CS 410 Technology Review

Author: Chuan Jiang(chuanj2@illinois.edu)

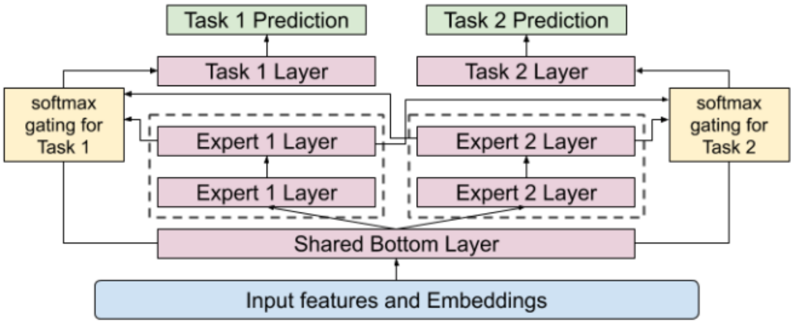
Date: 11/04/2022

Topic: The overview of Google's Multitask Ranking System

Recommendation systems are playing an important role in many applications from recommending videos in YouTube and Netflix, and to recommend hashtags in LinkedIn and Instagram. We want to recommend a product or video content to a user, and we want to see whether they like it or not. Youtube.com is the world's largest video sharing platform. at YouTube, sometimes they need recommend multiple objectives and it is not just that easy to combine all those objective together. Designing and developing a real-world large-scale video recommendation system is full of challenges. Firstly, there are often different and sometimes conflicting objectives which we want to optimize for. For example, we may want to recommend videos that users rate highly and share with their friends, in addition to watching [1]. Secondly, 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, not 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 [2]. Google YouTube team found a way to overcome these challenges such as Multi-date Mixture-of-Experts, it can quickly optimize for multiple ranking objectives, and improved the recommendation quality. Researchers at Google presented a solution for this multi-task ranking system. YouTube has billions of videos and content. If we want to rank all those billions of videos and show it to the user that is not going to be very efficient and that takes a lot of time.

What Engineers at Google did was to from the billions of videos corpus, they select come up with like 500 videos, then they use a very sophisticated model to only check those 500 candidates. It starts with candidate generation that from the billions of videos they just first select one million of them and that can be just as simple as SQL query. For example, if you have an index table there and we have a video just has been watched by the user, we can just write a simple query that select 1 million videos from this corpus. Then we can have logistic regression to narrow down even further mainly select 50000 from the 1 million, and again we can have a random forest to again make it even narrower. At the end, we end up with 500 candidates which are the most relevant candidates to the video. At this point, a ranker, which is the very sophisticated model, it can quickly rank those 500 candidates and recommend it to the user.

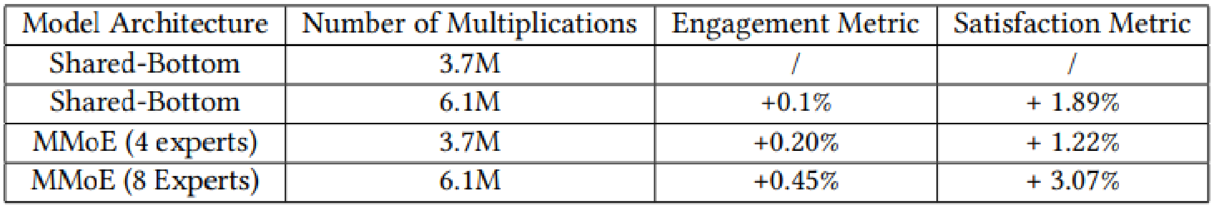
YouTube has 2 billion monthly active users, that is 66 million users per day, 2 million users per hour, 700 users per second. In order to handle such large group of concurrent users, the system must be very efficient to process minimum 700 users per second since it also may encounter peak times. Google engineers designed input features and embeddings that has some share bottom layers like below figure. It consists of two tasks, task 1 will handle engagement task, and task 2 will hand satisfaction task. They also designed the mixture of experts, which is like the sub network that are supposed to learn or become an expert on a specific part of the input information. This will greatly improve the recommendation system.



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

[3]

Google's multitask ranking system works as expected and does improve the ranking system. As below figure, we can see that it increased the result of both engagement metric and satisfaction Metric.



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%

[4]

REFERENCES

[1][2][3][4] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, Ed Chi. 2019. Recommending What Video to Watch Next: A Multitask Ranking System. 41-50