# BERT4Rec: Sequential Recommendation with BERT

Authors: Fei Sun, Jun Liu, Jian Wu, ... from Alibaba Group Presenter: Hong Fan Zhao

Sequential Recommendation

# Sequential Recommendation

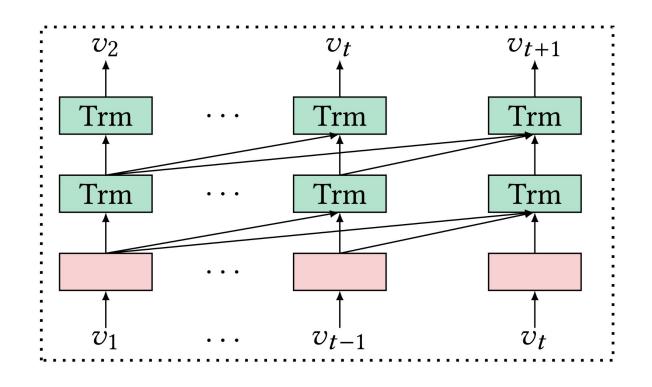


# Sequential Recommendation

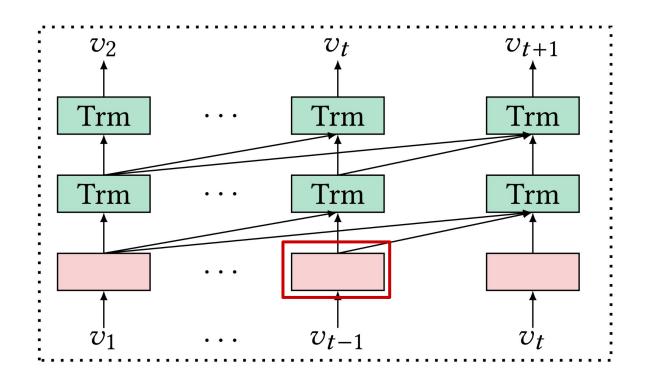


# Previous Works

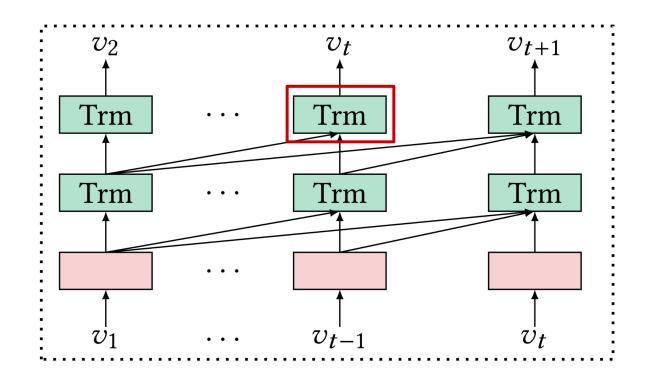
#### Previous Works: SASRec



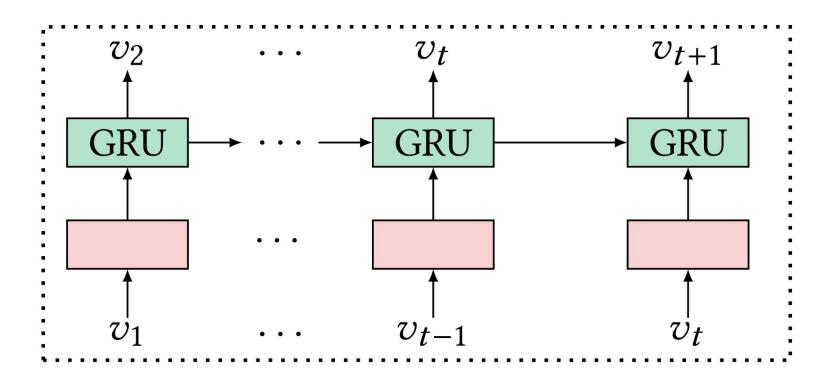
#### Previous Works: SASRec

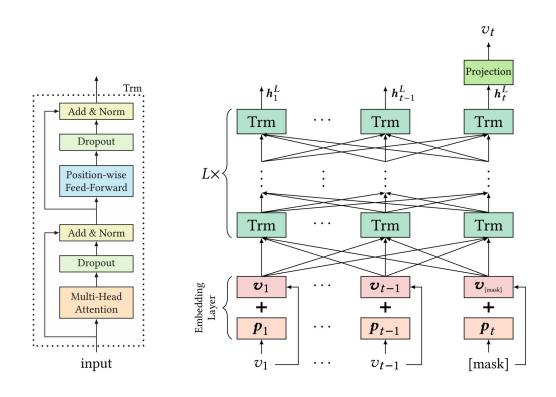


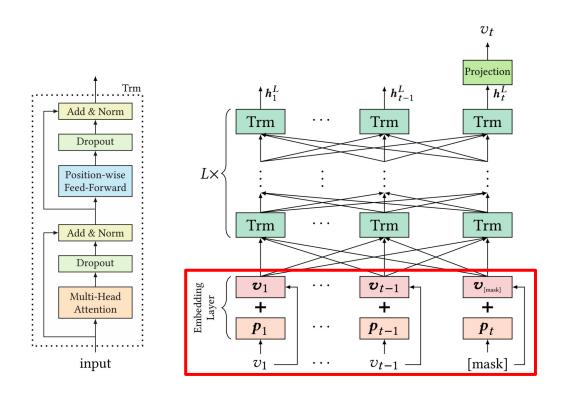
#### Previous Works: SASRec

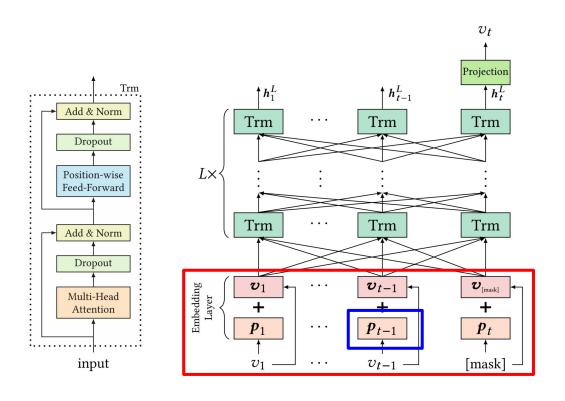


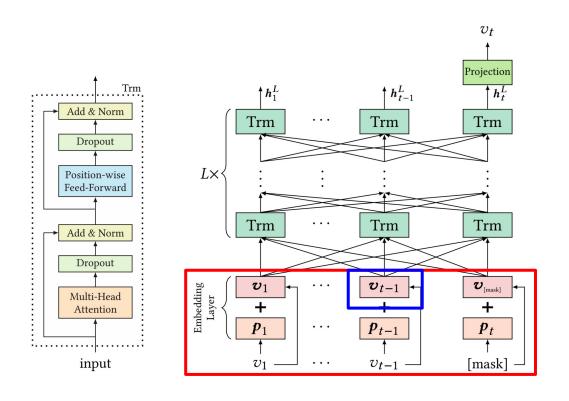
## Previous Works: RNN Based Sequential Recommendation

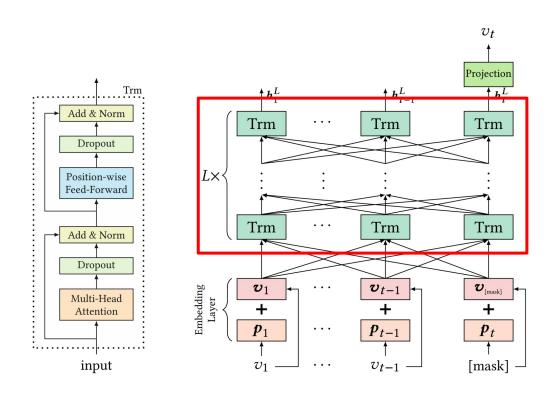


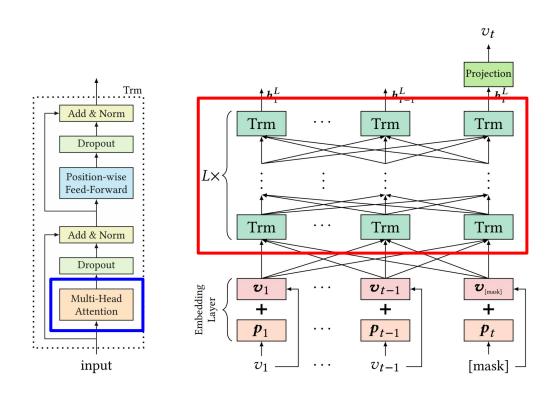


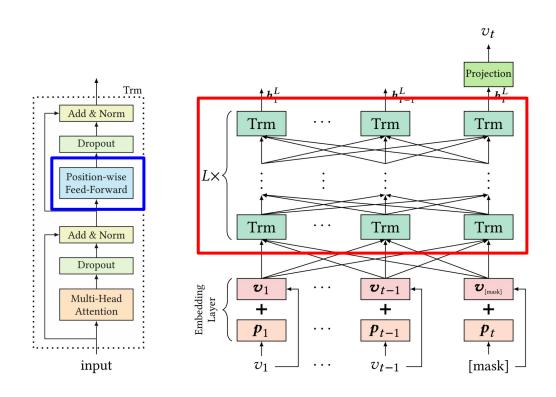


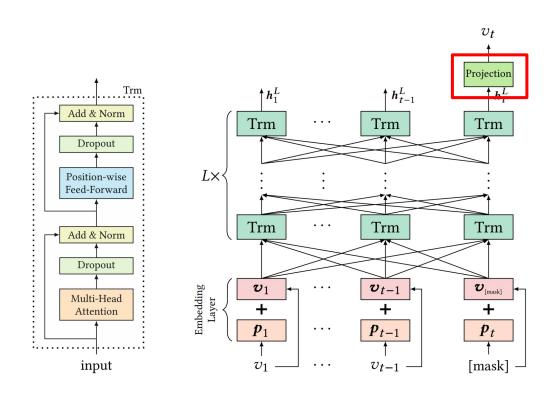












Cloze Task (Masked Language Model)

#### Cloze Task

I have bought a Big-Mac menu, containing a [mask], some french [mask] and a soft [mask].

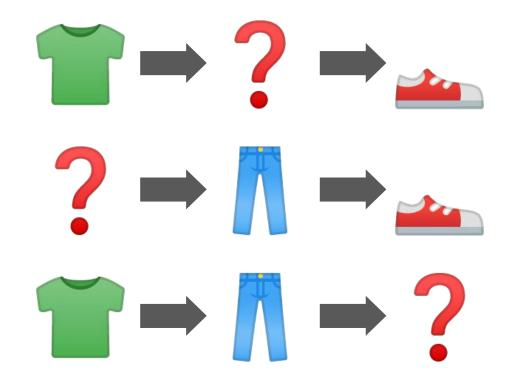
#### Cloze Task

I have bought a Big-Mac menu, containing a Big-Mac, some french fries and a soft drink.

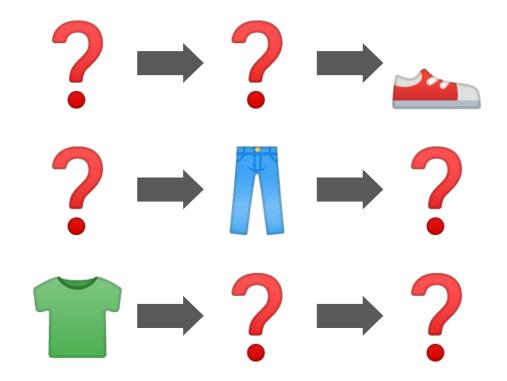
## Cloze Task



### Cloze Task: Training



### Cloze Task: Training



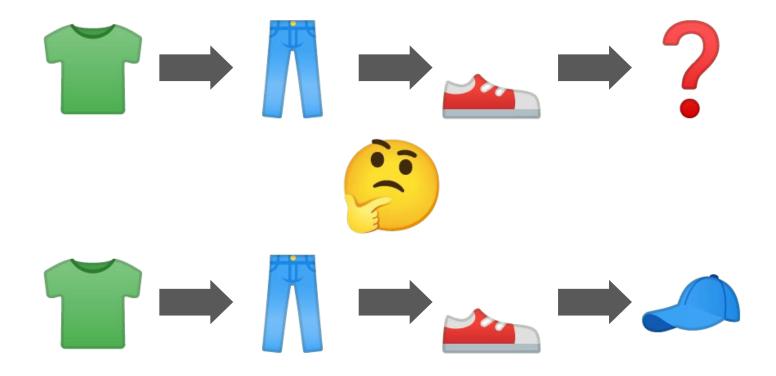
Cloze Task: Training

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

#### Cloze Task: Inference



#### Cloze Task: Inference



$$\mathcal{L} = \frac{1}{|\mathcal{S}_u^m|} \sum_{v_m \in \mathcal{S}_u^m} -\log P(v_m = v_m^* | \mathcal{S}_u')$$

$$\mathcal{L} = \sum_{v_m \in \mathcal{S}_u^m}^{1} -\log P(v_m = v_m^* | \mathcal{S}_u')$$

$$\mathcal{L} = \frac{1}{|\mathcal{S}_u^m|} \sum_{v_m \in \mathcal{S}_u^m} -\log P(v_m = v_m^* | \mathcal{S}_u')$$

$$\mathcal{L} = \frac{1}{|S_u^m|} \sum_{v_m \in S_u^m} -\log P(v_m = v_m^* | S_u')$$

# Experiments

#### Baselines

#### Matrix Factorization Based:

- POP
- BPR-MF
- NCF
- FPMC

#### RNN or CNN Based:

- GRU4Rec
- GRU4Rec<sup>+</sup>
- Caser

#### Transformer Based:

SASRec (Previous State of the Art)

#### Baselines

#### Matrix Factorization Based:

- POP
- BPR-MF
- NCF
- FPMC

#### RNN or CNN Based:

- GRU4Rec
- GRU4Rec<sup>+</sup>
- Caser

#### Transformer Based:

SASRec (Previous State of the Art)

# Datasets (d = 32, L = 2, h = 2)

Amazon Beauty: dataset crawled from Amazon containing users reviews in the Beauty category. ( $\rho$  = 0.6, N = 50)

**Steam**: dataset collected from **Steam**, which is an online video game distribution platform. ( $\rho$  = 0.4, N = 50)

**MovieLens**: a dataset for movie recommendation (**ML-1m** and **ML-20m** are used for the experiments). ( $\rho$  = 0.2, N = 200)

Table 1: Statistics of datasets.

Datasets	#users	#items	#actions	Avg. length	Density
Beauty	40,226	54,542	0.35m	8.8	0.02%
Steam	281,428	13,044	3.5m	12.4	0.10%
ML-1m	6040	3416	1.0m	163.5	4.79%
ML-20m	138,493	26,744	20m	144.4	0.54%

# Metrics: HR@k

$$HR@k = \frac{1}{N} \sum_{i}^{N} in\_top\_k$$

#### Prediction:

- <u>► (50%)</u> Ground Truth: **\*\***
- 15%)
- 👖 (5%)

With k = (1, 2, 3): **in\_top\_k** is **True** 

With k = 4: in\_top\_k is False

# Metrics: NDCG@k

$$NDCG@k = \frac{DCG@k}{IDCG}$$

$$DCG@k = \sum_{i=1}^{k} \frac{G_i}{\log_2(i+1)}$$

$$IDCG = \frac{1}{\log_2 2} = 1$$

#### Prediction:

$$DCG@1 = 0.5$$

Ground Truth:

$$DCG@2 = 0.5 + 0.3 / log(i + 2)$$

. . .

### Metrics: NDCG@k

$$NDCG@k = \frac{DCG@k}{IDCG}$$

$$DCG@k = \sum_{i=1}^{k} \frac{G_i}{\log_2(i+1)}$$

$$IDCG = \frac{1}{\log_2 2} = 1$$

### Prediction:

$$DCG@1 = 0.5$$

$$DCG@2 = 0.5 + 0.3 / log(i + 2)$$

. . .

### Metrics: NDCG@k

$$NDCG@k = \frac{DCG@k}{IDCG}$$

$$DCG@k = \sum_{i=1}^{\kappa} \frac{G_i}{\log_2(i+1)}$$

$$IDCG = \frac{1}{\log_2 2} = 1$$

### Prediction:

$$DCG@1 = 0.5$$

$$DCG@2 = 0.5 + 0.3 / log(i + 2)$$

. . .

### Metrics: NDCG@k

$$NDCG@k = \frac{DCG@k}{IDCG}$$

$$DCG@k = \sum_{i=1}^{k} \frac{G_i}{\log_2(i+1)}$$

$$IDCG = \frac{1}{\log_2 2} = 1$$

### Prediction:

$$DCG@1 = 0.5$$

$$DCG@2 = 0.5 + 0.3 / log(i + 2)$$

. . .

### Metrics: MRR

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

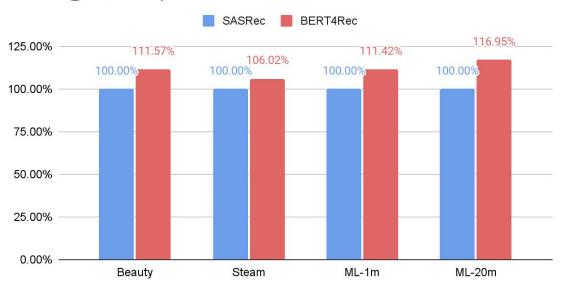
### Prediction:



rank\_i = 3

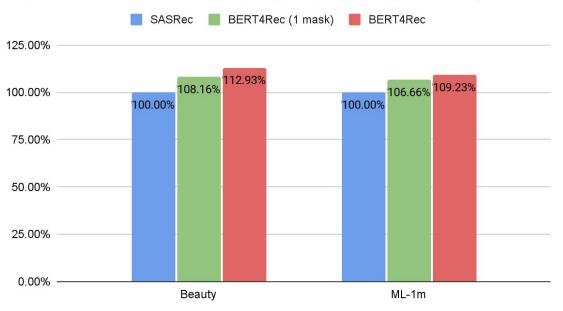
### Improvements

Average Improvements (HR@1, HR@10, HR@5, NDCG@5, NDCG@10, MRR)

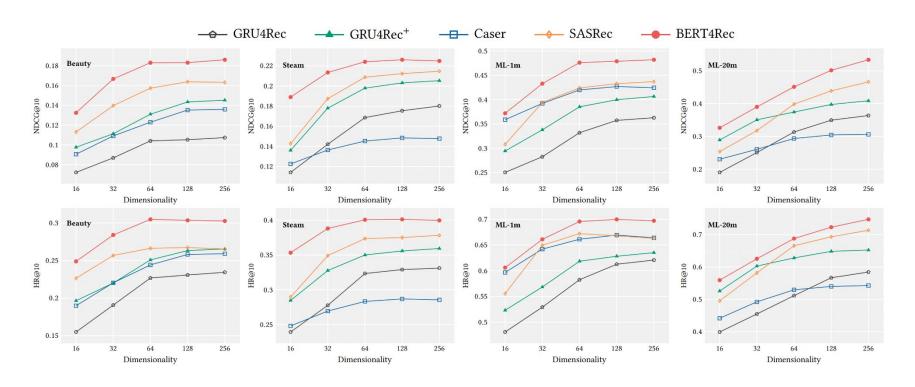


### Cloze Task or Bidirectional Encoding?

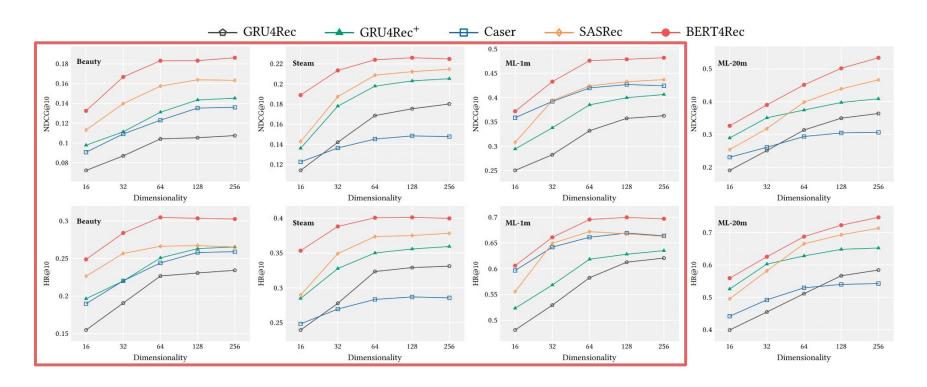
Improvements with 1 Mask (HR@10, NDCG@10, MRR)



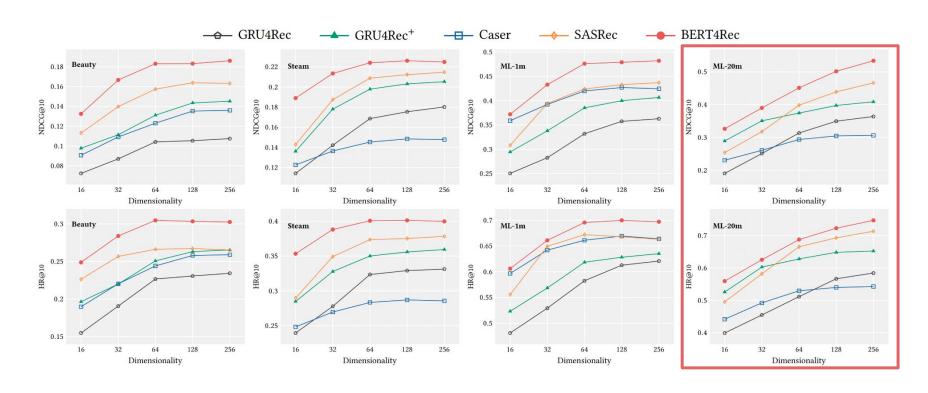
### Influence of Hidden Dimensionality d



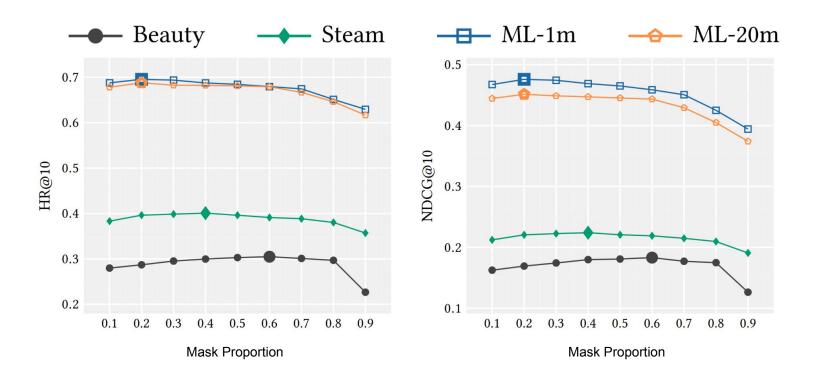
### Influence of Hidden Dimensionality d



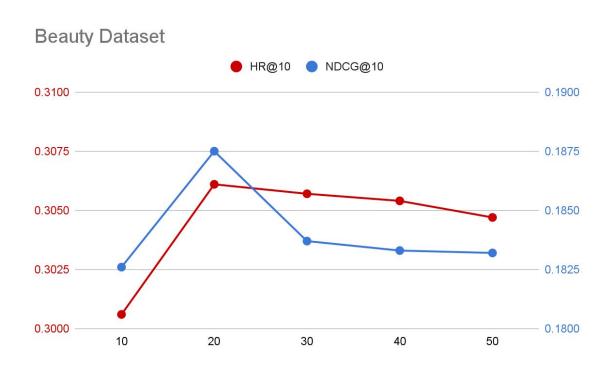
### Influence of Hidden Dimensionality d



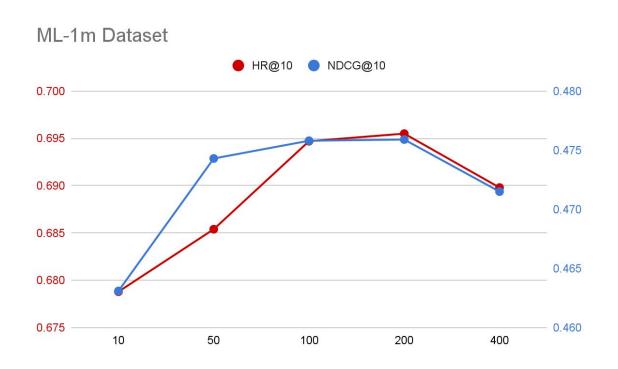
### Impact of Mask Proportion $\rho$



### Impact of Maximum Sequence Length N



### Impact of Maximum Sequence Length N



Ablation Study (NDCG@10)

L=2, h	= 2
w/o PE w/o PFF	N
w/o LN w/o RC w/o Dro	pout
1 layer 3 layers 4 layers	
1 head 4 heads	(h=1) $(h=4)$

8 heads (h = 8)

Architecture

	Dataset		
Beauty	Steam	ML-1m	
0.1832	0.2241	0.4759	
0.1741 0.1803	0.2060 0.2137	$0.2155 \downarrow \\ 0.4544$	
$0.1642 \downarrow \\ 0.1619 \downarrow \\ 0.1658$	0.2058 0.2193 0.2185	0.4334 0.4643 0.4553	
0.1782 <b>0.1859</b> <b>0.1834</b>	0.2122 <b>0.2262</b> <b>0.2279</b>	0.4412 <b>0.4864</b> <b>0.4898</b>	
<b>0.1853</b> 0.1830 0.1823	0.2187 <b>0.2245</b> <b>0.2248</b>	0.4568 <b>0.4770</b> 0.4743	

ML-20m

0.4513

 $0.2867 \downarrow$ 

0.4296

0.4186

0.4483

0.4471

0.4238

0.4661

0.4732

0.4402

0.4520

Ablation Study (NDCG@10)

L=2,h=2
w/o PE w/o PFFN
w/o LN w/o RC

w/o Dropout

1 layer (L=1)

3 layers (L=3)

4 layers (L=4)

1 head (h = 1)

4 heads (h = 4)

8 heads (h = 8)

Architecture

### Beauty 0.1832 0.1741 0.1803 $0.1642 \downarrow$ 0.1619

0.1658

0.1782

0.1859

0.1834

0.1853

0.1830

0.1823

0.2122

0.2262

0.2279

0.2187

0.2245

0.2248

$$0.4759$$
 $0.2155 \downarrow$ 
 $0.4544$ 
 $0.4334$ 
 $0.4643$ 
 $0.4553$ 

Dataset

ML-1m

0.4412

0.4864

0.4898

0.4568

0.4770

0.4743

$$\begin{array}{cccc}
0.2155 \downarrow & 0.2867 \downarrow \\
0.4544 & 0.4296 \\
\hline
0.4334 & 0.4186 \\
0.4643 & 0.4483 \\
0.4553 & 0.4471
\end{array}$$

ML-20m

0.4513

0.4296

0.4186

0.4483

0.4471

0.4238

0.4661

0.4732

0.4402

0.4520

Ablation 3	Study
(NDCG@	<u> </u>

# L = 2, h = 2w/o PE w/o PFFN w/o LN w/o RC w/o Dropout 1 layer (L = 1)3 layers (L = 3)

4 layers (L=4)

1 head (h = 1)

4 heads (h = 4)

8 heads (h = 8)

Architecture

I	Beauty	Steam
	0.1832	0.2241
	0.1741 0.1803	0.2060 0.2137
	).1642↓ ).1619↓ 0.1658	0.2058 0.2193 0.2185
(	0.1782 <b>0.1859</b>	0.2122 <b>0.2262</b>

0.2279

0.2187

0.2245

0.2248

0.1834

0.1853

0.1830

0.1823

Dataset

ML-1m

0.4759

0.2155

0.4544

0.4334

0.4643

0.4553

0.4412

0.4864

0.4898

0.4568

0.4770

0.4743

ML-20m

0.4513

 $0.2867 \downarrow$ 

0.4296

0.4186

0.4483

0.4471

0.4238

0.4661

0.4732

0.4402

0.4520

Ablation Study
(NDCG@10)

### L = 2, h = 2w/o PE w/o PFFN w/o LN w/o RC w/o Dropout 1 layer (L=1)3 layers (L=3)4 layers (L=4)

1 head (h = 1)

4 heads (h = 4)

8 heads (h = 8)

Architecture

Beauty	Steam	ML-1m
0.1832	0.2241	0.4759
0.1741 0.1803	0.2060 0.2137	0.2155↓ 0.4544
$0.1642 \downarrow \\ 0.1619 \downarrow \\ 0.1658$	0.2058 0.2193 0.2185	0.4334 0.4643 0.4553
0.1782 <b>0.1859</b> <b>0.1834</b>	0.2122 <b>0.2262</b> <b>0.2279</b>	0.4412 <b>0.4864</b> <b>0.4898</b>
<b>0.1853</b> 0.1830 0.1823	0.2187 <b>0.2245</b> <b>0.2248</b>	0.4568 <b>0.4770</b> 0.4743

Dataset

ML-20m

0.4513

 $0.2867 \downarrow$ 

0.4296

0.4186

0.4483

0.4471

0.4238

0.4661

0.4732

0.4402

0.4520

Ablation Study
(NDCG@10)

## L = 2, h = 2 w/o PE w/o PFFN

Architecture

### 0.1832 0.1741 0.1803

Steam

$$0.2155 \downarrow \\ 0.4544 \\ \hline 0.4334$$

ML-1m

0.4759

ML-20m

0.4513

0.2867

0.4296

0.4186

0.4520

0.4550

Dataset

w/o LN  
w/o RC  
w/o Dropout  
1 layer 
$$(L = 1)$$
  
3 layers  $(L = 3)$ 

4 layers (L=4)

1 head (h = 1)

4 heads (h = 4)

8 heads (h = 8)

0.1853

0.1830

0.1823

Beauty

0.2279

0.2187

0.2245

0.2248

0.2058

0.4770

Ablation \$	Study
(NDCG@	(01 ( <u>0</u>

# L = 2, h = 2 w/o PE w/o PFFN w/o LN

Architecture

$$0.1832$$
 $0.1741$ 
 $0.1803$ 
 $0.1642 \downarrow$ 

Beauty

0.1619

0.1658

0.1853

0.1830

0.1823

Steam

0.2193

0.2185

0.2187

0.2245

0.2248

Dataset

$$0.2155 \downarrow \\
0.4544 \\
\hline
0.4334$$

0.4643

0.4553

0.4568

0.4770

0.4743

ML-1m

0.4759

ML-20m

0.4513

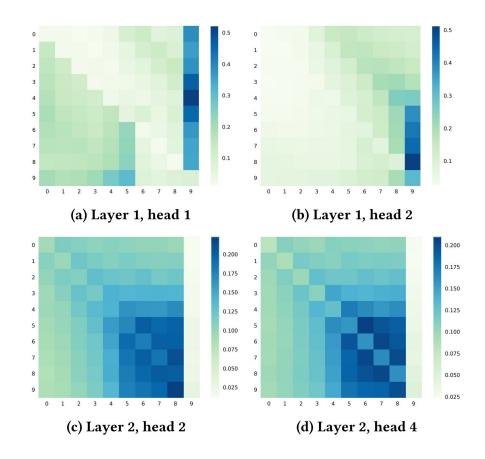
0.2867

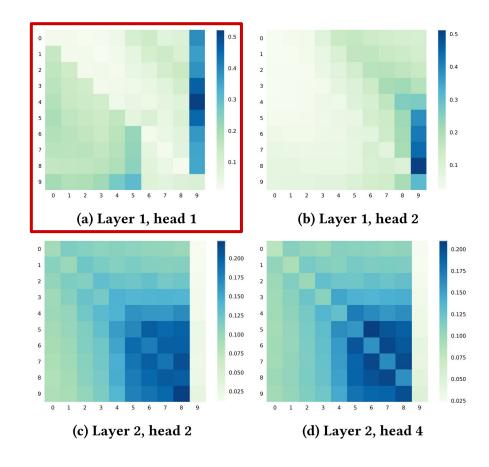
w/o RC  
w/o Dropout  
1 layer 
$$(L = 1)$$
  
3 layers  $(L = 3)$   
4 layers  $(L = 4)$ 

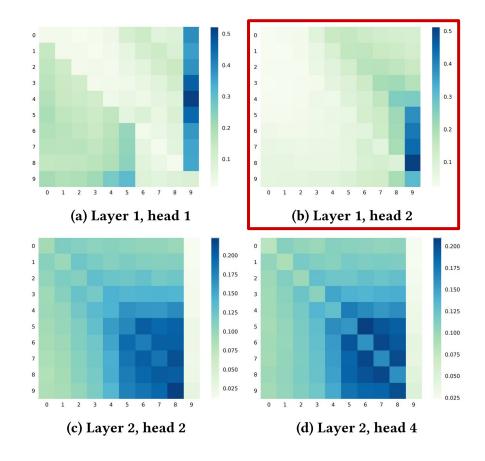
1 head (h = 1)

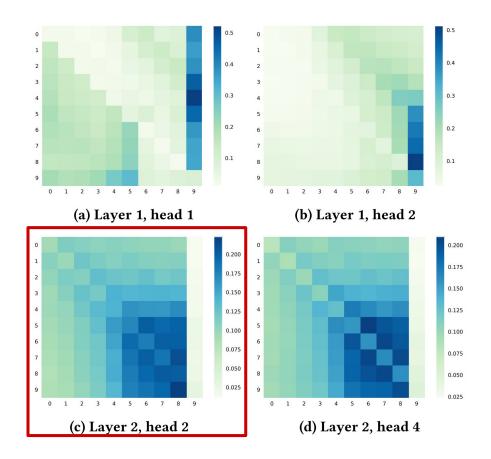
4 heads (h = 4)

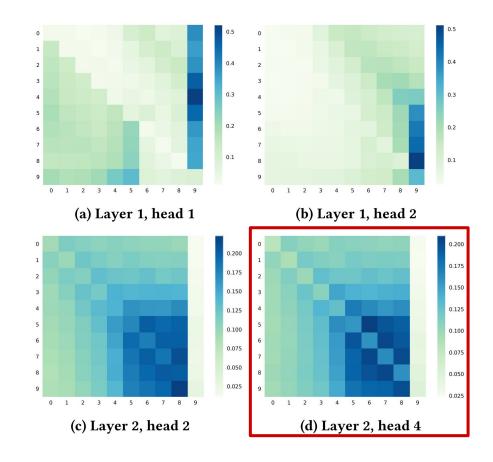
8 heads (h = 8)

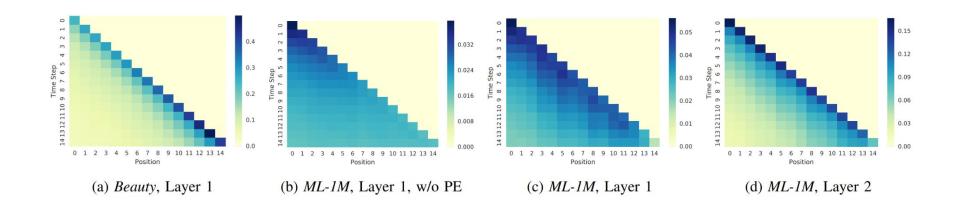






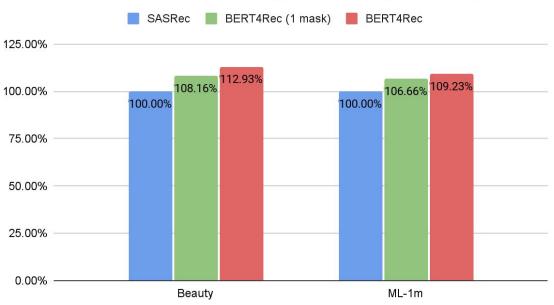






### Isolation from Cloze Task





Thanks for being here

### References

- <u>BERT4Rec: Sequential Recommendation with Bidirectional Encoder</u> <u>Representations from Transformer - Fei Sun</u>
- Self-Attentive Sequential Recommendation Wang-Cheng Kang