FID3018 Advanced Course in Data Mining and Analytics Opposition Report

Sina Sheikholeslami

May 2021

In this report, I provide my reflection on two presentations given by Tianze Wang and Stefanos Antaris in the FID3018 course.

1 Tianze's Presentation

Tianze discussed the following three papers:

1. A Hierarchical Model for Device Placement (Mirhoseini et al. 2018)
2. Spotlight: Optimizing Device Placement for Training Deep Neural Networks (Gao, Chen, and Li 2018b)
3. Post: Device Placement with Cross-Entropy Minimization and Proximal Policy Optimization (Gao, Chen, and Li 2018a)

The selected papers address the problem of device placement for model-parallel training of deep neural networks. We need model-parallelism when the size of a model exceeds the the memory of a GPU, so we have to partition the model across different devices, e.g., GPUs. This approach allows us to train much larger models, e.g., the state-of-the-art language models. However, the main question is how can we efficiently find an optimal placement, since this search for different placements incurs a lot of costs by itself. This is a very interesting line of research that has gained increased popularity in the past couple of years, as we have seen an explosion in the size of the state-of-the-art deep neural networks.

Tianze started the presentation with a very clear formulation of the device placement problem, and introduced the terms and concepts common across the three papers. He also explained relevant prior work in this domain. His presentation included lots of examples, and he provided his own account of the strengths and weaknesses of each of the proposed methods. He ended his presentation with a very interesting timeline of the research on device placement, which made it very clear for the audience to once more understand the relationship between the three papers. Before opening for discussion, he also provided a list of questions that he thought would be worth of discussing, and it really helped to spark the conversation.

2 Setfanos' Presentation

Stefanos presented the following three papers that revolve around temporal interaction networks, graph neural networks, node embeddings, and reinforcement learning (RL):

1. Streaming Graph Neural Networks (Ma et al. 2020)
2. Learning Temporal Interaction Graph Embeddings via Coupled Memory Networks (Zhang et al. 2020)
3. End-to-End Deep Reinforcement Learning based Recommendation with Supervised Embedding (Liu et al. 2020)

The first two papers provide approaches for creating node embeddings using temporal interactions of graphs, and the third paper investigates different RL architectures for updating the embeddings in real-time and exploiting the embeddings for recommender systems.

Stefanos started his talk by giving an overview of the three papers and their connection to each other, as well as temporal interaction networks as the common fundamental concept present in the papers. His explanation was very clear and he used examples to clarify the concepts, using figures from the papers. In the end, he

presented the shortcomings of each paper. He also engaged very good with the audience as they asked questions during his talk. One minor comment is that he spent some significant time on a number of slides, and he could maybe broke the content into several slides or use different animations or helper shapes; however, the time spent on those slides was well-deserved as they were mainly the fundamental points of the papers.

References

Gao, Yuanxiang, Li Chen, and Baochun Li (2018a). "Post: Device placement with cross-entropy minimization and proximal policy optimization". In: *Advances in Neural Information Processing Systems*, pp. 9971–9980.

— (2018b). "Spotlight: Optimizing device placement for training deep neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 1676–1684.

Liu, Feng et al. (2020). "End-to-end deep reinforcement learning based recommendation with supervised embedding". In: *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 384–392.

Ma, Yao et al. (2020). "Streaming graph neural networks". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 719–728.

Mirhoseini, Azalia et al. (2018). "A hierarchical model for device placement". In: *International Conference on Learning Representations*.

Zhang, Zhen et al. (2020). "Learning Temporal Interaction Graph Embedding via Coupled Memory Networks". In: *Proceedings of The Web Conference 2020*, pp. 3049–3055.