	7.94	74.34	7.75
	Attention [5]	76.99	3.55	76.99	3.53	76.99	3.51
	Pure FIA [1]	80.75	4.85	81.19	4.62	81.86	4.72
	FIA [1]	81.64	4.15	81.86	4.10	81.86	4.10
	ACCENT (proposed)	**81.86†**	**2.83*†**	**82.08†**	**2.75*†**	**82.08†**	**2.74*†**

Best values in each column are in bold. * and † denote statistical significance of ACCENT over FIA and Attention, respectively.

Table 1: Performance comparison of ACCENT with baselines on our sample of the MovieLens 100K benchmark.

Recommendation	ACCENT Explanation	Replacement
The Silence of the Lambs	Contact Fargo	Donnie Brasco
Titanic	True Romance The Basketball Diaries	East Of Eden
The Devil's Advocate	Speed Eraser It's A Wonderful Life	My Fair Lady
Contact	Forrest Gump Nell	Das Boot

Table 2: Counterfactual sets generated by ACCENT.

Results

- ACCENT can find **concise counterfactual explanations**
- Current methods based on **attention and influence fall short** of ACCENT in percentage and size of CF sets
- ACCENT is applicable to a **broad class of neural models**

References

1. Weiyu Cheng, Yanyan Shen, Linpeng Huang, and Yanmin Zhu. 2019. Incorporating Interpretability into Latent Factor Models via Fast Influence Analysis. In KDD.
2. Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. 2020. PRINCE: Provider-Side Interpretability with Counterfactual Explanations in Recommender Systems. In WSDM.
3. Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In WWW.
4. Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In ICML.
5. Xin Xin, Xiangnan He, Yongfeng Zhang, Yongdong Zhang, and Joemon Jose. 2019. Relational Collaborative Filtering: Modeling Multiple Item Relations for Recommendation. In SIGIR.

UNIVERSITÄT DES SAARLANDES

max planck institut informatik

Contact: rishiraj@mpi-inf.mpg.de

Data and Code: https://github.com/hieptk/accent