CCL4Rec: Contrast over Contrastive Learning for Micro-video Recommendation

Shengyu Zhang1, Bofang Li1, Dong Yao1, Fuli Feng3, Jieming Zhu6, Wenyan Fan1, Zhou Zhao1,2, Xiaofei He1, Tat-seng Chua5, Fei Wu1,2,3

1 Zhejiang University 2 Shanghai Institute for Advanced Study of Zhejiang University 3 Shanghai AI Laboratory 4 University of Science and Technology of China 5 National University of Singapore 6 Huawei Noah’s Ark Lab

{sy_zhang,yaodongai,zhaohou,wufeij}@zju.edu.cn fulifeng93@gmail.com,jamie.zhu@huawei.com,chuats@comp.nus.edu.sg wenyang.17@intl.zju.edu.cn,chuxuepsn@163.com,xiaofei_h@qq.com

1 INTRODUCTION

Micro-video online services, such as TikTok, Kuaishou, and Instagram, have developed rapidly in recent years, leading to the proliferation of micro-video production and communication. The tremendously grown volume of micro-videos has intensified the need of retrieval and recommender systems that can permit personalized content discovery and consumption in micro-video online services. In the literature of modern recommender systems, the recent emergence of deep learning techniques has provided enticing methods [3–5, 8, 11, 12, 15, 16, 22, 24, 28, 30–32, 35, 36] in learning user representations that capture the characteristics of users’ preferences and interests. More recently, in frameworks that are specifically designed for micro-video recommendation, we notice that Li et al., [15] encapsulate the temporally remote connections upon the behavior sequence based on visual similarity and devise the temporal graph-based LSTM as the preference encoder. Jiang et al., [12] explicitly perform group routing and assignment to capture group-level interest. Many of these works follow the sequential recommendation schema [7, 9, 10, 26, 44], which typically makes predictions based on the historical behavior sequence and a sequence modeling module.

Despite the fruitful progress in the literature, it is worth noting that there are still some common challenges for micro-video recommendation. One of the major challenges can be the sheer difficulty of learning discriminating and effective user representations. Specifically, users behaviors are mostly implicit feedbacks and the ubiquitous false-positive interactions [28] may render the learned user representation indiscriminating (i.e., towards an average/over-smoothing representation of all users). Recently, contrastive learning [39, 40] has been empirically successful at learning high-level features that are robust to low-level noises (e.g., image rotation, flip). As illustrated in Figure 1, a straightforward way to combine contrastive learning and recommendation is to construct positive/negative behavior sequences via random augmentation/sampling. Then a contrastive learning loss function, such as InfoNCE [23], helps to distinguish the relative difference between <query, positive>, and <query, negative>. However, such a straightforward solution might be sub-optimal for recommendation due to the following two problems: 1) random augmentations (e.g., item drop/replacement/re-order) on the user behavior sequence treat all items equally and cannot shield off the learned representation from false-positive interactions within implicit feedbacks; and 2) existing

ABSTRACT

Micro-video recommender systems suffer from the ubiquitous noises in users’ behaviors, which might render the learned user representation indiscriminating, and lead to trivial recommendations (e.g., popular items) or even weird ones that are far beyond users’ interests. Contrastive learning is an emergent technique for learning discriminating representations with random data augmentations. However, due to neglecting the noises in user behaviors and treating all augmented samples equally, the existing contrastive learning framework is insufficient for learning discriminating user representations in recommendation. To bridge this research gap, we propose the Contrast over Contrastive Learning framework for training recommender models, named CCL4Rec, which models the nuances of different augmented views by further contrasting augmented positives/negatives with adaptive pulling/pushing strengths, i.e., the contrast over (vanilla) contrastive learning. To accommodate these contrasts, we devise the hardness-aware augmentations that track the importance of behaviors being replaced in the query user and the relatedness of substitutes, and thus determining the quality of augmented positives/negatives. The hardness-aware augmentation also permits controllable contrastive learning, leading to performance gains and robust training. In this way, CCL4Rec captures the nuances of historical behaviors for a given user, which explicitly shields off the learned user representation from the effects of noisy behaviors. We conduct extensive experiments on two micro-video recommendation benchmarks, which demonstrate that CCL4Rec with far less model parameters could achieve comparable performance to existing state-of-the-art method, and improve the training/inference speed by several orders of magnitude.

CCS CONCEPTS

• Information systems → Recommender systems.

KEYWORDS

Contrast over Contrast; Controllable Contrastive Learning; Micro-video Recommendation
contrastive learning objectives that treat all positive/negative samples as equally important prevent the framework from modeling the nuances between augmented samples and thus making it inferior in learning discriminating representations.

To investigate the above challenges and effectively consolidate the merits of contrastive learning into recommendation, we propose the Contrast over Contrast framework, abbreviated as CCL4Rec. CCL4Rec follows the sequential recommendation schema to have the behavior sequence as input. In essence, compared to existing contrastive learning techniques that solely consider the contrast on <positive, negative> pairs, CCL4Rec further considers the contrasts on <positive, positive>, <negative, negative> pairs. To make the contrast even more distinguishing, we enhance all kinds of contrasts with adaptive pulling/pushing strengths for different items within the historical behavior sequence (for the first problem) and the nuances between different augmented samples (for the second problem).

In the experiments, we equip CCL4Rec with a naive sequence modeling module and validate CCL4Rec on two large-scale micro-video recommendation benchmarks. Quantitative experiments, including ablation studies, and in-depth architecture analysis, demonstrate the strengths of CCL4Rec in being simple, lightweight, and effective compared with existing state-of-the-art methods. Remarkably, we show that CCL4Rec could achieve several orders of magnitude improvement on the training/inference speed with comparable performance to the state-of-the-art methods. In a nutshell, the contributions of this work are listed as follows:

- We propose to learn discriminating user representations for micro-video recommendation via contrastive learning and further model the nuances of augmented samples in a contrast over contrast manner.
- We devise the novel CCL4Rec framework, which permits hardness-aware augmentation with trackable hardness scores for augmented samples, adaptive pulling/pushing strengths for different contrasting, and controllable contrastive learning that can choose augmented samples with the desired hardness scores for contrasting.
- We conduct extensive experiments which demonstrate the strengths of CCL4Rec in being simple, lightweight, and effective in learning discriminating user representations.

2 RELATED WORKS

2.1 Contrastive Learning

Recently, the contrastive learning objective has become advantageous for unsupervised representation learning and led to state-of-the-art results in various domains. The essence of contrastive learning lies in the infomax [17] principle, where mutual information of a query and its positive views should be maximized, and the strategy of obtaining positive pairs. Typical augmentation methods for computer vision [2] include random cropping and flipping [23]. Tian et al. [27] also use different views of the same scene as positive samples. Similarly, in natural language processing, Logeswaran and Lee [21] use the context sentences as the positive views of the query sentence. Laskin et al., [14] employ the contrastive objective to improve sample efficiency of reinforcement learning. However, most of the existing contrastive learning frameworks treat all positive/negative samples as equally important, and we argue this can hinder the learned representation from being discriminating. In this paper, we propose to model the nuances between positives/negatives using contrast over the traditional contrastive learning. Also, the proposed controllable augmentation strategy constructs negatives with the desired hardness/quality from the query sample, which empirically leads to robust training and better performance.

Due to its effectiveness, there are increasing research interests on the intersection of contrastive learning and recommendation [39, 40]. It can be challenging to design augmentation strategies for recommendation samples since we have neither easily accessible contexts like the language nor straightforward (while effective) tools such as cropping or rotating like the computer vision. Noteworthy, Zhou et al., [43] directly treat items clicked by one user as the positive samples for contrastive learning and demonstrate that such an objective can help reduce popularity bias. Xie et al.,
[33] construct positive views for one behavior sequence by cropping, masking, and rendering part of the sequence, and treat the other randomly sampled behavior sequences as negative samples. Their framework follows a typical pretraining strategy which constitutes a pretraining stage and a fine-tuning stage. Zhou [45] introduces the InfoNCE objective [23] into the self-supervised masked item/attribute/segment prediction for sequential recommendation. Liu et al., [20] propose a graph-based framework that encapsulates graph perturbation, which can be viewed as the higher-order version of the masking operation in [33] and a debiased contrastive learning objective with items as positive/negative samples, similar to [43]. Different from these works, we propose a hardness-aware augmentation strategy and construct positives/negatives directly from the original behavior sequence. The proposed CCL4Rec framework models the nuances between different behaviors and augmented samples, and thus learns discriminating user representations in the embedding hypersphere.

2.2 Micro-video Recommendation

Micro-video recommendation is a nascent research area in content-based recommender systems [34] and attracts increasing research interests [3, 18, 19] recently due to the booming of micro-video production and communication on the internet [37, 38]. Typically, Li et al., [15] propose to model the dynamic and multi-level interests using the temporal graph-based LSTM and a multi-level interest modeling layer. Wei et al., [32] explicitly model the modality-specific user preferences and use graph neural networks to leverage inter-dependence between users and micro-videos in multiple modality-specific graphs. Hao et al., [12] devise the time-aware parallel masks for leveraging multi-scale time effects and the group routing algorithm to assign historical behaviors to different groups. Different from these works, we propose to learn discriminating user representations using contrastive learning, which remains largely unexplored in the literature of both generic and micro-video recommendations.

3 METHODS

3.1 Overview

We now present the contrast over contrastive learning framework for recommendation, termed CCL4Rec. The critical ideas of CCL4Rec are illustrated in Figure 2. CCL4Rec follows a schema of sequential recommendation, i.e., explicitly modeling the historical behavior sequence for prediction. Such a schema has the advantage of dynamically leveraging newly added items for prediction without model re-training. CCL4Rec mainly encapsulates three components: 1) Hardness-aware Augmentation, which determines the importance of each historical behavior and the relatedness of each substitute. By jointly considering the importance and relatedness, we conduct replacement augmentation and obtain positive/negative user behavior sequences of various hardness scores. Controllable Augmentation means we can obtain augmented samples with the desired hardness scores. Based on the scores, we can explicitly choose augmented samples from specific range of hardness along with the contrastive learning processes, i.e., controllable contrastive learning. Such scores further permit contrast over contrastive learning. 2) Sequence Modeling, which encodes the original behavior sequence and the augmented sequences into vectorial representations. We adopt a simplified design for efficiency. 3) Contrast over Contrast, which explicitly models the nuances of different augmented samples. We achieve this by designing hardness-related objectives and contrasting either negative or positive samples besides the conventional <query, positive, negative> contrasting schema. Furthermore, based on the hardness of augmented samples, we learn from simple contrasts at the early stage of training and progressively learn from harder contrasts, which empirically leads to a more robust training process and better performance.

3.2 Behavior/Substitute Ranking

Let u and v denote a micro-video and an user, and let the bold symbols v and u denote the corresponding vectorial representations. We denote the behavior sequence of u as $X_u = \{v_{u,t}\}_{t=1}^{N_u}$, where $v_{u,t}$ denotes the tth micro-video interacted by user u and $N_u$ is the number of behaviors in the sequence. Since the following description is within the range of one user, we write $v_i$ in place of $v_{u,i}$ for brevity. Given $X_u$, we propose to determine the importance of each element $v_i$ in representing the interests of user u, i.e., behavior importance ranking. We firstly compute the relevance of each element with other items in the behavior sequence, i.e.,

$$a_{ij} = f_u(v_i, v_j) = (v_iW_1)^T(v_jW_2),$$

(1)

where $W_1$ and $W_2$ are trainable linear mapping matrices. We propose to sum the relevance scores to all items and treat the summed score as the importance for representing the user, i.e.,

$$a_i = f_u(v_i) = \sum_{j=1}^{N_u} a_{ij},$$

(2)

Intuitively, when a micro-video is more relevant to other micro-videos watched by the user, it has a higher chance of being an important one. The computation of importance introduces trainable parameters, i.e., $W_1$ and $W_2$. Since we have no groundtruth for the importance score, we encapsulate these parameters into the sequence modeling module and jointly optimize them by the final cross-entropy loss function. We note that the proposed contrastive objectives that explicitly rely on the importance scores will not optimize $W_1$, $W_2$, and we achieve this simply using the detach() function in pytorch when computing the importance scores. We will illustrate the objectives and training in Section 3.5 in detail.

For user u, all other items not in the behavior sequence will be treated as potential substitutes for replacement augmentation. To accommodate the hardness-aware augmentation described in Section 3.3, we propose to rank these substitutes and compute the relatedness scores, i.e., substitute relatedness ranking. However, it can be infeasible to rank the whole item gallery, which can easily reach a billion-scale in real-world recommender systems. Towards this end, we propose to approximate the above process and choose to rank a subset of randomly sampled substitutes, i.e., $Z_u = \{v_k\}_{k=1}^{N_z}$, where $v_k$ is a potential substitute and $N_z$ is the number of sampled substitutes to be ranked. For user behaviors, we compute the importance score at the user level, and for substitutes, we compute the relatedness score at a fine-grained behavior level:

$$\beta_{ik} = f_k(v_i, v_k^*) = (v_iW_1)^T(v_k^*W_2),$$

(3)
where \( \beta_{mk} \) denotes the relatedness of substitute \( z^2_k \) and behavior \( q_t \). We reuse the scoring matrices \( W_1 \) and \( W_2 \) to reduce parameters.

### 3.3 Hardness-aware Augmentation

Based on the behavior importance scores and substitute relatedness scores, we propose to determine the hardness of augmented sequences. The term hardness is borrowed from the term hard negative mining and refers to the quality of the mined samples w.r.t. the query sample. We note that we largely refrain from mining hard negatives from all other samples by augmenting the current sample to have positives/negatives. Augmentation has the potential to be efficient, controllable, and of high quality by manipulating samples at a fine-grained component level. We conduct replacement augmentation, i.e., \( X^u_+ = (X_u \setminus \tilde{X}_u) \cup \tilde{Z}_u \) by replacing \( N_r \) micro-videos \( \tilde{X}_u \) with substitutes \( \tilde{Z}_u \) since replacement is more trackable as compared to other alternatives such as sequence re-order.

**Constructing Negatives.** When we prefer to replace items with high relative importance score \( \frac{\exp(\alpha_i)}{\sum_{j \in X_u} \exp(\alpha_j)} \), the constructed behavior sequences are potential negatives. We compute the hardness score for a potential negative \( X^u_\neg \) as follows:

\[
a^- = \sum_{(m, n) \in \tilde{X}_u} \exp(\beta_m) \frac{\exp(\alpha_m)}{\sum_{i \in X_u} \exp(\alpha_i)} \sum_{i \in X_u} \exp(\beta_m),
\]

where \( \alpha_m \in \tilde{X}_u \) is a replaced behavior and \( \beta_m \in \tilde{Z}_u \) is the corresponding substitute. \( \alpha_m \) and \( \beta_m \) denote the importance score of \( \alpha_m \) for user \( u \) and the relatedness score of substitute \( \beta_m \) for \( \alpha_m \). We use softmax to transform the absolute importance and relatedness into the relative ones within the range of the behavior sequence and the substitute subset, respectively. The intuition behind the multiplication of relative importance and relatedness scores for constructing negatives is that replacing an important item with a highly related substitute will make the augmented negative harder. For example, when a user loves watching the cat micro-videos, replacing a cat micro-video in the behavior sequence with a dog micro-video (of high related score) will make the augmented behavior sequence much harder than replacing it with a cooking micro-video (of low related score). We essentially treat all items the user does not click as negatives, which is a common strategy in training deep candidate generation models for recommendation. Also, replacing unimportant behaviors will intuitively lead to easier augmented samples.

**Constructing Positives.** When we prefer to replace items with high relative unimportance score \( \frac{\exp(-\alpha_i)}{\sum_{j \in X_u} \exp(-\alpha_j)} \), the constructed behavior sequences are potential positives. We compute the hardness score for \( X^u_+ \) as follows:

\[
a^+ = \sum_{(m, n) \in \tilde{Z}_u} \exp(-\beta_m) \frac{\exp(-\alpha_m)}{\sum_{i \in X_u} \exp(-\alpha_i)} \sum_{i \in X_u} \exp(-\beta_m).
\]

The difference between positives and negatives lies in that we use the negative importance/relatedness score before softmax, which will result in a relative unimportance/unrelatedness score after softmax. The intuition behind the multiplication of relative unimportance and unrelatedness scores for constructing positives is that replacing an unimportant item with an unrelated substitute might make the augmented positive be of high quality. For example, when a user dislikes ads micro-videos, replacing an ads micro-video in the behavior sequence with a micro-video that is far from ads (of low related score) will have a higher chance of making the positive be of high quality than replacing it with another ads micro-video (of high related score). Replacing important behaviors will intuitively make the constructed positives be of low quality.

Based on the above computation, we can devise controllable augmentation by deliberately choosing the behaviors to be replaced and the substitutes to construct positives/negatives samples with the desired hardness scores, i.e., hardness-aware augmentation. The controllable augmentation permits many enticing contrastive techniques compared with the vanilla contrastive learning, of which the details are discussed in 3.5.

### 3.4 Sequence Modeling

Sequence modeling lies in the core of most sequential recommendation architectures, which transform the historical behavior sequence into a holistic (or multiple) user interest representation. To better demonstrate the effectiveness of the proposed framework, we propose to simplify the sequence modeling and introduce just one linear transformation layer \( W_3 \) with NO further trainable parameters. This design largely improves the efficiency of both training and inference, which is a critical merit for industrial systems. Specifically, we obtain the original user representation as:

\[
\hat{v}_i = \sum_{t=1}^{N_t} (v_i W_3) \frac{\exp(\alpha_{it})}{\sum_{j \in X_u} \exp(\alpha_{jt})},
\]

\[
u_i^q = \sum_{t=1}^{N_t} \hat{v}_i \frac{\exp(\alpha_{it})}{\sum_{j \in X_u} \exp(\alpha_{jt})},
\]

where \( v_i \) is updated by aggregating information from all other micro-video in the behavior sequence by re-using the correlation score \( \alpha_{it} \) as the weights after softmax. The final user representation \( \nu_i^q \) sums all micro-video features by re-using the importance score \( \alpha_{it} \) as the weights after softmax. This architecture has the strengths of being simple: efficient for training/serving, lower chance of over-fitting, and less space complexity. Augmented behavior sequences \( \{X_{1i}^{\tilde{+}}, X_{2i}^{\tilde{-}}, \ldots, X_{Ni}^{\tilde{-}}\} \) are also transformed into the corresponding user representations \( \{\nu_1^q, \nu_2^q, \ldots, \nu_N^q\} \).

### 3.5 Training and Objectives

#### 3.5.1 Contrast on Users.

Based on the augmented positive and negative user representations, we conduct contrastive learning to learn discriminating user representations. A straightforward way is to employ the margin triplet loss function [1]:

\[
\sum_{i=1}^{N_u} \sum_{j=1}^{N_u} \max\left( d(\nu_i^q, \nu_i^q) - d(\nu_i^q, \nu_j^q) + \delta, 0 \right),
\]

where \( d(\cdot, \cdot) \) denotes the distance measurement and \( \delta \) denotes the margin, which indicate the maximum relative distance between original, positive, and negative.
Figure 2: Schematic of the proposed CCL4Rec architecture, which mainly encapsulates three critical contributions: 1) **Hardness-aware augmentation**, that explicitly tracks the importance of behaviors and relatedness of substitutes in user behavior sequence augmentation, resulting in a hardness score of the augmented sample. 2) **Contrast over contrast** objective that explicitly models the nuances between positives/negatives based on the hardness scores. 3) **Controllable contrastive learning** that can deliberately choose augmented samples with the desired hardness scores.

**Contrast over Contrastive Learning.** However, the above formulation basically treats all positives and negatives as equally important. Failing to model the nuances between them might hinder contrastive learning from learning discriminating user representations. Towards this end, we propose to: 1) model the hardness of positives and negatives in the modeling; and 2) further contrast the negatives or positives in a contrast over contrast manner. Mathematically, the hardness can be incorporated into the objective as follows:

\[
\delta^+ = \max(\{a_i^+, a_j^-\} \times \delta^s, \delta^u, \delta^d),
\]

\[
\mathcal{L}_{ccl} = \sum_{i=1}^{N_p} \sum_{j=1}^{N_n} \max \left( d(u^q, u^+_i) - d(u^q, u^-_j) + \delta^s, 0 \right),
\]

where \(a_i^+, a_j^-\) denote the hardness scores of augmented samples \(u^+_i, u^-_j\), computed in Equation 5 and Equation 4, respectively. Intuitively, if a positive is of high quality, we should pull the corresponding user representation \(u^+_i\) closer to the original \(u^q\). Similarly, if a negative is of high hardness, we should push \(u^-_j\) farther from the original. Therefore, the maximum relative distance, i.e., the margin \(\delta\) should be larger to permit further contrasting than the other pairs. We use the sum of harness score \(a_i^+ + a_j^-\) to control the margin and employ hyper-parameters \(\delta^s, \delta^u, \delta^d\) to control the scale, upper bound, lower bound of the final margin. Besides the conventional way of contrasting positives with negatives, we further contrast positives or negatives as follows:

\[
\delta^{+, \cdot} = \max(\{a_i^+, a_j^-\} \times \delta^s, \delta^u, \delta^d),
\]

\[
\mathcal{L}_{ccl}^{+} = \sum_{i=1}^{N_p} \sum_{k=1}^{a_i^+} \max \left( d(u^q, u^+_i) - d(u^q, u^+_k) + \delta^{+, \cdot}, 0 \right).
\]

Unlike \(\mathcal{L}_{ccl}\), we use \(a_i^+ - a_i^\cdot\) to control the margin. Intuitively, when contrasting positives, if the qualities of two augmented samples largely differ from each other, the maximum relative distance should be large. Similarly in spirit, we could compute \(\mathcal{L}_{ccl}^-\) by contrasting the negatives. We note that the above objectives that explicitly rely on the hardness scores will not optimize the score-related parameters.

**Controllable Contrastive Learning.** As demonstrated in Section 3.3, with the computation of hardness scores, we can easily have hardness-aware augmentation, which further permits controllable contrastive learning by deliberately choosing augmented samples of the desired hardness scores for contrasting. In this paper, we propose multiple controllable contrast strategies and discuss their empirical performance in Section 4.3.1. We take the easy2hard strategy as an example for illustration. With the easy2hard strategy, we propose to learn from augmented samples with smaller hardness scores at the beginning and increase the hardness scores of the chosen samples as the learning procedure processes. Specifically, when constructing positives/negatives, we filter the half items
with the largest/smallest importance scores $a_i$, and further sample half items from the filtered set to be replaced, i.e., $\hat{X}_u$. The sampling probabilities are initially set to the relative unimportance scores, i.e., $\frac{\exp(-a_i)}{\sum_{j \in X_u} \exp(-\sigma_j)}$, which means we construct positives with lower hardness scores at the early stage of training. The sampling probabilities will be linearly transformed towards the importance scores $\frac{\exp(a_i)}{\sum_{j \in X_u} \exp(a_j)}$ along with the training processes. Similarly, we sample substitutes according to the unrelatedness scores $\frac{\exp(-\beta_i)}{\sum_{j \in X_u} \exp(-\beta_j)}$, which will be linearly transformed towards the relatedness score along with the training. Controllable negative augmentation shares similar processes as the above.

3.5.2 click-through-rate prediction. As a common practice, we incorporate a multi-layer perceptron (MLP) to predict the CTR and a cross-entropy loss by globally enlarging the gap between the summed predictions of positive pairs and that of negative pairs.

$$\ell_{ce} = - \sum_{y \in \{0, 1\}} y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}),$$

(14)

where $\lfloor \cdot \rfloor$ denotes the concatenation operation and $\ast$ denotes element-wise product. We also compute the corresponding cross-entropy loss for augmented samples. While $\ell_{ce}^+$ for positives share the same computation, $\ell_{ce}^-$ is computed with the groundtruth $y_u^\ast$ reverted, i.e., $1 \rightarrow 0$, and disregards target micro-videos with label 0.

3.5.3 Contrast on User-Item Pairs. Besides contrasting users, we also consider the contrast between the user and positive/negative items. The objective function can be written as:

$$\ell_{cu} = \sum_u \log \frac{\sum_{y_u, y_u^\ast} y_u^\ast \exp(\hat{y}_u)}{\sum_{y_u, y_u^\ast} y_u \exp(\hat{y}_u) + \sum_{y_u, y_u^\ast} (1 - y_u^\ast) \exp(\hat{y}_u)}.$$  

(15)

This contrast is an essential complement to the original cross-entropy loss by globally enlarging the gap between the summed predictions of positive pairs and that of negative pairs.

3.5.4 Training. During training, we naively sum the above loss functions and do not tune the relative weights with bells and whistles to better demonstrate the strengths of the proposed objectives:

$$\ell = \ell_{cc} + \ell_{ccl}^+ + \ell_{ccl}^- + \ell_{ce} + \ell_{ce}^+ + \ell_{ce}^- + \ell_{cu}.$$  

(16)

4 EXPERIMENTS

We are concerned with the following research questions:

- **RQ1**: How does CCL4Rec perform compared with the state-of-the-art micro-video recommender?
- **RQ2**: Do all contrastive objectives contribute to the effectiveness of CCL4Rec? How do different key hyper-parameters settings and controllable contrastive strategies affect the performance?
- **RQ3**: Does CCL4Rec learn discriminating user presentations?

4.1 Experimental Settings

Datasets As a common practice of micro-video recommendation research [12, 15], we conduct experiments on two micro-video datasets, i.e., MicroVideo-1.7M [3] and Kuaishou 1. These two datasets both provide pre-extracted video features and logged interactions in the form of <User ID, Video ID, Timestamp>. Each interaction is labeled as positive or negative according to whether the user clicked the exposed video when he/she saw the thumbnail. We exactly follow the data pre-processing of Chen et al. [3], and Li et al. [15] for a fair comparison with the existing state-of-the-art models. The statistics of these datasets are listed in Table 1.

| Dataset         | #Users | #Items | #Interactions | #Density |
|-----------------|--------|--------|---------------|----------|
| MicroVideo-1.7M | 10,986 | 1,704  | 880 12,737,619| 0.068%   |
| Kuaishou        | 10,000 | 3,239,534 | 13,661,383     | 0.042%   |

1https://www.kesci.com/home/competition/5ad306e633a98340e004f9d1.

Table 1: Statistics of the Datasets.
Table 2: Overall performance comparison between CCL4Rec and micro-video recommendation baselines. We highlight the best/next-best performance with bold/underline style. We show that CCL4Rec achieves comparable performance to the state-of-the-art methods but is of far more efficiency as illustrated in Table 3.

| Model    | MicroVideo-1.7M | Kuaishou |
|----------|-----------------|----------|
|          | AUC  | Precision@50 | Recall@50 | F1-Score@50 | AUC  | Precision@50 | Recall@50 | F1-Score@50 |
| BPR      | 0.583 | 0.241 | 0.181 | 0.206 | 0.595 | 0.290 | 0.387 | 0.331 |
| LSTM     | 0.641 | 0.277 | 0.205 | 0.236 | 0.713 | 0.316 | 0.420 | 0.360 |
| CNN      | 0.650 | 0.287 | 0.214 | 0.245 | 0.719 | 0.312 | 0.413 | 0.356 |
| NCF      | 0.672 | 0.316 | 0.225 | 0.262 | 0.724 | 0.320 | 0.420 | 0.364 |
| ATRank   | 0.660 | 0.297 | 0.221 | 0.253 | 0.722 | 0.322 | 0.426 | 0.367 |
| THACIL   | 0.684 | **0.324** | 0.234 | 0.269 | 0.727 | 0.325 | 0.429 | 0.369 |
| ALPINE   | **0.713** | 0.300 | 0.460 | 0.362 | 0.739 | 0.331 | 0.436 | 0.376 |
| MTIN     | **0.729** | 0.317 | **0.476** | **0.381** | **0.752** | **0.341** | **0.449** | **0.388** |
| CCL4Rec  | 0.722 | 0.312 | 0.472 | 0.376 | 0.750 | 0.340 | **0.449** | **0.387** |

Table 3: Computation Complexity Analysis. Training time concerns one epoch and testing concerns one batch. We choose two best-performing state-of-the-art methods and all models are with batch size of 32 and 1 NVIDIA V100 GPU. We use the official codebases of ALPINE and MTIN provided by the authors and run them with default settings on the Kuaisouh dataset. We highlight the speedup rate over two state-of-the-art methods.

| Model    | Training (s) | Inference (s) | #Parameters |
|----------|--------------|---------------|-------------|
| ALPINE   | 10,778.50    | 2.896.83      | 208,854,467 |
| MTIN     | 121.88       | 0.0076        | 10,849,706  |
| CCL4Rec  | 121.88       | 0.0076        | 758,341     |
| Efficiency | **88.44/23.8×** | **94.6/463.4×** | **275.4/14.3×** |

4.2 Comparison Methods

To demonstrate the effectiveness of CCL4Rec, we consider the following state-of-the-art recommenders as comparison methods.

- **BPR** [25]. Bayesian personalized ranking devises the maximum posterior estimator from the Bayesian perspective. The BPR loss is computed as the relative difference between positive pairs and negative pairs.
- **LSTM** [41]. An LSTM based sequential recommender. We use the final hidden state as the user interest vector and an MLP predictor for click-through-rate prediction.
- **CNN**. CNN extracts features from the behavior sequence, and we take the pooled feature as the user interest vector. We use various kernels with various window sizes to extract features from the behavior sequence and apply global pooling on the extracted features to obtain user interest vector. Similarly, we use an MLP predictor for prediction.
- **NCF** [6]. NCF is a collaborative filtering recommender equipped with deep neural networks, which can be of great representation power compared with the traditional inner product.
- **ATRank** [42]. ATRank is a ranking recommender that comprehensively leverages the power of attention mechanisms.
- **THACIL** [3]. THACIL divides the historical behaviors into multiple temporal blocks and captures both the intra-block and inter-block correlations. In addition, THACIL considers both category-level and item-level user interests for representing users’ interests.
- **ALPINE** [15]. The essence of ALPINE lies in the temporal graph-based LSTM module, which captures dynamic interests within the temporal behavior graph, and the multi-level interest modeling layer, which models multi-type behaviors.
- **MTIN** [12]. To investigate the multi-scale time effects and item-to-interest grouping problem, MTIN devises the time-aware parallel masks and the group routing algorithm.

We report the evaluation results of all methods on two micro-video recommendation benchmarks in Table 2. Since computation complexity is one of the critical factors considered by industrial recommender systems, we also report the training/inference/parameters of the strongest state-of-the-art methods (i.e., ALPINE, MTIN) and CCL4Rec in Table 3. By analyzing these results, we have the following observations:

- Although CCLRec cannot beat the strongest state-of-the-art method, i.e., MTIN, in most cases, it achieves comparably good results and outperforms many other existing micro-video recommendation state-of-the-art methods. In particular, the relative improvement over other simple designs (e.g., BPR, LSTM, CNN, NCF) and also some advanced approaches (e.g., ATRank, THACIL) reach nearly 100% and 8% with Recall@50 on the MicroVideo and Kuaishou benchmarks, respectively. We attribute these advantages to the following aspects: 1) The contrastive learning objective helps the sequence modeling module to learn more distinguishing user representations that can mitigate the effects of ubiquitous false-positive interactions by contrasting the learned representation with augmented samples. 2) The proposed hardness-aware augmentation considers the importance of replaced items and the relatedness of substitutes, and accordingly determining the hardness scores of augmented samples in
Table 4: Analysis on the number of augmented positives/negatives.

| $N_p$, $N_n$ | AUC   | Precision@50 | Recall@50  | F1-Score@50 |
|-------------|-------|--------------|------------|-------------|
| 2           | 0.7425| 0.3386       | 0.4469     | 0.3852      |
| 3           | 0.7471| 0.3402       | **0.4491** | **0.3871**  |
| 4           | 0.7503| 0.3401       | 0.4487     | 0.3869      |
| 5           | 0.7487| 0.3397       | 0.4479     | 0.3863      |

Table 5: Analysis on different contrastive learning objectives.

| Model       | AUC     | Precision@50 | Recall@50  | F1-Score@50 |
|-------------|---------|--------------|------------|-------------|
| CCL4Rec     | **0.7503** | 0.3401       | **0.4487** | **0.3869**  |
| $-\mathcal{L}^{+/\cdot}_{ccl}$ | 0.7388 | 0.3359       | 0.4439     | 0.3824      |
| $-\mathcal{L}_{ccl}$  | 0.7273 | 0.3353       | 0.4428     | 0.3816      |
| $-\mathcal{L}_{cui}$  | 0.7239 | 0.3327       | 0.4404     | 0.3790      |

a fine-grained manner. These designs help to model the nuances of different items in augmentation, and the nuances of different augmented samples in contrasting, leading to even more distinguishing user representations.

- While achieving competitive performance, CCL4Rec significantly improves the training efficiency (88.44/23.8 times faster than ALPINE/MTIN) and the inference efficiency (94.6/463.4 times faster than ALPINE/MTIN). These results demonstrate the strengths of CCL4Rec in being simple and effective, which are critical for industrial scenarios where the number of users/items can easily reach billions in scale. Furthermore, the number of parameters is orders of magnitude (275.4/14.3) less than the other two state-of-the-art methods. Considering the rather simplified sequence modeling module in the CCL4Rec framework, CCL4Rec has the potentials to enhance a tiny-size model that will be used for on-device/edge computing.

4.3 Study of CCL4Rec (RQ2)

4.3.1 Study of the controllable contrastive strategies. To demonstrate the merits of controllable contrastive learning empowered by the controllable augmentation, we devise multiple strategies as follows:

- The harder strategy, which means we prefer to choose the augmented samples with larger hardness scores. We achieve this by using importance/relatedness scores as weights to sample the micro-videos to be replaced and the substitutes;
- The easier strategy, which means we prefer ones with smaller hardness scores. We achieve this by using unimportance / unrelatedness scores as weights.
- The easy2hard strategy as illustrated in Section 3.5;
- The hard2easy strategy; and 5) The random strategy as a widely used baseline.

From the training curve depicted in Figure 3, we have the following observations: 1) compared to the baseline random strategy that has several high loss/performance jitters, other controllable contrastive strategies overall have significantly less and lower jitters; 2) among all the contrastive strategies, the easy2hard strategy has the most robust training process and yields the best performance in most times. We note that random strategy can be the most straightforward one that is widely used in many contrastive learning based methods. However, according to our analysis, it might lead to unstable training process and more training epochs to converge. These results jointly demonstrate the necessity of controllable contrastive learning and the effectiveness of our design in boosting performance and training robustness.

4.3.2 Study of the number of augmented positives/negatives. We vary the number of augmented positives/negatives to reveal its effect on the performance. The results in Table 4 indicate that: 1) the performance of CCL4Rec is overall insensitive to this hyperparameter and small $N_p$, $N_n$ will lead to competitive performance; 2) increasing $N_p$, $N_n$ (2 → 3, 4) will yield performance improvement since the contrast over contrast objective should learn more
distinguishing user representation by modeling the nuances between multiple positives/negatives; 3) Further increasing $N_p, N_n$ ($3 \rightarrow 4$) will lead to fewer gains. The reason for this might be that moderate $N_p, N_n$ is enough to reach the optimal, which can be computationally efficient.

4.3.3 Study of the contrastive objectives. (Ablation Study). The critical contributions of CCL4Rec are the contrastive objectives. To evaluate whether each of them contributes to the final performance, we ablate CCL4Rec by progressively removing them and test the resulting architectures. Specifically, we observe the following from Table 5: 1) By removing $L_{ce}^{-/+}$ which means we disregard the contrasts between positives or negatives, the performance gap between the resulting model and the full model demonstrates the necessity and effectiveness of the contrast over contrast framework; 2) By further removing $L_{cel}$, i.e., eliminating the effect of contrasting the learned representation with positive and negative ones, we can observe a clear performance drop. It is noteworthy that $L_{cel}$ also exhibits a contrast over contrast merit since the adaptive pulling/pushing strengths help CCL4Rec model the nuances between different augmented samples; 3) Further removing $L_{cut}$, which means the CCL4Rec framework totally lost the contrastive learning capability. The performance drop further shows the merits of the proposed contrastive learning framework.

4.4 Case Study (RQ3)
To evaluate whether we learn discriminating user representations via contrast over contrastive learning, we follow a widely used case study schema [29, 31] to visualize the t-SNE transformed embeddings. We adopt a contrastive baseline that is trained with $L_{ce}$ and $L_{cel}$ without adaptive pulling/pushing strengths, i.e., the architecture without contrast over contrast. We plot two randomly sampled users and each user has four augmented positives/negatives with the hardness scores predicted by the corresponding model. As shown in Figure 4, we have the following observations:

- Both results yield clear gap between positives and negatives, which reveal the effectiveness of contrastive learning.
- The positives/negatives are less distinguishable from each in the baseline hyper-embedding space. Since positives and negatives are obtained by transforming important/unimportant items in the behavior sequence, these results basically indicate that the contrastive baseline treat all items as equally important and thus learning indiscriminating representations.
- By further considering the hardness scores, CCL4Rec successfully pushes negatives with high hardness scores farther and pulls positives with high hardness scores closer (while the baseline can not). These results demonstrate the effectiveness of contrast over contrast objective in CCL4Rec.

5 CONCLUSION AND FUTURE WORK
In this work, we investigate the problem of learning discriminating user representation via contrastive learning. Different from existing contrastive learning approaches, we propose to better accommodate recommendation using hardness-aware augmentation and contrast over contrastive learning. By hardness-aware augmentation, we can easily obtain augmented samples with the desired hardness scores by deliberately manipulating the importance of behaviors being replaced and the relevance of substitutes. Based on the hardness scores, we devise the controllable contrastive strategies, contrasts on negatives or positives, and contrast with adaptively pulling/pushing strengths, which jointly help to model the nuances between different augmented samples and thus learning discriminating representations.

To the best of our knowledge, the contrast over contrastive learning framework is among the earliest attempts in research on recommendation, while also contributing to the literature of generic contrastive learning. We believe the insights of CCL4Rec are inspirational to future developments on learning effective user representations and may be beneficial to a broad range of research domains. We plan to further explore the strengths by designing better augmentation cost prediction modules and better objectives to model the nuances of augmented samples. Another future direction is to take a more in-depth analysis of controllable contrastive learning. We mainly analyze the effects on the training process and final performance. Carefully designed strategies and broader merits are ripe for exploration. Some previous works found that the contrastive learning objective can be beneficial for bias reduction. We thus plan to also explore whether the contrast over contrastive learning framework can further help to reduce biases in recommender systems (e.g., popularity bias) compared with the traditional one.
REFERENCES

[1] Vassileios Balntas, Edgar Ribas, Daniel Ponsa, and Krystian Mikolajczyk. 2016. Learning local feature descriptors with triplets and shallow convolutional neural networks. In Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event.

[3] Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, and Yan Li. 2018. Temporal Hierarchical Attention at Category- and Item-Level for Micro-Video Click-Through Prediction. In 2018 ACM Multimedia Conference on Multimedia, MM 2018, Seoul, Republic of Korea, October 22-26, 2018.

[4] Xiaoyu Du, Xiang Wang, Xiangnan He, Zechao Li, Jinhui Tang, and Tat-Seng Chua. 2020. How to Learn Item Representation for Cold-Start Multimedia Recommendation?. In MM ’20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-26, 2020.

[5] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2017. A Unified Personalized Video Recommendation via Dynamic Recurrent Neural Networks. In Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017.

[6] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017.

[7] Balazs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Dominos Kopf. 2016. Session-based Recommendations with Recurrent Neural Networks. In 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings.

[8] Shintami Chusnul Hidayati, Cheng-Chun Hsu, Yu-Ting Chang, Kai-Lung Hua, Shing-Ming Cheng, and Jen-Hao Jheng. 2018. What Fits Me Best?: Fashion Recommendation on the Clothing Style for Personal Body Shape.. In 2018 ACM Multimedia Conference on Multimedia, MM 2018, Seoul, Republic of Korea, October 22-26, 2018.

[9] Xiaowen Huang, Quan Fang, Shengsheng Qian, Jiao Song, Yan Li, and Chang-sheng Xu. 2019. Explainable Interaction-driven User Modeling over Knowledge Graph for Sequential Recommendation. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2019, Nice, France, October 21-25, 2019.

[10] Xiaowen Huang, Shengsheng Qian, Quan Fang, Jiao Song, and Changsheng Xu. 2018. CSAN: Contextual Self-Attention Network for User Sequential Recommendation. In 2018 ACM Multimedia Conference on Multimedia, MM 2018, Seoul, Republic of Korea, October 22-26, 2018.

[11] Hao Jiang, Wenjie Wang, Meng Liu, Liqiang Nie, Ling-Yu Duan, and Changsheng Xu. 2019. Market2Dish: A Health-aware Food Recommendation System.. In Video Recommendation via Dynamic Recurrent Neural Networks.. In Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence Conference 2022 Virtual Event / Seattle, WA, USA, October 12-16, 2020.

[12] Hao Jiang, Wenjie Wang, Meng Liu, Liqiang Nie, Ting-Teng Chen, and Changsheng Xu. 2020. Neural Graph Collaborative Filtering. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019.

[13] Yongqi Li, Meng Liu, Jianhua Yin, Chaoran Cui, Xin-Shun Xu, and Liqiang Nie. 2018. What Aspect Do You Like: Multi-scale Time-aware User Interest Model- ing for Micro-video Recommendation. In Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019.

[14] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. 2020. CURL: Contrastive Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 13-18, 2018.

[15] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive Multiview Coding. In Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI.

[16] Steffen Rendle. 2016. Session-based Recommendations with Recurrent Neural Networks.. In Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, 3772–3782.

[17] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In WSDM 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009.

[18] Shengyu Zhang, Ziqi Tan, Zhou Zhao, Jin Yu, Kun Kuang, Tan Jiang, Jingren Zhou, Zhiyong Cheng, Xuzheng Yu, Zhou Zhao, Lei Zhu, and Liqiang Nie. 2019. Personalized Hashtag Recommendation for Micro-videos.. In Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019.

[19] Shengyu Zhang, Ziqi Tan, Zhou Zhao, Jin Yu, Kun Kuang, Jie Liu, Jingren Zhou, Zhiyong Cheng, Xuzheng Yu, Zhou Zhao, Lei Zhu, and Liqiang Nie. 2019. Personalized HashTag Recommendation for Micro-videos.. In Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019.

[20] Yuan Nie, Shengyu Zhang, Zhong Zhao, Jieming Zhu, Qi Zhang, Jingjie Li, Xiaoguang He, Xiaofei He, Tat-seng Chua, and Fei Wu. 2021. Why Do We Click: Visual Impression-aware New Recommendation. In MM ’21: ACM Multimedia Conference, ACM, 3881–3890.

[21] Xinwen Yang, Dongliang Xie, Xin Wang, Jiangbo Yuan, Wanqing Ding, and Pengyun Yan. 2020. Learning Tuple Compatibility for Conditional Outfit Recommendation. In MM ’20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-26, 2020.

[22] Xingzheng Cheng, and Jun Gao. 2018. ATRank: An Attention-Based User Behavior Modeling Framework for Sequential Recommendation of Micro-video.. In Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019.

[23] Xu Xie, Fei Sun, Zhaoyang Liu, Jinyang Gao, Bolin Ding, and Bin Cui. 2020. Contrastive Pre-training for Sequential Recommendation. CoRR (2020).

[24] Jiahao Xun, Shengyu Zhang, Zhou Zhao, Jieming Zhu, Qi Zhang, Jingjie Li, Xiaoguang He, Xiaofei He, Tat-seng Chua, and Fei Wu. 2021. Why Do We Click: Visual Impression-aware New Recommendation. In MM ’21: ACM Multimedia Conference, ACM, 3881–3890.

[25] Shengsheng Qian, Shengyu Zhang, Zhong Zhao, Jieming Zhu, Qi Zhang, Jingjie Li, Xiaoguang He, Xiaofei He, Tat-seng Chua, and Fei Wu. 2021. Why Do We Click: Visual Impression-aware New Recommendation. In MM ’21: ACM Multimedia Conference, ACM, 3881–3890.
Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18). New Orleans, Louisiana, USA, February 2-7, 2018.
[43] Chang Zhou, Jianxin Ma, Juanwei Zhang, Jingren Zhou, and Hongxia Yang. 2020. Contrastive Learning for Debiased Candidate Generation at Scale. CoRR (2020).
[44] Guorui Zhou, Xiaqiang Zhu, Chengru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep Interest Network for Click-Through Rate Prediction. In KDD.
[45] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-Rec: Self-Supervised Learning for Sequential Recommendation with Mutual Information Maximization. In CIKM.