Text Is All You Need: Learning Representations for Sequential Recommendation

method

task

Advisor : Jia-Ling, Koh

Speaker : Hsuan Lu

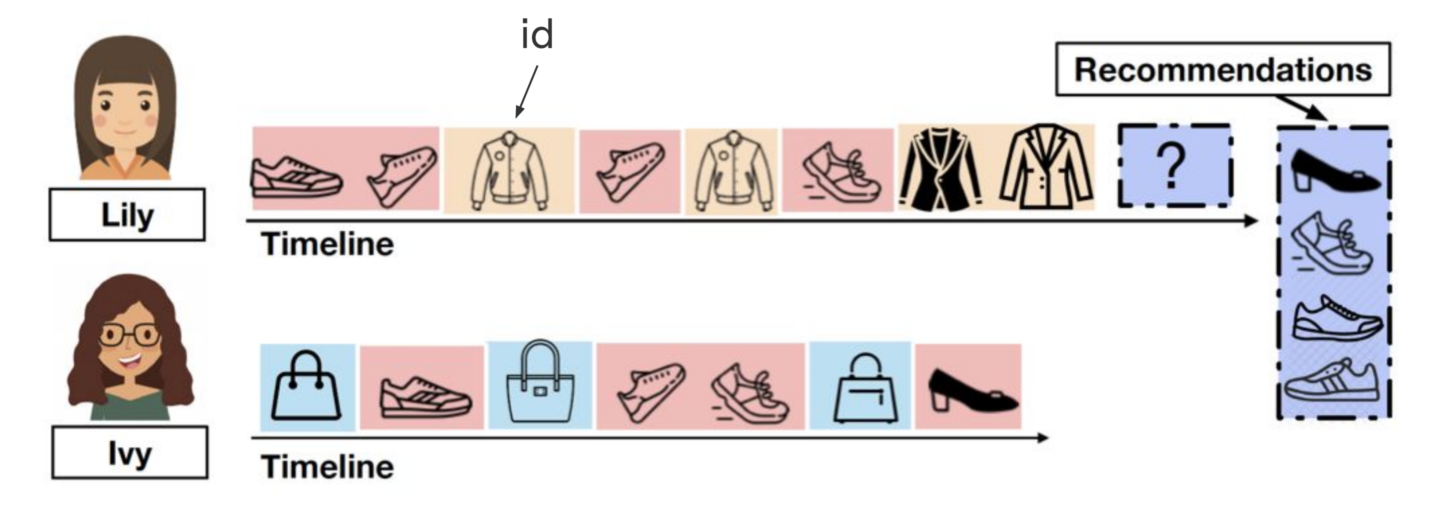
Source : KDD'2023

Date : 2024/04/02

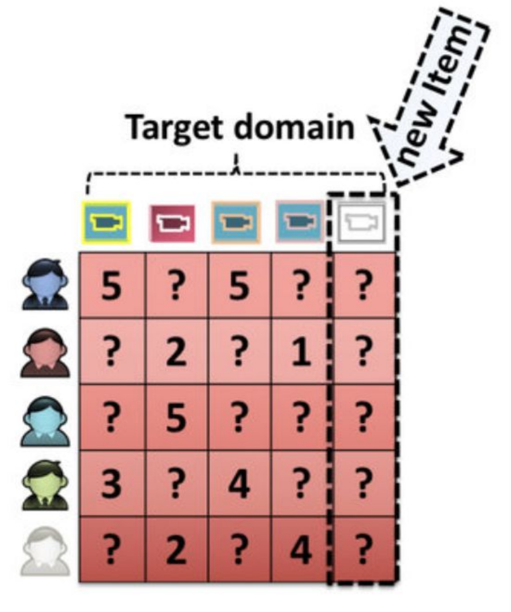
Outline

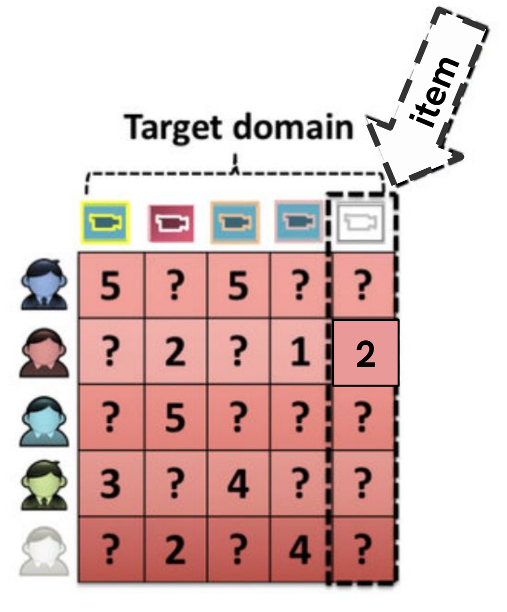
- Introduction
- Method
- Experiment
- Conclusion

Sequential Recommendation

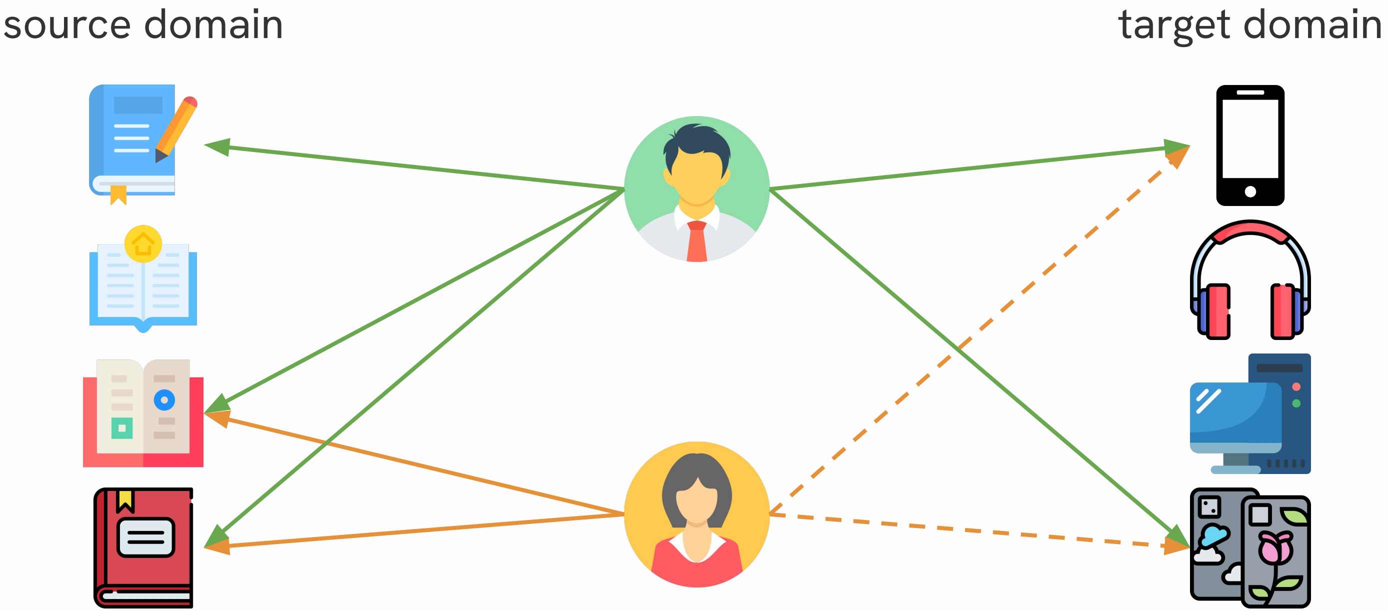


Cold-Start Items

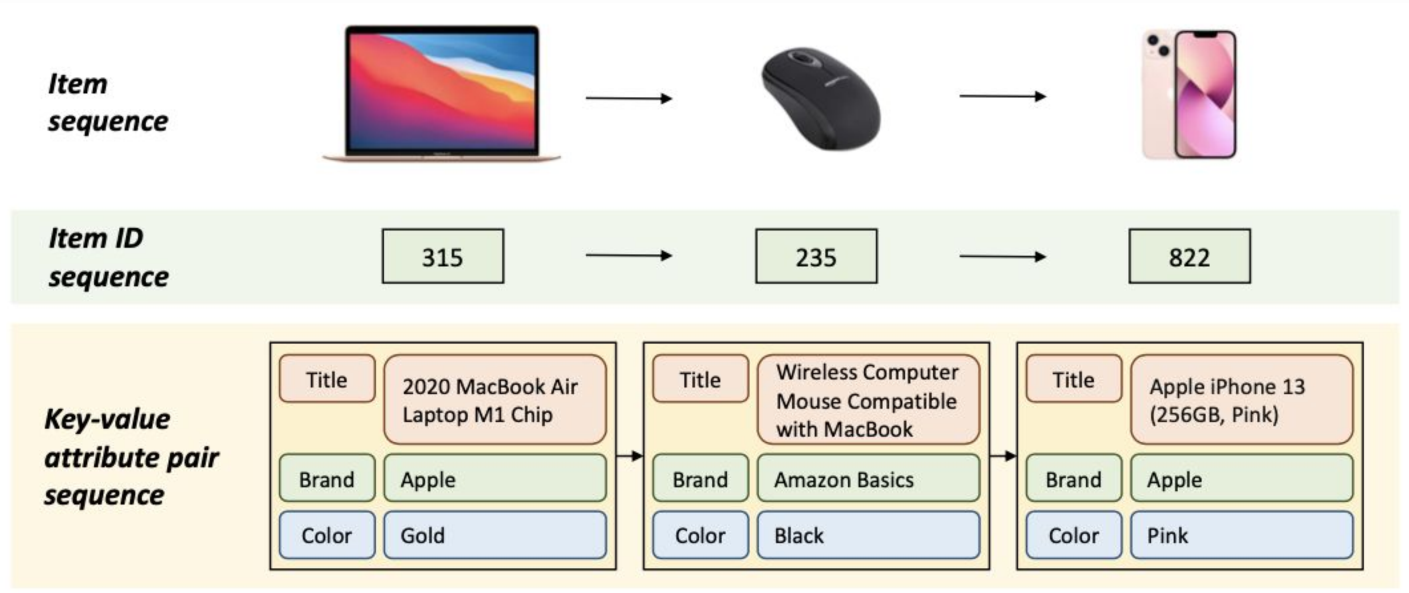




Cross-Domain Recommendation

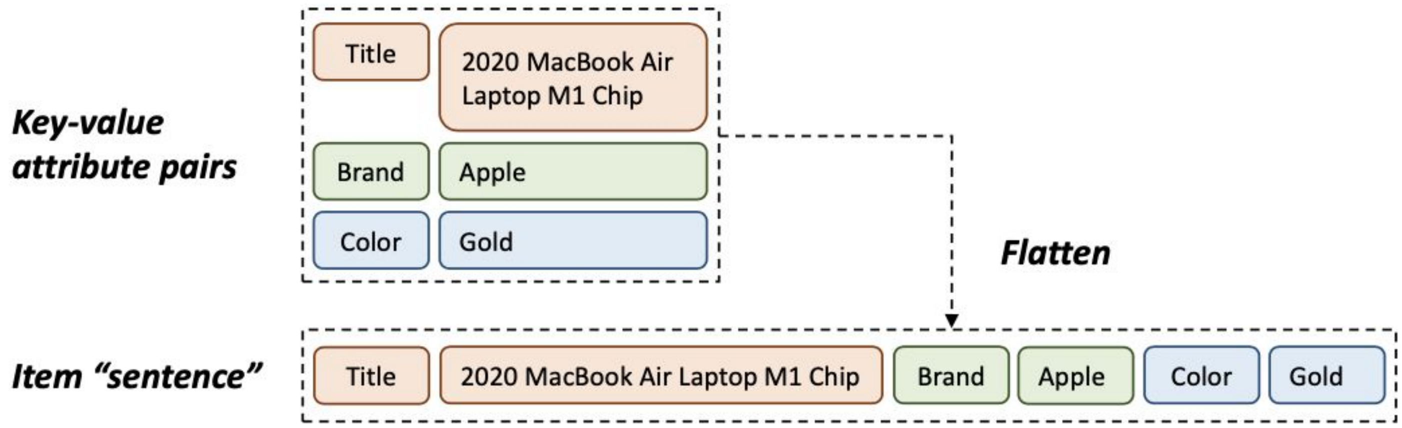


Input Data

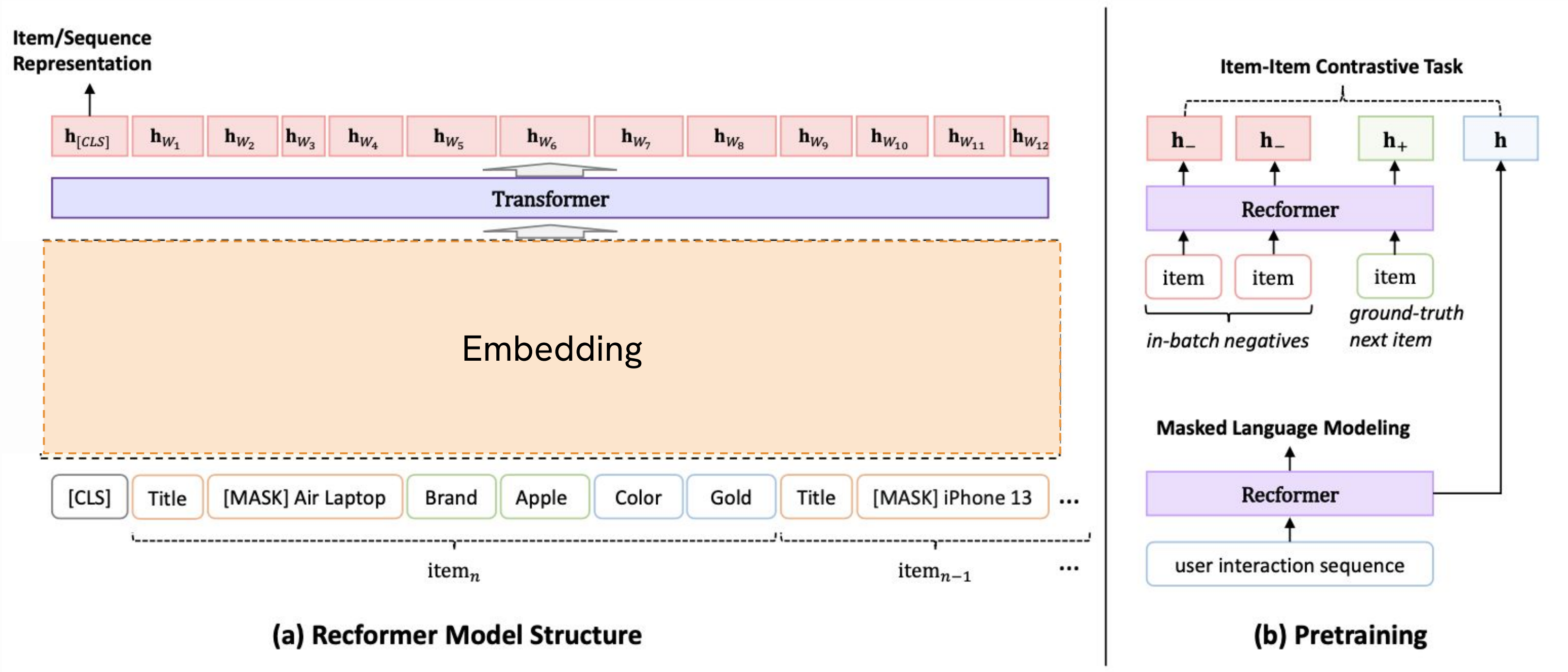


Item sequence
Item ID sequence
315
235
822
Key-value attribute pair sequence
Title
2020 MacBook Air Laptop M1 Chip
Brand
Apple
Color
Gold
Title
Wireless Computer Mouse Compatible with MacBook
Brand
Amazon Basics
Color
Black
Title
Apple iPhone 13 (256GB, Pink)
Brand
Apple
Color
Pink

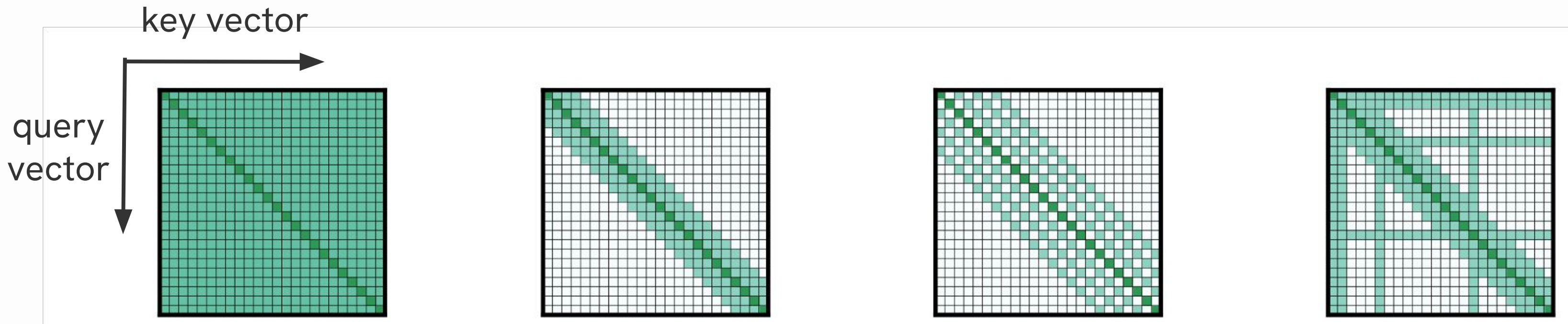
Input Data


Key-value
attribute pairs
Title
2020 MacBook Air
Laptop M1 Chip
Brand
Apple
Color
Gold
Flatten
Item "sentence"
Title
2020 MacBook Air Laptop M1 Chip
Brand
Apple
Color
Gold

Overview of Recformer



Longformer



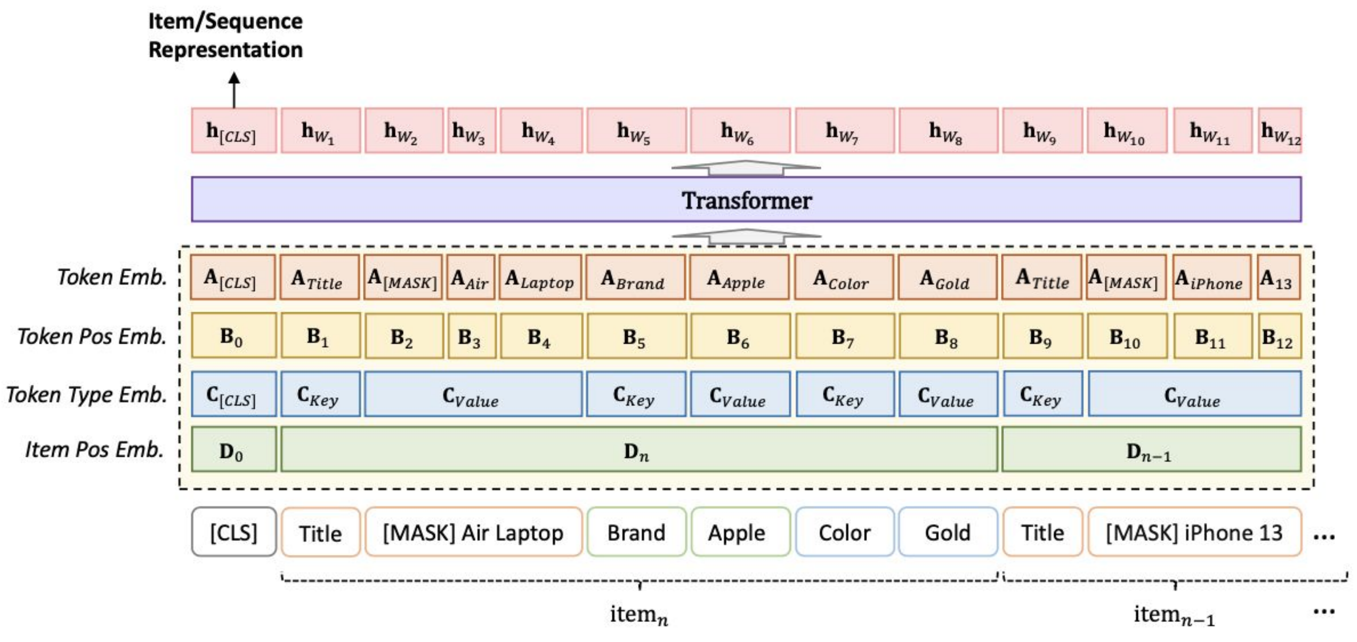
(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

Figure 2: Comparing the full self-attention pattern and the configuration of attention patterns in our Longformer.

Outline

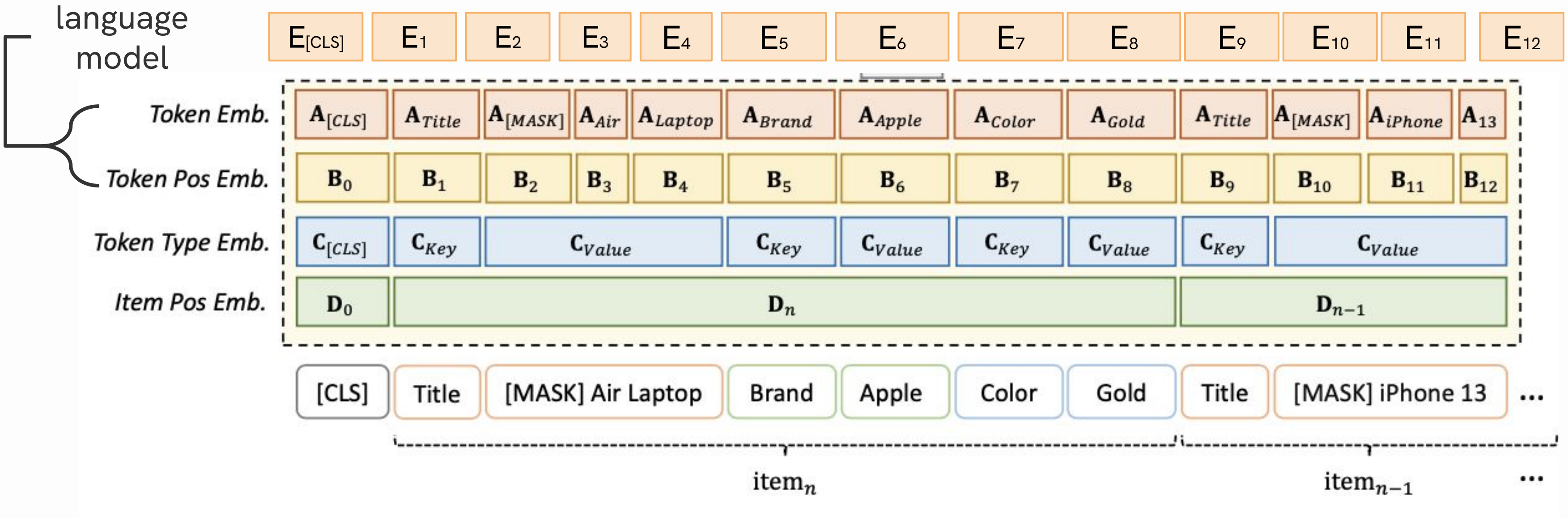
- Introduction
- Method
- Experiment
- Conclusion

Overview of Recformer



(a) Recformer Model Structure

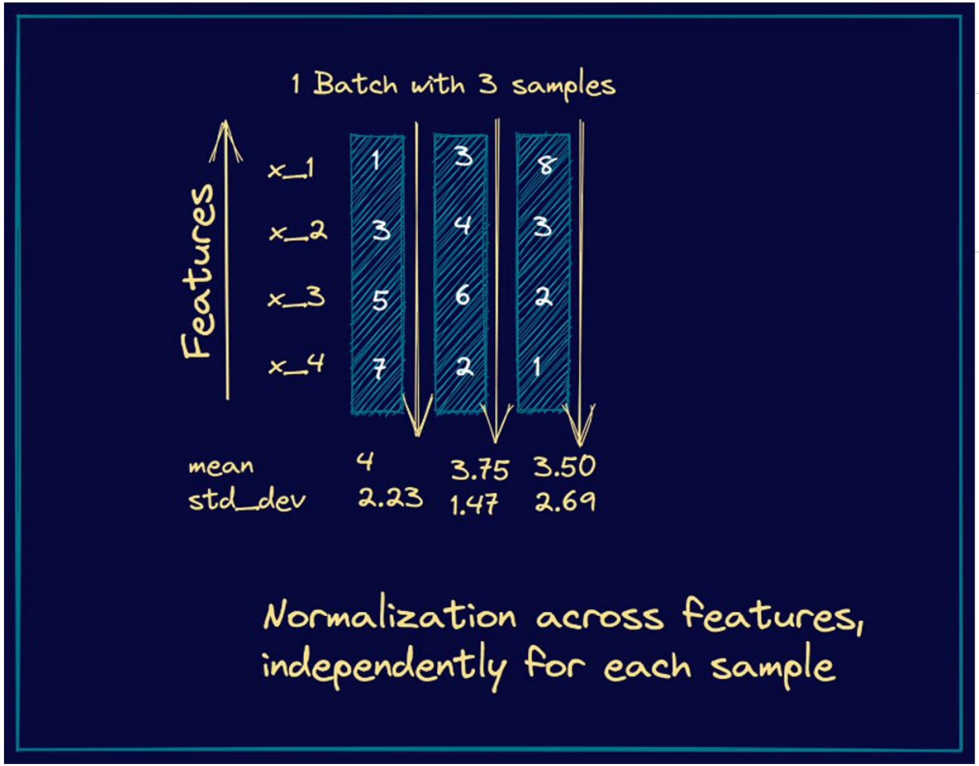
Embedding Layer



$$\mathbf{E}_w = \mathrm{LayerNorm}(\mathbf{A}_w + \mathbf{B}_w + \mathbf{C}_w + \mathbf{D}_w) \qquad \mathbf{E}_w \in \mathbb{R}^d$$

$$\mathbf{E}_X = [\mathbf{E}_{[\mathrm{CLS}]}, \mathbf{E}_{w_1}, \ldots, \mathbf{E}_{w_l}] \qquad \mathbf{E}_X \in \mathbb{R}^{(l+1)\times d}$$

Layer Normalization

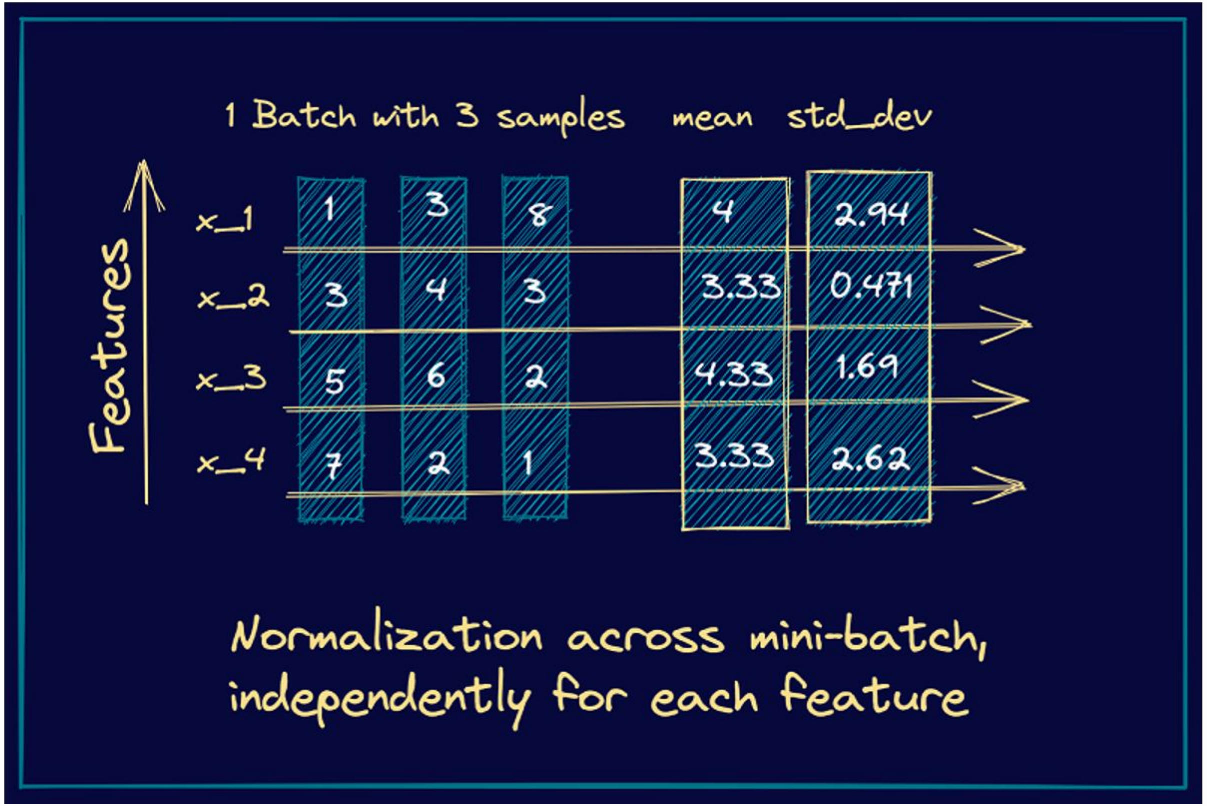


$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \text{ and } \sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2.$$

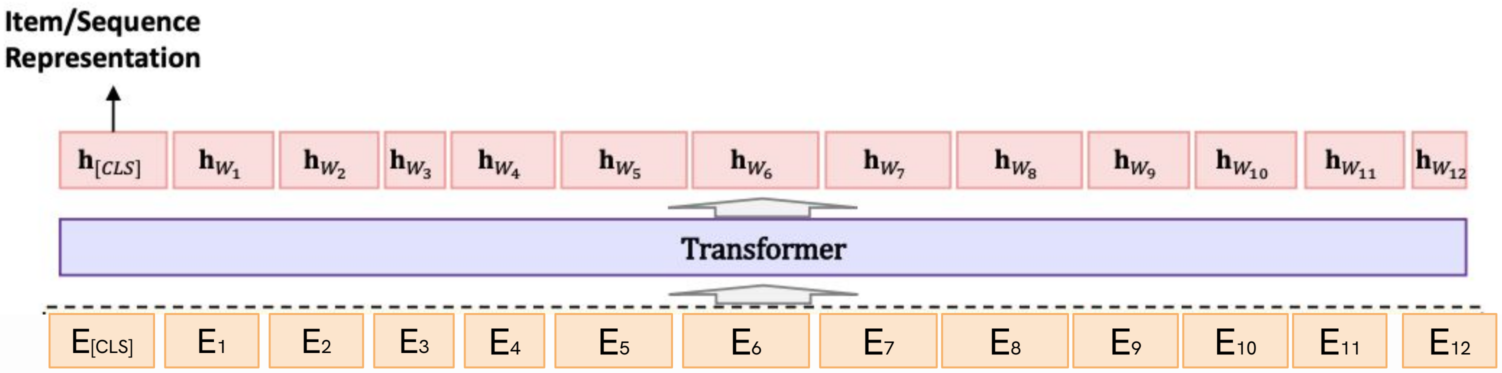
$$\hat{x}_i^{(k)} = \frac{x_i^{(k)} - \mu_B^{(k)}}{\sqrt{\left(\sigma_B^{(k)}\right)^2 + \epsilon}}$$

$$y_i^{(k)} = \gamma^{(k)}\hat{x}_i^{(k)} + \beta^{(k)},$$

Batch Normalization



Item or Sequence Representation



$$[\mathbf{h}_{[\mathtt{CLS}]}, \mathbf{h}_{w_1}, \ldots, \mathbf{h}_{w_l}] = \mathrm{Longformer}([\mathbf{E}_{[\mathtt{CLS}]}, \mathbf{E}_{w_1}, \ldots, \mathbf{E}_{w_l}])$$

$$\mathbf{h}_w \in \mathbb{R}^d$$

Prediction

sequence representation

item representation

$$r_{i,s} = \frac{\mathbf{h}_i^\top \mathbf{h}_s}{\|\mathbf{h}_i\| \cdot \|\mathbf{h}_s\|}$$

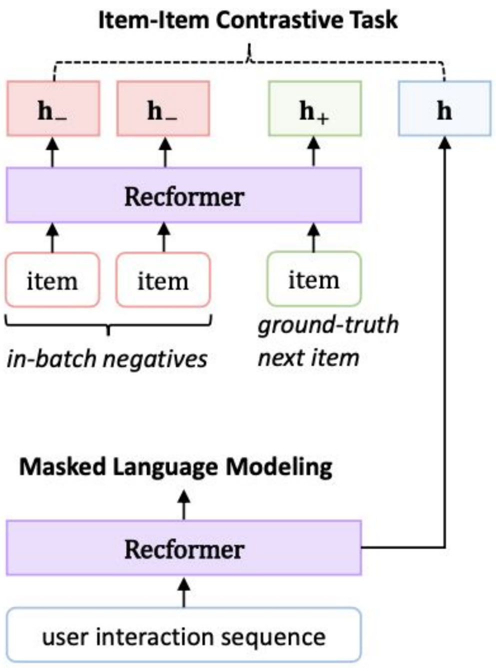
$$\hat{i_s} = \mathrm{argmax}_{i \in \mathcal{I}} (r_{i,s})$$

the predicted item given
user sequence s

all item

s: user's interaction sequence
i: item i

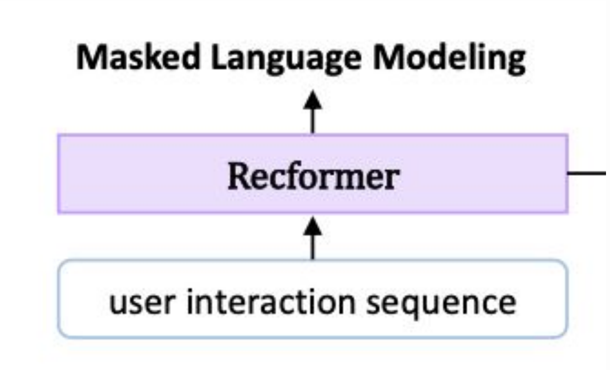
Learning Framework



(b) Pretraining

Pre-training

- Masked Language Modeling
  - 隨機取15%的位置進行預測
    - 80% [MASK]
    - 10% 隨機替換
    - 10% 不變



$$\mathbf{m} = \mathrm{LayerNorm}(\mathrm{GELU}(\mathbf{W}_h \mathbf{h}_w + \mathbf{b}_h))$$
$$p = \mathrm{Softmax}(\mathbf{W}_0 \mathbf{m} + \mathbf{b}_0)$$

$$\mathbf{W}_h \in \mathbb{R}^{d \times d}, \mathbf{b}_h \in \mathbb{R}^d, \mathbf{W}_0 \in \mathbb{R}^{|\mathcal{V}| \times d}, \mathbf{b}_0 \in \mathbb{R}^{|\mathcal{V}|}$$

vocabulary

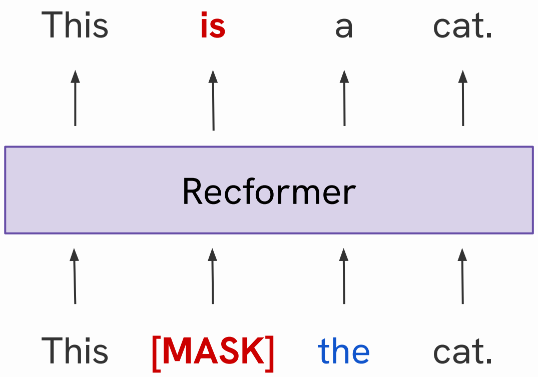
Pre-training

- Masked Language Modeling

$$\mathcal{L}_{\mathrm{MLM}} = -\sum_{i=0}^{|\mathcal{V}|} y_i \log(p_i)$$

對1個MASK的位置,
最後平均所有mask的loss

Yis = 1
Pis = 0.7



Pre-training

- Item-Item Contrastive Task
  - positive instance
    - ground-truth next item
  - negative instances
    - in-batch next items

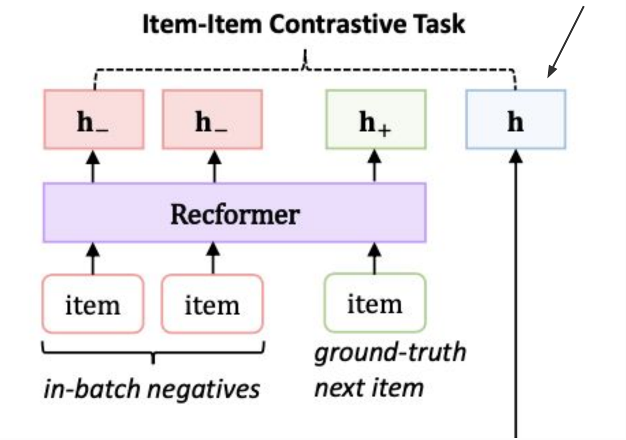
(**<u>User 1</u>**, Item 3) Item 3 ← positive instance

(User 2, Item 1) Item 1 ← negative instance

(User 3, Item 4) Item 4 ← negative instance

(User 4, Item 7) Item 7 ← negative instance

sequence representation

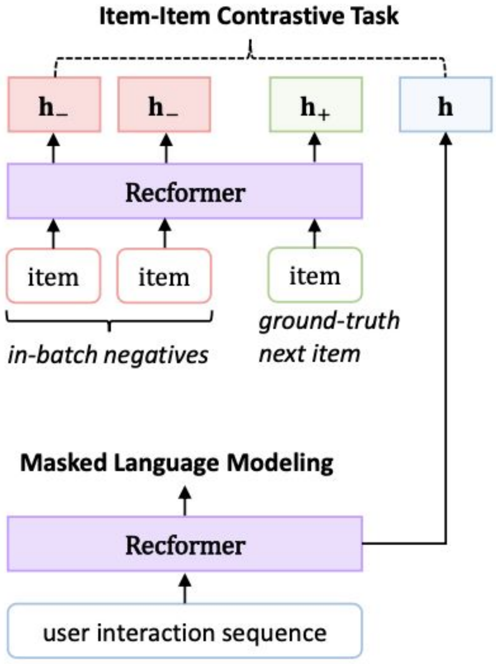


$$\mathcal{L}_{\mathrm{IIC}} = -\log \frac{e^{\mathrm{sim}(\mathbf{h}_s, \mathbf{h}_i^+)/\tau}}{\sum_{i \in \mathcal{B}} e^{\mathrm{sim}(\mathbf{h}_s, \mathbf{h}_i)/\tau}}$$

ground truth item set in one batch

Multi-Task Training Strategy

$$\mathcal{L}_{\mathrm{PT}} = \mathcal{L}_{\mathrm{IIC}} + \lambda \cdot \mathcal{L}_{\mathrm{MLM}}$$



(b) Pretraining

Two-Stage Finetuning

- $\mathcal{I}$: item set
- M: pre-trained language model
- p: 模型準確率
- **I**: item feature matrix

Algorithm 1: Two-Stage Finetuning

Input: D_{train}, D_{valid}, $\mathcal{I}$, M
Hyper-parameters: n_{epoch}
Output: M', $\mathbf{I}'$

$M \leftarrow$ initialized with pre-trained parameters
$p \leftarrow$ metrics are initialized with 0
Stage 1
for n in n_{epoch} **do**
  $\mathbf{I} \leftarrow \text{Encode}(M, \mathcal{I})$
  $M \leftarrow \text{Train}(M, \mathbf{I}, D_{\text{train}})$
  $p' \leftarrow \text{Evaluate}(M, \mathbf{I}, D_{\text{valid}})$
  if $p' > p$ **then**
    $M', \mathbf{I}' \leftarrow M, \mathbf{I}$
    $p \leftarrow p'$
  end if
end for

Two-Stage Finetuning

$$\mathcal{L}_{\mathrm{FT}} = -\log \frac{e^{\mathrm{sim}(\mathbf{h}_s, \mathbf{I}_i^{+})/\tau}}{\sum_{i \in \mathcal{I}} e^{\mathrm{sim}(\mathbf{h}_s, \mathbf{I}_i)/\tau}}$$

item set

Stage 2

$M \leftarrow M'$
for n in n_{epoch} **do**
    $M \leftarrow \mathrm{Train}(M, \mathbf{I}', D_{\text{train}})$
    $p' \leftarrow \mathrm{Evaluate}(M, \mathbf{I}', D_{\text{valid}})$
    if $p' > p$ **then**
        $M' \leftarrow M$
        $p \leftarrow p'$
    end if
end for
return $M', \mathbf{I}'$

Outline

- Introduction
- Method
- Experiment
- Conclusion

Dataset

- Amazon review datasets

the average length of item sequences

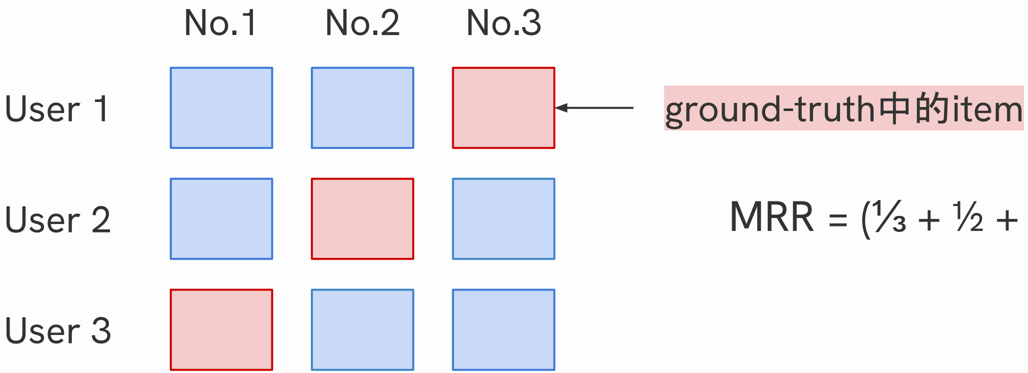
Datasets	#Users	#Items	#Inters.	Avg. n	Density
Pre-training	3,613,906	1,022,274	33,588,165	9.29	9.1e-6
-Training	3,501,527	954,672	32,291,280	9.22	9.0e-6
-Validation	112,379	67,602	1,296,885	11.54	1.7e-4
Scientific	11,041	5,327	76,896	6.96	1.3e-3
Instruments	27,530	10,611	231,312	8.40	7.9e-4
Arts	56,210	22,855	492,492	8.76	3.8e-4
Office	101,501	27,932	798,914	7.87	2.8e-4
Games	11,036	15,402	100,255	9.08	5.9e-4
Pet	47,569	37,970	420,662	8.84	2.3e-4

Evaluation

N: user總數量
Pi: 對第i個user，推薦列表中第一個在ground-truth結果中的item所在的排列位置

- MRR (Mean Reciprocal Rank):

$$\mathrm{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{p_i}$$

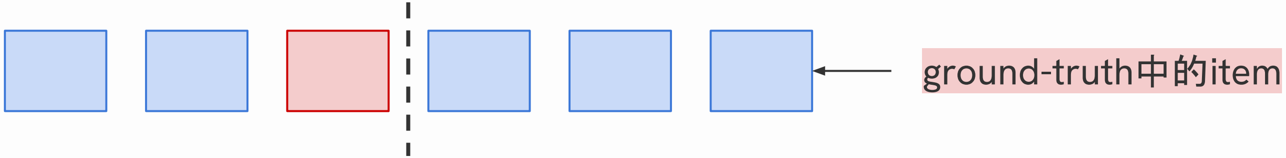


MRR = (⅓ + ½ + 1) / 3 = 0.61

Evaluation

- Recall: $$Recall = \frac{TP}{TP + FN}$$

	「模型預測」為真 (positive)	「模型預測」為非 (negative)
「真實情況」為真	true positive (TP)	false negative (FN)
「真實情況」為非	false positive (FP)	true negative (TN)



Recall@3 = 1 (or 0)

Evaluation

- NDCG (Normalized Discounted Cumulative Gain):

$$\mathrm{DCG}_v = \sum_{i=1}^{v} \frac{g(rel_i)}{log(i+1)}$$

$$\mathrm{IDCG}_v = \sum_{k \in \mathrm{REL}_v} \frac{g(rel_k)}{log(k+1)}$$

$$\mathrm{nDCG}_v = \frac{\mathrm{DCG}_v}{\mathrm{IDCG}_v}$$

No.1	No.2	No.3
0	1	0

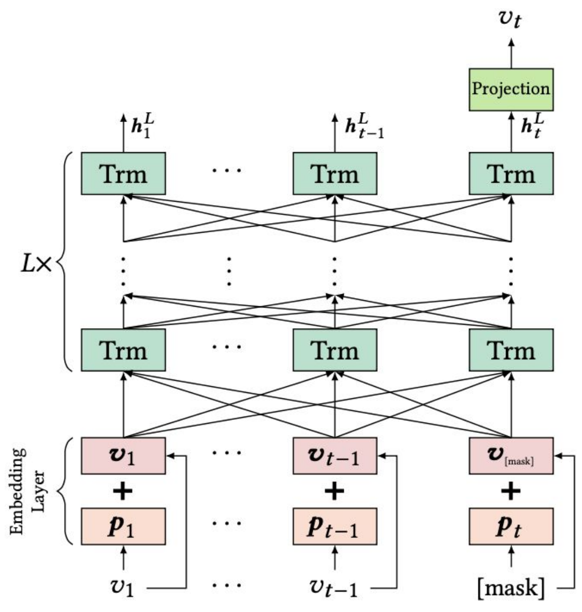
$$DCG_{@3} = \frac{0}{log(1+1)} + \frac{1}{log(2+1)} + \frac{0}{log(3+1)}$$

$$IDCG_{@3} = \frac{1}{log(1+1)} + \frac{0}{log(2+1)} + \frac{0}{log(3+1)}$$

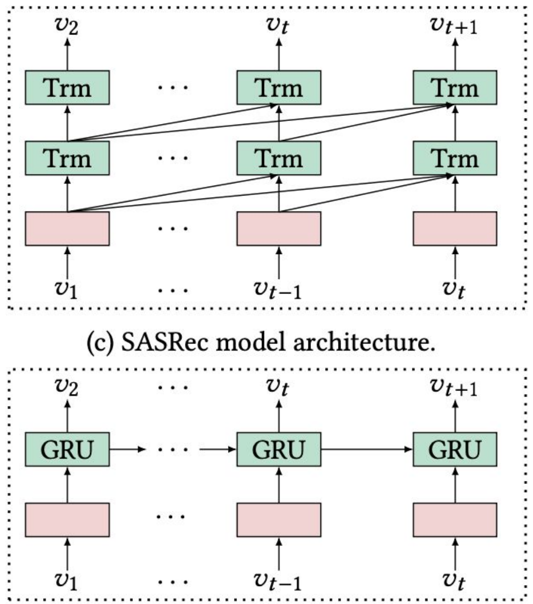
$$NDCG_{@3} = \frac{0.63}{1} = 0.63$$

Baseline - ID-Only methods

- GRU4Rec
- SASRec
- BERT4Rec



(b) BERT4Rec model architecture.

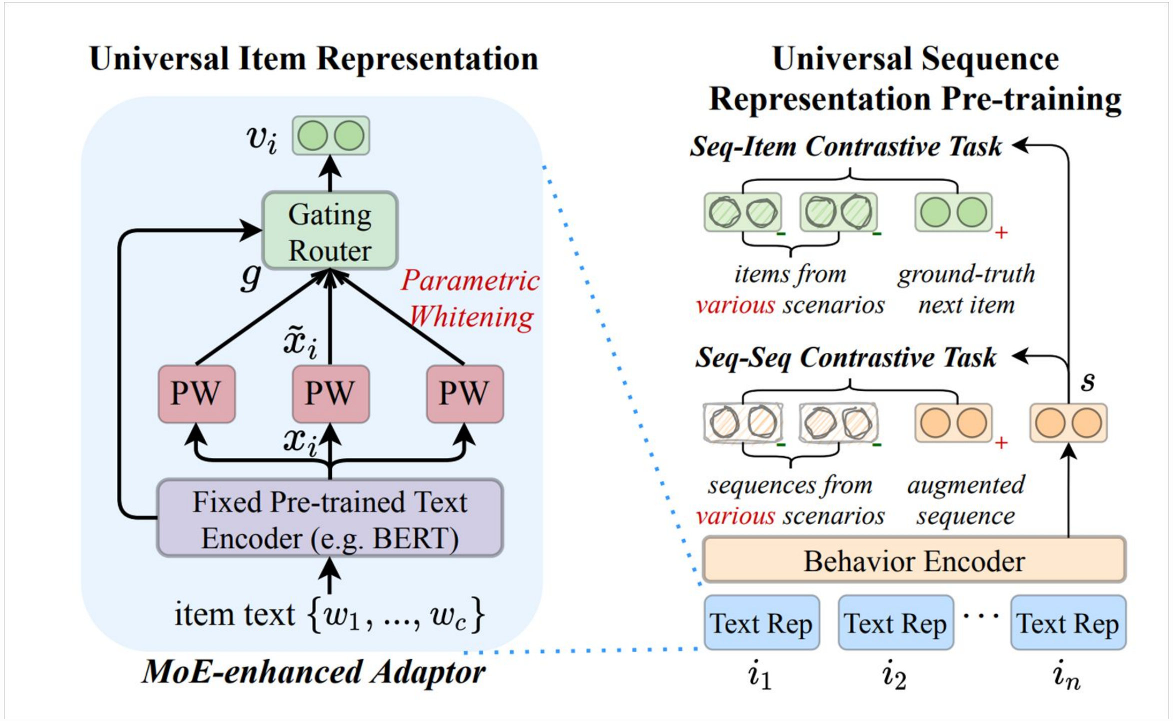


(c) SASRec model architecture.

(d) RNN based sequential recommendation methods.

Baseline - Text-Only Method

- UniSRec



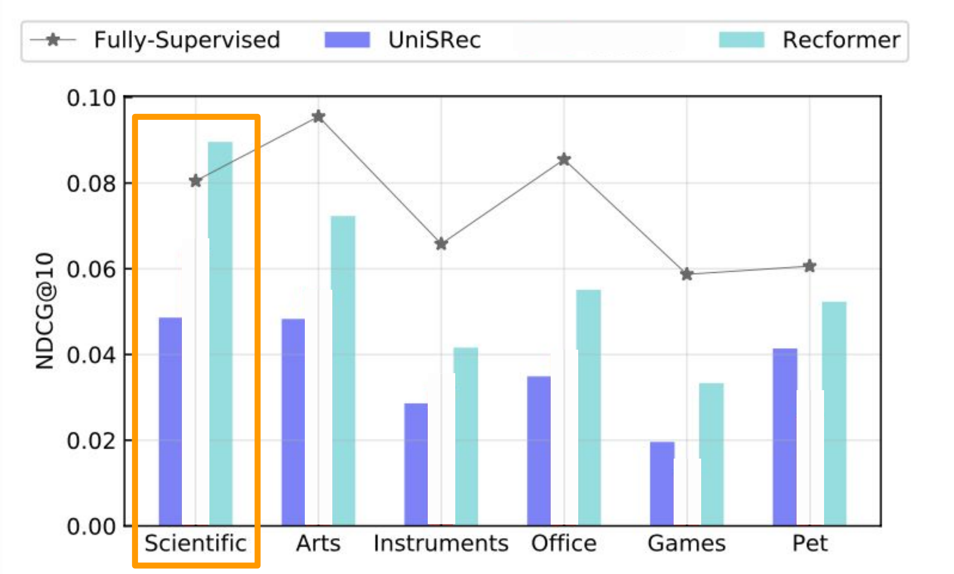
Experiment

Dataset	Metric	ID-Only Methods			Text-Only Methods		Improv.
		GRU4Rec	SASRec	BERT4Rec	UniSRec	RECFORMER	
Scientific	NDCG@10	0.0826	0.0797	0.0790	0.0862	**0.1027**	19.14%
	Recall@10	0.1055	0.1305	0.1061	0.1255	**0.1448**	10.96%
	MRR	0.0702	0.0696	0.0759	0.0786	**0.0951**	20.99%
Instruments	NDCG@10	0.0633	0.0634	0.0707	0.0785	**0.0830**	4.14%
	Recall@10	0.0969	0.0995	0.0972	**0.1119**	0.1052	-
	MRR	0.0707	0.0577	0.0677	0.0740	**0.0807**	6.89%
Arts	NDCG@10	0.1075	0.0848	0.0942	0.0894	**0.1252**	16.47%
	Recall@10	0.1317	0.1342	0.1236	0.1333	**0.1614**	15.37%
	MRR	0.1041	0.0742	0.0899	0.0798	**0.1189**	12.49%
Office	NDCG@10	0.0761	0.0832	0.0972	0.0919	**0.1141**	17.39%
	Recall@10	0.1053	0.1196	0.1205	0.1262	**0.1403**	9.18%
	MRR	0.0731	0.0751	0.0932	0.0848	**0.1089**	12.04%
Games	NDCG@10	0.0586	0.0547	0.0628	0.0580	**0.0684**	8.92%
	Recall@10	0.0988	0.0953	0.1029	0.0923	**0.1039**	0.97%
	MRR	0.0539	0.0505	0.0585	0.0552	**0.0650**	11.11%
Pet	NDCG@10	0.0648	0.0569	0.0602	0.0702	**0.0972**	28.91%
	Recall@10	0.0781	0.0881	0.0765	0.0933	**0.1162**	11.84%
	MRR	0.0632	0.0507	0.0585	0.0650	**0.0940**	32.39%

Experiment - Zero-Shot

ID-Only

Text-Only



Experiment - Low-Resource

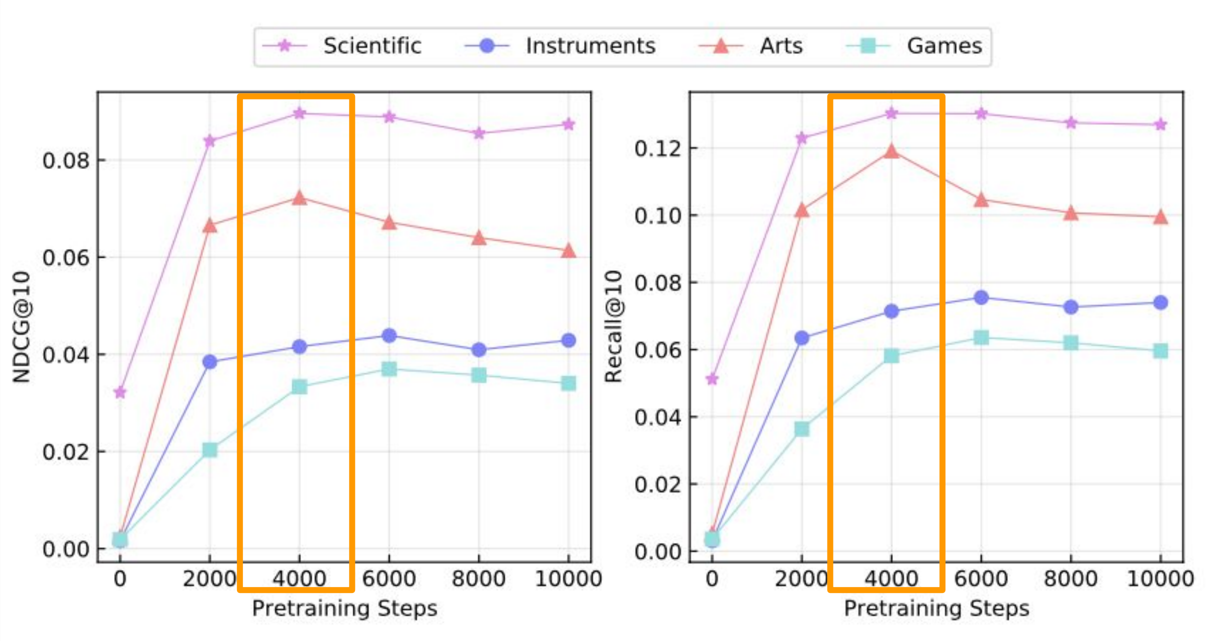
Experiment - Cold-Start Item

		ID-Only		Text-Only			
		SASRec		**UniSRec**		**Recformer**	
Dataset	**Metric**	**In-Set**	**Cold**	**In-Set**	**Cold**	**In-Set**	**Cold**
Scientific	N@10	0.0775	0.0213	0.0864	0.0441	0.1042	0.0520
	R@10	0.1206	0.0384	0.1245	0.0721	0.1417	0.0897
Instruments	N@10	0.0669	0.0142	0.0715	0.0208	0.0916	0.0315
	R@10	0.1063	0.0309	0.1094	0.0319	0.1130	0.0468
Arts	N@10	0.1039	0.0071	0.1174	0.0395	0.1568	0.0406
	R@10	0.1645	0.0129	0.1736	0.0666	0.1866	0.0689
Pet	N@10	0.0597	0.0013	0.0771	0.0101	0.0994	0.0225
	R@10	0.0934	0.0019	0.1115	0.0175	0.1192	0.0400

Experiment - Ablation Study

Variants	Scientific			Instruments		
	NDCG@10	Recall@10	MRR	NDCG@10	Recall@10	MRR
(0) RECFORMER	**0.1027**	**0.1448**	**0.0951**	**0.0830**	**0.1052**	**0.0807**
(1) w/o two-stage finetuning	0.1023	0.1442	0.0948	0.0728	0.1005	0.0685
(5) w/o pre-training	0.0722	0.1114	0.0650	0.0598	0.0732	0.0584
(6) w/o item position emb. & token type emb.	0.1018	0.1427	0.0945	0.0518	0.0670	0.0501

Experiment - Pre-training Steps



Outline

- Introduction
- Method
- Experiment
- Conclusion

Conclusion

- **Recformer** can effectively learn language representations for sequential recommendation
  - formulate items as key-value attribute pairs instead of item IDs
  - design a learning framework including pre-training and finetuning