Generalizable MoE based Data Pruning in Recommendation

Abstract

Existing recommendation algorithms are often model-centric relying on a large quantity of data with intensive augmentation. The importance of data quality and the cost of big data, however, is often overlooked.

Recently, a data-centric approach called data pruning addressed the problem by selecting a small subset of high-quality data that produces a comparable model performance. This helps alleviate the expensive data storage and model training costs. Despite the success in the vision or language tasks, the sparse data nature and the high number of item classes in recommendation bring difficulties in sampling a diverse set of item sessions. Besides, the model bias problem in sampling is further exacerbated by the unbalanced item distribution.

To bridge this gap, we 1) leverage a mixture of experts (MoE) model to provide a fair data importance score and 2) a graph-based sampling to obtain high data diversity.

First, to reduce model bias, we construct a powerful lightweight MoE model based on GNN, RNN and transformer with only the gating model being trainable. Then, we rank each data point based on sample difficulty using model.

To promote diversity, we treat the recommendation dataset as an undirected graph and apply a message-passing based data pruning. Specifically, the difficulty scores are updated by aggregation and propagation on a neighbour graph, where sessions that share the same items are connected. Subsequently, both popular and cold items can be sampled with minimal duplicates.

Our results show that we can achieve x% performance with only y % data and can transfer well across various architectures. This observation questions the necessity of intensive data augmentation that is commonly applied in session recommendation. Additionally, our graph-based sampling can naturally provide an explainable data pruning process and a model-driven data visualization.

1 1

2 Exp

References