BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer

CIKM 2019

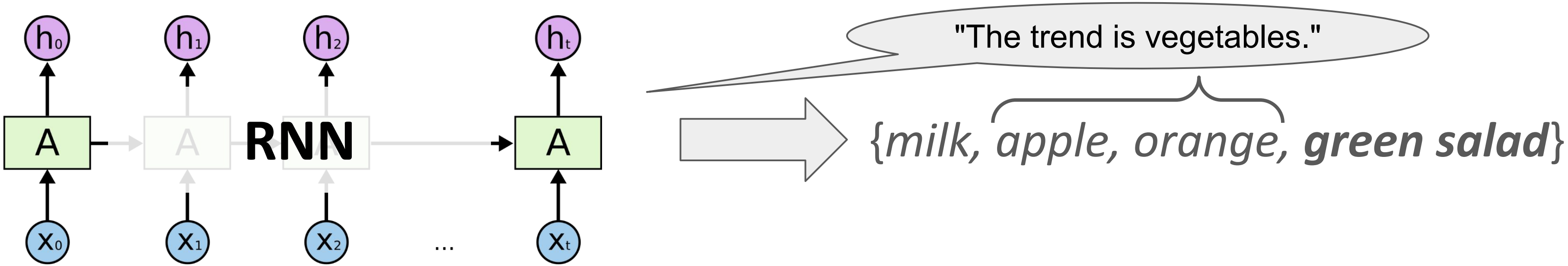
Authors: Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang

Presenter: Hyunwoo Jung

SEOUL NATIONAL UNIVERSITY

Motivation [1]

- A user purchased {*milk, apple, orange*}
  What is the next product the user will buy?



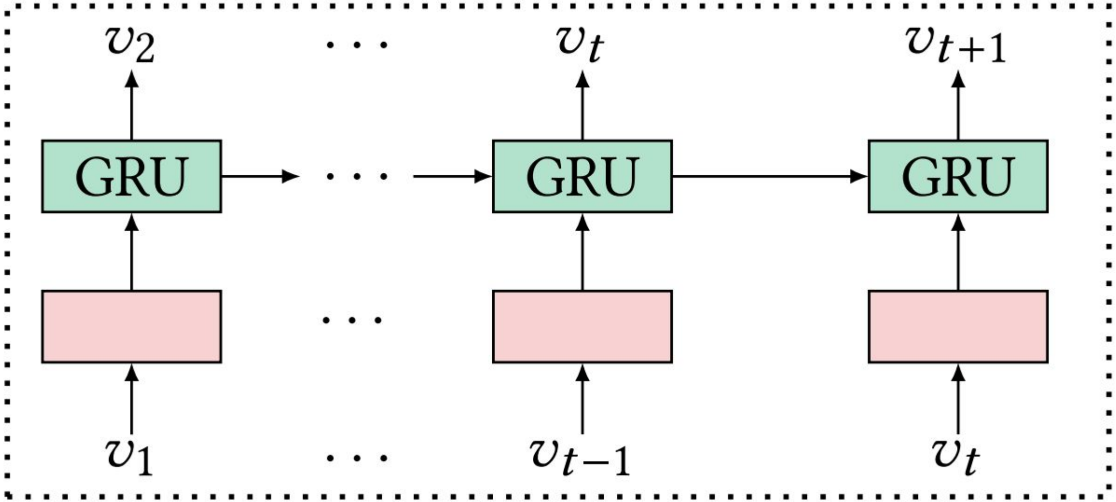
⇒ The answer is ***bread*** to eat with *milk*.

<u>Relevance between items is more important than the rigid order.</u>

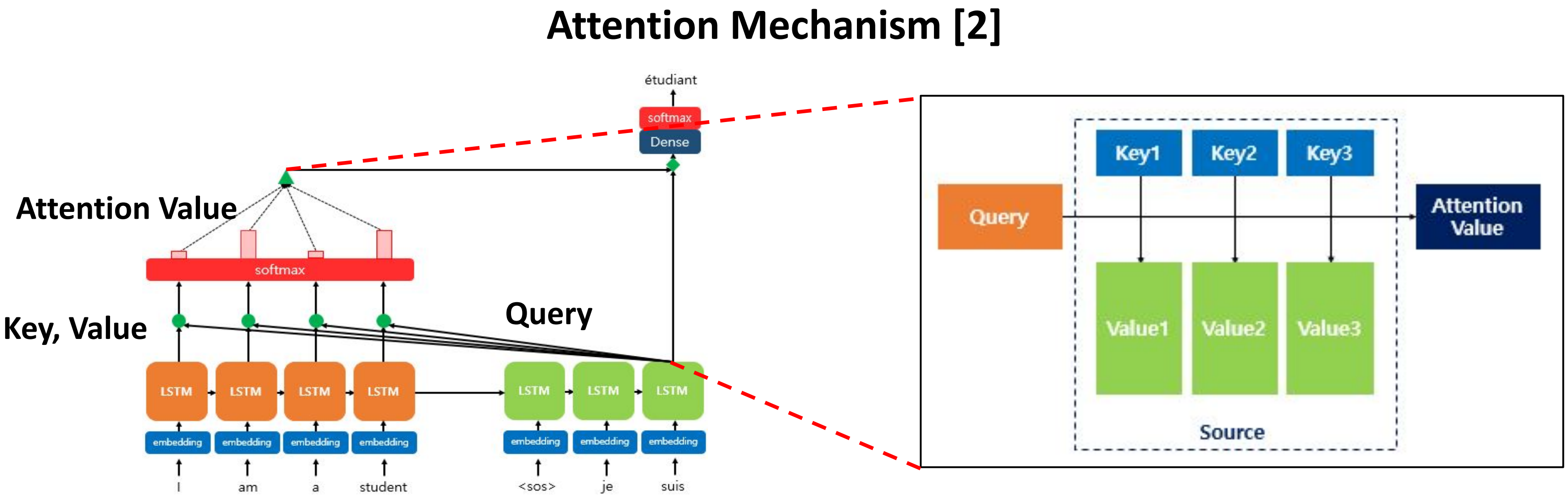
Prior Work 1: Unidirectional Models

- Originally proposed for NLP tasks.
- Unsuitable for noisy sequences such as transactions.

RNN based Model Architecture

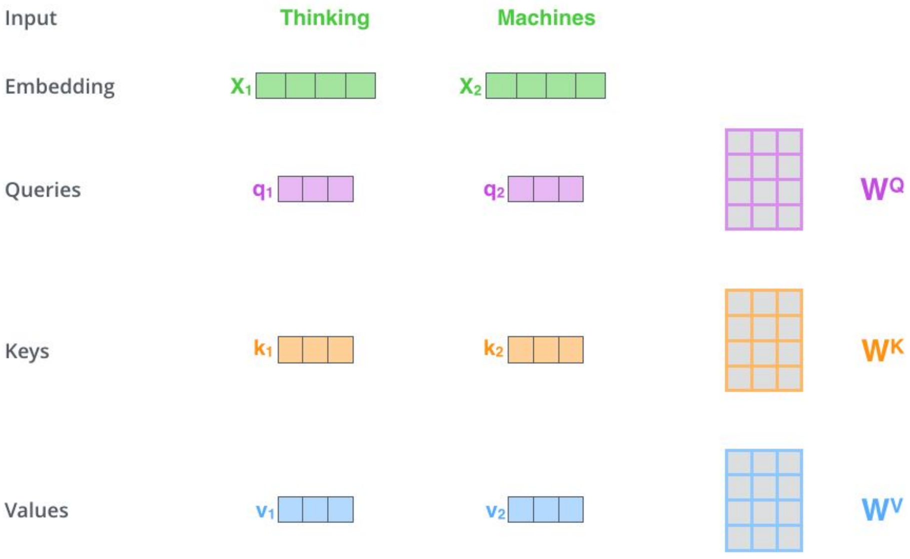


Attention Mechanism

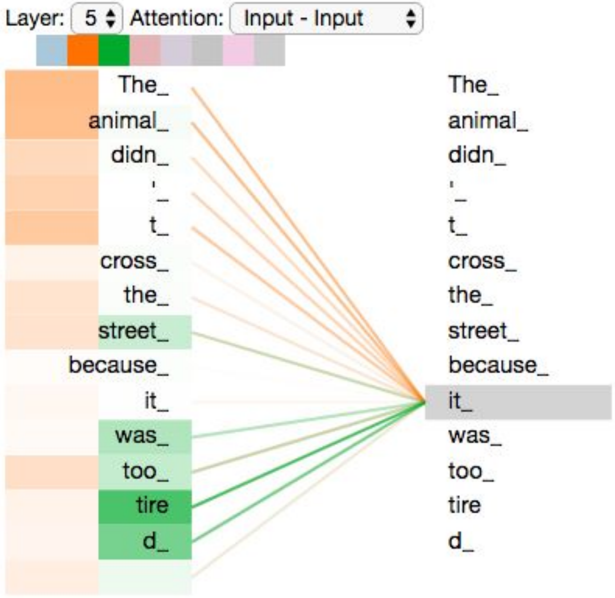


Self-Attention

Query, Key, Value from the Same Vector



Relevance between Words Captured by Self-Attention

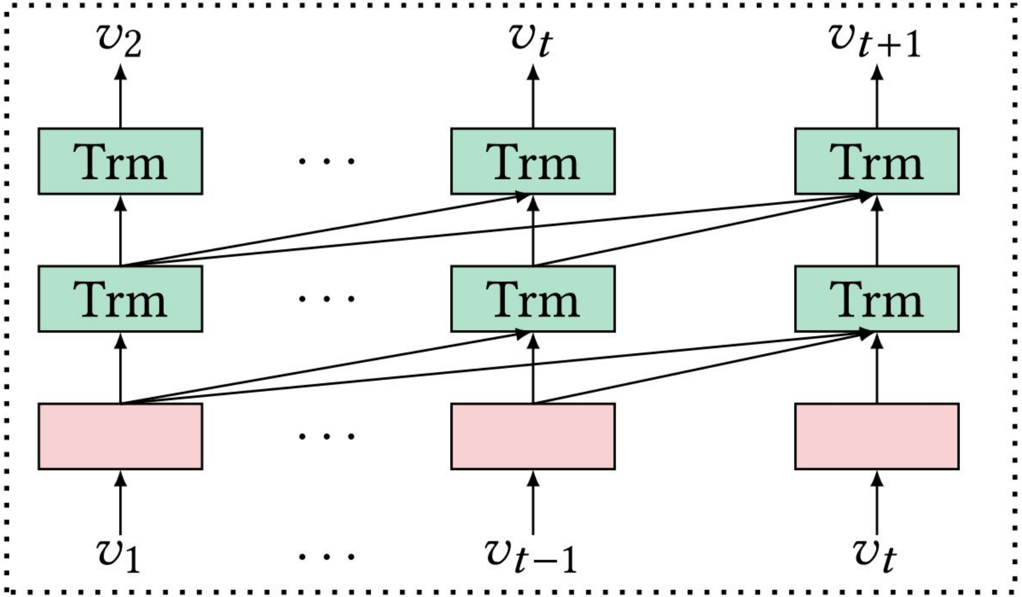


[3]

Prior Work 2: SASRec

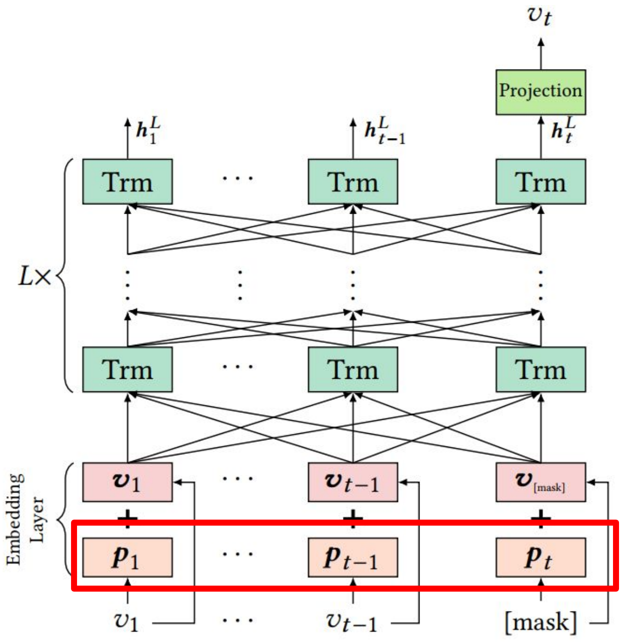
- Transformer-based unidirectional sequential recommendation model

SASRec Model Architecture



BERT4REC Architecture

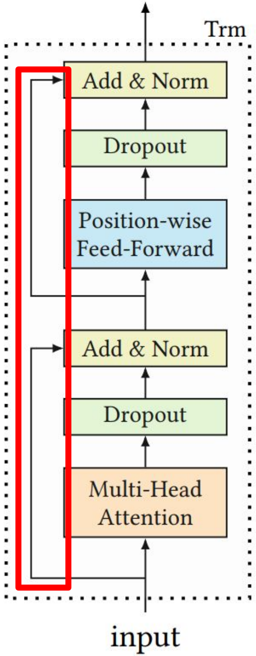
BERT4Rec Model Architecture



learnable positional embedding

Transformer Layer Architecture

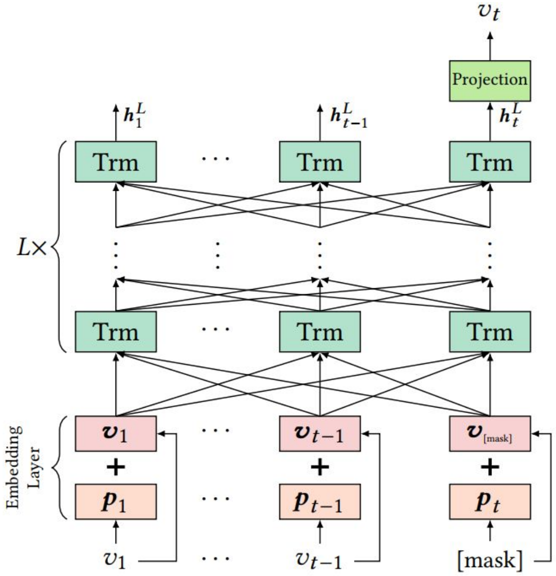
residual connection



Training

- Bidirectional architecture makes predicting the future trivial.
- Leverage ***Cloze*** task objective (*Masked Language Model*)

BERT4Rec Architecture



Cloze Task Objective

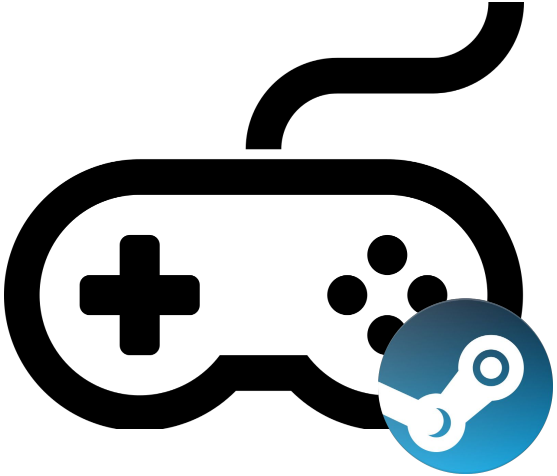
Input: $[v_1, v_2, v_3, v_4, v_5] \xrightarrow{\text{randomly mask}} [v_1, [\mathtt{mask}]_1, v_3, [\mathtt{mask}]_2, v_5]$

Labels: $[\mathtt{mask}]_1 = v_2, \quad [\mathtt{mask}]_2 = v_4$

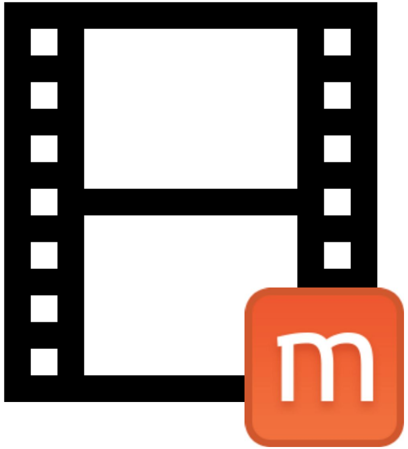
Datasets

Amazon Beauty
Product review

Steam
Game



MovieLens
Movie ratings



Metrics

- Hit Ratio (HR) ≡ Recall

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

- Normalized Discounted Cumulative Gain (NDCG)

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$$

- Mean Reciprocal Rank (MRR)

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}.$$

Baselines

Method	Approach
POP	popularity-based
BPR-MF	Matrix Factorization + pairwise ranking loss
NCF	MLP-based
FPMC	Matrix Factorization + First-order Markov Chain
GRU4Rec	GRU-based
GRU3Rec+	GRU-based
Caser	CNN + high-order Markov Chain
SASRec	left-to-right Transformer model

Performance Comparison

Datasets	Metric	POP	BPR-MF	NCF	FPMC	GRU4Rec	GRU4Rec$^+$	Caser	SASRec	BERT4Rec	Improv.
Beauty	HR@1	0.0077	0.0415	0.0407	0.0435	0.0402	0.0551	0.0475	0.0906	**0.0953**	5.19%
	HR@5	0.0392	0.1209	0.1305	0.1387	0.1315	0.1781	0.1625	0.1934	**0.2207**	14.12%
	HR@10	0.0762	0.1992	0.2142	0.2401	0.2343	0.2654	0.2590	0.2653	**0.3025**	14.02%
	NDCG@5	0.0230	0.0814	0.0855	0.0902	0.0812	0.1172	0.1050	0.1436	**0.1599**	11.35%
	NDCG@10	0.0349	0.1064	0.1124	0.1211	0.1074	0.1453	0.1360	0.1633	**0.1862**	14.02%
	MRR	0.0437	0.1006	0.1043	0.1056	0.1023	0.1299	0.1205	0.1536	**0.1701**	10.74%
Steam	HR@1	0.0159	0.0314	0.0246	0.0358	0.0574	0.0812	0.0495	0.0885	**0.0957**	8.14%
	HR@5	0.0805	0.1177	0.1203	0.1517	0.2171	0.2391	0.1766	0.2559	**0.2710**	5.90%
	HR@10	0.1389	0.1993	0.2169	0.2551	0.3313	0.3594	0.2870	0.3783	**0.4013**	6.08%
	NDCG@5	0.0477	0.0744	0.0717	0.0945	0.1370	0.1613	0.1131	0.1727	**0.1842**	6.66%
	NDCG@10	0.0665	0.1005	0.1026	0.1283	0.1802	0.2053	0.1484	0.2147	**0.2261**	5.31%
	MRR	0.0669	0.0942	0.0932	0.1139	0.1420	0.1757	0.1305	0.1874	**0.1949**	4.00%
ML-1m	HR@1	0.0141	0.0914	0.0397	0.1386	0.1583	0.2092	0.2194	0.2351	**0.2863**	21.78%
	HR@5	0.0715	0.2866	0.1932	0.4297	0.4673	0.5103	0.5353	0.5434	**0.5876**	8.13%
	HR@10	0.1358	0.4301	0.3477	0.5946	0.6207	0.6351	0.6692	0.6629	**0.6970**	4.15%
	NDCG@5	0.0416	0.1903	0.1146	0.2885	0.3196	0.3705	0.3832	0.3980	**0.4454**	11.91%
	NDCG@10	0.0621	0.2365	0.1640	0.3439	0.3627	0.4064	0.4268	0.4368	**0.4818**	10.32%
	MRR	0.0627	0.2009	0.1358	0.2891	0.3041	0.3462	0.3648	0.3790	**0.4254**	12.24%
ML-20m	HR@1	0.0221	0.0553	0.0231	0.1079	0.1459	0.2021	0.1232	0.2544	**0.3440**	35.22%
	HR@5	0.0805	0.2128	0.1358	0.3601	0.4657	0.5118	0.3804	0.5727	**0.6323**	10.41%
	HR@10	0.1378	0.3538	0.2922	0.5201	0.5844	0.6524	0.5427	0.7136	**0.7473**	4.72%
	NDCG@5	0.0511	0.1332	0.0771	0.2239	0.3090	0.3630	0.2538	0.4208	**0.4967**	18.04%
	NDCG@10	0.0695	0.1786	0.1271	0.2895	0.3637	0.4087	0.3062	0.4665	**0.5340**	14.47%
	MRR	0.0709	0.1503	0.1072	0.2273	0.2967	0.3476	0.2529	0.4026	**0.4785**	18.85%

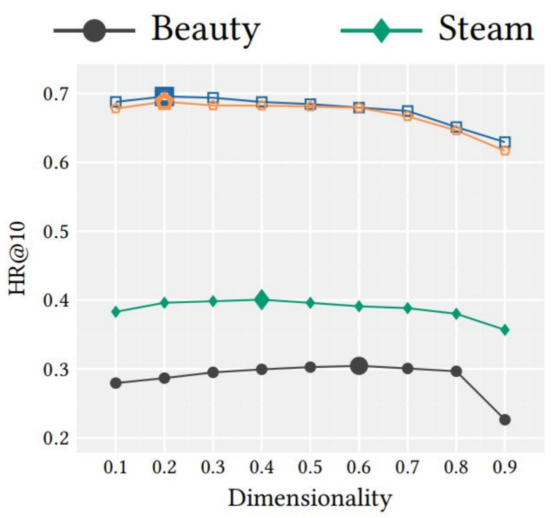
Gain of Cloze Objective

- Cloze objective improves the performances.
- The mask proportion should not be too small or big.

Performance with/without Cloze Objective

Model	Beauty			ML-1m		
	HR@10	NDCG@10	MRR	HR@10	NDCG@10	MRR
SASRec	0.2653	0.1633	0.1536	0.6629	0.4368	0.3790
BERT4Rec (1 mask)	0.2940	0.1769	0.1618	0.6869	0.4696	0.4127
BERT4Rec	0.3025	0.1862	0.1701	0.6970	0.4818	0.4254

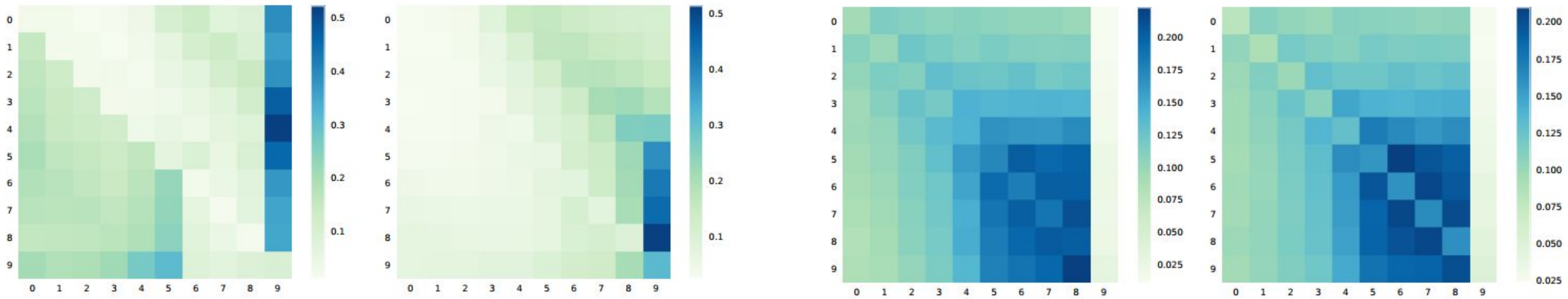
Performance with Different Mask Proportion



Effect of Bidirectional Model Architecture

- Attention varies across different heads/layers.
- BERT4Rec can attend on the items at both sides.

Heat-maps of Average Attention Weights on Amazon Beauty Dataset



(a) Layer 1, head 1

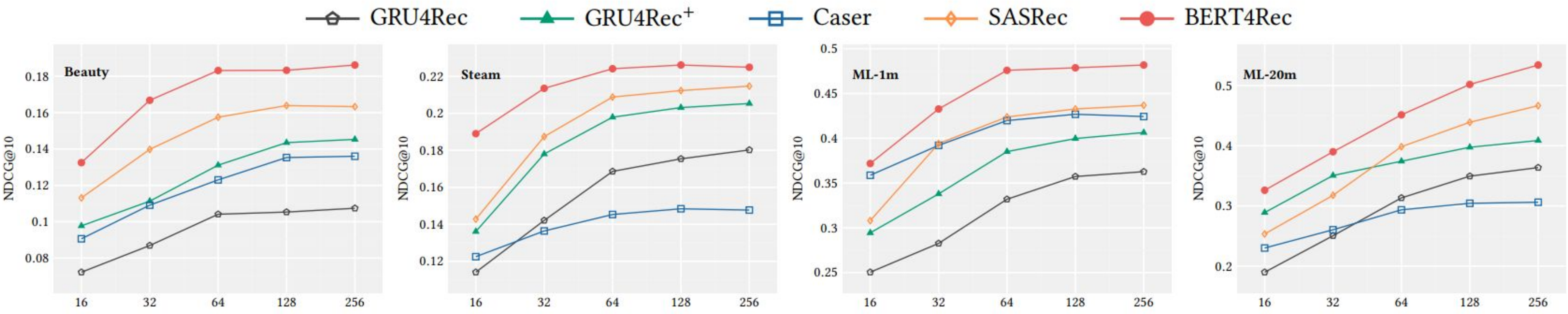
(b) Layer 1, head 2

(c) Layer 2, head 2

(d) Layer 2, head 4

Effect of the Hidden Dimensionality

- The performance converge as the dimensionality increases.



Impact of Maximum Sequence Length

- A user's behavior is affected by
  - more recent items (short sequence)
  - less recent items (long sequence)

			10	20	30	40	50
		#samples/s	5504	3256	2284	1776	1441
short sequence	Beauty	HR@10	0.3006	0.3061	0.3057	0.3054	0.3047
		NDCG@10	0.1826	0.1875	0.1837	0.1833	0.1832
			10	50	100	200	400
		#samples/s	14255	8890	5711	2918	1213
long sequence	ML-1m	HR@10	0.6788	0.6854	0.6947	0.6955	0.6898
		NDCG@10	0.4631	0.4743	0.4758	0.4759	0.4715

Summary

- Deep bidirectional self-attention architecture shows high performance on sequential recommendation.

- Cloze task improves the performance.

References

- [1] Attention-Based Transactional Context Embedding for Next-Item Recommendation, AAAI '18
- [2] https://wikidocs.net/22893
- [3] https://jalammar.github.io/illustrated-transformer/

Thank You
Q & A