# Language-Enhanced Session-Based Recommendation with Decoupled Contrastive Learning

Zhipeng Zhang*
zp.zhang@std.uestc.edu.cn
University of Electronic Science
and Technology of China
Chengdu, China

Piao Tong*
piaot@std.uestc.edu.cn
University of Electronic Science
and Technology of China
Chengdu, China

Yingwei Ma*
yingwei.ywma@gmail.com
National University of
Defense Technology
Changsha, China

Qiao Liu✉
qliu@uestc.edu.cn
University of Electronic Science
and Technology of China
Chengdu, China

Xujiang Liu
575694112@qq.com
Pangang Group
Research Institute
Chengdu, China

Xu Luo
39440938@qq.com
Pangang Group
Research Institute
Chengdu, China

## ABSTRACT

Session-based recommendation techniques aim to capture dynamic user behavior by analyzing past interactions. However, existing methods heavily rely on historical item ID sequences to extract user preferences, leading to challenges such as popular bias and cold-start problems. In this paper, we propose a hybrid multimodal approach for session-based recommendation to address these challenges. Our approach combines different modalities, including textual content and item IDs, leveraging the complementary nature of these modalities using CatBoost. To learn universal item representations, we design a language representation-based item retrieval architecture that extracts features from the textual content utilizing pre-trained language models. Furthermore, we introduce a novel Decoupled Contrastive Learning method to enhance the effectiveness of the language representation. This technique decouples the sequence representation and item representation space, facilitating bidirectional alignment through dual-queue contrastive learning. Simultaneously, the momentum queue provides a large number of negative samples, effectively enhancing the effectiveness of contrastive learning. Our approach yielded competitive results, securing a 5th place ranking in KDD CUP 2023 Task 1. We have released the source code and pre-trained models associated with this work[1].

## CCS CONCEPTS

• **Information systems → Recommender systems**.

## KEYWORDS

session-based recommendation, e-commerce, language models

[1] https://github.com/gaozhanfire//KDDCup2023

* The first four authors contributed equally to this work.

✉ Corresponding Author.

*KDDCup '23, August 9, 2023, Long Beach, CA, USA*

**ACM Reference Format:**
Zhipeng Zhang, Piao Tong, Qiao Liu, Yingwei Ma, Xujiang Liu, Xu Luo. 2023. Language-Enhanced Session-Based Recommendation with Decoupled Contrastive Learning. In *KDD Cup 2023 Workshop: Multilingual Session Recommendation Challenge, August 9, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 4 pages.

## 1 INTRODUCTION

With the progress of data mining and machine learning, session-based recommendation, which employs customer session data to forecast their next click, is gaining momentum. Sequential recommender [3, 7] has garnered popularity owing to its capability of capturing users' transient and enduring preferences, rendering it widely applicable across diverse recommendation contexts.

Numerous methodologies have been proposed to model sequential recommendation scenarios, spanning from early matrix factorization approaches (e.g., FPMC [9]) to more recent sequence neural networks (such as GRU4Rec [3], STAMP [7], and Transformer [10]). While they may differ, their underlying principles are akin: capturing behavioral patterns that mirror user preferences based on the sequential access structure and establishing a mapping between user preferences and specific items grounded in their interaction relationships. However, the dependence on item interaction relationships curtails their efficacy when faced with the cold-start predicament of insufficient user interactions with certain items.

Recent research has explored the use of natural language text, such as item titles and descriptions, to gain domain knowledge and address the challenges mentioned earlier [4, 6]. This approach involves utilizing pre-trained language models like BERT to obtain text representations and learning how to transform these representations into item representations. While previous attempts have shown promise, there are still unresolved challenges that need to be addressed: (1)*Insufficient integration of multi-modal information*: Existing methods [4] combine the text representation of an item with its corresponding item representation, but this approach can introduce interference between modalities. For example, if an item has dense interaction data and the item embedding's representation capability surpasses that of the text representation, incorporating the text representation may negatively impact the expressive power

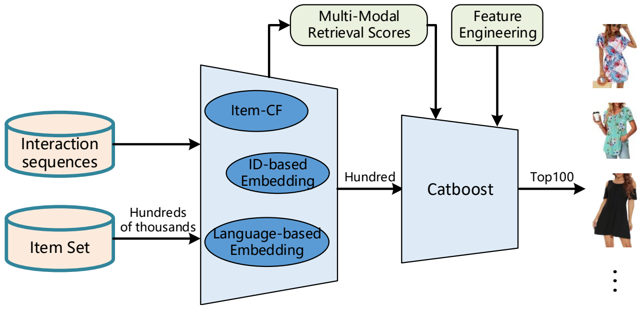

**Figure 1: The Holistic Framework of the Proposed Adaptive Hybrid Multimodal Approach.**

of the item representation. (2)*Partial Semantic Synchronization*: Current approaches often treat candidate items as targets, resulting in parameter updates that only align the sequential representation with item representations that match sequence preferences. As a result, the alignment between item representations and sequential representations is considered, while the alignment in the reverse direction is overlooked.

In order to address the aforementioned challenges, we propose an adaptive hybrid multimodal framework for session-based recommendation. Figure 1 provides an illustrative overview of the framework. This innovative framework employs multiple distinct models to encode autonomous representations of item IDs and their corresponding text sequences derived from a user's historical items. These representations are then leveraged to generate separate predictions for the subsequent click. To achieve dynamic fusion of the predictions, we employ the CatBoost algorithm, leveraging its capability to compute the conditional probability distribution of a class based on specific feature conditions. Additionally, we introduce Decoupled Contrastive Learning to enhance the representation capability of item text. This method separates the negative sample space into two independent subspaces: sequential representation and item representation. By decoupling these two spaces and utilizing a dual-momentum queue [1], a substantial number of negative samples are updated and stored. Two types of contrastive learning losses are then computed. This process ultimately enables bidirectional deep alignment between sequential representations and item representations. Furthermore, our team actively participated in the KDD CUP 2023, where we conducted extensive evaluations of our proposed solution using the Amazon-M2 dataset [5]. Through our rigorous experiments, we achieved competitive results, securing a 5th place ranking in task 1.

## 2 METHODOLOGY

### 2.1 Multimodal Approach to Item Retrieval

*2.1.1 ItemCF.* Despite the remarkable success of deep-learning-based approaches in recommender systems, traditional methods like collaborative filtering (CF) continue to find extensive applications on online platforms, owing to their irreplaceable strengths. These strengths include interpretability, minimal reliance on hardware, and high efficacy in both training and deployment.

In this competition, we incorporated several crucial factors into the construction of the item similarity matrix, including position distance information, directional cues in interactions, and normalizations based on item popularity. Formulas 1-4 encapsulate the fundamental concepts we employed to achieve these outcomes.

$$\mathrm{sim}(x,y) = \sum_{x,y \in Session} \frac{W_{\mathrm{dist}} * W_{\mathrm{dire}} * W_{\mathrm{pos}}}{|x|^{0.8} * |y|^{0.15}} \tag{1}$$

$$W_{dist}(x,y) = \frac{1}{(|\ \mathrm{Pos}_x - \mathrm{Pos}\ _y\ | + 1)^2} \tag{2}$$

$$W_{dire}(x,y) = \begin{cases} 1 & \text{if } Pos_x < Pos_y \\ 1/3 & \text{if } Pos_x > Pos_y \end{cases} \tag{3}$$

Our research presents a unique approach by incorporating positional weights. It stems from the observation that the last item clicked by users, after multiple interactions, often reflects their meticulously chosen ultimate preference.

$$W_{pos}(x,y) = \begin{cases} 1.8 & \text{if } Pos_y = |Session| - 1 \\ 1 & \text{if } Pos_y < |Session| - 1 \end{cases} \tag{4}$$

*2.1.2 ID-based Embedding : GRU.* Recurrent neural networks are the most popular choice for modeling sequential data. We designed a customized RNN which achieves a 0.379 public leaderboard score. The model architecture is shown in Figure 2.

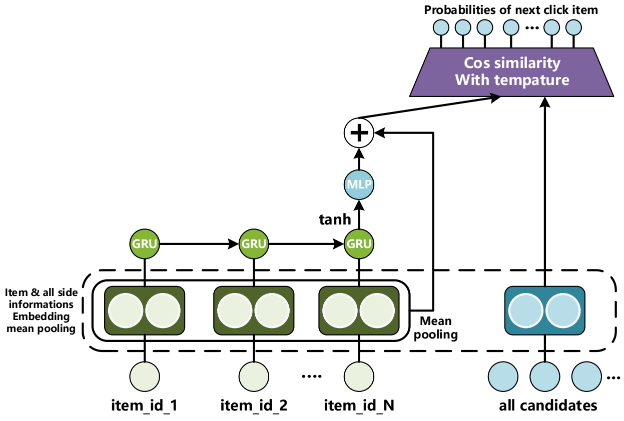

**Figure 2: The customized GRU architecture.**

During the phase of shared embedding lookup on the user's last N clicked items, we utilize mean pooling on the embeddings of the items as well as their associated side information (such as price and brand) to derive fused embeddings. Subsequently, these embeddings are fed into a GRU layer for the extraction of sequential features. To address the challenge of vanishing gradients, we introduce a sequence-level residual connection block. This block conducts mean pooling on the initial fused embeddings prior to the GRU layer. It is then added to the output of the GRU layer to generate the final sequential representation. Differing from the approach adopted in previous studies [10], which involves mean-pooling the GRU layer outputs at each time step, our method preserves the original fused embeddings, ensuring facilitating smoother gradient propagation.

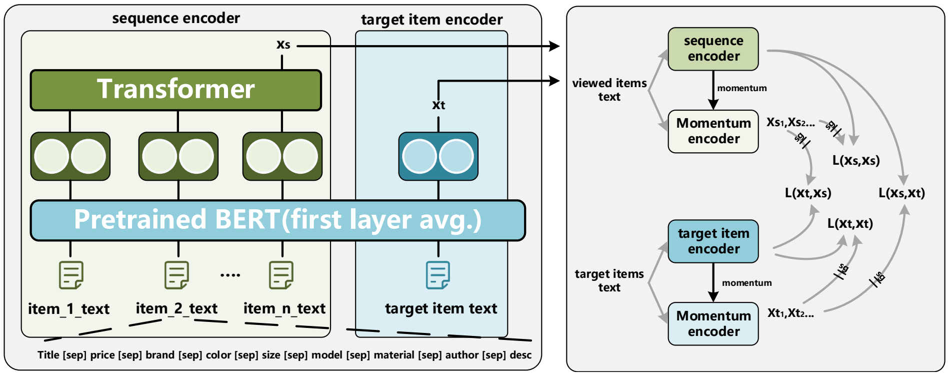

**Figure 3: Visualization of Language Embedding Generation and Training Flow of Decoupled Contrastive Learning.**

*2.1.3 Language Embedding.* Contrastive learning has found wide applications in information retrieval problems [2, 8]. Recently, its potential has also been explored in the field of session-based recommendation [6]. Motivated by these studies, this paper proposes a vector-based approach for item representation. The model architecture is shown in Figure 3.We segment the attribute texts of items using [SEP] and encode them using a pre-trained BERT model. This encoding process includes both previously browsed items and the target item. Finally, we perform average pooling on the outputs of the first layer and apply L2 normalization to obtain the representation vector for each item.

Each BERT encoder has a momentum encoder, whose parameters $\theta$ are updated by exponential moving average of the base encoder as follows: $\Theta_t \leftarrow \beta\Theta_{t-1} + (1-\beta)\Theta_{\text{base}}$ . In each iteration step t, we maintain two memory queues: one for sequence representations and the other for target item representations. These queues store $K$ vectors, which are encoded by the momentum encoder from the most recent batches. With each optimization step, the oldest vectors in the memory queues are replaced by the vectors from the current batch. The momentum coefficient $\beta \in [0, 1]$, usually set close to 1 (e.g., 0.999), ensures consistency across batches. Additionally, setting $K$ to a large value (>10000) ensures an abundant supply of negative samples, thereby promoting the learning of robust representations.

To train the encoder, we employ the ArcCon Loss [11], a loss function that introduces a hyperparameter $m$ to add a decision boundary, promoting compactness among sentence representations with the same semantics and increasing the dissimilarity between representations of sentences with different semantics:

$$\mathcal{L}(x, y) = -\log \frac{e^{s\cos\left(\theta_{x,y_i}+m\right)}}{e^{s\cos\left(\theta_{x,y_i}+m\right)} + \sum_{j\neq i} e^{s\cos\left(\theta_{x,y_j}\right)}} \tag{5}$$

In Equation 5, the variable $s$ denotes the temperature hyperparameter. In essence, Equation 5 can be perceived as a (K+1)-way softmax classification, where $y = y_i$ represents the positive sample, and the items stored in the queue $\{y_i\}_{i\neq j}^{K}$ represent the negative samples. It is important to note that gradients do not propagate through the momentum encoder or the storage queue during backpropagation.

An important contribution in our study is the introduction of the training method called Decoupled Contrastive Learning. We divide the set of negative samples into two separate subsets, each containing all sequential representations and target item representations, respectively. By doing so, we can construct a symmetrical contrastive learning loss with the following form:

$$min\ \lambda_1\mathcal{L}(x_s, x_t) + \lambda_2\mathcal{L}(x_t, x_s) + \lambda_3\mathcal{L}(x_s, x_s) + \lambda_4\mathcal{L}(x_t, x_t) \tag{6}$$

The hyperparameter $\lambda$ serves to harmonize the losses of each task. Upon completion of the training process, we discard the momentum encoders and memory queues, retaining solely the base encoders for the computation of sentence representations.

## 2.2 Multimodal Fusion Rerank

In this section, we retrieve items for each session using three retrieval methods, resulting in three sets of retrieved items. To combine the scores from the retrieval methods, we multiply them together, represented as $y_i = \prod_{t=1}^{3} y_t$, where $y_t$ is the score for each retrieval method. From the fused scores, we select the top 120 items as the preliminary retrieval results. Next, we integrate the scoring of retrieval results from different modalities and perform fine-grained feature engineering. We employ the CatBoost algorithm, which is based on decision trees and can effectively merge the retrieval results from different modalities by adaptively combining features.

*2.2.1 Item Hot Features.* The fundamental characteristics of item view frequencies in the dataset are considered. We have also incorporated the sort orders feature to enhance the model's ability to learn the ranking task with greater precision.

*2.2.2 Session Features.* In harmony with item features, we conduct composite statistics based on the item characteristics that users engage with within each session. This includes computing aggregate metrics such as the mean popularity and average price of all items within a session.

*2.2.3 Graph Features.* This study constructs a co-occurrence graph based on itemcf's co-occurrence relationship matrix. Each item is a node in the graph, and the weights between items serve as edge weights. By examining the degree relationships between items and their neighboring nodes, these features describe the importance and connectivity of candidate items within the co-occurrence graph. Our analysis includes node importance measures like PageRank, centrality features including betweenness centrality, Katz centrality, and degree centrality, as well as neighbor node relationship characteristics such as mean, count, maximum, and standard deviation of edge weights. Integrating these features enhances the model's understanding of candidate item characteristics.

# 3 EXPERIMENTS

## 3.1 Datasets

We applied data augmentation techniques referenced in the literature [3,8] to enhance the Amazon-M2 dataset across the UK, DE, and JP regions. Through a five-fold cross-validation method, the data was partitioned into training and validation sets. Notably, the validation set exclusively comprised sessions without data augmentation, with the final item serving as the label. The validation set not only served for evaluating the model's performance but also facilitated training set generation during the reranking stage. To ensure efficiency, the validation was conducted exclusively on the UK region's data.

## 3.2 Overall Performance

Our approach achieved considerably performance gain over the baseline solution, and ranked top-5 in task-1. The main results are shown in Table 1.

| Dataset | Metric | Ranking |
|---|---|---|
| leaderboard | mrr@100=0.4033 | 5th |
| validation | mrr@100=0.3394 | - |

**Table 1: Performance of our approach.**

## 3.3 Ablation Studies

We have delved into exploring the impact of applying Decoupled Contrastive Learning on model performance. Table 2 presents the performance of the model in terms of the MRR@100 metric after three epochs of training.It is important to note that the default hyperparameters utilized in this study were set as $\lambda_1$:$\lambda_2$:$\lambda_3$:$\lambda_4$=0.35:0.35: 0.15:0.15.

| Variants | CV-MRR@100 |
|---|---|
| w/o Decoupled Contrastive Learning | 0.2610 |
| $\lambda_1$:$\lambda_2$:$\lambda_3$:$\lambda_4$=0.5:0.5:0:0 | 0.2687 |
| $\lambda_1$:$\lambda_2$:$\lambda_3$:$\lambda_4$=0.25:0.25:0.25:0.25 | 0.2676 |
| $\lambda_1$:$\lambda_2$:$\lambda_3$:$\lambda_4$=0.45:0.3:0.125:0.125 | 0.2718 |
| Language Emb | **0.2725** |

**Table 2: The influence of Decoupled Contrastive Learning.**

We observed that by decoupling within the negative sample space and employing symmetric contrastive loss, the model's performance can be enhanced by 0.011. Additionally, adjusting the weight ratios of each loss function also influences the model's effectiveness.

As illustrated in Table 3, we evaluated the performance of the Multimodal Fusion Rerank methodology. Initially, we presented the performance metrics of various standalone models for comparative purposes against Multimodal Fusion. The experimental findings demonstrate that incorporating the catboost algorithm and conducting meticulous feature engineering both contribute significantly to elevating the performance of Multimodal Fusion.

| Variants | CV-MRR@100 | Leaderboard |
|---|---|---|
| GRU | 0.3073 | 0.3792 |
| ItemCF | 0.2941 | - |
| Language Emb | 0.2725 | - |
| Catboost Fusion | 0.3175 | - |
| ++Feature Engineering | **0.3394** | **0.4034** |

**Table 3: Performance of the Multimodal Fusion Rerank.**

# 4 CONCLUSION AND FUTURE WORK

This paper introduces a hybrid multimodal recommendation system that combines itemID and language modalities for item retrieval. We incorporate the catboost algorithm and employ fine-grained feature engineering to merge the retrieval results from both modalities. Our future work includes investigating the applicability of our algorithm in real-world e-commerce scenarios at a large scale.

# ACKNOWLEDGMENTS

This work is jointly supported by the National Natural Science Foundation of China (U19B2028), the National Science and Technology Major Project of the Ministry of Science and Technology of China (2022YFB4300603), the Sichuan Science and Technology Program (2023YFG0151) and the Development of a Big Data-based Platform for Analyzing the Coupling Relationship of Strip Production Processes Project.

# REFERENCES

[1] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).

[2] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

[3] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[4] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 585–593.

[5] Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, Haoyu Han, Hanqing Lu, Zhengyang Wang, Ruirui Li, Zhen Li, Monica Xiao Cheng, Rahul Goutam, Haiyang Zhang, Karthik Subbian, Suhang Wang, Yizhou Sun, Jiliang Tang, Bing Yin, and Xianfeng Tang. 2023. Amazon-M2: A Multilingual Multi-locale Shopping Session Dataset for Recommendation and Text Generation. (2023).

[6] Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian McAuley. 2023. Text Is All You Need: Learning Language Representations for Sequential Recommendation. *arXiv preprint arXiv:2305.13731* (2023).

[7] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: short-term attention/memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1831–1839.

[8] Yingwei Ma, Yue Yu, Shanshan Li, Zhouyang Jia, Jun Ma, Rulin Xu, Wei Dong, and Xiangke Liao. 2023. MulCS: Towards a Unified Deep Representation for Multilingual Code Search. In *2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 120–131.

[9] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.

[10] Benedikt Schifferer, Jiwei Liu, Sara Rabhi, Gilberto Titericz, Chris Deotte, Gabriel De Souza P. Moreira, Ronay Ak, and Kazuki Onodera. 2022. A Diverse Models Ensemble for Fashion Session-Based Recommendation. In *Proceedings of the Recommender Systems Challenge 2022*. 10–17.

[11] Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4892–4903.