DeepMix: A Recommender System with CrossNetMix and Improved Deep Neural Network Structures

Junyan Li
Shanghai Jiao Tong University
521030910098

Jiasheng Yun
Shanghai Jiao Tong University
521030910093

Shuwen Wu
Shanghai Jiao Tong University
521030910087

ABSTRACT

Modern recommender systems face growing demands in terms of scalability, efficiency, and adaptability. This paper introduces DeepMix, an improved recommender system architecture that builds upon the foundations of DEEP CROSSING.

At the core of DeepMix is a flexible deep neural network (DNN) structure that enhances the modeling of complex feature interactions. It is coupled with a novel CrossNetMix layer, which learns feature interactions across different subspaces through a gating mechanism. This synergistic combination of architectural innovations and deep learning-based optimization aims to overcome the limitations of existing recommender system solutions.

Through extensive evaluations, we demonstrate that DeepMix achieves superior performance in terms of recommendation accuracy, latency, and resource utilization compared to traditional architectures. DeepMix's deep learning components enable proactive recommendation management, dynamic feature engineering, and intelligent user-item matching, fostering a more efficient, personalized, and adaptable recommender system infrastructure. The proposed architecture paves the way for the next generation of recommender systems.

1. INTRODUCTION

The rapid growth of online services and the increasing demands placed on modern recommender systems have driven the need for innovative system architectures and optimization techniques. Traditional recommender system designs often struggle to keep up with the scale, complexity, and dynamism of today's user preferences and item characteristics, leading researchers and industry leaders to explore new approaches to address these challenges.

One notable development in this space is Alibaba's DCN V2 (Data Center Network Version 2), which represents the second generation of the tech giant's recommender system design. DCN V2 introduces a hierarchical network topology, leverages advanced techniques like Clos and Segment Routing, and incorporates software-defined networking principles to enhance scalability, efficiency, and programmability.

Building upon the advancements made by DCN V2, researchers have also explored ways to boost the feature interaction modeling capabilities of recommender systems through the DEEP CROSSING framework. DEEP CROSSING seamlessly integrates various deep learning modules, enabling these powerful models to better understand and reason about the complex relationships between user profiles, item attributes, and their interactions.

Inspired by these groundbreaking works, we have developed an improved recommender system architecture called DeepMix. DeepMix builds upon the foundations laid by DEEP CROSSING and introduces several key enhancements, including a novel CrossNetMix layer and a flexible DNN structure that enhances the modeling of complex feature interactions. This synergistic combination of architectural innovations and deep learning-based optimization aims to overcome the limitations of existing recommender system solutions.

The objective of this paper is to present the DeepMix architecture and demonstrate its potential to address the evolving challenges faced by modern recommender systems. By seamlessly integrating deep learning, flexible feature interaction modeling, and sophisticated user-item matching, DeepMix aims to push the boundaries of recommender system optimization and management, ultimately enabling more efficient, personalized, and adaptable online services.

Representative examples of LLMs include OpenAI's GPT (Generative Pre-trained Transformer) series models, such as GPT-3, and Google's BERT (Bidirectional Encoder Representations from Transformers) model. These models leverage large-scale unsupervised learning pre-training techniques, automatically learning patterns and language rules from text data, achieving impressive performance across a wide range of natural language processing tasks.

The emergence of LLMs has had profound impacts on various fields, including information retrieval, intelligent customer service, automatic summarization, sentiment analysis, knowledge graph construction, and more. Their robust language understanding capabilities enable more efficient processing and comprehension of large-scale text data, driving advancements and innovations in the field of natural language processing. This project aims to explore content comprehension using LLMs and other deep learning models, focusing on regression and classification tasks on the rel-movielens 1M dataset.

2. RELATED WORKS

2.1 DCN V2

One of the notable developments in data center networking architectures is Alibaba's DCN V2 (Data Center Network Version 2). Introduced by the tech giant to support their rapidly growing cloud computing infrastructure, DCN V2 represents the second generation of Alibaba's data center network design.

DCN V2 builds upon the original DCN architecture by introducing several key improvements. The new design employs a hierarchical network topology and leverages techniques like Clos topology, Segment Routing, and In-Band Network Telemetry to enhance scalability, efficiency, and programmability. These advancements enable DCN V2 to support a larger number of hosts and network devices within the data center while optimizing performance and reducing operational complexity.

Furthermore, DCN V2 incorporates software-defined networking (SDN) principles, allowing for centralized control and programmability of the network through APIs and controllers . This increased programmability, coupled with advanced telemetry and analytics capabilities, provides network administrators with real-time visibility and fine-grained control over the data center network.

Additionally, DCN V2 focuses on improving resilience and security by incorporating redundancy, failover mechanisms, access control, encryption, and anomaly detection features. These enhancements help to ensure high availability and protect the data center network against potential threats.

The development and deployment of DCN V2 by Alibaba have been instrumental in supporting the company's rapid growth and the expansion of their cloud computing services. The architecture's improved scalability, efficiency, programmability, and security features have enabled Alibaba to manage their data centers more effectively, meeting the increasing demands of their cloud customers.

2.2 DEEP CROSSING

Another notable work in the field of data center network architectures is DEEP CROSSING, a solution developed by researchers at Microsoft. DEEP CROSSING aims to address the challenges of traditional data center network designs, which often struggle to keep up with the growing demands of modern cloud workloads.

The DEEP CROSSING architecture introduces a novel approach that leverages deep learning techniques to optimize network resource allocation and improve overall performance. By modeling the complex relationships between various network parameters, such as traffic patterns, resource utilization, and application requirements, DEEP CROSSING can make intelligent decisions to dynamically adjust the network configuration.

One of the key features of DEEP CROSSING is its ability to perform end-to-end network path optimization. The system uses deep neural networks to predict the optimal paths for data flows, considering factors like congestion, latency, and bandwidth availability. This proactive path selection can help to reduce network congestion, improve application performance, and enhance the overall efficiency of the data center network.

Additionally, DEEP CROSSING incorporates advanced monitoring and analytics capabilities, leveraging in-band telemetry and machine learning techniques to gain deep insights into network behavior. This visibility enables network administrators to quickly identify and address performance bottlenecks, optimize resource utilization, and make informed decisions about network management.

The development of DEEP CROSSING by Microsoft's research team demonstrates the potential of incorporating deep learning and data-driven approaches into the design of data center networks. By leveraging the power of machine learning, DEEP CROSSING aims to overcome the limitations of traditional network management strategies and provide a more intelligent, adaptive, and efficient data center network solution.

2.3 GraphLLM

The limitations of Large Language Models (LLMs) in graph reasoning tasks have been a subject of growing interest in the research community. Recent studies have highlighted the underwhelming performance of LLMs on fundamental graph reasoning problems, even with specialized prompting (Wang et al., 2023a; Guo et al., 2023; Ye et al., 2023).

A prevalent approach to applying LLMs to graph data has been the Graph2Text strategy, which involves converting graphs into natural language descriptions. While this method allows LLMs to process graph information, it introduces inherent drawbacks. LLMs trained on Graph2Text representations must infer implicit graph structures from sequential text, which can be less efficient compared to dedicated graph learning models. Additionally, the Graph2Text approach often results in lengthy contexts, posing challenges for LLMs to identify the essential information required for graph reasoning tasks (Liu et al., 2023).

To address these limitations, a previous work introduced GraphLLM (Chai et al., 2023), a novel approach that synergistically integrates graph learning models with LLMs. GraphLLM takes an end-to-end approach, harnessing the strengths of both graph learning and language modeling to enhance the graph reasoning ability of LLMs. By incorporating a graph transformer module, GraphLLM can capture both node-level information and graph-level structure, and then distill this knowledge into a concise, graph-enhanced prefix for the LLM.

The key innovations of GraphLLM include:
Collaborative Synergy: GraphLLM combines graph learning models and LLMs within a single, cohesive system, allowing LLMs to leverage the superior expressive power of graph representations for enhanced graph reasoning performance.
Context Condensation: GraphLLM condenses graph information into a concise, fixed-length prefix, circumventing the need for lengthy graph descriptions required by the Graph2Text strategy.

Empirical evaluations of GraphLLM on a range of fundamental graph reasoning tasks have demonstrated significant

improvements in accuracy and efficiency compared to LLMs trained using the Graph2Text approach (Chai et al., 2023). These results highlight the potential of synergistic integration of graph learning and language modeling to advance the state-of-the-art in AI systems capable of robust graph reasoning.

3. MODELS

Our model is an improved version based on DeepCrossing, named DeepMix. Deep Crossing is a recommendation model based on deep learning, which can be divided into four layers: Embedding layer, Stacking layer, Multiple Residual Units layer, and Scoring layer. Our improvements mainly include three aspects:

3.1 Transformer Data Processor

In our data preprocessing phase, we focus on encoding the textual information associated with the nodes in the classification dataset. This step is inspired by GraphLLM's approach, where the goal is to extract essential information from node descriptions for the classification task.

We utilize a textual transformer encoder-decoder architecture to process the node descriptions. The encoder applies self-attention mechanisms to capture semantic meaning, generating context vectors that represent the relevant information for the classification task. These context vectors serve as compressed representations of the input text.

Subsequently, the transformer decoder produces node representations by attending over the context vectors and newly-initialized trainable query embeddings. We save these node representations in a npy file.

3.2 Improved Multiple Residual Units Layer

The original Multiple Residual Units layer adopts a ResNet structure to capture the nonlinear relationships and high-order feature representations of input features. We have made modifications to this layer by replacing the ResNet section with a more flexible DNN structure. The purpose of doing so is to enhance the modeling ability of the model for complex interactions between features, and to improve the model's expressive and learning abilities.

3.3 CrossNetMix Layer

In addition to improving the Multiple Residual Units layer, we have also introduced a new layer called CrossNetMix for learning feature interactions in different subspaces. At this level, we define multiple experts, each responsible for learning specific feature interaction patterns.

To effectively combine the outputs of different experts, we introduce a gating mechanism. This mechanism calculates the weight of each expert's output and combines the outputs of different experts based on their contribution level. Then, the combined output is subjected to nonlinear transformation in a low-dimensional space, thereby achieving flexible modeling of input feature interaction.

3.4 DeepMix Structure

The architecture of our DeepMix model comprises the following key components:

- **Embedding Layer**: This layer maps discrete input features (e.g., categorical features) into a low-dimensional dense vector space, known as embedding vectors. Continuous features (e.g., numerical features) retain their original values.
- **Stacking Layer**: This layer consists of multiple neural network layers stacked together to gradually extract abstract representations from the input features.
- **Improved Multiple Residual Units Layer**: This layer replaces the original ResNet structure with a more flexible deep neural network (DNN) design. The purpose is to enhance the model's ability to capture complex interactions between input features and improve its expressive and learning capabilities.
- **CrossNetMix Layer**: This is a novel layer introduced in DeepMix, which learns feature interactions across different subspaces. It defines multiple "experts," each responsible for learning specific feature interaction patterns. A gating mechanism is used to effectively combine the outputs of these experts based on their contributions.
- **Scoring Layer**: This final layer maps the output of the model to the desired prediction result, whether it's a classification or regression task.

The key innovations in the DeepMix architecture are the Improved Multiple Residual Units Layer and the CrossNetMix Layer. These components aim to enhance the model's ability to capture complex feature interactions and learn diverse interaction patterns, ultimately improving the overall performance of the DeepMix model.

4. RESULTS AND FINDINGS

To evaluate the performance of our DeepMix model, we conducted experiments on both classification and regression tasks using the MovieLens 1M dataset. We utilized a set of diverse evaluation metrics, including micro F1 score, macro F1 score, and Mean Absolute Error (MAE) loss.

4.1 Classification Task

For the classification task, our goal was to predict the genre of movies. We evaluated the model's performance using micro F1 score and macro F1 score.

The experimental results for the classification task are as follows:

- Micro F1 score: 0.3498
- Macro F1 score: 0.3108

The micro F1 score of 0.3498 indicates that our DeepMix model achieved a moderate overall performance in classifying movie genres. This score considers the balance between precision and recall across all classes, which is particularly important in scenarios with imbalanced class distributions.

The macro F1 score of 0.3108 suggests that the model's performance varies across different movie genres. Some genres may have lower F1 scores compared to others, indicating potential areas for improvement in accurately classifying certain genres. Further analysis of the class-wise performance can provide insights into the specific challenges the model faces, allowing for targeted improvements.

4.2 Regression Task

For the regression task, our goal was to predict the ratings given by users to movies. We evaluated the performance using Mean Absolute Error (MAE) loss.

The experimental result for the regression task is as follows:

- MAE Loss: 0.91176736

The MAE loss of 0.9118 suggests that, on average, our DeepMix model's predictions differ from the true ratings by approximately 0.9118. This performance demonstrates a certain level of predictive rating ability, but there is still significant room for improvement in further reducing the prediction errors.

Overall, the results indicate that the DeepMix model has achieved moderate performance on both the classification and regression tasks. The findings also highlight areas for potential improvement, such as enhancing the model's ability to accurately classify certain movie genres and reducing prediction errors in the regression task. Further investigation and refinement of the model architecture and training strategies may lead to enhanced performance in these tasks.