Google's Multitask Ranking system for Video Recommendation

Introduction:

An ideal video recommendation system would provide a list of suggested videos based on the user's current video being watched. In real time, there are challenges in providing personalized video recommendation for users. In this article, we will understand the challenges and the solution available using Multitask Ranking system.

Challenges:

Real-time video recommendation in any video sharing platform is challenging either in terms of correctness or scalability. Recommendation videos for any user is determined using certain ranking algorithms. Each ranking algorithm has its own objective and sometimes they are conflicting. Also, user's feedback is considered as a critical factor in determining the ranking. User's feedback turned out to be biased at times when the user watched the video for just being ranked high and shown up, not because user liked it. User's satisfaction index is very low though user has watched it. It would not be the right decision to recommend videos based on this case. Moreover, user behavior can change every day. It is expensive to get high volume of training data for modeling.

Also, learning from the user data and predicting is challenging due to different behaviors of users. Content-based or Collaborative based filtering is not as effective as models based on neural networks. Whereas, neural based networks might not have features to support mixture of spaces like text from video titles. Certain multi-objective ranking system that supports such features are not scaling up.

Google's Multitask ranking system is considered to provide solutions to address the challenges.

Solution:

Google's Multitask ranking system primarily uses the state-of-the-art multitask learning model architecture called Multi-gate Mixture-of-Experts (MMoE) along with a shallow-tower model. Multitask learning model is a neural network-based model which uses multiple learning techniques each to support various ranking objectives.

Training Data:

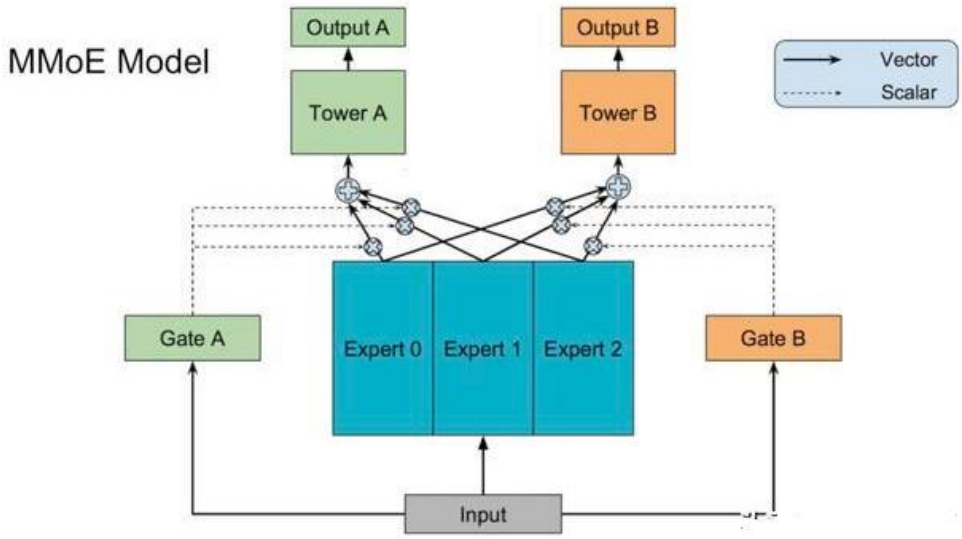
For training data, since getting explicit feedback from users is costly, implicit feedback is used. Their ranking system uses user's clicks and watches (engagement behavior) and their likes and dismissals (satisfaction behavior) as training data. But Implicit Feedback is considered biased.
It is selection biased when the user watches the video because it was shown up by the existing ranking system. It is position biased when the user watches the video because it is placed in the higher position among the recommended videos by the existing system. New models trained on these data would only give biased data towards the existing system. In order to reduce the position bias, position can be included as an input while training the model which is the most common approach.

Proposed Ranking System:

In the proposed system, first they generate the candidates using several candidate generating algorithms. Each candidate has one detail of similarity between the user typed query and the video. All the generated candidates are added into a set. Ranking system which is created using advanced machine learning techniques leveraging neural network architecture scores the candidates based on the relevancy. Ranking system supports multiple objectives and each objective will predict one type of user behavior like click, watch, like or dislike.

Multi-gate Mixture-of-Experts (MMoE):

Multi-gate Mixture-of-Experts is a multi-task learning approach that can perform multi task learning by sharing the experts across all tasks with a gating network trained to optimize each task.



In the proposed model, experts are added on the top of a shared hidden layer so that Mixture-of-Experts layer can learn modularized information from its input. The gating networks takes input from the input layer directly. This has helped the input features to select the experts directly. It also uses shallow tower to avoid the biases from implicit feedback. A shallow tower is trained with attributes contributing to biases like position feature for position bias. It is then added to the main model. It has helped to reduce the position bias.

This proposed architecture is experimented on YouTube. Both the proposed model and the baseline model are trained everyday with all incoming data. This has helped to get the most recent data as frequent changing of user behavior is one of the challenges we have seen before. Based on the experiment, it is concluded that MMoE's gating networks have effectively helped to modularize input information into experts. Gating network polarization is eliminated by applying a 10% probability of setting utilization of experts to 0 and re-normalizing the softmax outputs. Also position bias is being reduced by using shallow tower. This is proved by comparing the output from the proposed method with baseline methods which uses position feature to eliminate position bias.

YouTube live experiment results for MMoE:

Model Architecture	*Number of Multiplications*	*Engagement Metric*	*Satisfaction Metric*
Shared-Bottom	*3.7M*	*/*	*/*
Shared-Bottom	*6.1M*	*+0.1%*	*+ 1.89%*
MMoE (4 experts)	*3.7M*	*+0.20%*	*+ 1.22%*
MMoE (8 Experts)	*6.1M*	*+0.45%*	*+ 3.07%*

Future Recommendations:

There are still ways to make the proposed system much better. This system can be enhanced to have stability, trainability and expressiveness altogether. Also, different compression methods can be applied to reduce the cost.

Conclusion:

In this article, we have gone through the challenges in making real-time video recommendations. Challenges include scalability, accuracy, biases based on implicit feedback like position bias and selection bias along with dynamically changing user behaviors and access patterns. Google's multitask ranking system uses Multi-gate Mixture-of-Experts (MMoE) along with a shallow-tower model to address these challenges and provide a much better real-time video recommendation.

References:

- ✓ https://daiwk.github.io/assets/youtube-multitask.pdf
- ✓ https://www.kdd.org/kdd2018/accepted-papers/view/modeling-task-relationships-in-multi-task-learning-with-multi-gate-mixture-
- ✓ https://dl.acm.org/doi/10.1145/3240323.3240330