Marcus Hwai Yik Tan Oct 28, 2021

Tech Review – Recommending What Video to Watch Next: A Multitask Ranking System

Introduction

This paper reviews a scalable video recommender system proposed and experimented on Youtube by Zhe Zhao et al. [1]. The work is important because the system addresses the challenges of large-scale video recommender systems including scalability, different types of features, multiple conflicting objectives and bias.

Method and Contribution

The general flow of the recommender system is the following. Based on a query video (defined as a video currently being watched), a list of candidates is obtained by multiple matching algorithms. A deep neural network (DNN) ranking model called Multi-gate Mixture-of-Experts (MMoE) model is then used to score or rank the candidates for multiple user tasks/behaviors. User tasks are divided into two categories: (1) engagement such as clicking and watching, and (2) satisfaction such as liking and rating a video. A model called shallow tower is introduced and combined with the MMoE to reduce selection bias due to the position of the video.

Candidates are generated with 4 existing/known techniques: (1) matching video topics to the query video topic, (2) using degree of association of a video with the query video, i.e., how often the video is watched together with the query video, (3) analyzing a user's watch history, (4) taking into consideration the context (for example, the time of the day when the video is watched). The candidate generation system is significantly less computation intensive than the ranking system. By limiting the videos chosen for ranking, it provides scalability to the recommender system.

A key contribution of the paper is the extension of MMoE (originally proposed by the same authors [2]) for the multi-objective/task ranking of candidate videos. Conventional multi-objective ranking DNN (Fig 1(a)) contains multiple bottom/hidden layers immediately after the input layer shared across all tasks, the purpose of which is to construct representative features and reduce the number of features. However, such an architecture harms learning as it assumes that all tasks depend on the same set of features. MMoE overcomes the issue by reducing the number of shared layers to one and introducing smaller expert layers for each task (Fig 1(b)). Features relevant to a task are selected from the expert layers and the shared layer using a gating softmax dedicated to the task. Removing the shared layer and directly feeding the input features and embedding to the expert layers would significantly increase model training and inference costs. The authors claimed that the live experiment they had conducted did not show significant difference between 1 shared layer and no shared layer.

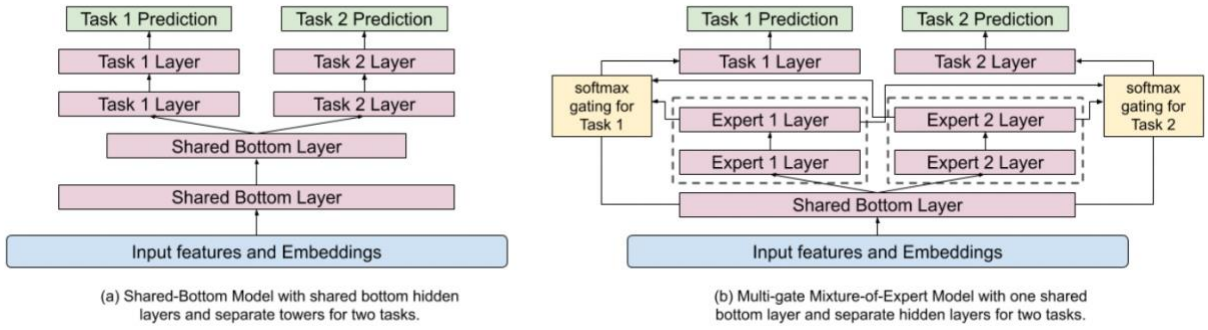

Figure 1: (a) Conventional DNN and (b) MMoE ranking systems

The ultimate goal of the video recommender system is to recommend videos that maximize user utility, which is assumed to be measurable by user actions/tasks. However, it has been demonstrated that user actions can be biased by the interaction between the system and the user [3]. For example, the user is more likely to click on a video that is ranked higher by the system or rated highly by others though the video may not be the most relevant or interesting to the user. The authors address positional bias by training a side model (shallow tower) that takes the click position as an input feature and set the position to a fix value during inference to remove position bias. The side model is then combined with the main model (MMoE) described earlier to rank the video for user engagement (Fig. 2).

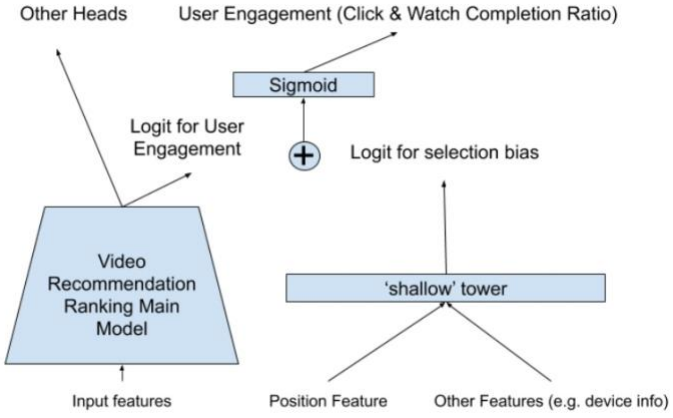

Figure 2: Adding a shallow tower (right) to reduce position bias

Result

The authors conducted live experiment on Youtube and measured the performance of the proposed ranking system with a production system using A/B testing. Their system outperforms a similar size (defined by the number of multiplications) conventional shared bottom model on both user engagement and satisfaction metrics (Fig. 3).

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%

Figure 3: Performance metrics of conventional shared bottom and MMoE ranking models

It was also shown that the shallow tower can predict the position bias of video clicks. For example, the model predicts that the position with the most click through rates has the most bias. With the addition of the shallow tower, the ranking system outperforms other existing methods in prediction user engagement (Fig. 4)

Method	Engagement Metric
Input Feature	-0.07%
Adversarial Loss	+0.01%
Shallow Tower	+0.24%

Figure 4: Performance of shallow tower and other existing methods

Conclusion

A video ranking system that addresses the challenges of large-scale ranking was reviewed in this report. The general flow of the system and key components were described. On Youtube, key performance metrics of the system shows significant improvements over the production system.

References

1. Zhe Zhao et al. Recommending what video to watch next: a multitask ranking system, In Proceedings of the 13th ACM Conference on Recommender Systems, 2019
2. Jiaqi Ma et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts, In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2018
3. Aman Agarwal et al. Estimating position bias without intrusive interventions. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, ACM, 2019