CS410 – Tech Review

Intro

This paper[1] addresses how to improve the current recommendation system, especially recommending what video to watch next on YouTube, one of the world's largest video sharing platforms. The authors first identify the challenges, propose the solutions to tackle the challenges, and then evaluate the results to prove the improvement.

The identified challenges include conflicting objectives the recommendation system intends to optimize for, implicit bias in the system, learning representation from multimodal feature space for recommendation, scalability for building a recommendation system for billions of users and videos that can be efficient during serving.

High-Level Proposed Framework

For the proposed solution, the authors extend the Wide & Deep model architecture by adopting Multi-gate Mixture-of-Experts (MMOE) for multitask. Besides, they also include a shallow tower to model and remove selection bias.

To tackle the conflicting objectives the recommendation system intends to optimize for, the authors group multiple objectives into two categories, engagement objectives and satisfaction objectives. Then, use MMOE to learn parameters to share across potentially conflicting objectives. The Mixture-of-Experts architecture modularizes input layer into experts, each of which focuses on different aspects of input.

To model and reduce the implicit bias, such as selection bias, from training data, the authors add a shallow tower to the main model. The shallow tower takes input related to the selection bias and outputs a scalar serving as a bias term to the final prediction of the main model. The model factorizes the label in training data into two parts: the unbiased user utility learned from the shallow tower, and the estimated propensity score learned from the shallow tower.

To learn representation from multimodal feature space for recommendation, the authors extract features such as video meta-data and video content signals for each video as its

[1] Recommending What Video to Watch Next: A Multitask Ranking System https://daiwk.github.io/assets/youtube-multitask.pdf

representation. The authors also use features such as user demographics, device, time, and location for context.

To deal with scalability, the authors propose different solutions for two stages in the recommendation system. They retrieve a few hundred candidates from a huge corpus at the candidate generation stage, and provides a score for each candidate and generates the final ranked list at ranking system stage.

Evaluations

The authors design and conduct offline and live experiments to verify the effectiveness of multitask learning and removing a common type of selection bias. By directly using position feature as input feature, it eliminates position bias so that the lower the position, the smaller the learned bias is. Besides, the proposed methods increase engagement metrics by 0.01% and 0.24% by using adversarial loss methods and adding shallow tower to the main model respectively. Eventually, the proposed framework show significant improvements comparing with state-of-the-art baseline methods.

Conclusion

Even though the final evaluations show significant improvements, this proposed framework still face several challenges, such as tradeoff between effectiveness and efficiency, other types of biases apart from selection bias, evaluation challenges, etc. The improved model might be complicated resulting in increasing latency in generating recommended items in real time, which indirectly affects user experiences and serving costs; apart from selection bias that was captured in this proposed framework, there might be other unpredictable, potential biases in the framework that the authors don't know; using offline evaluation to evaluate each online prediction task might have misalignment to online metrics.

In the future improvement, there is another recent work that might improve current model stability without hurting prediction performance but to balance stability, trainability and expressiveness in multi-objective ranking; in order to reduce negative impact on user experiences, it is required to compress the recommendation model to decrease the serving cost. Thus, exploring different types of model compression is crucial for the advanced

[1] Recommending What Video to Watch Next: A Multitask Ranking System https://daiwk.github.io/assets/youtube-multitask.pdf

ranking and recommendation models; in addition to model harmful, unknown biases from the framework, exploring and developing model architectures which can automatically identify potential biases from the training data would be the next big step.

[1] Recommending What Video to Watch Next: A Multitask Ranking System https://daiwk.github.io/assets/youtube-multitask.pdf