Tech Review: Multitask Ranking Systems

Sam Song
samsong2@illinois.edu

1. INTRODUCTION

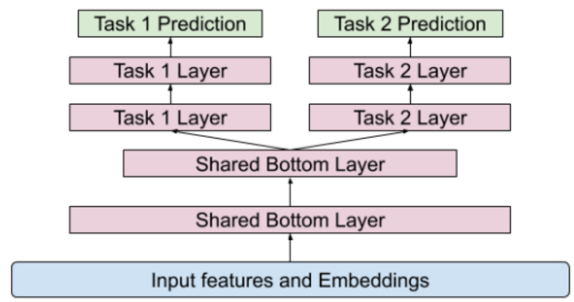
In this tech review we take a look at the Multigate-Mixture of Experts architecture and how it is being used in industrial scale recommender systems.

A recommender system is composed of two parts. The first part consists of candidate generation. Based upon the generated user profile the recommender system then returns a list of items that may interest the user. This second part of returning a list of relevant items that interest the user can also be seen as a ranking problem. To create the ranked list Deep Neural Networks(DNN) are often used in today's industrial systems.

Sometimes though, it can be desirable to optimize for multiple tasks in recommender systems. For example, user satisfaction and user engagement. For this, multitask learning techniques can be used.

2. Previous Multitask Ranking System

One of the most common multitask learning techniques used is the Shared-Bottom multi-task DNN structure.



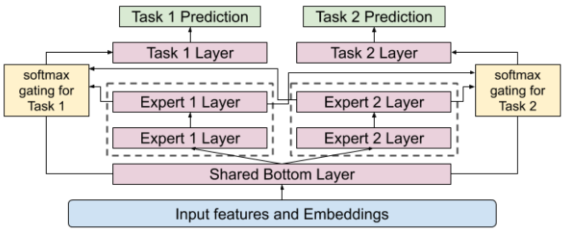
(a) Shared-Bottom Model with shared bottom hidden layers and separate towers for two tasks.

This consists of a bottom network with several layers that are shared across all neural networks. The output from these bottom layers are then fed into individual "tower" networks for each task. The results from each task tower can then be used to compute a final ranking score.

One of the problems with this architecture is that the Shared-Bottom layer needs tasks to have a high degree of correlation. When there is only a low degree of correlation between the tasks the shared layer may end up harming learning of the tasks instead.

3. Multigate-Mixture of Experts

To solve this issue the Multigate-Mixture of Experts(MMoE) [3] architect was proposed. The MMoE replaces a Shared Bottom Layer with a collection of "expert" submodels. These experts are then shared across all tasks. Each task then has a gating network that adjusts the parametrization of the experts to suit that particular task.



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

After going through a gate the features from the experts are then passed to the associated task tower. The results show that this method is easier to train and results in less loss[3].

4. Example

In the provided case[1] the MMoE is used as part of the recommender system to recommend what Youtube videos a user should watch next. The MMoE in this case is implemented as part of using Deep and Wide Neural[2] network model.

The Deep part of the neural network consists of a DNN utilizing MMoE. The ranking system utilizes multiple objectives with each objective predicting one type of user behavior related to user utility. The objectives can be split into those related to user engagement and those related to user satisfaction.

Since there are often some amounts of implicit user bias the wide part of the neural network is used to account for any sort of user selection bias. The wide part of the neural network consists of a single shallow tower that takes in features for any sort of possible selection bias. The results from the shallow tower then feed into the towers for user engagement.

5. CONCLUSION

Overall the MMoE architecture is a generic multi-task learning architecture that through further modularization of neural networks, has allowed for the optimization of tasks with much lower correlation than before. As such it could possibly be utilized for any sort of analysis or ranking where multiple objectives may need to be taken into account, not just for recommender systems. For example, this structure has been used to analyze user activity streams[4].

6. REFERENCES

[1] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. 2019. Recommending what video to watch next: a multitask ranking system. In Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19). Association for Computing Machinery, New York, NY, USA, 43–51. DOI:https://doi.org/10.1145/3298689.3346997

[2] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & Deep Learning for Recommender Systems. In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems (DLRS 2016). Association for Computing Machinery, New York, NY, USA, 7–10. DOI:https://doi.org/10.1145/2988450.2988454

[3] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. 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 & Data Mining (KDD '18). Association for Computing Machinery, New York, NY, USA, 1930–1939. DOI:https://doi.org/10.1145/3219819.3220007

[4] Zhen Qin, Yicheng Cheng, Zhe Zhao, Zhe Chen, Donald Metzler, and Jingzheng Qin. 2020. Multitask Mixture of Sequential Experts for User Activity Streams. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). Association for Computing Machinery, New York, NY, USA, 3083–3091. DOI:https://doi.org/10.1145/3394486.3403359