Analysis of Google's multitask ranking system vs. Amazon.com' s Item-to-Item Collaborative Filtering

Introduction

In this paper, I would like to provide an overview, challenges, solution to the challenges and further exploration on the mechanism of the systems on both the Google's multitask ranking system and Amazon.com' s Item-to-Item Collaborative Filtering. In the end of the paper, I also will include some comments on the comparison of the two system.

Summary of Google's multitask ranking system used for video recommendation

Google's multitask ranking system is a large scale multi-objective ranking system for recommending their users what video they might want to watch next on an industrial video sharing platform. There are some issues with the existing system and their researchers have found a solution to address them.

Challenges in current Google's recommending system

Two main challenges in the current multitask ranking system that Google uses have been identified by their researchers. The first one challenge different and occasional conflicting objectives that need optimization [1]. For instance, a video that is watched by a user should be considered lower ranking compares to a video that is rated highly or shared often by the user. The second challenge is systematic implicit bias [1]. For instance, a video might be clicked and/or viewed by a user merely due to it is highly ranked rather than liked by the user. This is creating a feedback loop effect and shows the models trained using data generated from the existing system is biased.

Solution to address the challenges

An efficient multitask neural network architecture was created by the Google researchers for the existing ranking system to improve the said challenges. The general concept behind it is to adopt Multi-gate Mixture-of-Experts (MMoE) to the Wide & Deep model architecture for multitask learning [1]. To address the systematic implicit bias challenge, a shallow tower was introduced in order to remove the selection bias [1]. The mechanism of the ranking system architecture is to consume user logs as training data. Multi-gate Mixture-of-Experts layer then is built for user behaviors' predictions. There are two types of user behaviors to be predicted by the system, engagement and satisfaction. Finally, the ranking selection bias is removed by a shallow tower and the final resulting ranking score is generated after combining with the multiple predictions [1].

Deep dive into the system and performance analysis

The workflow of the system can be summarized as two main parts: the main tower which is responsible for user-utility component and a shallow tower which is responsible for removing bias component. MMoE is basically composed of Multi-Layer Perceptron and rectified linear units (ReLU). Input features are learned by each Mixture-of-Experts and output is generated by the MMoE layer to feed into a gating network. Gating network's output will become inputs of different types of objective functions for training, and each of the function communicates with the Mixture-of-Experts with different input features via the gating network to decide how many and which one(s) they think are relevant for deciding the objective function. On the side, a shallow tower, that is trained to discount bias like position of the recommendation and to predict whether a bias

exists or not, is also generating a score and feed it to the objective functions to generate an overall score [1].
The experiments were setup using Tensor Processing Units (TPUs) to build the training model and TFX Servo to serve it. Offline experiments were conducted via monitoring AUC for classification task and regression tasks of squared error. A/B testing with the production system was used for live experiment. The results showed that this model was able to perform better in these experiments and introduction of a shallow tower indeed helped to improve the engagement metric [1].

Summary of Amazon.com' s Item-to-Item Collaborative Filtering
Amazon.com uses Item-to-Item Collaborative Filtering as their recommendation system to be used as a targeted marketing tool. The concept behind it is to match the purchases and rated items of each user to similar items and combines those similar items into a recommendation list [2]. The general theory of the algorithm that decides which items are the most similar matched for a given item is build a similar-item table by finding items that customer might potentially purchase as a bundle [2].

Challenge in the initial design
The team in Amazon.com initially designed the algorithm as to build a product-to-product matrix first via iterating through all available item pairs so that a similarity metric for each pair can be computed. There is one challenge in it, that is many of the computed product pairs that they computed using the algorithm do not share common customers which makes the method inefficient specifically regarding processing time and memory usage [2].

Solution to address the Challenges
The team in Amazon.com found another method to do the iteration to improve the inefficiency. First, the algorithm will iterate through each item in the product catalog, ad within this iteration, it will also iterate through all the customers who have bought the current item in the catalog. For the current customer who purchased the current item, it will check if the same customer also purchased a similar item. Finally, the similarity between the two items will be computed. The method is executed using the cosine measure and computed offline since it is an extreme time intensive algorithm. Another benefit of this method is scalability. Since Amazon.com has nearly a hundred million users [3], designing a system that can scale is crucial for them. For scalability, the item-to-item collaborative filtering system allows the expensive computations to be completed offline while keeping the online portion of the system simple which is looking up similar items for the previous purchases and ratings of each user. The online portion is dependent on the number of titles the user has purchased or rated. Therefore, the system can achieve high performance even if the data set is extremely large [2].

Conclusion
To compare Google' s multitask ranking system with Amazon.com' s Item-to-Item Collaborative Filtering, I think both have good performance. Since I am a user of both platforms, I think both provide good and relevant recommendations for me. Based on the number of active users, I think Google's system perhaps is even better at scalability because they have over a billion of users [1] whereas Amazon.com do not have similar number of users.

References:

[1] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, Ed Chi. “Recommending What Video to Watch Next: A Multitask Ranking System”. RecSys ’19, September 16–20, 2019. Copenhagen, Denmark ACM ISBN 978-1-4503-6243-6/19/09.

[2] G. Linden, Brent Smith, J. York. “Amazon.com Recommendations: Item-to-Item Collaborative Filtering”. 2003. Computer Science. IEEE Internet Comput. DOI:10.1109/MIC.2003.1167344. Corpus ID: 14604122.

[3] Statista. 2021. Amazon’s app is the second-most popular shopping app in the US, used by 98 million users every month.