Learning to Rank with Multi-gate Mixture-of-Experts

Ilya (Ilia) Andreev
iandre3@illinois.edu
Department of Computer Science,
University of Illinois at Urbana-Champaign

1 Introduction

It is a common situation in deep learning when multiple tasks need to be performed on input data, potentially with conflicting objectives. Instead of creating independent architectures for each of the tasks, both efficiency and accuracy can be increased by building a shared architecture that can optimize for each of the tasks, thus letting the model learn input data similarities and differences pertinent to each of the objectives.

For instance, a movie recommender system can have two different objectives: suggest videos that maximize time spent watching and suggest videos that maximize enjoyment gained from watching. One might imagine that the first objective translates to users spending money and watching the videos, whereas the second objective translates to users wanting to come back the next day for more videos. Some famous approaches to doing this are the Shared-Bottom architecture and its various descendants such as Mixture-of-Experts and Multi-gate Mixture-of-Experts models.

This technology review summarizes some of the recent approaches to multi-task learning, particularly in the context of video recommendation. Large-scale video recommendation services are very sensitive to model efficiency as well as accuracy. Additionally, recommendation platforms are subject to bias in training due to users watching videos that are ranked highly instead of videos that are the best choice for the user globally. All these challenges can be addressed with a proper multi-task ranking system.

2 Overview

Multi-task learning is a methodology of solving for multiple tasks at a time by leveraging input data features that are relevant to several tasks. Instead of training multiple entirely separate networks (figure 1), one way to leverage commonalities in data is by sharing a single feed-forward network that nevertheless has multiple outputs, each for a separate task (figure 2) (Caruana 44). This architecture is commonly referred to as Shared-Bottom Multi-task architecture. However, this architecture is prone to exhibiting interference between the weights responsible for one task and the weights learned for a different task.

In situations where we know in advance which of the training cases are to be used for which specific task, a different architecture called mixture models can be used. First mixture models were proposed in the late 1980s. In 1991, Jacobs et al. described a system composed of multiple "expert" networks and a single gating network that decides which of the experts is responsible for handling each particular input (81). Earlier works had assumed that the final system output is a linear combination of outputs of each of the experts, but Jacobs et al. suggested that the gating network's error measure should represent a stochastic decision regarding which single expert is to be used on each example (80). As such, the weights of each expert don't interfere with other experts' weights, allowing for the experts to be truly "local" (figure 3). Still, the experts leverage each other since a change in the weights of a single expert will affect the gating network's "preferences" over all experts (Jacobs et al. 81).

Later works proposed improvements upon the single-gate Mixture-of-Experts model. One such improvement is the Multi-gate Mixture-of-Experts architecture by Ma et al. They too use a group of bottom networks, each called an "expert" (1932). However, instead of using a single gate, they create a dedicated gate for each of the learned tasks, thus creating "mixed" inputs produced by different weighted combinations of experts that are then fed into task-specific feed-forward "tower networks" (Ma et al. 1932). Such architectures successfully model task relationships while not requiring that we limit ourselves to a single task at a time or constrain multiple tower networks to a single gate (figure 4) (Ma et al. 1938).

The Multi-gate Mixture-of-Experts (further, MMoE) technique has proven to be performant and accurate as shown by Zhao et al. in their paper on the YouTube recommender system. Their system has ranking objectives of two forms: engagement and satisfaction (44). Each objective corresponds to a particular type of implicit user feedback (49). Additionally, in their paper Zhao et al. focus on addressing the misalignment between implicit feedback and true user utility caused by selection bias (45). Below we summarize the system described by Zhao et al.

Based on the video the user is currently watching, the YouTube candidate generation system uses multiple algorithms focused on different contextual aspects to retrieve several hundred candidate recommendations (Zhao et al. 46). Then, the ranking system is tasked with a point-wise learn-to-rank goal on these candidates; point-wise and not pair- or list-wise ranking is used mainly for efficiency reasons (46). Predicted user behaviors on retrieved items map to either binary classification tasks (e.g. clicks) for which cross-entropy loss is computed or regression tasks (e.g. ratings) for which squared loss is computed (47). The final ranking score is calculated as a weighted product of these subscores (47). Instead of a naive Shared-Bottom architecture, Zhao et al. use MMoE (figure 5). Their choice is partly due to having to run their network on multimodal input, which is better modeled by Mixture-of-Experts architectures than Shared-Bottom architectures (47). However, they do add a shared hidden layer between the

experts and the input in order to decrease the dimensionality of the experts and decrease the model training and serving costs (47).

To overcome selection bias in training, Zhao et al. add a separate shallow tower (figure 6) trained on features related to bias, such as position and device type which tends to affect the position bias (48). The result is then added to the final logit of the main model. At serving time, the position feature is treated as missing (a different approach to be considered is to always treat position as 1 at serving time).

In their experiments, Zhao et al. found the MMoE architecture to outperform the Shared-Bottom model on multitask ranking (48). Additionally, they found that the shallow tower approach to handling bias performs better than adversarial learning approaches or approaches that use the position feature directly as an input feature (49).

3 Conclusion

Mixture-of-Experts is a powerful deep learning paradigm that proves to be a performant and accurate approach to solving multi-task ranking problems. It can be extended to address subtle issues such as selection bias in ranking and can be adapted to handle complex multimodal inputs at scale.

References

Caruana, Rich. “Multitask Learning.” *Machine Learning*, vol. 28, no. 1, 1997, pp. 41–75., https://doi.org/10.1023/a:1007379606734.

Jacobs, Robert A., et al. “Adaptive Mixtures of Local Experts.” *Neural Computation*, vol. 3, no. 1, 1991, pp. 79–87., https://doi.org/10.1162/neco.1991.3.1.79.

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, https://doi.org/10.1145/3219819.3220007.

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

Figures

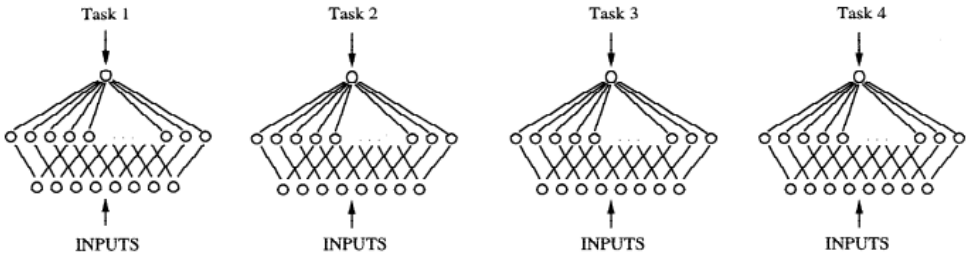

Figure 1, Caruana 43. Multi-task learning with independent feed-forward networks.

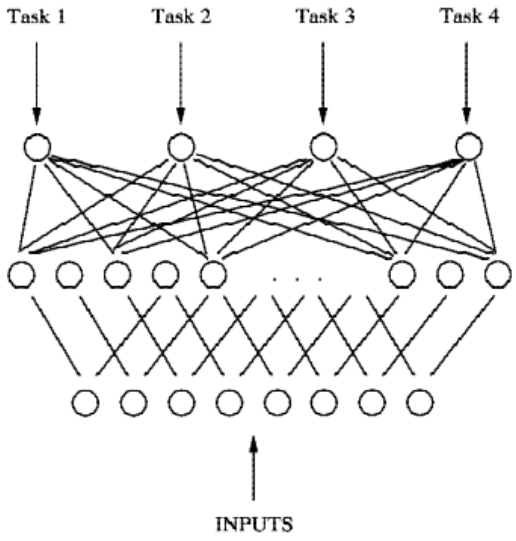

Figure 2, Caruana 44. Multi-task learning with a single feed-forward network with multiple outputs.

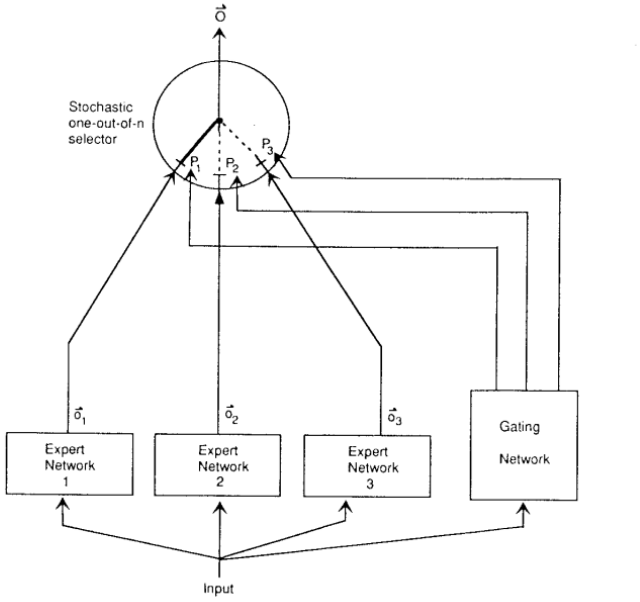

Figure 3, Jacobs et al. 81. Mixture-of-Experts model with a single mutually exclusive gate.

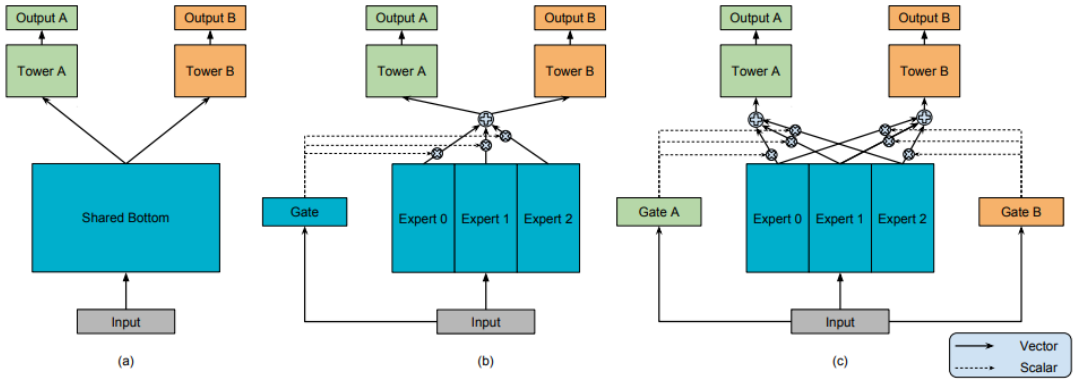

Figure 1: (a) Shared-Bottom model. (b) One-gate MoE model. (c) Multi-gate MoE model.

Figure 4, Ma et al. 1931. (a) is a Shared-Bottom model, (b) is a one-gate Mixture-of-Experts model, and (c) is a Multi-gate Mixture-of-Experts model.

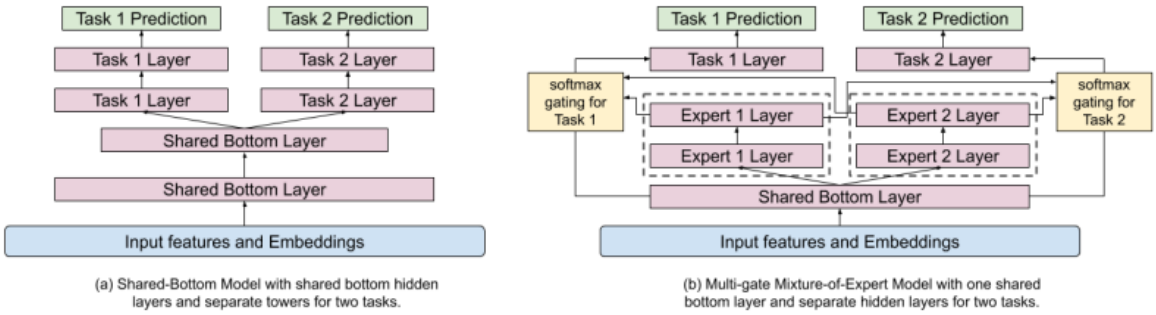

Figure 5, Zhao et al. 47. (a) is a Shared-Bottom model with 2 towers, and (b) is a Mixture-of-Experts model with 2 towers with a Shared-Bottom layer. Both models represent an architecture with multimodal input space.

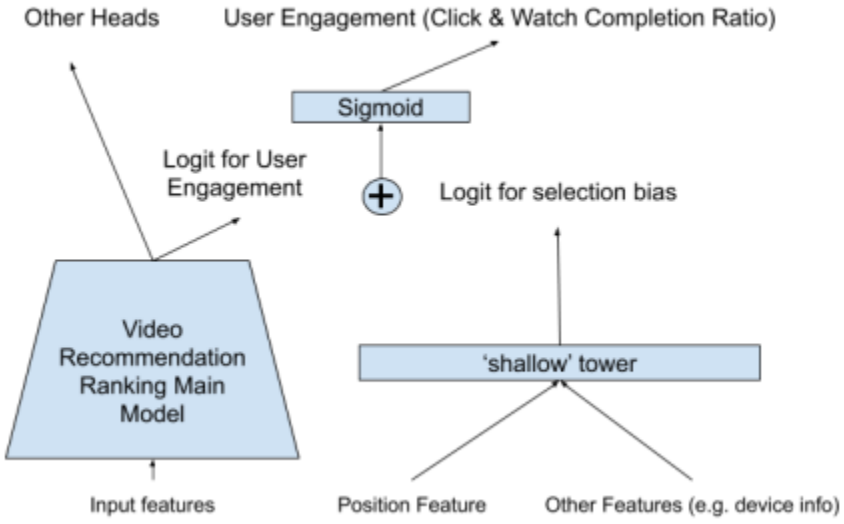

Figure 6, Zhao et al. 47. Shallow tower for position bias.