YouTube’s Multitask Ranking System

Introduction

This review is relating to YouTube’s video recommendation system that creates a list of videos to recommend to the viewer to watch after the current video. It consists of candidate generators (which generate several hundred candidates to be ranked), a ranking system, and post-ranking adjustments (deduplication, diversification, etc.). This tech review focuses on the ranking part of the system, and evaluates its design goals and its Multi-gate Mixture-of-Experts design.

Body

Challenges:

Designing and developing a real-world large-scale video recommendation system is full of challenges, including:

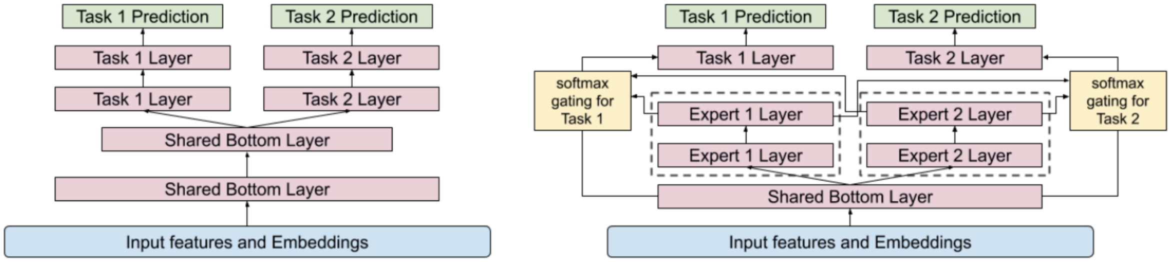
- There are often different and sometime conflicting objectives which we want to optimize for. For example, we may want to recommend videos that users rated highly and shared with their friends, in addition to watching.
- There is often implicit bias in the system. For example, a user might have clicked and watched a video simply because it was being ranked high, no because it was the one that the user liked the most. Therefore, models trained using data generated from the current system will be biased, causing a feedback loop effect. How to effectively and efficiently learn to reduce such biases is an open question.

Solution:

YouTube’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. Shallow Tower is an embedded feature to minimize the position bias.

Multi-gate Mixture-of-Experts (MMoE) and Shallow Tower:

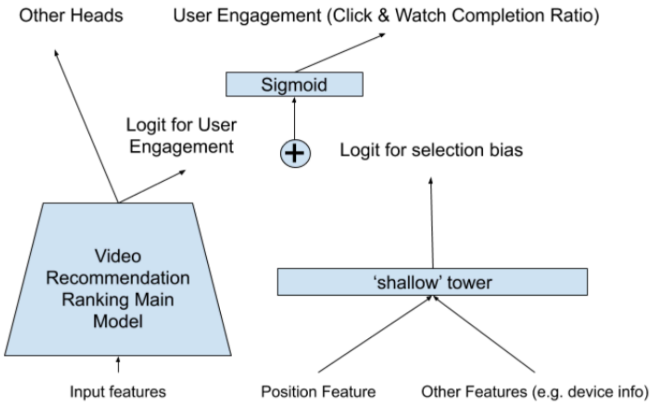
MMoE is a soft-parameter sharing model structure designed to model task conflicts and relations. It adapts the Mixture-of-Experts (MoE) structure to multitask learning by having the experts shared across all tasks, while also having a gating network trained for each task. The MMoE layer is designed to capture the task differences without requiring significantly more model parameters compared to the shared-bottom model. The key idea is to substitute the shared ReLu layer with the MoE layer and add a separate gating network for each task.



(a) Shared-Bottom Model with shared bottom hidden layers and separate towers for two tasks.

(b) Multi-gate Mixture-of-Expert Model with one shared bottom layer and separate hidden layers for two tasks.

The ranking system mitigates the position bias (videos getting more clicks not because they are more relevant to the users but because they are ranked higher by the current ranking system) with shallow tower next to the main ranking model. Note that the shallow tower is in the same neural network as the main ranking model, so that it learns the position bias from the production system, as opposed to random experiments. The input to the shallow tower includes device and user session information because a certain position may be on the first page in a device with a large screen (hence smaller bias) but not on the first page in a device with a smaller screen (hence larger bias).


Other Heads
User Engagement (Click & Watch Completion Ratio)
Sigmoid
Logit for User Engagement
Logit for selection bias
Video Recommendation Ranking Main Model
'shallow' tower
Input features
Position Feature
Other Features (e.g. device info)

Conclusions

The YouTube's multitask ranking system is a large-scale multi-objective ranking system and an effective, scalable, and flexible system for identifying engaging and quality contents for individual viewers. To efficiently optimize multiple ranking objectives, it uses Multi-gate Mixture-of-Experts model architecture to utilize soft-parameter sharing. The system also contains a lightweight and effective method to model and reduce the selection biases, especially position bias. Furthermore, via live experiments on one of the world's largest video sharing platforms, YouTube, the proposed techniques have led to substantial improvements on both engagement and satisfaction metrics.

References

https://daiwk.github.io/assets/youtube-multitask.pdf