DIF-SR

Decoupled Side Information Fusion for Sequential Recommendation

김태공 | 12.10.2022

Contents

1. Introduction

2022 SIGIR conference Accept

Decoupled Side Information Fusion for Sequential Recommendation

Yueqi Xie*
HKUST
yxieay@connect.ust.hk

Peilin Zhou*
Upstage
zhoupalin@gmail.com

Sunghun Kim
HKUST
hunkim@ust.hk

ABSTRACT

Side information fusion for sequential recommendation (SR) aims to effectively leverage various side information to enhance the performance of next-item prediction. Most state-of-the-art methods build on self-attention networks and focus on exploring various solutions to integrate the item embedding and side information embeddings before the attention layer. However, our analysis shows that the early integration of various types of embeddings limits the expressiveness of attention matrices due to a *rank bottleneck* and constrains the flexibility of gradients. Also, it involves mixed correlations among the different heterogeneous information resources, which brings extra disturbance to attention calculation. Motivated by this, we propose **D**ecoupled Side **I**nformation **F**usion for **S**equential **R**ecommendation (DIF-SR), which moves the side information from the input to the attention layer and decouples the attention calculation of various side information and item representation. We theoretically and empirically show that the proposed solution allows higher-rank attention matrices and flexible gradients to enhance the modeling capacity of side information fusion. Also, auxiliary attribute predictors are proposed to further activate the beneficial interaction between side information and item representation learning. Extensive experiments on four real-world datasets demonstrate that our proposed solution stably outperforms state-of-the-art SR models. Further studies show that our proposed solution can be readily incorporated into current attention-based SR models and significantly boost performance. Our source code is available at https://github.com/AIM-SE/DIF-SR.

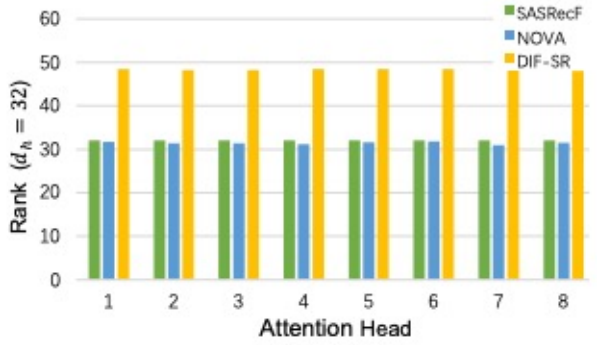


Figure 1: Rank of attention matrices: A comparison of the average rank of attention score matrices of early-integrated embedding based solutions, i.e., SASRecF and NOVA, and our proposed DIF-SR. The early-integration of embeddings leads to lower rank of the attention matrices and limits the expressiveness.

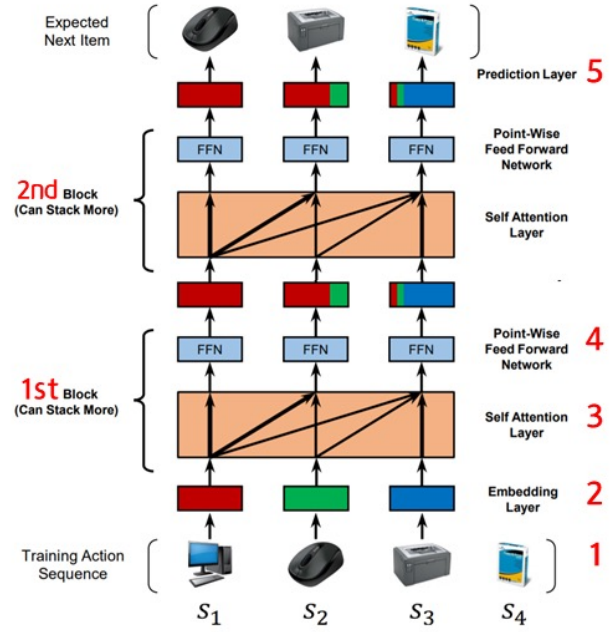
International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3531963

1 INTRODUCTION

2. Background

- Recommendation System
  - Ex) SASRec, BERT4Rec
  - Basic Suppose: buying behavior in the **Session** -> closely related to each other
  - Input Data: click item sequence

  - Output Data: shift of input data



3. History

- Question List

  Q1. what is **side information**??

  Q2. What is the effect of **side information**??

  Q3. How do add **side information**??

3. History

- Define side information (Paper)
  - additional information about the item
  - Item related
    - Brand, category, price
  - behavior-related
    - Position, rating
- effect of side information
  - Use more information to understand your purchase intentions!
  - Related Features Provide Additional Information
  - Recommended performance improvement as a **constraint**

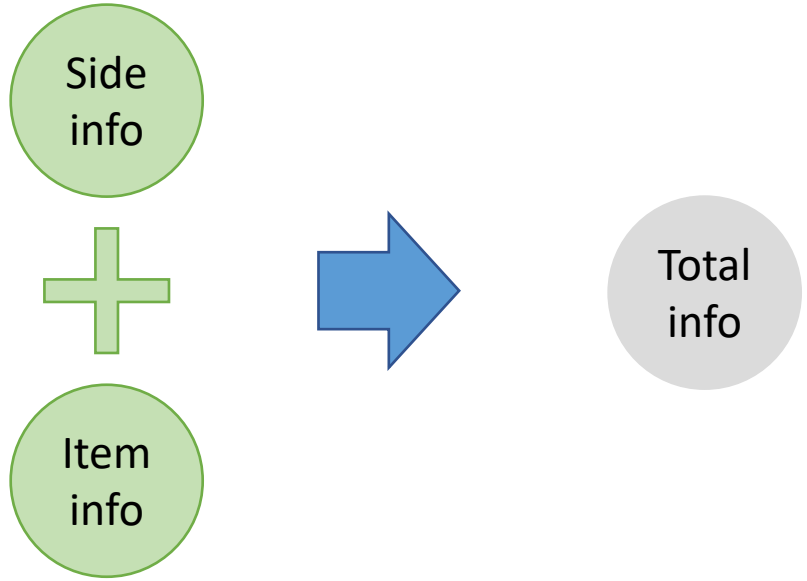
[WHY Side – information]

Session 에서 구매자는 무엇을 찾고 있을까?

Fig 1: Example Session and Purchase Data

3. History

- How do add side information??
  - Fusion: side information + input sequence information
  - Categorized into 3 categories
    1. Early fusion
       - ICAI-SR, NOVA
    2. Final stage fusion
       - FDSA, S3-Rec
    3. **Decoupled fusion**
       - **DIF-SR**



3. History

- Early fusion
  - ICAI-SR, NOVA
  - Fusion in model **input step**!
  - Weakness
    - Rank bottleneck
    - Share the same gradient

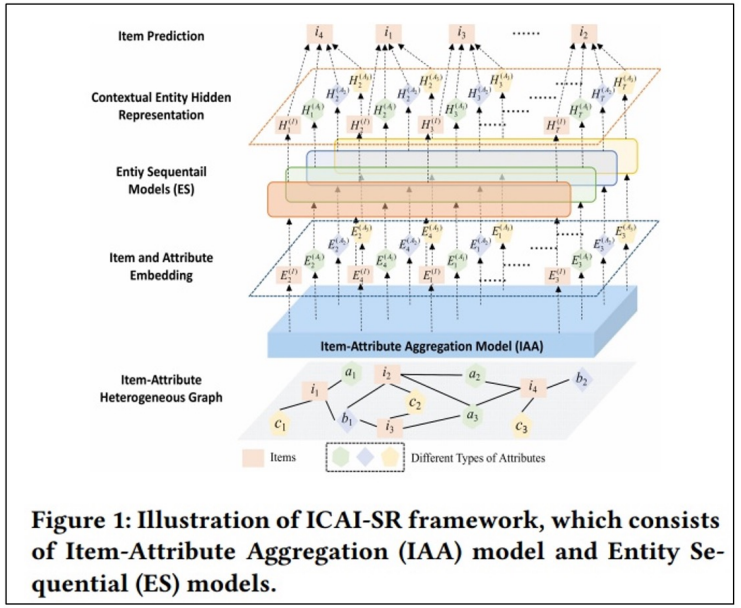
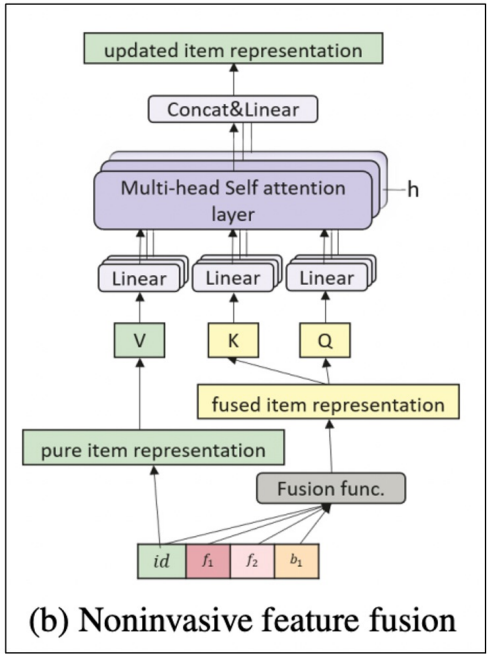


Figure 1: Illustration of ICAI-SR framework, which consists of Item-Attribute Aggregation (IAA) model and Entity Sequential (ES) models.

<2021. ICAI-SR>



(b) Noninvasive feature fusion

<2021. NOVA>

3. History

- Final stage fusion
  - FDSA, S3-Rec
  - Fusion in model **output step**!
  - Weakness
    - Reflect directly on item sequence information X

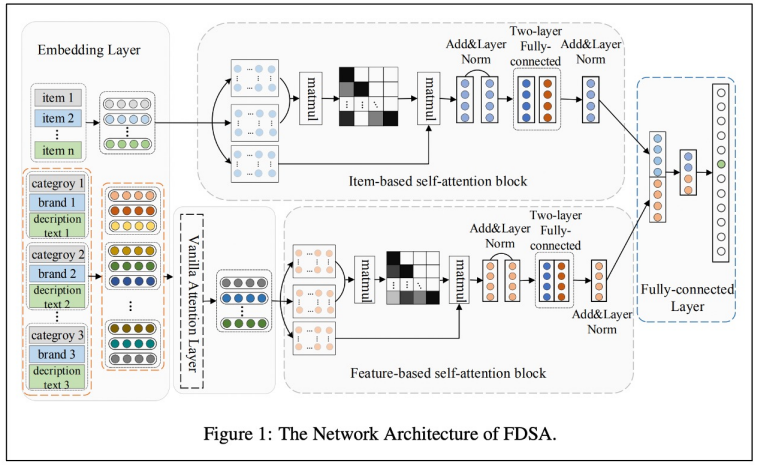


Figure 1: The Network Architecture of FDSA.

<2019. FDSA>

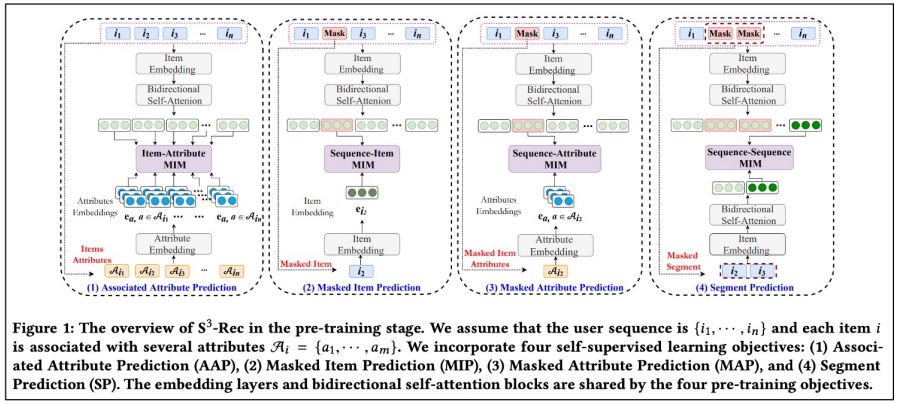


Figure 1: The overview of S^3-Rec in the pre-training stage. We assume that the user sequence is $\{i_1, \cdots, i_n\}$ and each item i is associated with several attributes $\mathcal{A}_i = \{a_1, \cdots, a_m\}$. We incorporate four self-supervised learning objectives: (1) Associated Attribute Prediction (AAP), (2) Masked Item Prediction (MIP), (3) Masked Attribute Prediction (MAP), and (4) Segment Prediction (SP). The embedding layers and bidirectional self-attention blocks are shared by the four pre-training objectives.

<2020. S3-Rec>

3. History

- Decoupled fusion
  - DIF-SR
  - Fusion in model **middle step(attention)!!**
  - advantage
    - **Rank bottleneck** clear!
      - Reflect directly on item sequence information
      - Enables **independent** gradient learning

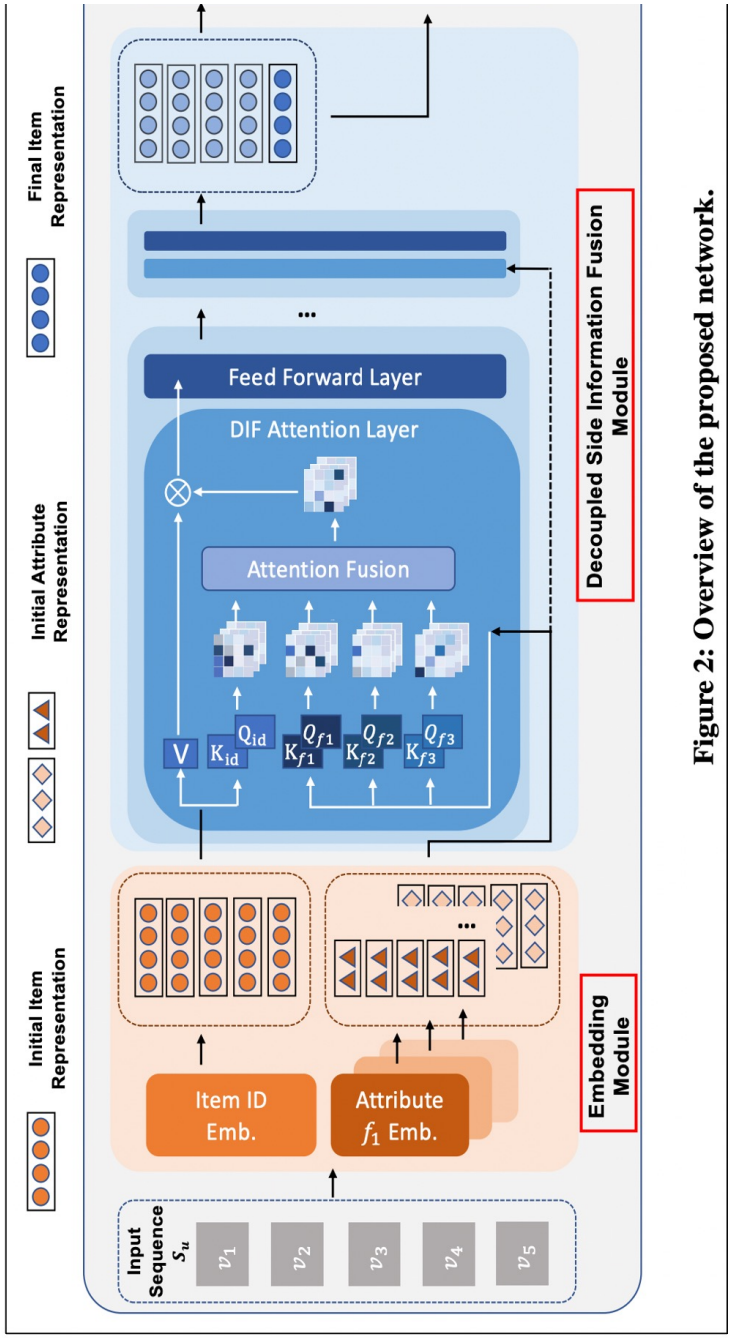


Figure 2: Overview of the proposed network.

4. DIF-SR

- DIF-SR Model Overview
  1. Embedding Module
  2. Decoupled Side Information Fusion Module: Fusion in model middle step(attention)
  3. Prediction Module: result prediction (next item, side information)

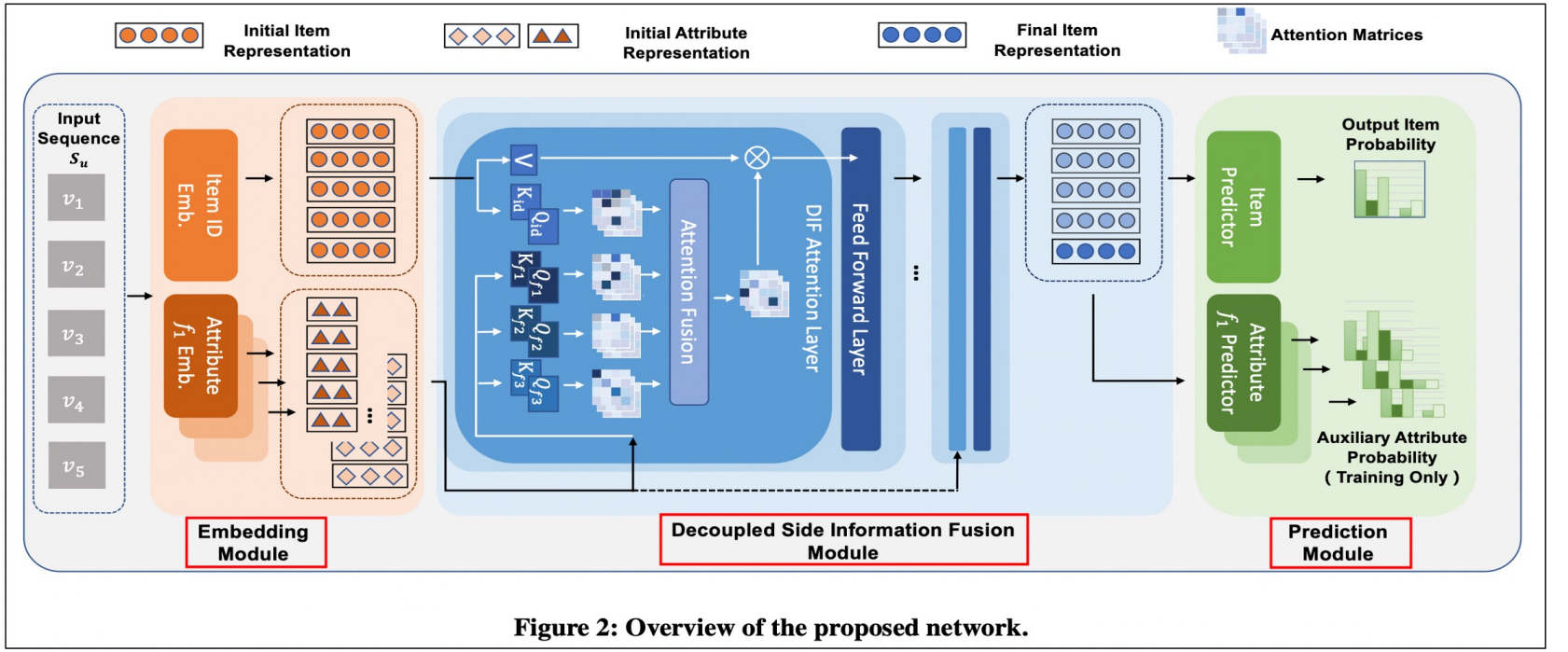
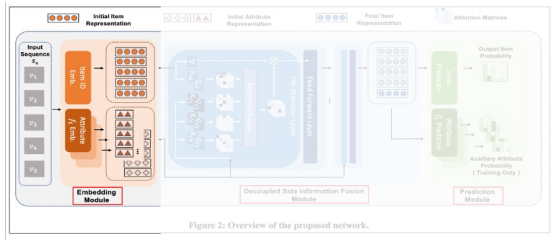
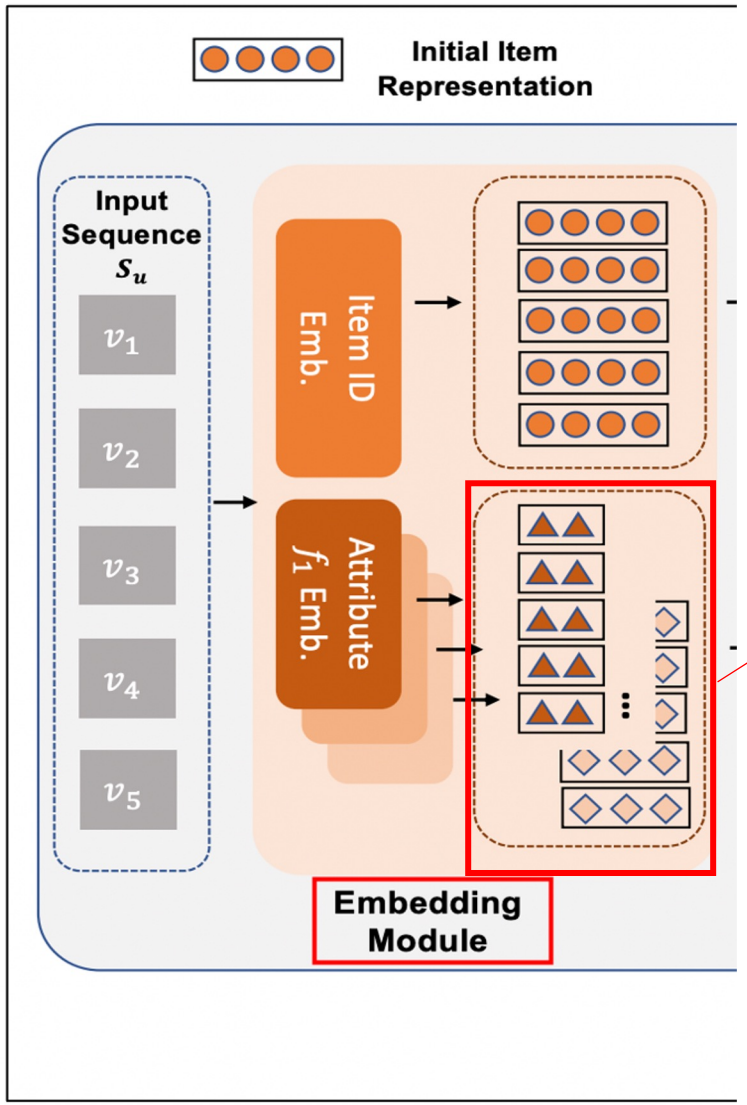


Figure 2: Overview of the proposed network.

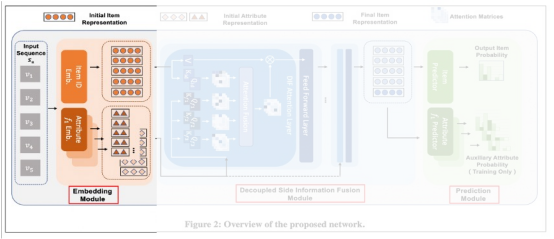
4. DIF-SR





1. Embedding Module
   - Embedding: Look-up in Embedding Table
   - Item Embedding: $N \times d$
   - Attribute k: $N \times d_k$
   - embedding dimension is not important (embedding size is not the same)
   - Item embedding size small: memory, performance ++

4. DIF-SR



1. Embedding Module
   - Embedding: Look-up in Embedding Table
   - Item Embedding: $N \times d$
   - Attribute k: $N \times d_k$
   - embedding dimension is not important (embedding size is not the same)
   - Item embedding size small: memory, performance ++

Input Sequence — Next Item prediction Loss

Items	Item 5	Item 2	Item 3	Item 2	Item 10	Item 4	Val Item	Test Item
시점(t)	1	2	3	4	···	t-2	t-1	t
Attribute 1	A1	A1	A2	A1	···	A4	A?	A?
Attribute 2	B2	B1	B3	B3	···	B2	B?	B?

Item Emb Matrix ← Look up → **Positional Encoding**

index	Look up table
1	$\mathbb{R}^d$
2	$\mathbb{R}^d$
3	$\mathbb{R}^d$
...	...
#item num	$\mathbb{R}^d$

D — Item seq(E) (5, 2, 3, 2, ..., 4) — **Session item**

\+

D — Item seq(E) (1, 2, 3, 4, ..., max) — **Position (Learnable)**

index	Look up table
1	$\mathbb{R}^d$
2	$\mathbb{R}^d$
3	$\mathbb{R}^d$
...	...
max len	$\mathbb{R}^d$

attribute prediction Loss (only train)

Attribute1 Emb Matrix

index	Look up table
1	$\mathbb{R}^{d_1}$
2	$\mathbb{R}^{d_1}$
3	$\mathbb{R}^{d_1}$
...	...
Att1 num	$\mathbb{R}^{d_1}$

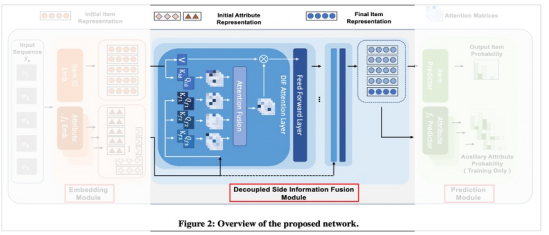
N — D_1 — Att1 seq(E) (A1, A1, A2, A1, ..., A4) — **Sequence Attribute 1**

N — D_2 — Att2 seq(E) (B2, B1, B3, B3, ..., B2) — **Sequence Attribute 2**

Attribute2 Emb Matrix

index	Look up table
1	$\mathbb{R}^{d_2}$
2	$\mathbb{R}^{d_2}$
...	...
Att2 num	$\mathbb{R}^{d_2}$

4. DIF-SR



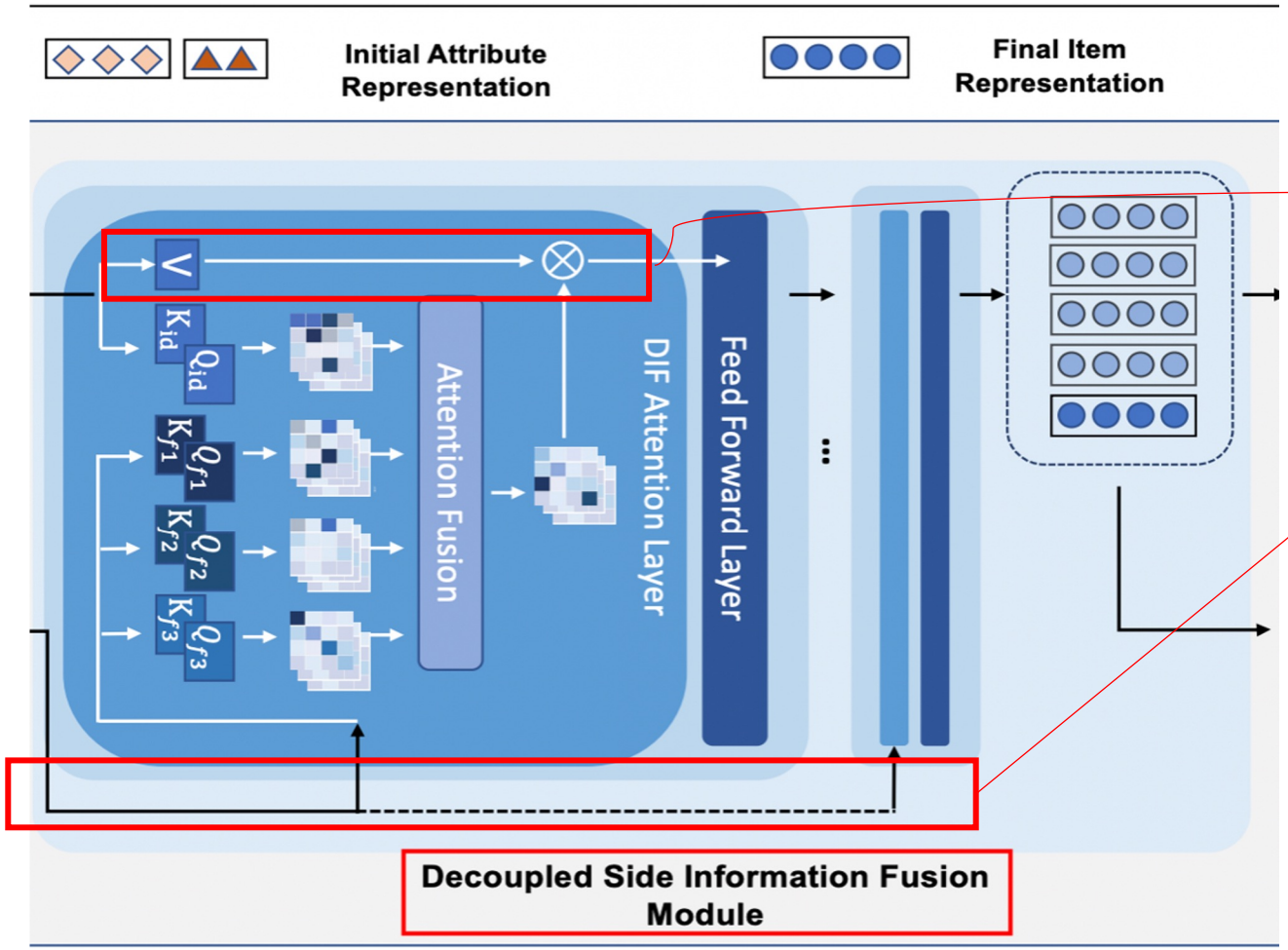


Figure 2: Overview of the proposed network.

2. Decoupled Side Information Fusion Module
   - Self Attention
   - Value * (fusion of item seq and side information)
   - Directly reflect side information on each layer (reduce computation and avoid overfitting)

4. DIF-SR

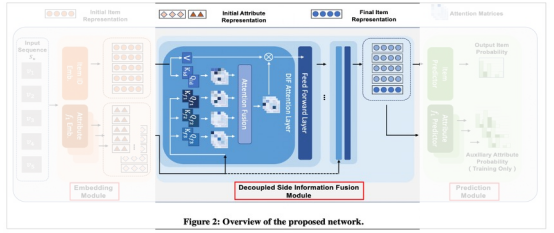
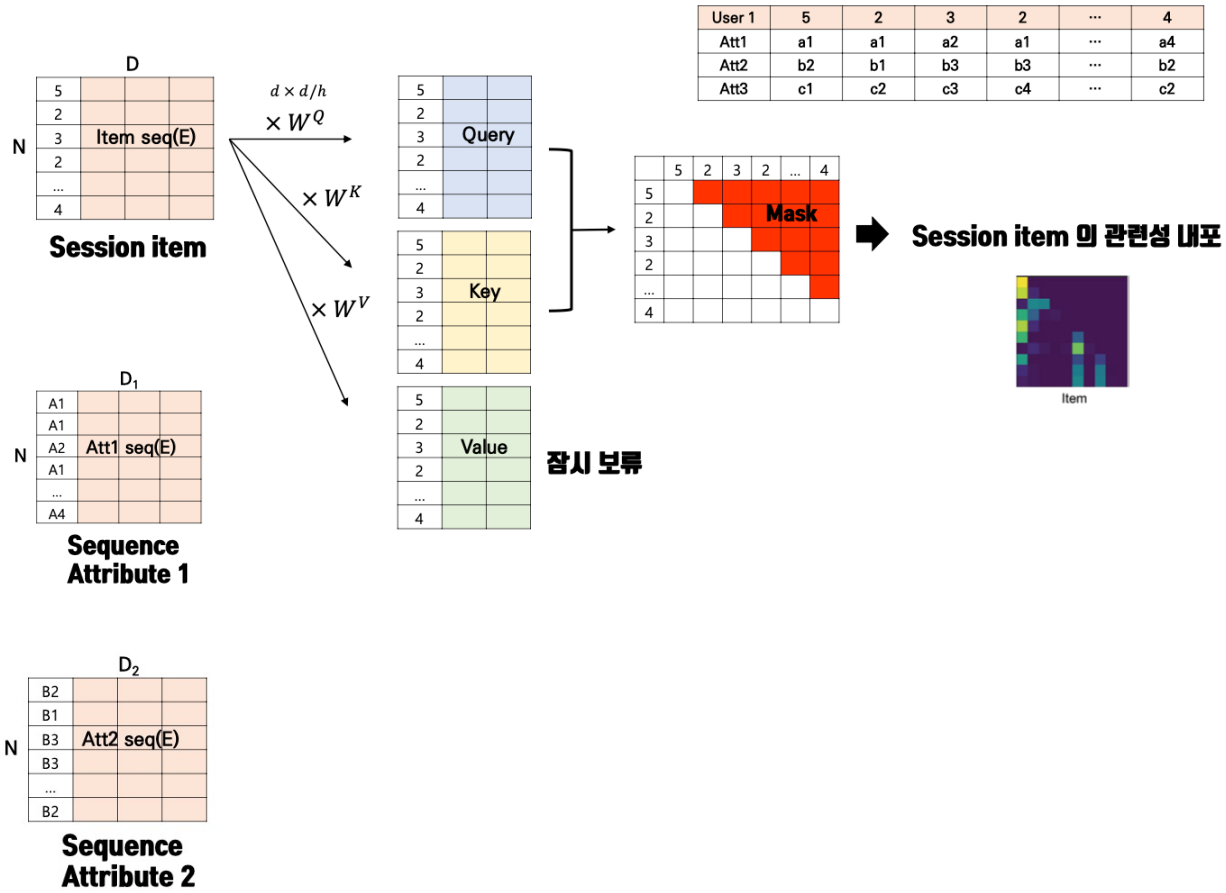


Figure 2: Overview of the proposed network.



- Decoupled Side Information Fusion (item)
  - Generate a Query, Key matrix by head count of heads
  - Query * Key
  - Masking (SASRec)
  - Value Matrix Holding

4. DIF-SR

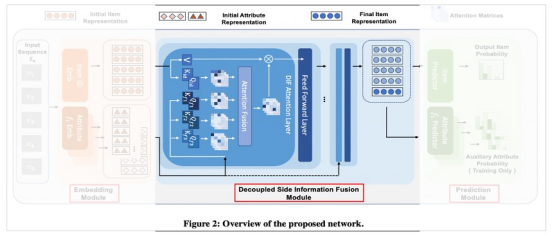
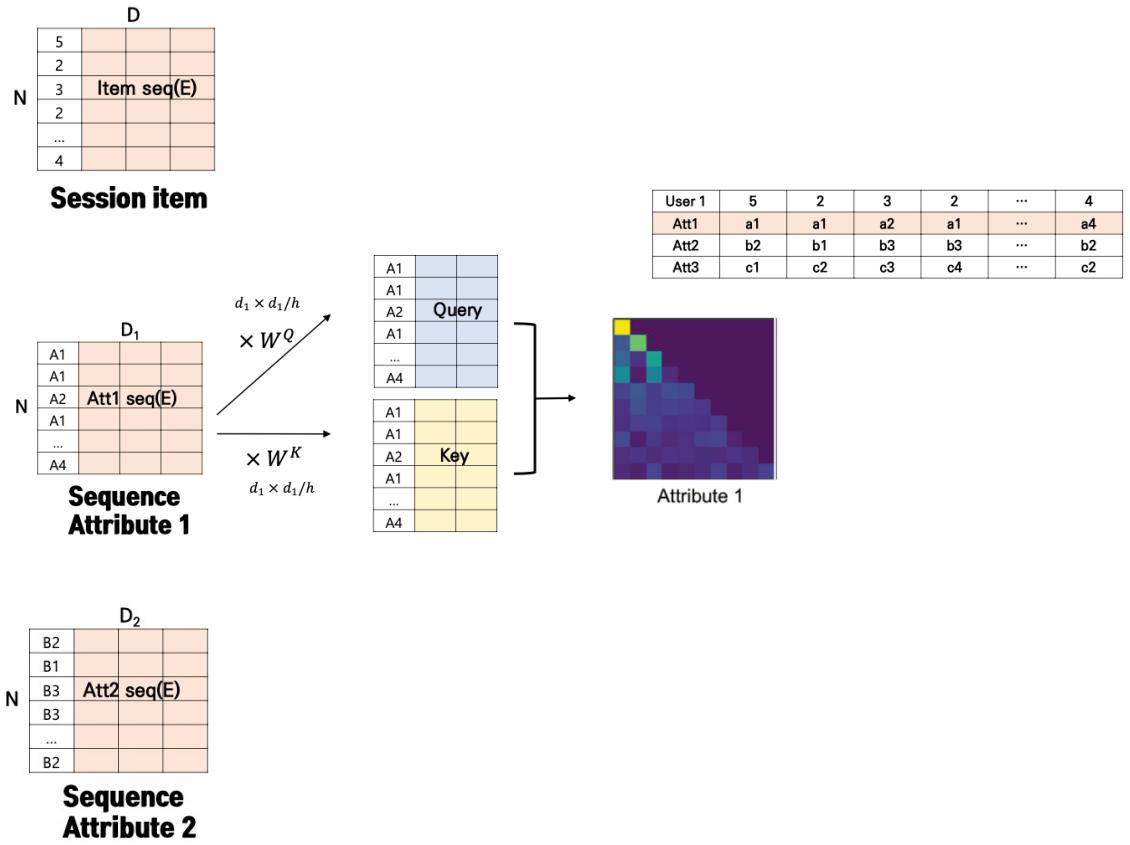


Figure 2: Overview of the proposed network.



- Decoupled Side Information Fusion (side info)
  - Generate a Query, Key matrix by head count of heads
  - Query * Key
  - Masking (SASRec)
  - embedding size is not the same $(d_h \neq d_{1h})$

4. DIF-SR

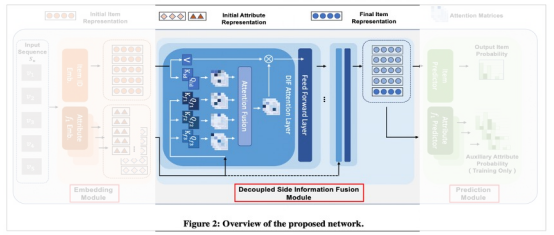
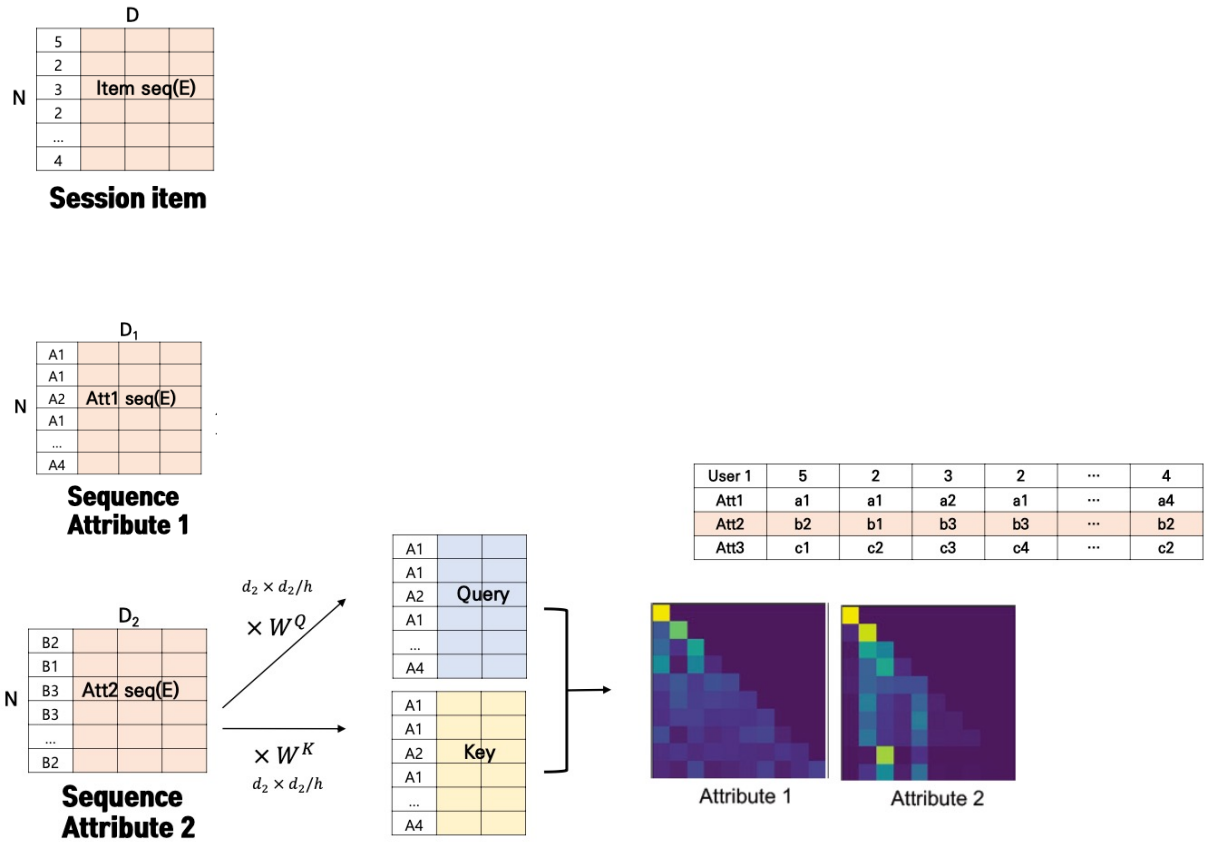


Figure 2: Overview of the proposed network.



- Decoupled Side Information Fusion (side info)
  - Generate a Query, Key matrix by head count of heads
  - Query * Key
  - Masking (SASRec)
  - embedding size is not the same $(d_h \neq d_{2h})$

4. DIF-SR

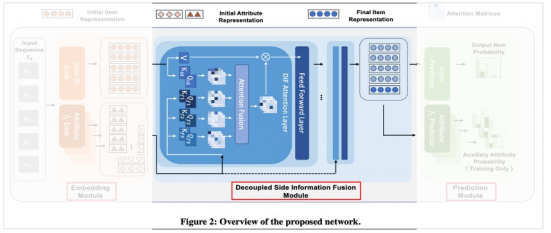
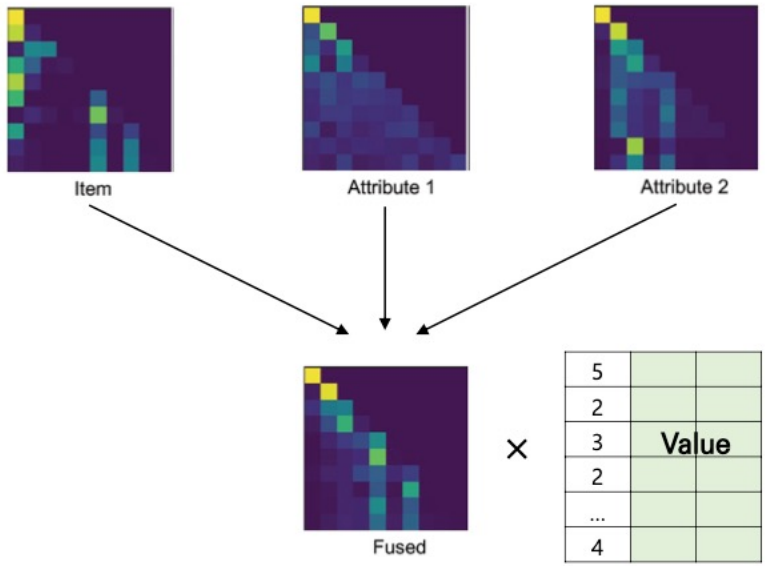


Figure 2: Overview of the proposed network.

개별적인(Decoupled) attention map 을 통한 fusion

User 1	5	2	3	2	···	4
Att1	a1	a1	a2	a1	···	a4
Att2	b2	b1	b3	b3	···	b2
Att3	c1	c2	c3	c4	···	c2



- Decoupled Side Information Fusion
  - fusion of Attention results (Fused)
    - Fusion(att(item), att(att1), att(att2), …)
  - F = Fused * Value
  - $F_1, F_2, \dots, F_h$
  - Concat($F_1, F_2, \dots, F_h$) and FFL through
- Perform as many Layer
- Directly reflect side information on each layer (reduce computation and avoid overfitting)

4. DIF-SR

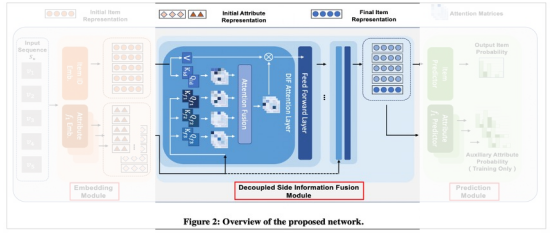


Figure 2: Overview of the proposed network.

- Fusion method
  1. sum
  2. concat
  3. gating (weighted sum)

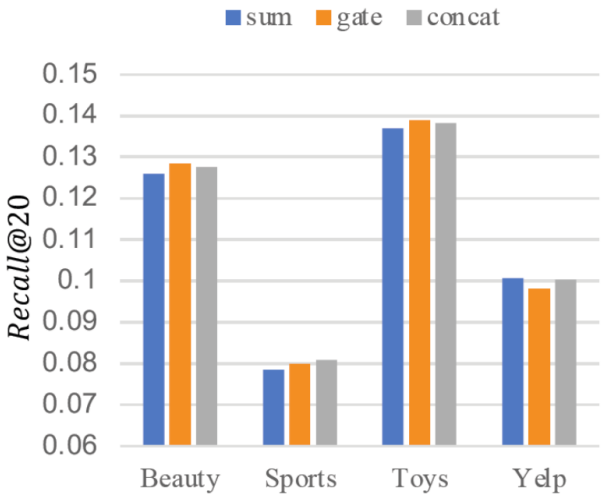
$$\mathcal{F}_{\text{add}}(f_1, \ldots, f_m) = \sum_{} f_i$$

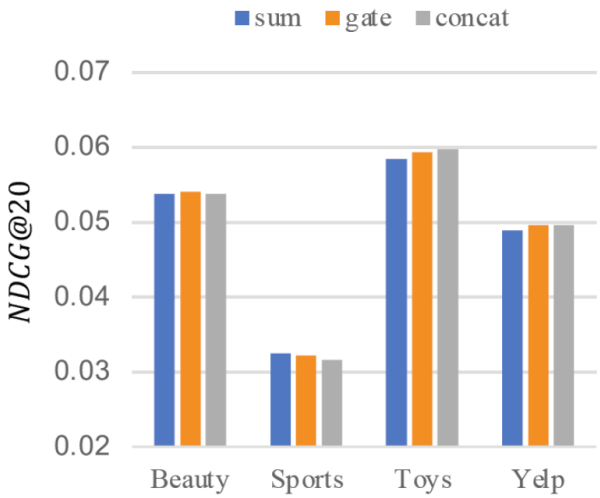
$$\mathcal{F}_{\text{concat}}(f_1, \ldots, f_m) = \boldsymbol{FC}(f_1 \odot \cdots \odot f_m)$$

$$\mathcal{F}_{\text{gating}}(f_1, \ldots, f_m) = \sum_{i=1}^{m} G^{(i)} f_i$$

$$G = \sigma(FW^F)$$

Feature 개수만큼의 차원을 가지는 Gate 생성





✓ Model 의 cost 가 증가하는 Concat / Gate 방식과 sum 을 하는 방식의 차이가 거의 없음

→ 이전 다른 방법론(NOVA) 에서도 Fusion 방식에서 따라서는 큰 차이가 없었음

→ *DIF 구조 때문에 이런 현상이 일어나는 건 아님*

4. DIF-SR

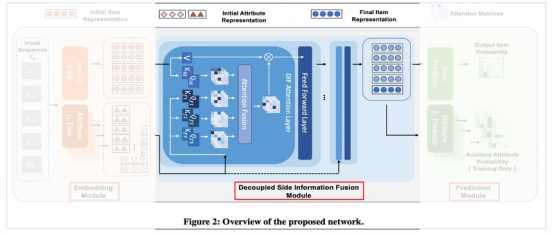


Figure 2: Overview of the proposed network.

- **Wait!!!!!**
- So Why use this Model??
  - The Problem of Early fusion (ICAI-SR, NOVA)
  - Rank bottleneck
  - Share the same gradient

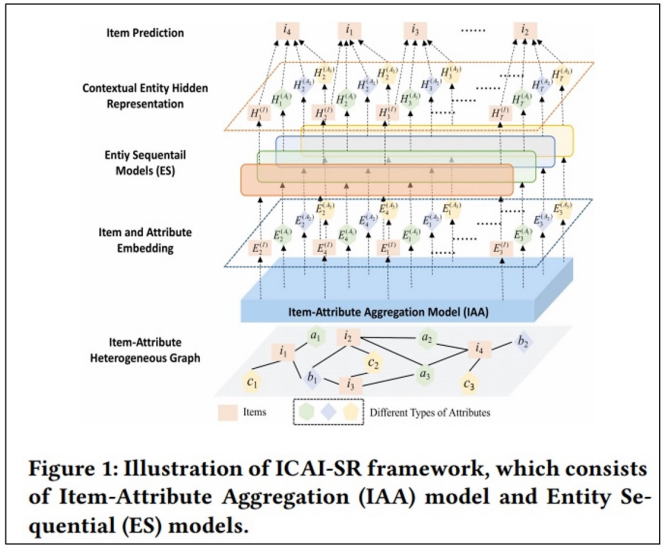
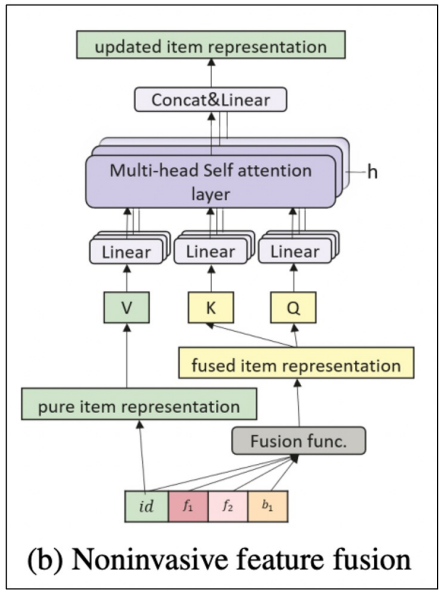


Figure 1: Illustration of ICAI-SR framework, which consists of Item-Attribute Aggregation (IAA) model and Entity Sequential (ES) models.

<2021. ICAI-SR>



(b) Noninvasive feature fusion

<2021. NOVA>

4. DIF-SR

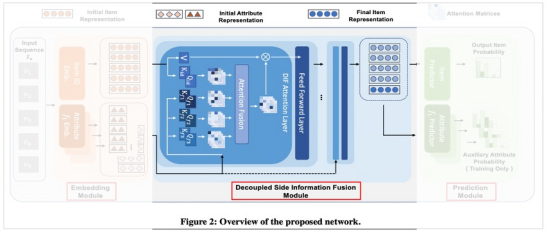


Figure 2: Overview of the proposed network.

- Rank bottleneck (Problem 1)
  - Rank: The dimension of the vector space that can be created by a column(row) of matrices.
  - Typical attention are **limited in size to the maximum rank** -> **weakening of expression**
  - DIF-attention solves this problem (reason for fusion independent attention results)

Query **Key**

일반 attention

$$rank(att) = rank(RW_Q W_K^\top R^\top)$$
$$\leq \min(rank(R), rank(W_Q), rank(R), rank(W_K))$$
$$\leq d_h,$$

head dim

rank(AB) ≤ min{rank(A), rank(B)}

DIF attention

$$rank(DIF\_att) = d_h + \Sigma_{j=1}^{p} d_{h_j} > d_h.$$

head dim + 개별 attribute 의 head dim

Low-Rank Bottleneck in Multi-head Attention Models (ICML 2020)

고정된 크기의 Multi -head dim 이 너무 작으면 rank bottle neck 으로 표현력이 낮아짐을 제기함

4. DIF-SR

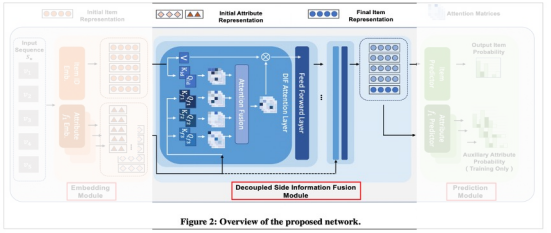
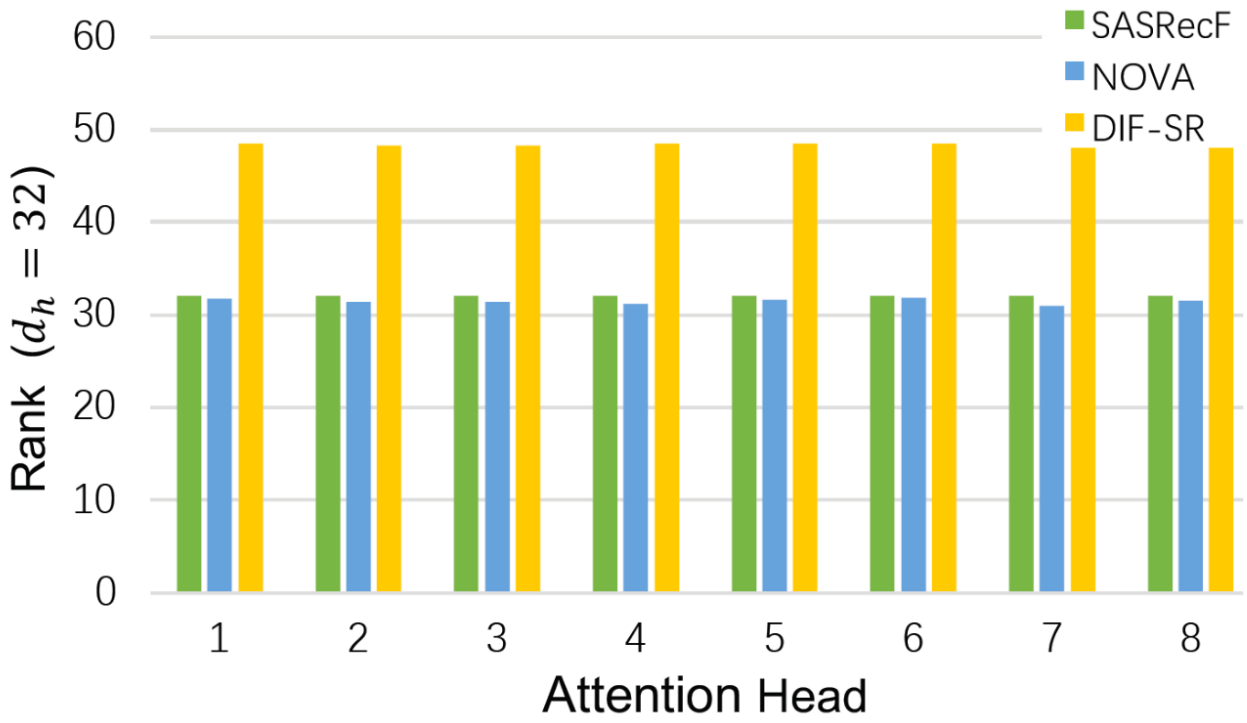


Figure 2: Overview of the proposed network.

- Rank bottleneck (Problem 1)
  - Rank comparison result (under figure)
  - Early Fusion rank (32) -> **weakening of expression** (rank bottleneck) << DIF-SR (48)



4. DIF-SR

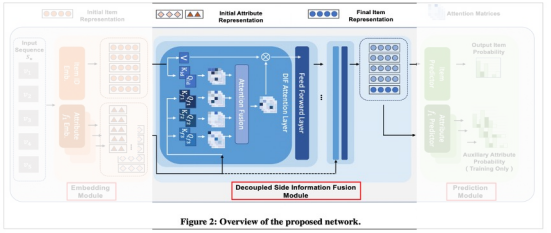
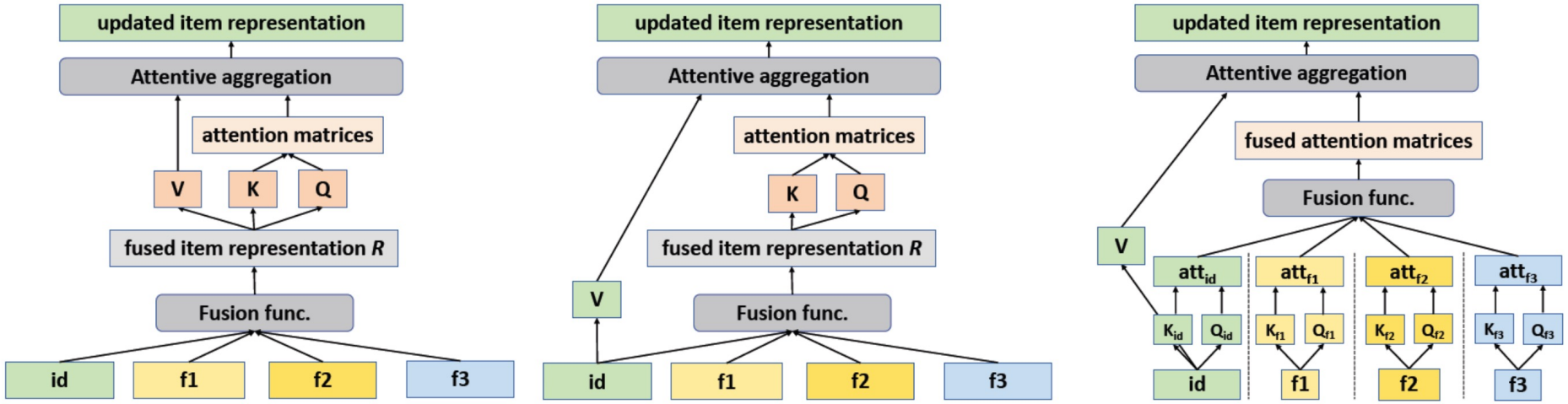


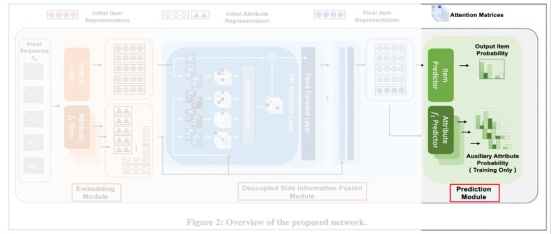
Figure 2: Overview of the proposed network.

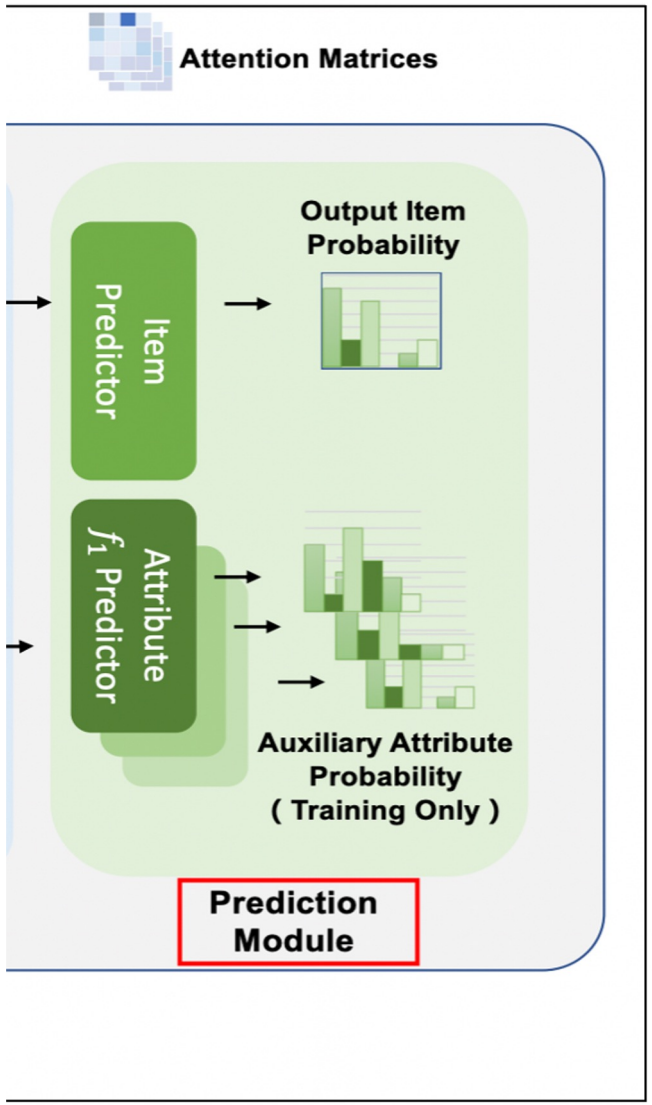
- Same gradient (Problem 2)
  - shares the same gradient (item, side information)



(a) SASRec$_F$. (b) NOVA-SR. (c) DIF-SR.

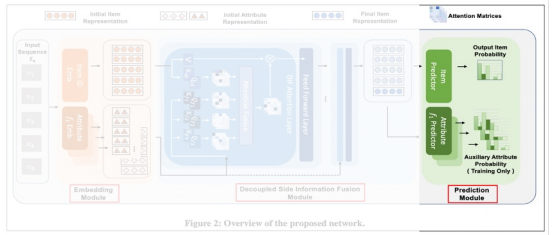
4. DIF-SR



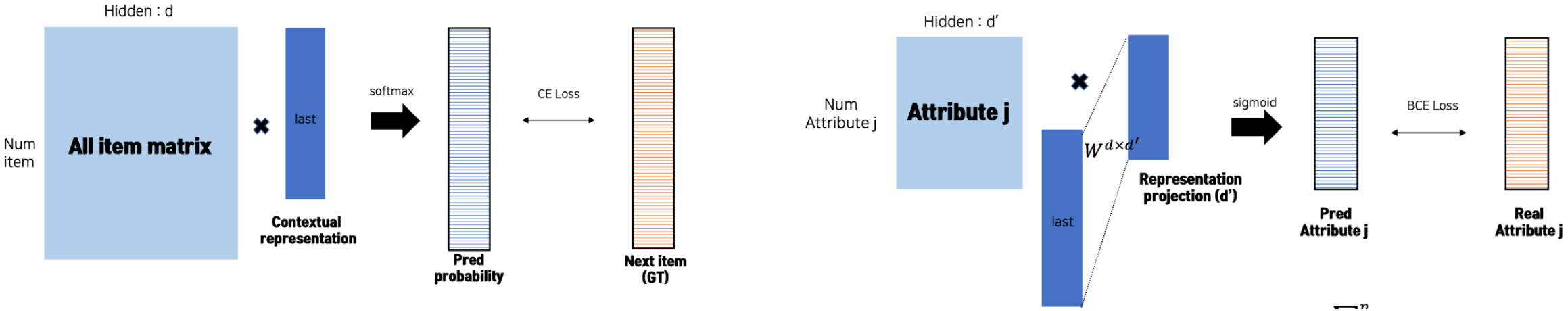


3. Prediction Module
   - Next item prediction
   - Auxiliary Attribute Probability

4. DIF-SR



- Loss Define
  - Next item Prediction Loss + Auxiliary Attribute Predictors Loss
  - Softmax and Cross Entropy



$$Total\ Loss = Loss_{next} + \lambda \sum_{j=1}^{p} Loss_{AAP}$$

5. Experiments and Conclusion

- **Improved performance**
- Simple **early fusion may not significantly affect** model performance

Dataset	Metric	GRU4Rec	Caser	BERT4Rec	GRU4Rec$_F$	SASRec	SASRec$_F$	S^3Rec	NOVA	ICAI	DIF-SR
Beauty	Recall@10	0.0530	0.0474	0.0529	0.0587	0.0828	0.0719	0.0868	0.0887	0.0879	**0.0908**
	Recall@20	0.0839	0.0731	0.0815	0.0902	0.1197	0.1013	0.1236	0.1237	0.1231	**0.1284**
	NDCG@10	0.0266	0.0239	0.0237	0.0290	0.0371	0.0414	0.0439	0.0439	0.0439	**0.0446**
	NDCG@20	0.0344	0.0304	0.0309	0.0369	0.0464	0.0488	0.0531	0.0527	0.0528	**0.0541**
Sports	Recall@10	0.0312	0.0227	0.0295	0.0394	0.0526	0.0435	0.0517	0.0534	0.0527	**0.0556**
	Recall@20	0.0482	0.0364	0.0465	0.0610	0.0773	0.0640	0.0758	0.0759	0.0762	**0.0800**
	NDCG@10	0.0157	0.0118	0.0130	0.0199	0.0233	0.0235	0.0249	0.0250	0.0243	**0.0264**
	NDCG@20	0.0200	0.0153	0.0173	0.0253	0.0295	0.0286	0.0310	0.0307	0.0302	**0.0325**
Toys	Recall@10	0.0370	0.0361	0.0533	0.0492	0.0831	0.0733	0.0967	0.0978	0.0972	**0.1013**
	Recall@20	0.0588	0.0566	0.0787	0.0767	0.1168	0.1052	0.1349	0.1322	0.1303	**0.1382**
	NDCG@10	0.0184	0.0186	0.0234	0.0246	0.0375	0.0417	0.0475	0.0480	0.0478	**0.0504**
	NDCG@20	0.0239	0.0238	0.0297	0.0316	0.0460	0.0497	0.0571	0.0567	0.0561	**0.0597**
Yelp	Recall@10	0.0361	0.0380	0.0524	0.0361	0.0650	0.0413	0.0589	0.0681	0.0663	**0.0698**
	Recall@20	0.0592	0.0608	0.0756	0.0578	0.0928	0.0675	0.0902	0.0964	0.0940	**0.1003**
	NDCG@10	0.0184	0.0197	0.0327	0.0182	0.0401	0.0216	0.0338	0.0412	0.0400	**0.0419**
	NDCG@20	0.0243	0.0255	0.0385	0.0236	0.0471	0.0282	0.0416	0.0483	0.0470	**0.0496**

Early fusion

Reference

- Paper
  - https://arxiv.org/pdf/2204.11046.pdf
- Blog
  - https://dhgudxor.tistory.com/entry/%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0-Decoupled-Side-Information-Fusion-for-Sequential-Recommendation
  - https://jonhyuk0922.tistory.com/229
- Youtube
  - https://www.youtube.com/watch?v=5Ftg8PpJi5A