John (Ziyuan) Zhou

Week 4 Summary

UW ID: 2150115

In the article “Session-based Recommendation with Self-Attention Networks”, the authors propose a session-based recommendation with self-attention networks (SR-SAN). This architecture allows session-based recommendation captures the global dependencies among all items of a session regardless of their distance. Compared to the conventional approaches which are content-based and collaborative filtering based, it is much easier to learn item dependencies from long distance – the attention base model name Transformer achieved excellent performance in the WMT2014, but its encoder-decoder architecture is not suit for session-based recommendation which output size is not equal to input size. This happened because the dependencies among all items may lost during the aggregation of adjacent items with the graph neural network; thus, leads to inaccurate item embedding learning. The methodology of SR-SAN consists of two parts: obtaining item latent vectors with self-attention networks and making recommendation with prediction layers. It will compute each head attention by softmax function, to find the multi-head attention. The multi-head attention will be passed through a two layer fully connected feed-forward network. For the prediction layer, it is composed of score compute with dot product and softmax function. The model is trained by minimizes the cross-entropy between ground-truth and the prediction. For the experiments, there are two datasets (Yoochoose and Diginetica) are selected to test the performance of the proposed method. The experimental result shows that the SR-SAN outperforms some state-of-the-arts, compared to the baseline methods.