Iterative Sparse Attention for Long-sequence Recommendation

Guanyu Lin[1,2], Jinwei Luo[3], Yinfeng Li[1] Chen Gao*[1] Qun Luo[4] Depeng Jin[1]

[1]BNRist, Tsinghua University
[2]Carnegie Mellon University
[3]Shenzhen University
[4]Tencent Inc.
guanyul@andrew.cmu.edu, chgao96@gmail.com

Abstract

Longer historical behaviors often improve recommendation accuracy but bring efficient problems. As sequences get longer, the following two main challenges have not been addressed: (1) efficient modeling under increasing sequence length and (2) interest drifting within historical items. In this paper, we propose **I**terative **S**parse **A**ttention for Long-sequence Recommendation (ISA) with Sparse Attention Layer and Iterative Attention Layer to efficiently capture sequential pattern and expand the receptive field of each historical items. We take the pioneering step to address the efficient and interest drifting challenges for the long-sequence recommendation simultaneously. The theoretical analysis illustrates that our proposed iterative method can approximate full attention efficiently. Experiments on two real-world datasets show the superiority of our proposed method against state-of-the-art baselines.

Code — https://github.com/tsinghua-fib-lab/ISA

Introduction

With the exploration of modern personalized platforms such as Wechat's Channels and Taobao e-commerce etc., it is essential to model the users' preferences and engage them. One of the most important tasks to model users' preference is the sequential recommendation (Zhou et al. 2019; Kang and McAuley 2018; Chang et al. 2021; Qiao et al. 2024) that leverages behavior sequence to promote the performance, especially when the sequence is longer (Pi et al. 2020; Lin et al. 2022a; Cao et al. 2022). However, when the sequence becomes longer, an efficient problem will arise. Besides, it is hard to capture the interest drift within the long sequence. Indeed, there are two key challenges for the long-sequence recommendation.

- **User sequence is too long to achieve efficient modeling.** In the long-sequence recommendation, users often click many items, resulting in long sequence and make the efficient modeling difficult.
- **User interest drifts dramatically in a long sequence.** As time goes on and the accumulation of observed items, user interest will drift dramatically between distant items in the historical item sequence. For example, a user may prefer ice cola in summer but prefer hot tea in winter.

Figure 1: Illustration of attention distance for sequential recommendation where the last item attends to neighborhood items by different steps. Here the first and last items are both about the car, and thus the first item with three steps away is the most important item for the last item.

To achieve efficient modeling, DIN (Zhou et al. 2018) leverages the target item to aggregate the related historical items. Moreover, some search-based works (Pi et al. 2020; Chen et al. 2021; Cao et al. 2022) leverage the target item to query and filter the historical items. However, these works ignore the interest drifting within the historical item sequence. To further capture the interest drifting within historical items, SAM (Lin et al. 2022a) exploits the aggregated historical items to query themselves. Such a way is implicit and, indeed, will only aggregate the historical items that are indirectly related to the target item.

To address the efficiency and interest drifting challenges at the same time, we propose a method named ISA, a novel iterative sparse attention for long-sequence recommendation with Sparse Attention Layer to reduce the cost and Iterative Attention Layer to expand the interest drifting. Specifically, to elegantly address the efficient challenge, we leverage three sparse attention mechanisms (Zaheer et al. 2020) from NLP with local window attention, global attention, and random attention. Specifically, we leverage (1) the global attention to capture the global drifted interests of short-term items, (2) the window attention to extract the local drifted interests of historical items, and (3) the random attention to randomly connect the drifted interests of pair items. However, these sparse attention methods may fail to capture some distant relations when correlated items are not neighboring. For example, as the sequence is shown in Figure 1, short attention distance with less than two steps may be insufficient when the really correlated items are three steps away. That is to say, each single sparse attention layer can not guarantee suf-

*Corresponding Author.

ficient attention distance for all items. Indeed, even for full self-attention, the shallow layers still mainly attend to the neighboring items (Dosovitskiy et al. 2020). Though stacking the layers can help our proposed sparse attention layer to extend the receptive field of interest drifting for distant items (Child et al. 2019; Dai et al. 2019), it will unavoidably introduce more costs, which may be the bottleneck for the large-scale recommendation. Thus, our sparse attention approach still meets the interest drifting challenge for a long sequence. Besides, to address the interest drifting challenge, we further propose the Iterative Attention Layer, which tends to introduce a iterative process on the proposed Sparse Attention Layer for approximating efficient full attention and expanding attention distance for interest drifting. The Iterative Attention Layer first calculates attention scores as the Sparse Attention Layer. Then it extends the attention distance between historical items through the iterative process to capture and expand the interest drifting. Specifically, we leverage power iteration method for the Personalized PageRank, which can extend the local interest drifting to the entire sequence iteratively. We also theoretically and experimentally analyze that ISA can expand the interest drifting for the distant items and approximate the full attention efficiently. The contributions of this paper are as below.

- To the best of our knowledge, we take the pioneering step to address the efficient and interest drifting challenges for the long-sequence recommendation simultaneously.
- We develop a ISA, which consists of two parts: 1) Sparse Attention Layer that reduces cost and achieves efficient modeling, and 2) Iterative Attention Layer that expands the interest drifting between distant items and improves the sparse attention.
- We evaluate our method on the long-sequence recommendation. Experiments on two real-world datasets show that ISA can effectively improve attention mechanisms and reduce cost.

Problem Definition Before introducing our solution, we tend to define the problem of Long-sequence Recommendation. Supposing i_t is the clicked item of user u at time step t. Then if we have collected click sequence $\mathcal{I}_u = (i_1, i_2, \ldots, i_t)$, the sequential recommendation task aims to predict the click probability of user u towards target item i_{t+1} at time step $t+1$. Typically, the task with long behavior is often with item sequence longer than hundreds, which can be formulated as follows.

Input: Click item sequence $\mathcal{I}_u = (i_1, i_2, \ldots, i_t), t \geq 100$ for user u.

Output: Click probability of user u to target item i_{t+1}.

Methodology

Our proposed ISA is illustrated in Figure 2, which consists of five parts, where the Sparse Attention Layer and Iterative Attention Layer are designed for long-term modeling while the Target Attention Layer is designed for short-term modeling. These two layers are our main contributions to the two key challenges. (1) **Sparse Attention Layer**: We propose global attention for global drifted interests of short-term items, window attention for locally drifted interests of historical items, and random attention for casual drifted interests of pair items to reduce the quadratic cost and approximate to self-attention efficiently. (2) **Iterative Attention Layer**: We leverage Personalized PageRank (PPR) to expand the attention distance and connect distant items for the sparse attentions, which is also proven to approximate self-attention.

Long-short-term Embedding Layer

To be time-aware about the historical items, we build a positional embedding matrix $\mathbf{P} \in \mathbb{R}^{T \times D}$ and an item embedding matrix $\mathbf{M} \in \mathbb{R}^{m \times D}$. Here m is the number of items, T is the maximum number of items in one sequence for long-term behaviors, and D is the dimension of embedding. Then we generate the time-aware input embedding by adding the positional embedding and item embedding. Therefore, for click item sequence $\mathcal{I}_u = (i_1, i_2, \ldots, i_t)$, we generate the time-aware input embeddings i.e. $\mathbf{E} \in \mathbb{R}^{T \times D}$, as: $\mathbf{E} = [\mathbf{M}_{i_1}, \mathbf{M}_{i_2}, \ldots, \mathbf{M}_{i_T}] + [\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_T]$, where $[\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_T]$ are positions of items $\{i_1, i_2, \cdots, i_T\}$. Notice that we will pad those sequences with click items short than length T (Kang and McAuley 2018). Similarly, we can also get short-term item embeddings $\mathbf{E}^s \in \mathbb{R}^{(T^s) \times D}$ as:$\mathbf{E}^s = [\mathbf{M}_{i_x}, \mathbf{M}_{i_{x+1}}, \ldots, \mathbf{M}_{i_T}] + [\mathbf{P}_x, \mathbf{P}_{x+1}, \ldots, \mathbf{P}_T]$, where x is the index of the first short-term item and $T^s = T - x + 1$.

Sparse Attention Layer

To address the first challenge, as shown in the middle part of Figure 2, we leverage a combination of global attention, window attention, and random attention to reduce the quadratic cost.

Global Attention The most recent short-term items are often representative of the user's current interests (Pal et al. 2020). These short-term interests indeed drifted from historical items long-term ago. Thus we tend to extend the attention distance of each short-term item to the whole sequence (Zaheer et al. 2020; Beltagy, Peters, and Cohan 2020). In other words, the short-term items are relatively less but more important. So we introduce global attention to short-term items which will attend to the whole historical sequence and capture the global drifted interests for them. Specifically, for the selected short-term global items $(i_x, i_{x+1}, \ldots, i_t)$, we have:

$$(\mathbf{A}_{x,:x}, \mathbf{A}_{x+1,:x+1}, \ldots, \mathbf{A}_{t,:t}) = 1, \tag{1}$$

which is at $O(T^s T)$ complexity. Such complexity can be treated as linear dependency on long-term sequence length because of $T^s \ll T$.

Here we also compare our global attention with two existing typical works about global attention, i.e., Longformer (Beltagy, Peters, and Cohan 2020) and BigBird (Zaheer et al. 2020). Different from both of them (Beltagy, Peters, and Cohan 2020; Zaheer et al. 2020), we further consider the precursor and successor items of sequence, also known as causality (Kang and McAuley 2018). Specifically, when predicting the target item i_{t+1}, we only could consider the first t items, neither item i_{t+1} nor items after it. However, the traditional global attention contains embeddings of subsequent items, which indeed will cause the leakage of future items

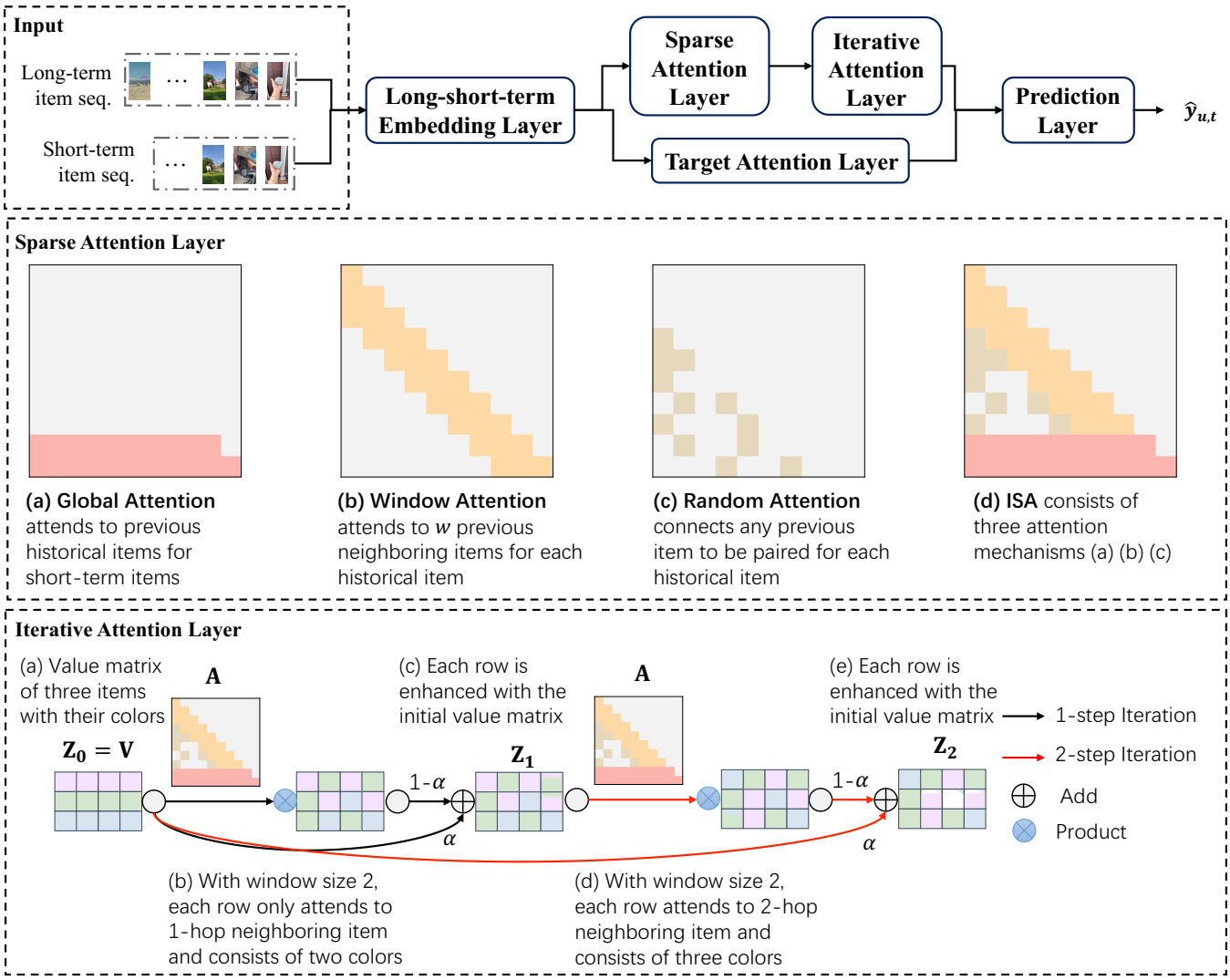

Figure 2: Illustration of ISA. (i) **Long-short-term Embedding Layer** aims to inject long-term, and short-term historical item embeddings with a positional embedding according in a time-aware manner; (ii) **Sparse Attention Layer** reduces the complexity of full-attention by Global Attention, Window Attention, and Random Attention; (iii) **Iterative Attention Layer** extends the attention distance of sparse attentions and further approximate to the full-attention efficiently; (iv) **Target Attention Layer** attends to the short-term items with target item and aggregates them based on the attention score; (v) **Prediction Layer** evolves long-term and short-term interests with prediction towers and performs logloss to train the model. Best viewed in color.

and break causality. Hence, we improve the global attention by masking the attention score of item i_{t_1} towards item i_{t_2} when $t_1 > t_2$. Different from Longformer (Beltagy, Peters, and Cohan 2020) with different parameters for each attention type, our global short-term items leverage the shared parameters to avoid overfitting for sparse recommendation data. Besides, we construct global attention with the individual short-term items, which is different from BigBird (Zaheer et al. 2020) where items will be grouped.

Window Attention Interest often drifts gradually across neighboring items. To capture the local drifted interest within neighboring items, inspired by existing works (Beltagy, Peters, and Cohan 2020; Zaheer et al. 2020), we propose window attention which will attend to neighboring items for each historical item. Specifically, given a window size w, each item only attends to w items on the left side (for causality). We also overlap $\frac{w}{2}$ items to reduce the cost into linear complexity at $O(Tw^2)$.

However, the attention distance of window attention is limited by size w, which facilitates us to expand it by iteration process in the subsequent Section . After our iteration, the distance of the window attention can attend will increase by K times with iteration steps K. Thus our ISA can attend to distant items even under narrow window sizes and shallow attention layers, based on the expanded iterative attention.

Length	SASRec	ISA			
		global	window	random	total
256	50.20%	6.07%	3.84%	0.68%	10.59%
300	50.17%	5.20%	3.28%	0.59%	9.08%

Table 1: Attention comparison between SASRec and our ISA. Here SASRec is the model with causality to mask the future items, while initial self-attention will be with 100% attention. Most importantly, our ISA, with global, window, and random attention, reduces the attention by around 40% v.s. SASRec.

Random Attention Two distant items may have potential or casual relation, i.e., casual drifted interest. To extract such interest, we exploit random attention to randomly connect any pair of items. Besides, the learning intuition behind random attention is to connect items at any distance for approximating full attention. Theoretically speaking, a random graph with expanding properties like Erdős-Rényi graph (Erdős, Rényi et al. 1960) is a decent approximation of a fully-connected graph for self-attention spectrally (Friedman 2008; Feng et al. 2022a). Specifically, for each historical item i_{t_1}, we randomly select r items ($O(\log(T)/T) < r \ll T$). For each selected item i_{t_2}, we have: $\mathbf{A}_{t_1,t_2} = 1, t_1 < t_2$, which is at $O(rT)$ complexity. Different from BigBird (Zaheer et al. 2020) whose random attention will group the items, ISA constructs random attention with individual items. The reason is that the individual item selection will generate more uniform distributions than the group item selection, which improves the expanding properties (Feng et al. 2022b).

To show the efficiency of the proposed Sparse Attention Layer, we also present the comparison of ISA with the full attention as Table 1. It is obvious that our proposed sparse attentions reduce the number of attentions totally by around 40%, which illustrates their efficiency.

Iterative Attention Layer

To address the second challenge and expand the attention distance of interest drifting, we will first diffuse the receptive field of sparse attention by multiple steps, as shown in Figure 3. Then we leverage the Personalized PageRank (PPR) to implement the attention process. Besides, to further reduce the computational complexity, we propose the power iteration to approximate PPR.

Multi-step Attention In the Sparse Attention Layer, the attention matrix $\mathbf{A}$ is calculated by global attention, window attention, and random attention as:

$$\mathbf{A}_{i,j} = \frac{\exp\left(\mathbf{Q}_i \mathbf{K}_j / \sqrt{d}\right)}{\sum_{j \in \mathcal{N}_i} \exp\left(\mathbf{Q}_i \mathbf{K}_j / \sqrt{d}\right)}, \tag{2}$$

where $j \in \mathcal{N}_i$ refers to the neighboring item of item i, connected by the sparse attentions. The attention matrix $\mathbf{A}$ represents the score between any connected item pair by initial sparse attention graph as Figure 3 (a), where some item pairs such as node 1 and node 3 are actually not connected. Due to such unavoidable disconnection, sparse attention of Figure 3 (a) may ignore some important relations. Thus, we further propose Iterative Attention Layer to diffuse the initial sparse attention graph by connecting the item pair multiple steps away. For example, Figure 3 (b) is the graph after two steps of iteration, where nodes 1 and 3 two steps away, are connected. Besides, Figure 3 (c) further connects nodes 1 and 4, which are three steps away. Formally speaking, the multi-step attention scores are calculated below.

$$\mathbf{Z} = \left(\sum_{k=0}^{\infty} \theta_k \mathbf{A}^k\right) \mathbf{V}, \tag{3}$$

where $\mathbf{A}$ is the sparse attention matrix, with the weighting parameter θ_k satisfies $\sum_{k=0}^{\infty} \theta_k = 1, \theta_k \in [0, 1]$. The initial attention distance based on the sparse attention graph will be longer as k becomes larger. The resulting attention score $\sum_{k=0}^{\infty} \theta_k \mathbf{A}^k_{i,j}$ will finally explore all paths between any item pair $< i, j >$, weighted by the parameter θ_k. Besides, each value vector $\mathbf{V}$ will be multiplied and weighted by this calculated attention score.

Personalized PageRank Iteration To efficiently implement the attention mechanism, we leverage Personalized PageRank (PPR), which is a widely used method for graph node proximity (Page et al. 1999). In PPR, the attention weight θ_k will be specified as $\alpha(1-\alpha)^k$, which keeps the initial attention at strength α while diffusing the attention at strength $(1-\alpha)$ for each step. Based on PPR, the implemented output will be as:

$$\mathbf{Z} = \left(\sum_{k=0}^{\infty} \alpha(1-\alpha)^k \mathbf{A}^k\right) \mathbf{V}, \tag{4}$$

where α is the teleport probability hyper-parameter. We then further adopt the power iteration method (Page et al. 1999) to reduce the matrix exponential operation of matrix $\mathbf{A}$ at Eqn. (3) into linear complexity and approximate PPR within K iteration steps. Specifically, each power iteration step is as below.

$$\mathbf{Z}_{(0)} = \mathbf{V} = \mathbf{E}\mathbf{W}_V, \mathbf{Z}_{(k+1)} = (1-\alpha)\mathbf{A}\mathbf{Z}_{(k)} + \alpha\mathbf{V}, \tag{5}$$

where $0 \le k \le K$. Here $\mathbf{Z}_K$ is the output after K steps iteration process, which will approximate to the real output $\mathbf{Z}$ as $K \to \infty$. The proposition is as below.

Proposition 0.1. $\lim_{K\to\infty} \mathbf{Z}_{(K)} = \mathbf{Z}$

Here $\mathbf{Z}_K$ is also the long-term representation of our ISA.

Target Attention Layer and Prediction Layer

In this section, we first illustrate the target attention for the target attention layer, then evolve the long-term and short-term representations by MLP for the prediction layer. Finally we leverage the widely used logloss (Zhou et al. 2018, 2019) to train the model.

Target Attention The target item often relates to historical items. Inspired by existing works (Zhou et al. 2018; Pi et al. 2020; Chen et al. 2021), we leverage target attention to extract the related short-term items. Specifically, we multiply the query matrix $\mathbf{Q}^s$ and key matrix $\mathbf{K}^s$ to obtain the attention

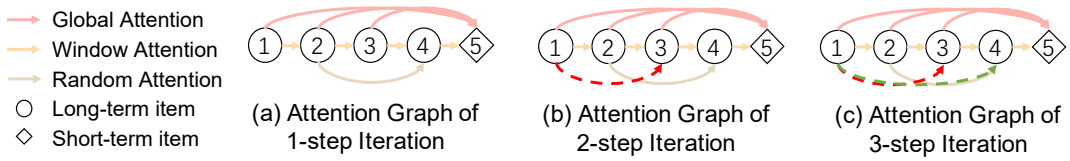

Figure 3: Diagram of Iterative Attention Layer at graph perspective: (a) the graph of 1-step iteration refers to the initial sparse attention graph; (b) the graph of 2-step iteration extends an edge from node 1 to node 3; (c) the graph of 3-step iteration extends an edge from node 1 to node 4 and approximate to full-attention.

score, which is used to weigh and sum the value matrix $\mathbf{V}^s$, as:

$$\mathbf{z}^s = \text{softmax}\left(\frac{\mathbf{Q}^s \mathbf{K}^{s\top}}{\sqrt{d}}\right)\mathbf{V}^s, \quad (6)$$
$$\mathbf{Q}^s = \mathbf{M}_{i_{t+1}}\mathbf{W}_Q^s, \mathbf{K}^s = \mathbf{E}^s\mathbf{W}_K^s, \mathbf{V}^s = \mathbf{E}^s\mathbf{W}_V^s,$$

where $\mathbf{W}_Q^s, \mathbf{W}_K^s, \mathbf{W}_V^s \in \mathbb{R}^{D\times D}$ are parameters to be learned. Here $\mathbf{M}_{i_{t+1}}$ is the target item embedding, and $\mathbf{E}^s$ is the historical short-term item embedding matrix. We aim to project the target item into query space with matrix $\mathbf{Q}^s$, and historical items into key and value spaces with matrices $\mathbf{W}_K^s$ and $\mathbf{W}_V^s$ here. Most importantly, $\mathbf{z}^s$ is the output of target attention and the short-term representation of our ISA. As there is one target item, it is at only $O(T)$ linear complexity, which is also the reason why the target attention is widely adopted for recommendation (Zhou et al. 2018; Pi et al. 2020; Lin et al. 2022a; Cao et al. 2022).

Interest Evolution To evolve the long-term and short-term interests, we fuse the long-term representation, short-term representation, and target item embedding into the prediction tower. Before feeding different representations into the final prediction towers, we first aggregate long-term representation by the sum pooling as: $\mathbf{z} = \sum_{t=1}^{T}\mathbf{Z}_t$, which is then finally fed into the prediction tower with the short-term representation and target item embeddings as:

$$logit_{u,t+1} = \mathbf{MLP}\left(\mathbf{z}^s \mid \mathbf{z} \mid \mathbf{M}_{i_{t+1}} \mid \mathbf{z} \odot \mathbf{M}_{i_{t+1}}\right), \quad (7)$$

where $logit_{u,t}$ is the predicted logit for user u on time step t, aiming to capture the long-term and short-term interests.

Optimization In the optimization step, we tend to optimize the next item prediction logloss (Zhou et al. 2018, 2019) based on the output of the prediction tower as:

$$-\frac{1}{|\mathcal{R}|}\sum_{(u,i_t)\in\mathcal{R}}(y_{u,t}\log\hat{y}_{u,t} + (1-y_{u,t})\log(1-\hat{y}_{u,t})) + \lambda \mid \Theta \mid, \quad (8)$$

where $\hat{y}_{u,t} = \sigma(logit_{u,t})$ is the predicted probability and $\mathcal{R}$ is the training set. Here Θ is the model parameter to be learned, and λ is the hyper-parameter for regularization.

Experiments

In this section, we experiment on two real-world datasets and explore the answers to the research questions (RQs): **RQ1**: How does the proposed ISA outperform state-of-the-art sequential recommendation models? **RQ2**: What is the impact of our sparse attention components? Are the high-level components, Sparse Attention Layer and Iterative Attention Layer, effective? **RQ3**: Does the proposed ISA still outperform state-of-the-art sequential models when varying sequence lengths?

Dataset	#Users	#Items	#Records	Mean Length
Taobao	37,293	64,996	1,505,878	40.38
Short Video	37,497	128,758	6,413,396	171.04

Table 2: Data statistics after 10-core setting filtering.

Experimental Settings

Datasets and Metrics As for datasets, we conduct evaluations of recommendation performance on two large-scale datasets. The data statistics after 10-core filtering are as Table 2, where mean length is the average of sequence length for users.

In evaluation metrics, we follow existing work (Gunawardana and Shani 2015) and adopt two popular accuracy metrics, AUC and GAUC. Besides, MRR and NDCG@10 are used for ranking evaluation (Chang et al. 2021).

Baselines The existing sequential recommendation models for long behaviors are based on target attention (Pi et al. 2020; Lin et al. 2022a) that aggregates historical items via target item. But there are also some historical attention models (Kang and McAuley 2018; Hidasi et al. 2016) that further consider the relation between historical items, which may be ignored for the efficient problem. Thus we compare our ISA with two types of competitive baselines: (1) Target Attention models like DIN (Zhou et al. 2018), SIM (Pi et al. 2020), ETA (Chen et al. 2021), UBR4CTR (Qin et al. 2020), SAM (Lin et al. 2022a), and SDIM (Cao et al. 2022); (2) Historical Attention models like SASRec (Kang and McAuley 2018), Long4Rec and Bigbird4Rec. Note that Long4Rec and Bigbird4Rec are adapted from Longformer (Beltagy, Peters, and Cohan 2020) and Bigbird (Zaheer et al. 2020) of NLP fields by us.

Overall performance (RQ1)

- **Our ISA performs best.** Our proposed method outperforms other baselines in the two datasets under four metrics. For the most important metric, GAUC, our ISA outperforms other baselines by 1.96% and 0.56%, respectively, in

Type		Target Attention						Historical Attention			
Model		DIN	SIM	ETA	UBR4CTR	SAM	SDIM	SASRec	Long4Rec	Bigbird4Rec	Ours
Taobao	AUC	0.6898	0.8574	0.8691	0.6105	0.6787	0.8548	0.8544	0.8601	0.8655	**0.8849**
	MRR	0.2977	0.3683	0.4064	0.2164	0.3215	0.3418	0.3645	0.3775	0.3799	**0.4314**
	NDCG	0.2246	0.3053	0.3492	0.1471	0.2487	0.2783	0.3035	0.3167	0.3201	**0.3751**
	GAUC	0.8429	0.8639	0.8678	0.7750	0.8525	0.8514	0.8576	0.8642	0.8653	**0.8848**
Short Video	AUC	0.5841	0.7489	0.7490	0.5889	0.5884	0.7747	0.7961	0.8014	0.8001	**0.8079**
	MRR	0.8833	0.8791	0.8716	0.8737	0.8830	0.8882	0.9063	0.9052	0.9049	**0.9088**
	NDCG	0.9138	0.9108	0.9053	0.9068	0.9136	0.9175	0.9308	0.9292	0.9290	**0.9327**
	GAUC	0.7618	0.7542	0.7438	0.7618	0.7771	0.7727	0.8079	0.8102	0.8098	**0.8147**

Table 3: Overall performance comparison for ISA against baselines under Taobao and Short Video datasets on four metrics. Here Bold is the highest result, and underline is the second highest result.

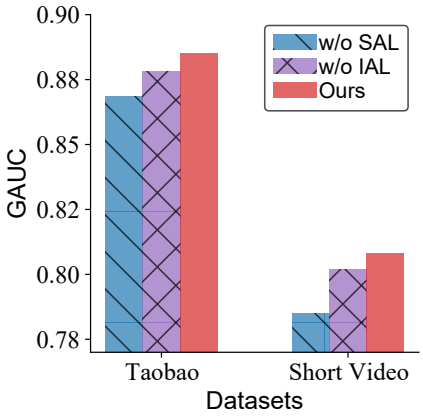

Figure 4: Ablation study of our proposed components, where SAL means Sparse Attention Layer and IAL means Iterative Attention Layer.

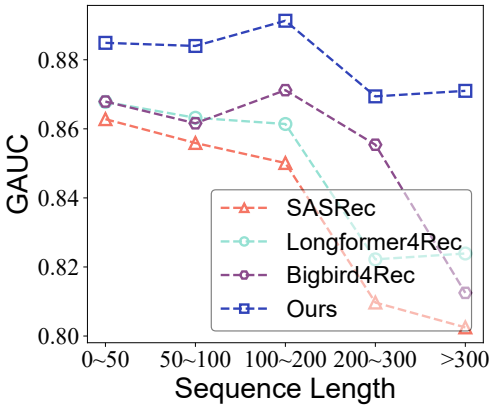

Figure 5: GAUC comparisons under different sequence lengths on the Taobao dataset.

the Taobao dataset and Short Video dataset. It is worth mentioning that existing works (Zhou et al. 2018; Pi et al. 2020; Cao et al. 2022) often claim 0.5% as a significant improvement for the GAUC metric. Moreover, Taobao dataset with more sparse behaviors sees a sharper improvement, which illustrates diffusing the sparse attention graph is beneficial to the performance of sequential recommendation.

- **Modeling the historical relationship is essential.** From the results, the models with historical attention generally outperform the target attention models, which means it is essential to capture the historical relation. However, as the historical items are very long for existing popular recommender systems, the quadratic cost of a self-attention based model like SASRec prevents it from real-time recommendation service. Besides, the full attention of SASRec may be redundant for sequential recommendation with sparse behaviors, which will result in overfitting actually. Thus, the efficient and overfitting problems facilitate us to adapt the sparse attention method from NLP.

Ablation Study (RQ2)

In this part, we tend to study the impact of two attention layers as Figure 4. From the ablation experiments, we can observe that:

- **Sparse Attention Layer is fundamental**. From Figure 4, removing SAL shows the most performance drops, which is because this component is also fundamental to IAL. Specifically, removing SAL means it is impossible to further apply IAL to diffuse the sparse. At the same time, removing SAL will make it fail to extract long-term relations for historical items and degenerate to even as poor as some decent target attention models on the two datasets.
- **Iterative Attention Layer further improves Sparse Attentions**. As shown in Figure 4, indeed, removing IAL refers to the model with solely sparse attention for long-term modeling. The observation that the model without IAL outperforms the model without SAL but underperforms our ISA illustrates sparse attention is effective but can be further improved by our proposed iteration process. Thus it is essential to extend the short attention distance by iteration.

Study on Sequence Length (RQ3)

For most popular personalized applications, there may be users surfing many items but also users with seldom usage. Subsequently, these two types of users will result in dramatically different lengths of historical item sequences, long or short, respectively. There is more learning bonus in a long sequence, but however, the noise and gradient vanishing

problems will also arise. That is to say, the increase in sequence length may be like a coin with two sides, bringing us both learning bonuses and challenges. Unlike long sequence, the short sequence is with limited learning signal and often results in overfitting. To further study the performance on various sequence lengths, we investigate the GAUC metric at five length categories. The results in the Taobao dataset are as illustrated in Figure 5, where we can have the following observations.

- **ISA performs the best throughout various lengths**. ISA achieves the best performance under these five sequence length categories in both the Taobao dataset. In the Taobao dataset, it is obvious that there is always a significant performance gap between ISA and other methods. When the sequence gets longer, the gap becomes more significant. This may result from the extension of the iterative process.
- **Full attention is redundant**. In the Taobao dataset, sparse attention methods like Long4Rec and Bigbird4Rec even outperform full attention methods like SASRec, which may be because full attention is redundant for Taobao dataset with shorter and sparser behaviors, resulting in overfitting.

Related Work

Sequential Recommendation In sequential recommendation, early approaches primarily centered on utilizing Markov Chain-based techniques to model the temporal dependencies within sequences (He and McAuley 2016). In recent years, there has been a shift towards using deep learning techniques. Recurrent neural networks (RNNs) have been employed to model the temporal dependencies within sequences (Hidasi et al. 2016; Zhou et al. 2019). Convolutional neural networks (CNNs) have also been used to capture skip connections in sequences (Tang and Wang 2018). Attention mechanisms have also gained traction in the field of sequential recommendation, like natural language processing (Vaswani et al. 2017; Devlin et al. 2019) and computer vision (Dosovitskiy et al. 2020; Liu et al. 2021). In the context of the click-through rate (CTR) prediction tasks, DIN (Zhou et al. 2018) proposed the utilization of target attention to calculate the importance of different historical behaviors given a candidate item. When dealing with long-term user behavior sequences in the CTR prediction task, SIM (Pi et al. 2020) proposed the utilization of hard search or soft search techniques with the target item as the query to filter a subsequence, which can also be seen as a type of coarse target attention mechanism. SDIM (Cao et al. 2022) proposed a simhash-based sampling method to approximate the target attention on the entire long sequence of historical behaviors. However, utilizing only target attention fails to capture the interactions and relationships between historical behaviors. In the context of the next item prediction task, SASRec (Kang and McAuley 2018) exploited the self-attention with a causality mask to capture the historical context. Furthermore, combining self-attention and target attention can improve performance (Lin et al. 2022b).

Efficient Transformer and Self-Attention Although self-attention has demonstrated effectiveness in various tasks, its computational and memory requirements grow quadratically with sequence length, rendering its application to longer sequences infeasible or excessively computationally costly (Tay et al. 2023). In response to the aforementioned challenges, researchers have proposed different approaches (Rae et al. 2020; Choromanski et al. 2021) to reduce the computational complexity of transformer models while maintaining their performance. For instance, Linformer (Wang et al. 2020) employs a low-rank decomposition of the attention matrix, which reduces the computational time complexity to linear. Additionally, one common technique is the use of sparse attention, which selectively attends to a subset of the input elements rather than the entire sequence. Longformer (Beltagy, Peters, and Cohan 2020) proposed to combine local window attention and global attention to provide different attention patterns while reducing the computational cost of self-attention. Similarly, BigBird (Zaheer et al. 2020) improves upon Longformer by introducing an efficient sparse attention mechanism that further considers random attention. Nonetheless, the aforementioned low-rankness and sparsity priors of self-attention may not always hold (Tay et al. 2023), resulting in suboptimal performance in some cases.

Iterative Diffusion Diffusion processes, a class of Markov stochastic processes with continuous trajectories, are widely applied in various fields (Ikeda and Watanabe 2014). In natural language processing, graph diffusion methods efficiently utilize multi-hop neighbors in graph-structured data while maintaining computational efficiency (Atwood and Towsley 2016). PPNP (Klicpera, Bojchevski, and Günnemann 2019) firstly use a GNN to make node-level predictions and then leverage PPR to propagate these predictions through the graph. The graph convolution diffusion model (GDC) (Klicpera, Weißenberger, and Günnemann 2019) is also proposed to incorporate the general diffusion method. Recently, MAGNA (Wang et al. 2021) proposed to incorporate multi-hop context information into attention calculation, thus increasing the receptive field of each layer of GNN.

Conclusion

In this work, we approach the problem of long-term sequential recommendation that is highly related to the industrial personalized service and propose a method named Iterative Sparse Attention (ISA) to achieve long-short-term modeling. ISA first leverages Sparse Attention Layer (SAL) to reduce the cost and address the efficient challenge, which consists of global, windows, and random attentions. Then, the Iterative Attention Layer (IAL) expands the receptive field of the proposed sparse attentions to address the interest drifting challenge and further approximate to full attention effectively. Our method is effective for long-sequence recommendation and can efficiently approximate full attention. As for future work, we plan to verify the performance when rich attributes of users/items are available.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants 62272262 and 72342032. This work is also supported by National Key Research and Development Program of China under Grant 2022YFB3104702.

References

Atwood, J.; and Towsley, D. 2016. Diffusion-Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 1993–2001.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Cao, Y.; Zhou, X.; Feng, J.; Huang, P.; Xiao, Y.; Chen, D.; and Chen, S. 2022. Sampling Is All You Need on Modeling Long-Term User Behaviors for CTR Prediction. *arXiv preprint arXiv:2205.10249*.

Chang, J.; Gao, C.; Zheng, Y.; Hui, Y.; Niu, Y.; Song, Y.; Jin, D.; and Li, Y. 2021. Sequential Recommendation with Graph Neural Networks. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 378–387.

Chen, Q.; Pei, C.; Lv, S.; Li, C.; Ge, J.; and Ou, W. 2021. End-to-End User Behavior Retrieval in Click-Through RatePrediction Model. *arXiv preprint arXiv:2108.04468*.

Child, R.; Gray, S.; Radford, A.; and Sutskever, I. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Choromanski, K. M.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlós, T.; Hawkins, P.; Davis, J. Q.; Mohiuddin, A.; Kaiser, L.; Belanger, D. B.; Colwell, L. J.; and Weller, A. 2021. Rethinking Attention with Performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q. V.; and Salakhutdinov, R. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Erdős, P.; Rényi, A.; et al. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1): 17–60.

Feng, A.; Li, I.; Jiang, Y.; and Ying, R. 2022a. Diffuser: Efficient Transformers with Multi-hop Attention Diffusion for Long Sequences. *arXiv preprint arXiv:2210.11794*.

Feng, A.; Li, I.; Jiang, Y.; and Ying, R. 2022b. Diffuser: Efficient Transformers with Multi-hop Attention Diffusion for Long Sequences. *arXiv preprint arXiv:2210.11794*.

Friedman, J. 2008. *A proof of Alon's second eigenvalue conjecture and related problems*. American Mathematical Soc.

Gunawardana, A.; and Shani, G. 2015. Evaluating Recommender Systems. In Ricci, F.; Rokach, L.; and Shapira, B., eds., *Recommender Systems Handbook*, 265–308. Springer US. ISBN 978-1-4899-7637-6.

He, R.; and McAuley, J. 2016. Fusing Similarity Models with Markov Chains for Sparse Sequential Recommendation. In *Proceedings of the 16th IEEE International Conf. on Data Mining*, 191–200.

Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; and Tikk, D. 2016. Session-based recommendations with recurrent neural networks. In *ICLR*.

Ikeda, N.; and Watanabe, S. 2014. *Stochastic differential equations and diffusion processes*. Elsevier.

Kang, W.-C.; and McAuley, J. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, 197–206. IEEE.

Klicpera, J.; Bojchevski, A.; and Günnemann, S. 2019. Predict then Propagate: Graph Neural Networks meet Personalized PageRank. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Klicpera, J.; Weißenberger, S.; and Günnemann, S. 2019. Diffusion Improves Graph Learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 13333–13345.

Lin, Q.; Zhou, W.-J.; Wang, Y.; Da, Q.; Chen, Q.-G.; and Wang, B. 2022a. Sparse Attentive Memory Network for Click-through Rate Prediction with Long Sequences. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3312–3321.

Lin, Q.; Zhou, W.-J.; Wang, Y.; Da, Q.; Chen, Q.-G.; and Wang, B. 2022b. Sparse Attentive Memory Network for Click-through Rate Prediction with Long Sequences. In *CIKM*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pal, A.; Eksombatchai, C.; Zhou, Y.; Zhao, B.; Rosenberg, C.; and Leskovec, J. 2020. Pinnersage: Multi-modal user embedding framework for recommendations at pinterest. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2311–2320.

Pi, Q.; Zhou, G.; Zhang, Y.; Wang, Z.; Ren, L.; Fan, Y.; Zhu, X.; and Gai, K. 2020. Search-based user interest modeling with lifelong sequential behavior data for click-through rate prediction. In *Proceedings of the 29th ACM International*

Conference on Information & Knowledge Management, 2685–2692.

Qiao, S.; Gao, C.; Li, Y.; and Yin, H. 2024. LLM-assisted Explicit and Implicit Multi-interest Learning Framework for Sequential Recommendation. *arXiv preprint arXiv:2411.09410*.

Qin, J.; Zhang, W.; Wu, X.; Jin, J.; Fang, Y.; and Yu, Y. 2020. User behavior retrieval for click-through rate prediction. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2347–2356.

Rae, J. W.; Potapenko, A.; Jayakumar, S. M.; Hillier, C.; and Lillicrap, T. P. 2020. Compressive Transformers for Long-Range Sequence Modelling. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tang, J.; and Wang, K. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *WWW*, 565–573.

Tay, Y.; Dehghani, M.; Bahri, D.; and Metzler, D. 2023. Efficient Transformers: A Survey. *ACM Comput. Surv.*, 55(6): 109:1–109:28.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Wang, G.; Ying, R.; Huang, J.; and Leskovec, J. 2021. Multi-hop Attention Graph Neural Networks. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, 3089–3096. ijcai.org.

Wang, S.; Li, B. Z.; Khabsa, M.; Fang, H.; and Ma, H. 2020. Linformer: Self-Attention with Linear Complexity. *CoRR*, abs/2006.04768.

Zaheer, M.; Guruganesh, G.; Dubey, K. A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33: 17283–17297.

Zhou, G.; Mou, N.; Fan, Y.; Pi, Q.; Bian, W.; Zhou, C.; Zhu, X.; and Gai, K. 2019. Deep interest evolution network for click-through rate prediction. In *AAAI*, 5941–5948.

Zhou, G.; Zhu, X.; Song, C.; Fan, Y.; Zhu, H.; Ma, X.; Yan, Y.; Jin, J.; Li, H.; and Gai, K. 2018. Deep interest network for click-through rate prediction. In *KDD*, 1059–1068.