Your Softmax Needs CPR

in Sequential Recommendation.

Achieving around 20% Improvement by just Switching Your Output Softmax Layer!

To Copy, or not to Copy; That is a Critical Issue of the Output Softmax Layer in Neural Sequential Recommenders

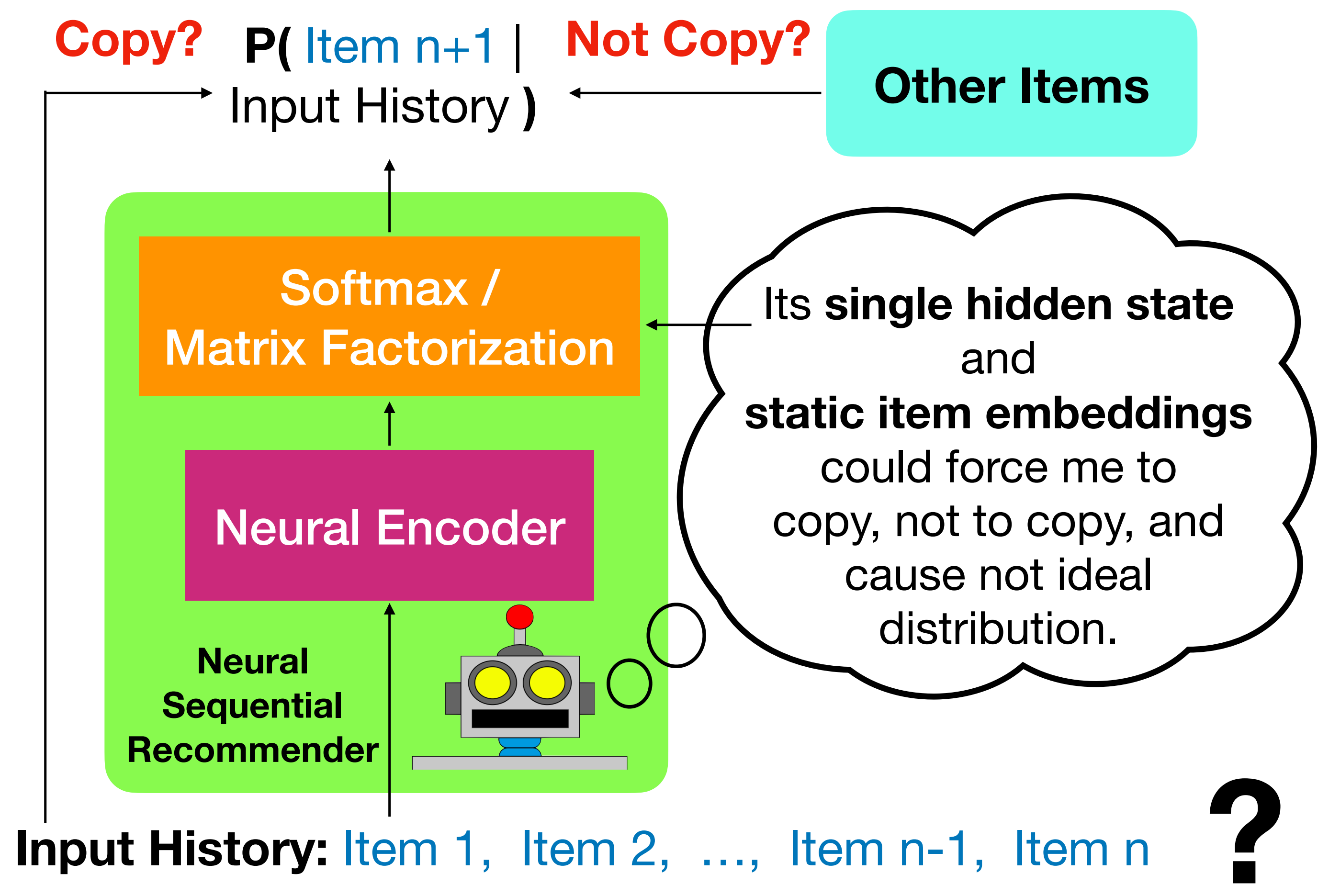
amazon alexa

Haw-Shiuan Chang, Nikhil Agarwal, and Andrew McCallum

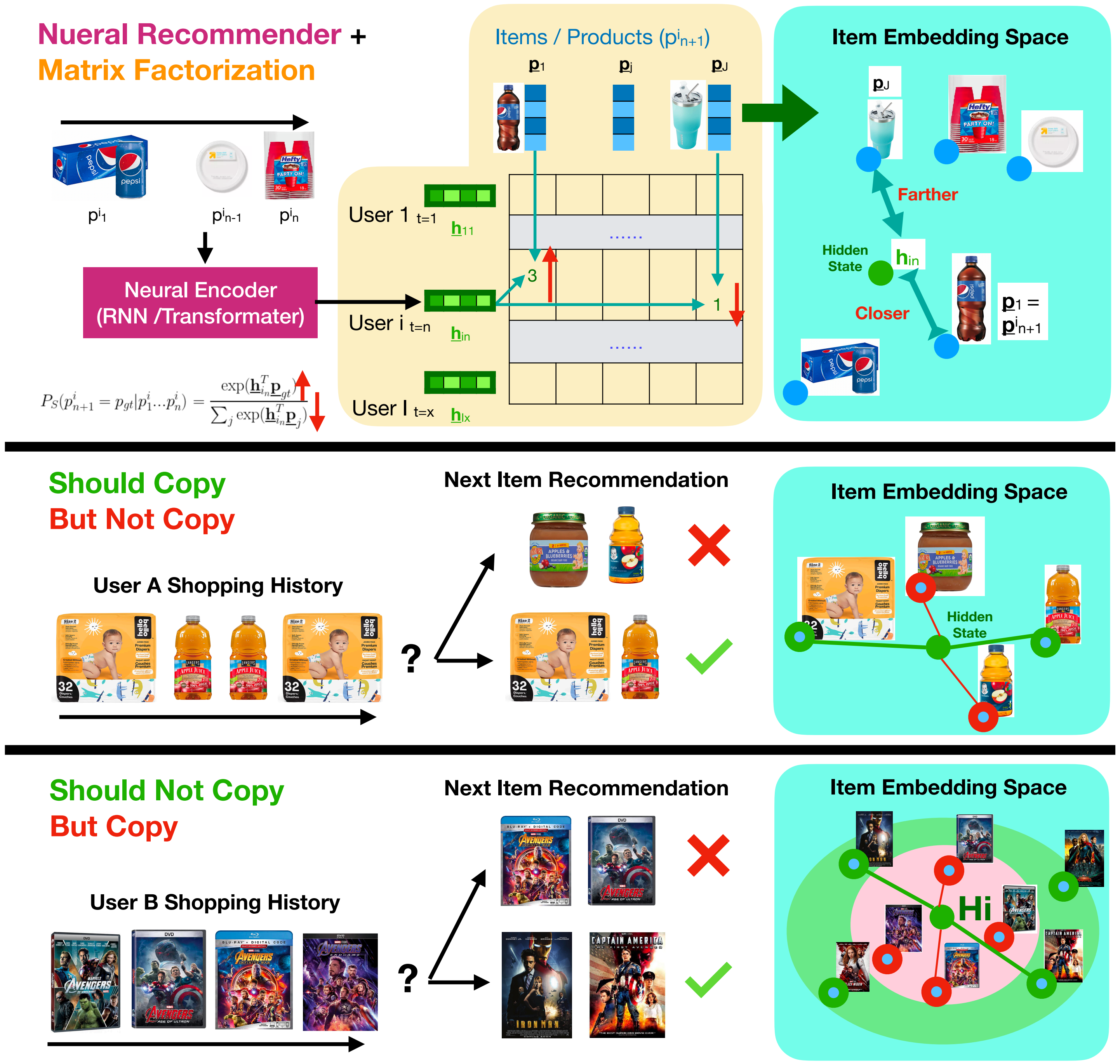
UMASS AMHERST

Introduction

- Did you ever experience this:
  - Training on a dataset with lots of repeated items, your STOA recommenders cannot learn to copy the items properly ?
  - Training on a dataset without repeated items, your STOA recommenders still keeps copying items ??????
- In this work, we find that the problem comes from the universally-used output softmax layer !!!!!!

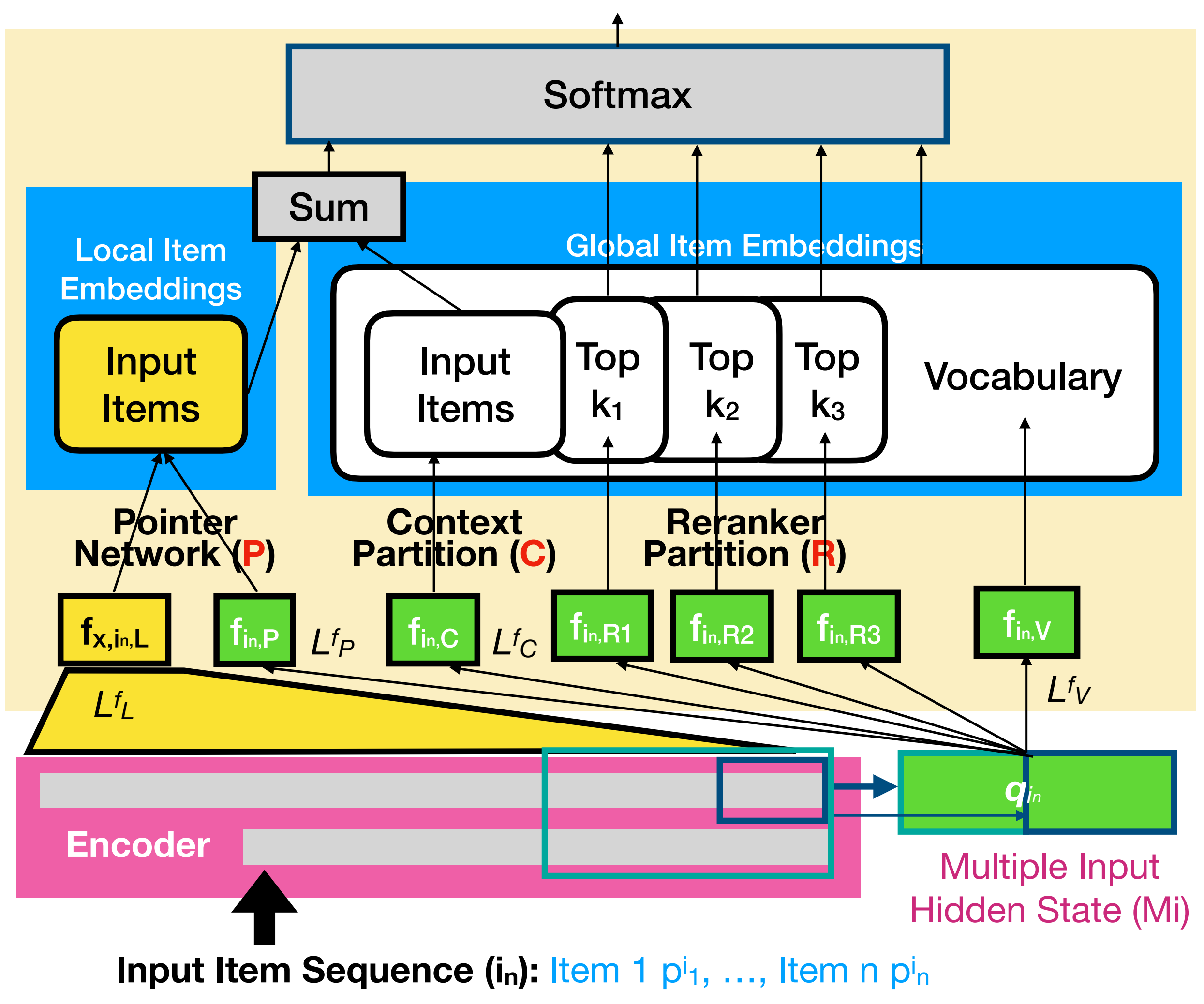

Copy?
P(Item n+1 | Input History)
Not Copy?
Other Items
Softmax / Matrix Factorization
Neural Encoder
Neural Sequential Recommender
Its single hidden state and static item embeddings could force me to copy, not to copy, and cause not ideal distribution.
Input History: Item 1, Item 2, …, Item n-1, Item n
?

Softmax Bottleneck Problems


Nueral Recommender + Matrix Factorization
p^i_1
p^i_{n-1}
p^i_n
Neural Encoder (RNN /Transformater)
Items / Products (p^i_{n+1})
$\mathbf{p}_1$
$\mathbf{p}_j$
$\mathbf{p}_J$
User 1 t=1 $\mathbf{h}_{11}$
User i t=n $\mathbf{h}_{in}$
User I t=x $\mathbf{h}_{Ix}$
……
3
1
$P_S(p^i_{n+1} = p_{gt}|p^i_1...p^i_n) = \frac{\exp(\mathbf{h}^T_{in}\mathbf{p}_{gt})}{\sum_j \exp(\mathbf{h}^T_{in}\mathbf{p}_j)}$
Item Embedding Space
$\mathbf{p}_J$
Farther
Hidden State
$\mathbf{h}_{in}$
Closer
$\mathbf{p}_1 = \mathbf{p}^i_{n+1}$
Should Copy
But Not Copy
Next Item Recommendation
User A Shopping History
?
Item Embedding Space
Hidden State
Should Not Copy
But Copy
Next Item Recommendation
User B Shopping History
?
Item Embedding Space
Hi

Solutions

1. Softmax (Original SASRec or GRU4Rec)
2. RepeatNet [2] (i.e., Pointer Network)
3. **MoS** (Mixture of Softmax) [3]
4. Softmax w/o Duplication [4]
5. Softmax + **C** (Context Partition)
6. Softmax + C**P** (Pointer Network)
7. Softmax + CP**R** (Reranker Partition)
8. Softmax + CPR:k_1,k_2,k_3 + **Mi** (Multiple Input Hidden State) [1]



Experiments

		Amazon-2014						MovieLens				Twitch-100k		Yelp-2018	
		Beauty		Books		Video Games		10m		1m					
		NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR
SASRec	Softmax	1.16	2.19	3.30	5.81	4.12	7.97	15.72	26.67	16.75	29.45	8.41	15.51	1.66	3.36
	Softmax + Mi	1.18	2.20	3.23	5.77	3.79	7.48	15.80	26.69	16.67	29.06	8.08	15.03	1.67	3.36
	Softmax + C	1.41	2.41	3.83	6.46	4.41	8.27	19.12	31.13	20.70	34.19	9.14	16.39	1.94	3.82
	Softmax + CP	**1.45**	**2.52**	3.94	6.71	4.54	8.59	18.62	30.51	20.69	**34.67**	**9.45**	**16.93**	2.04	3.91
	Softmax + CPR:100	1.38	2.42	4.15	6.89	4.57	**8.69**	**19.32**	**31.32**	20.79	34.25	9.11	15.94	**2.22**	4.24
	Softmax + CPR:100 + Mi	1.37	2.41	**4.30**	**7.20**	**4.47**	8.40	18.90	30.73	**20.82**	34.49	9.06	15.91	2.21	4.24
	Softmax + CPR:20,100,500 + Mi	1.39	2.43	3.93	6.60	4.46	8.58	19.19	30.93	20.48	33.61	8.58	14.88	2.20	**4.27**
	Mixture of Softmax (MoS)	1.19	2.24	3.24	5.75	3.74	7.35	15.88	26.82	17.05	29.83	8.17	15.19	1.69	3.42
	Softmax w/o Duplication [22]	1.34	2.42	3.73	6.27	4.42	8.35	18.35	30.19	20.06	33.81	9.01	16.13	1.85	3.64
GRU4Rec	Softmax	1.43	2.67	3.09	5.70	4.45	8.64	14.19	24.17	16.05	28.03	8.36	15.55	1.68	3.42
	Softmax + Mi	1.47	2.69	3.30	5.92	4.58	8.79	14.58	25.04	16.55	28.94	8.03	14.98	1.76	3.52
	Softmax + C	1.59	2.88	3.97	6.66	4.95	9.36	17.78	29.24	20.01	32.86	9.25	16.50	2.02	3.92
	Softmax + CP	1.61	2.94	4.07	6.83	**5.10**	9.41	17.46	28.64	19.63	32.91	9.14	16.09	2.00	3.85
	Softmax + CPR:100	1.78	**3.22**	4.28	7.06	5.05	**9.49**	17.78	29.01	20.35	33.73	9.04	15.82	2.27	4.35
	Softmax + CPR:100 + Mi	1.72	3.15	**4.42**	**7.23**	5.07	9.43	**18.09**	**29.43**	**21.00**	**34.52**	**9.32**	**16.20**	**2.37**	**4.51**
	Softmax + CPR:20,100,500 + Mi	**1.73**	3.11	4.37	7.14	5.02	9.33	17.87	29.09	20.44	33.63	8.80	15.20	2.31	4.39
	Mixture of Softmax (MoS)	1.46	2.73	3.15	5.76	4.06	8.00	14.40	24.50	16.14	28.06	7.90	14.69	1.73	3.50
	Softmax w/o Duplication [22]	1.60	2.91	3.71	6.26	4.83	9.09	16.85	27.68	18.54	31.72	8.94	16.03	1.94	3.80
RepeatNet	-	1.75	2.88	3.94	6.36	4.47	8.36	18.09	29.20	18.71	31.08	8.52	14.91	2.02	3.88

Table 2: We compare the test performance (%) of NDCG@10 and HR@10 in 7 datasets without duplicated items. C, P, R means context partition, pointer network, and reranker partition, respectively. 20,100,500 refers to $k_1 = 20$, $k_2 = 100$ and $k_3 = 500$; Mi means the multiple input hidden state enhancement. The best values given the same neural encoder are highlighted.

		Bridge to Algebra		Gowalla		Steam		Tmall-buy		Yoochoose-clicks	
		NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR
SASRec	Softmax	85.66	90.42	29.28	40.39	15.67	20.28	22.44	26.60	35.74	57.28
	Softmax + Mi	85.68	89.72	29.72	40.72	15.77	20.47	22.64	26.80	36.62	57.93
	Softmax + C	86.25	91.15	32.23	45.15	16.32	21.13	25.29	30.36	37.26	58.93
	Softmax + CP	85.60	89.75	32.88	45.68	16.30	21.05	25.58	30.50	37.43	59.02
	Softmax + CPR:100	87.40	91.09	33.03	46.17	16.43	21.31	25.73	**30.70**	37.79	59.15
	Softmax + CPR:100 + Mi	88.19	**92.19**	33.41	46.29	**16.48**	**21.39**	**25.74**	30.58	39.03	**59.69**
	Softmax + CPR:20,100,500 + Mi	**88.81**	92.07	**33.92**	**46.64**	16.34	21.15	25.58	30.22	**39.26**	59.68
	Mixture of Softmax (MoS)	84.77	89.78	29.74	40.87	15.90	20.49	23.07	27.28	35.59	57.07
	Softmax w/o Duplication [22]	80.13	82.89	3.92	7.00	4.89	9.15	4.29	6.28	17.00	27.84
GRU4Rec	Softmax	85.10	89.23	28.37	39.48	15.35	19.88	22.06	26.42	36.19	56.97
	Softmax + Mi	84.68	89.01	27.99	39.06	15.69	20.26	21.76	26.05	36.39	57.15
	Softmax + C	85.86	89.75	32.23	45.18	16.29	21.04	25.18	30.25	37.46	58.54
	Softmax + CP	86.24	91.06	32.48	45.43	16.32	21.06	25.45	30.36	37.90	58.76
	Softmax + CPR:100	88.56	**92.35**	33.01	46.08	16.36	21.15	**25.77**	30.34	38.35	59.15
	Softmax + CPR:100 + Mi	88.81	92.19	**33.22**	**46.09**	**16.49**	**21.35**	25.54	30.01	**38.72**	**59.42**
	Softmax + CPR:20,100,500 + Mi	**89.46**	92.29	33.18	45.93	16.41	21.19	25.72	**30.43**	38.54	59.20
	Mixture of Softmax (MoS)	86.11	90.30	27.91	38.60	15.89	20.41	21.50	25.75	36.39	56.97
	Softmax w/o Duplication [22]	79.06	81.67	3.93	7.05	4.65	8.72	4.24	6.32	16.80	27.44
RepeatNet	-	77.44	81.70	33.83	45.88	16.28	20.90	25.67	30.17	38.00	58.53

Table 3: The test performance (%) in 5 datasets with duplicated items. The notations are the same as Table 2.

1. The improvements of Softmax + CPR + Mi are consistent across 12 large datasets and hyperparameters
   A. 10% (4%-17%) in 5 datasets with duplicated items
   B. 24% (8%-39%) in 7 datasets **without** duplicated items
2. The best method increases the model size very slightly.
3. Softmax + C ≈ RepeatNet [2] ≈ Softmax w/o duplication [4]

Conclusions

1. Stop using softmax in your models!

2. Selection of the softmax layer is much more important than selection of the encoder on average in our experiments.

3. RepeatNet / pointer network improves the performance due to the softmax bottleneck instead of the attention mechanism.

Applications

1. Try Softmax-CPR in RecBole to achieve new STOA
2. Softmax-CPR also reduces hallucination in LLM
3. **Future work**: other matrix factorization models?

References

[1] Haw-Shiuan Chang*, Zonghai Yao*, Alolika Gon, Hong Yu, and Andrew McCallum. 2023. "Revisiting the Architectures like Pointer Networks to Efficiently Improve the Next Word Distribution, Summarization Factuality, and Beyond". In Findings of ACL 2023

[2] Pengjie Ren, Zhumin Chen, Jing Li, Zhaochun Ren, Jun Ma, and Maarten De Rijke. 2019.
Repeatnet: A repeat aware neural recommendation machine for sessionbased recommendation. In AAAI 2019

[3] Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. 2018. Breaking the Softmax Bottleneck: A High-Rank RNN Language Model. In ICLR 2018

[4] Ming Li, Ali Vardasbi, Andrew Yates, and Maarten de Rijke. 2023. "Repetition and Exploration in Sequential Recommendation". In SIGIR 2023

https://commons.wikimedia.org/wiki/File:Stop_hand-octogon-red.svg