Google's multitask ranking system

Introduction

The paper introduces an efficient way of recommending users the next video they can watch based on the current video the user is watching and also based on some user data background knowledge. The current state-of-the-art recommendation systems face multiple real-world challenges such as selection biases in user feedback and too many ranking objectives. Thus, the paper aims to solve these above challenges by extending the Wide & Deep framework and using the Multi-gate Mixture-of-Experts (MMoE) technique.

Body

There are generally two tasks that any recommendation system follows. The first task is generating possible candidates and the second task is ranking these candidates. The paper focuses on the second task, i.e ranking. The authors have grouped the multiple objectives into two categories. The first category is engagement objectives which include videos clicked on by the user, the degree of engagement of the user with recommended videos, etc. The second category is satisfaction objectives which include things like the user liking a video, leaving a rating, commenting, etc. To estimate the different types of the above parameters the paper uses the Multi-gate Mixture-of-Experts model followed by multiple gating networks. To solve the second major problem of user implicit biases, the authors propose the shallow tower.

The paper discusses three categories of previous related work namely, industrial recommendation systems, multi-objective learning for recommendation systems and understanding and modeling biases in training data. The paper also considers other features like multimodal feature space and scalability in addition to the two challenges mentioned above. The approach the authors have chosen for making predictions is the pointwise approach instead of the pairwise approach as the pointwise approach allows to efficiently scale to a large number of candidates and is also simple to implement.

The Multi-gate Mixture-of-Experts model architecture has experts which are distributed across the different objectives. For the proposed ranking system, the authors have chosen to add the experts on top of a shared hidden layer instead of the main input layer to reduce the training and serving costs. The implementation of the MMoE structure is nothing but a combination of multilayer perceptrons with ReLU activations. The outputs obtained from the MMoE model are then set as inputs to a network of gates. The gating networks are essentially linear transformations of the input with a sigmoid activation function. Then, by utilizing multiple gating networks, each of the objectives can choose one or more experts that are relevant for deciding that objective function.

The next step is introducing the shallow tower for handling the implicit bias problem. This issue arises as the user clicks the top-most video that is recommended to them but this recommendation may always not be the right one as the user might click on the video because it was suggested on the top instead of them being really interested in it or it could even happen that the video recommended is irrelevant. The paper is essentially trying to eliminate this position bias from the recommendation model. There are two components that are used for this purpose - a user-utility component which is a part of the main tower, and a bias component which is a part of the shallow tower. The shallow tower is trained using features that contribute to the selection bias. After that, the trained model tries to predict whether a bias was involved in the recommendation or not.

The results presented in the paper are tested on Youtube which is one of the largest and most popular video sharing platforms. The live experimental results are compared on factors including both 4 and 8 experts in the MMoE model. The MMoE model is then compared with the shared-bottom model. The results show that the MMoE architectural structure shows significant improvement in performance on both engagement and satisfaction factors.

Although the proposed recommendation system in the paper has achieved and tackled a lot of challenges in the existing systems, the paper could explore a new architecture which enables improved stability, trainability and expressiveness. Further additions could include a more efficient model that reduces costs even more, and also models that can identify potential biases automatically.

Conclusion

This paper has brilliantly solved the two major challenges faced by existing systems like the multiple competing ranking objectives and implicit selection biases. The paper solved these challenges by introducing the Multi-gate Mixture-of-Experts (MMoE) model architecture and the shallow tower. The results of the experiments showed that these proposed methods increased the performance substantially on both the engagement and satisfaction metrics.

References

[1] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi and Ed Chi 2019. Recommending what video to watch next: a multitask ranking system. RecSys '19: Thirteenth ACM Conference on Recommender Systems, 43–51