Tech Review

Nick Chun

Current standards for natural language understanding tasks usually rely on tuning a new model for every task. However, there are several consequences to this approach. For instance, serving multiple models requires a higher overall parameter cost, as well as higher general technical maintenance. The three papers addressed in this technology review address this issue by introducing a single multi-task model. Each paper proposes a different version of multi-task learning networks using unique transformer architecture to address specific issues related to training individual models for different natural language processing tasks. Although multi-task learning is still underdeveloped, there have been many important contributions by researchers at Google in this field, paving the way for the transition to a single multi-task learning model that will dramatically reduce the cost and increase the effectiveness of recommendation systems. All these systems apply an extensive array of experiments to test their new propositions, using both synthetic and real-world data to prove the increased effectiveness of their models.

The first paper (Tay, Yi, et al. 2021) introduces a new transformer architecture called HyperGrid Transformers, which leverages task-conditioned hyper networks for controlling its feed-forward layers. The researchers particularly focus on a decomposable hypernetwork that learns grid-wise projections that help to specialize regions in weight matrices for different tasks. The central idea to HyperGrid Transformers is the factorization of local and global components for weight generation; by applying a grid-wise weight generation the researchers imbue the model with a concrete structural layout. These grid-wise projections dynamically control the parameters through a segmentation of its feed-forward layers rather than using the standard row-wise weight generation methodology. Furthermore, the decomposable hyper-projections enable the model to learn deep contextual and pairwise interactions between two hypernetworks. The

results of their experiments testing against the GLUE/SuperGLUE benchmarks is a performance that matches the state-of-the-art systems like Text-to-Text Transformers (T5) while increasing parameter efficiency by 16 times. Summarily, the HyperGrid Transformer architecture proposes grid-wise decomposable hyper projections, which matches the performance of multiple finely tuned models while increasing parameter efficiency.

Google researchers additionally proposed a novel multi-task learning approach called Multi-gate Mixture-of-Experts (MMoE), which explicitly learns to model task relationships from data (Ma, Jiaqi, et al. 2018). This approach addresses the issue that recommendation systems face regarding the need to optimize multiple objectives at the same time. For example, when recommending movies to users, the company may want the users to not only purchase and watch the movies, but to also like the movies afterwards so that they will come back to watch more movies. In this case, the company aims to create models to predict both users' purchases and ratings simultaneously. MMoE is an extension of the original Mixture-of-Experts (MoE) structure to multi-task learning, implemented through the sharing of submodels across all tasks with a gating network trained to optimize each task. It's fundamentally based on the more common Shared-Bottom multi-task DNN structure, where several bottom layers following the input layer are shared across all the tasks and then each task has an individual network tower on top of the bottom networks. The MMoE model has a group of bottom networks called experts, where each of which is a feed-forward network. The researchers show the benefits of this new architecture by experimenting on both synthetic data and real large-scale recommendation systems, proving that MMoE is easier to train and can effectively handle scenarios where tasks are less related. It also considers the computational efficiency of machine learning production systems, noting that their model preserves the computational advantage since their gating

networks are typically lightweight and the expert networks are shared across all the tasks. To summarize, this paper introduced a multi-task learning approach called Multi-gate Mixture-of-Experts, which is inspired by the Mixture-of-Experts model and has opened doors for new developments in neural-based multi-task learning.

Researchers at Google further expand on the previous MMoE model by proposing a new framework called Mixture of Sequential Experts (MoSE), which explicitly models sequential user behavior using Long Short-Term Memory (LSTM) in the state-of-art MMoE modeling framework (Qin, Zhen, et al. 2020). According to these researchers, most of the multi-task model architectures proposed recently have focused on non-sequential input queries and context. However, input data is largely sequential in real-world data science scenarios. Thus, they aim to use this new MoSE framework for studying the problem of multi-task learning when the model consumes sequential user activity data. This framework is composed of a shared-bottom LSTM module that consumes sequential input data, a mixture of sequential expert layers specializing in different aspects for each task, gating networks to select a subset of experts based upon the input, and a multi-tower network with one tower per task. The benefits of MoSE are verified through a comparison with seven alternatives, including (Sequential) Multi-Model, (Sequential) Multi-head, (Sequential) Shared-bottom, and MMoE, on both a synthetic dataset and a real-world dataset involving millions of users in Google's G Suite. It consistently outperforms all other alternatives, illuminating MoSE's ability to effectively handle challenges in user activity streams such as sparse variables and complicated relations between heterogeneous data sources. Overall, this paper continues to explore the issue of learning multiple objectives in user activity streams by proposing a mixture of sequential experts framework that outperforms other common alternatives on a wide variety of datasets.

In recent years, there has been an ever-expanding number of researchers exploring the topic of deep neural multi-task learning models. They aim to replace the expensive and inefficient current standard of training individual models for each task with a flexible single model that can perform at the same level, if not better, to the individual models. However, many recent contributions to this field have proposed novel modeling techniques that rely on a greater computation cost due to the vast number of parameters to accommodate the task differences. This is highly unpopular in real production settings, so researchers at Google have proposed several different solutions to this problem. The three frameworks covered in this review are HyperGrid Transformers utilizing grid-wise decomposable hyper projections, Multi-gate Mixture-of-Experts model as an expansion of the original MoE architecture, and the Mixture of Sequential Experts framework that integrates MMoE and LSTM. In general, their work tends to share similarities and they often build upon each other's findings. Moreover, the researchers contend that their novel architecture improves on the existing frameworks by performing comparative experiments using common benchmarks like GLUE/SuperGLUE and both synthetic and real-world datasets. Multi-task learning is still gaining popularity, so there will continue to be further investment and research into this topic in the upcoming years. For instance, one direction could be improving on the MoSE architecture by creating more opportunities for flexible sharing between the expert networks. Even though multi-task models are still under development, there have been many significant advancements in this domain by researchers at Google, demonstrating that the future of deep learning architecture is headed towards a more streamlined single model approach.

References

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. Accessed 2022.

Qin, Zhen, et al. “Multitask Mixture of Sequential Experts for User Activity Streams.” *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* 2020, https://doi.org/10.1145/3394486.3403359. Accessed 2022.

Tay, Yi, et al. “HyperGrid Transformers: Towards a Single Model for Multiple Tasks.” *Google Research,* 2021, https://research.google/pubs/pub49971/. Accessed 2022.