Technology Review on Google Multitask Ranking System

By Tianli Ding (netId: td2)

INTRODUCTION

The current existing ranking systems have lots of issues and limitations, including scalability limitation, multi-object learning inefficiency, and selection bias. Recently, there is a newly-invented ranking system algorithm proposed by Google, using a deep neural network-based ranking model, both improving the efficiency and quality of existing models. To deal with the problems mentioned earlier, this algorithm also makes plenty of innovations and adjustments on existing algorithms.

In this review paper, we will introduce the detailed techniques that this ranking system collaborates with, and the limitations or drawbacks it might have. We will also introduce a couple of existing algorithms for better comparison.

ALGORITHMS INTRODUCTION AND COMPARISON

The new system uses a Wide & Deep model as a basic framework. Wide & Deep model was to combine wide linear models and deep neural networks, in order to maximize both the advantages of linear models -- memorization of feature interactions, and advantages of deep neural networks -- generalization of a recommendation system, within a single system. (Heng-Tze et al.)

I. Deep neural networks model

In a typical multi-objective learning process, there are two stages: first to retrieve and generate documents given a query and second to assign scores on each document and rank them. On the “Deep” side of the model, the new system improves the first stage of multi-object learning process by applying an revised Multi-Gated Mixture-of-Experts layer to modularize the input information into small chunks, each analyzed by one expert. (Zhao et al.)

The original Multi-Gated Mixture-of-Experts was developed based on deep neural networks to improve the performance when model task correlations have very little relationship. (Ma et al.) The new system modified the MMoE by removing the bottom layer of the ReLu layer, and instead using a MoE layer with a gating network for each layer. In order to offset the dimension augment which applying MoE layer directly on input layer causes, it decreased the number of experts to keep the time efficiency.

II. Wide linear model

The system improves the multi-object learning efficiency by applying MMoE, another problem it solves is selection bias during implicit feedback. Selection biases have many types, and there are also many that we have not detected. One of the biggest challenges is known as position bias. Position bias is when we choose recommended videos, we tend to choose the ones that are on the top, although they might not be the ones that are most attractive to us. In order to reduce the effect of bias, many experts have proposed different approaches. For example, Thorsten and his group proposed a method to add a normalizer for the bias by using a counterfactual inference framework and a Propensity-Weighted Ranking SVM to learn from implicit feedback, such as clicking. (Joachims et al.) However, the scalability of this method is not very well, in that it is not feasible to be used in a large platform with billions of clicks and variations.

Therefore, the new system extends the “Wide” part in Wide & Deep model to take the position as an input in the ranking process, uses a shallow tower together with the main model to learn and process the position bias instead of calculating propensity score.

FURTHER IMPROVEMENTS

There are also some further improvements that can be studied in this algorithm. As we discussed previously, it modified the original Multi-gated Mixture-of-Experts and reduced the number of experts used to reduce cost. To further prove and analyze the effectiveness and efficiency of this algorithm, the system is tested on one of the largest video platforms, YouTube. From the evaluation result provided in the paper, the more number of experts added, the better the performance, but the slower the process. Therefore, there can be raised a question whether they could further improve the approach, by adding more number of experts to improve more effectiveness while at the same time, keeping the efficiency.

Besides, as mentioned above, position bias can hinder the learning process of the system. From the result of evaluation, we can see the shallow tower method can improve the adversary learning algorithm by 0.31%. It can be a huge progress within such a platform with billions of users and videos. However, despite such an improvement, there are also many other types of bias, many of which are not discovered. How to identify and eliminate those biases can also be an interesting topic to continue digging.

CONCLUSION

This newly published ranking algorithm extends the Wide & Deep algorithm by adding a modified Multi-gated Mixture-of-Experts methods in deep neural networks and adding a shallow tower to mitigate the position bias in a wide linear model. After testing on one of the largest video platforms, YouTube, this algorithm is improved to improve the effectiveness of Multi-objects learning and reduce the impact of position bias. There are also some further improvements that can be investigated, including adding more experts while keeping the time in efficiency and eliminating the influence of other types of biases while performing implicit feedback during ranking.

REFERENCES

Heng-Tze Cheng, Levent, et al. “Wide & Deep Learning for Recommender Systems.” Wide & Deep Learning for Recommender Systems | Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, 1 Sept. 2016, dl.acm.org/doi/10.1145/2988450.2988454.

Joachims, Thorsten, et al. “Unbiased Learning-to-Rank with Biased Feedback.” Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, doi:10.24963/ijcai.2018/738.

Ma, Jiaqi, et al. “Modeling Task Relationships in Multi-Task Learning with Multi-Gate Mixture-of-Experts.” Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, doi:10.1145/3219819.3220007.

Zhao, Zhe, et al. “Recommending What Video to Watch Next.” Proceedings of the 13th ACM Conference on Recommender Systems, 2019, doi:10.1145/3298689.3346997.