Future Data Helps Training: Modeling Future Contexts for Session-based Recommendation

Fajie Yuan（Tencent）, Xiangnan He (National University of Singapore), Haochuan Jiang (University of Edinburgh), Guibing Guo (Northeastern University, China), {Jian Xiong, Zhezhao Xu, Yilin Xiong}（Tencent）

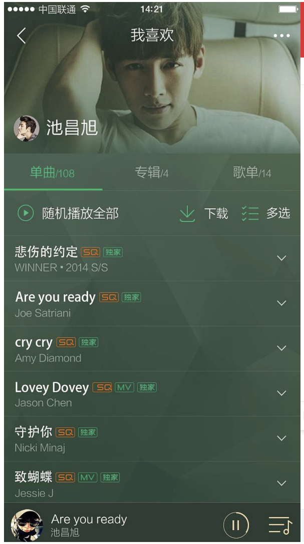
Future Data Helps Training:

- Session-based (aka Sequential) recommendation Apps:

  Short-videos (Tik Tok, Weishi, Kuaishou)

  Music (Tencent music, Yahoo! Music) & News

  Movie clips (You Tube, Netflix)



Future Data Helps Training:

- Related work:
- Markov chain: Long long ago
- RNN/LSTM: GRURec[1, 2]   2016-2018
- CNN : Caser [3], NextItNet [4] 2018-2019
- Attention : Transformer [5] 2018-2019

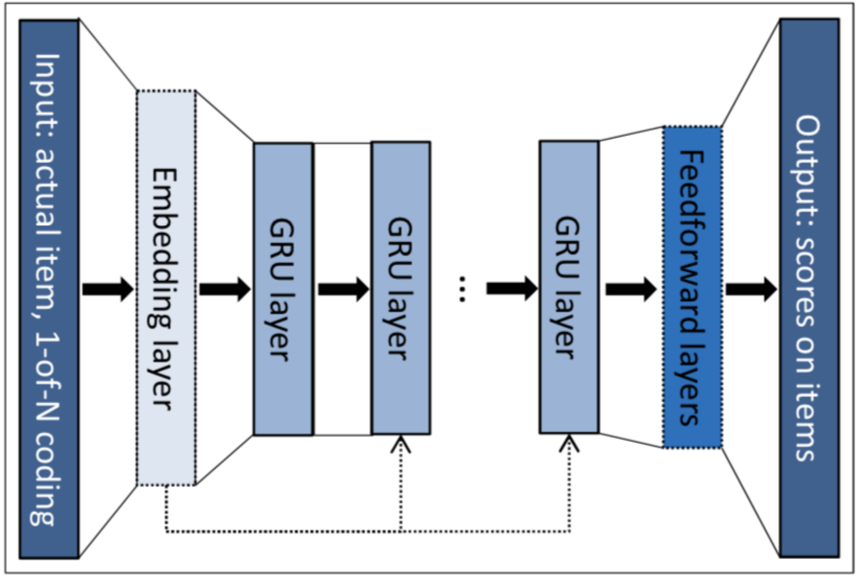
[1] session-based recommendations with recurrent neural networks. ICLR 2016
[2] Personalizing Session-based Recommendations with Hierarchical Recurrent Neural Networks. arXiv2017
[3] Personalized Top-N Sequential Recommendation via Convolutional Sequence Embedding. WSDM2018
[4] A Simple Convolutional Generative Network for Next Item Recommendation. WSDM2019
[5] Next Item Recommendation with Self-Attention. ICDM2018

Future Data Helps Training:

- Session-based rec: Top-n item recommendation
- Offline: NDCG, MRR, MAP, Pre@10,Rec@10
- Online: UV, VV, PV, CTR, DAU

Future Data Helps Training:

- RNN/LSTM: GRU4Rec[1], Improved GRU4Rec

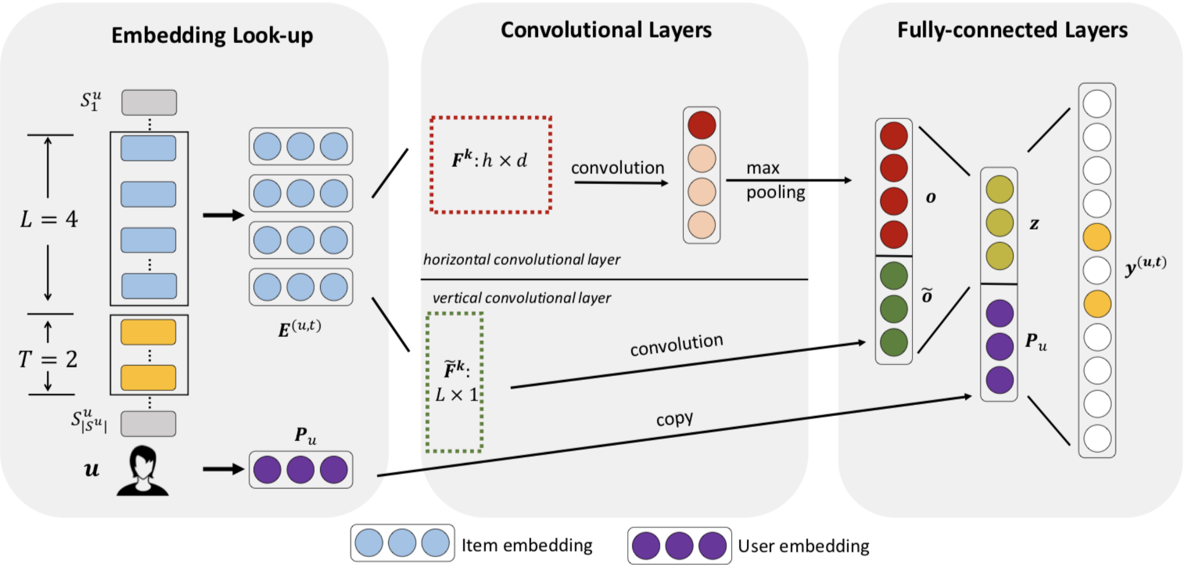


pros: good for modelling seq
cons: bad for utilizing GPU

[1] session-based recommendations with recurrent neural networks. ICLR 2016

Future Data Helps Training:

- CNN: Caser[1]

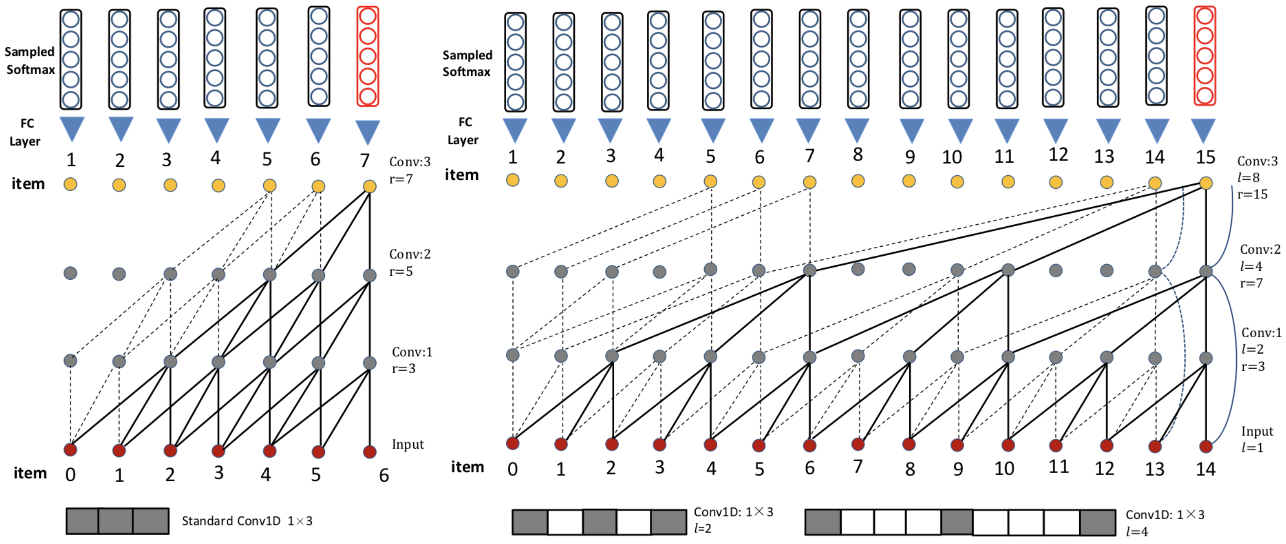


Pros: good for using GPU
Cons: max pooling loses some information, shallow layers

[1] Personalized Top-N Sequential Recommendation via Convolutional Sequence Embedding. WSDM2018

Future Data Helps Training:

- Dilated CNN: NextItNet[1]



(a) Standard CNN

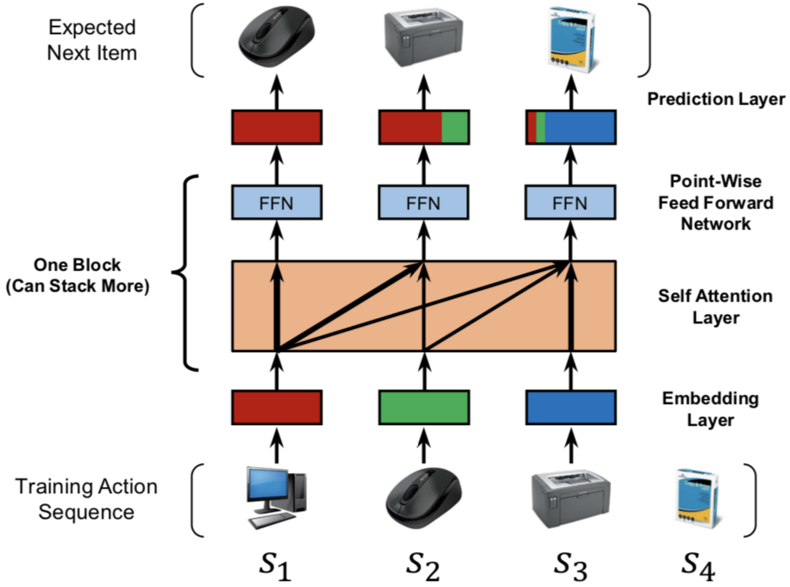
(b) Dilated CNN with 'Holes'

Pros:
CNN structure-model parallelism
Residual block: deeper & stronger
Dilated CNN: longer and better

[1] A Simple Convolutional Generative Network for Next Item Recommendation. WSDM2019

Future Data Helps Training:

- Attention[1]

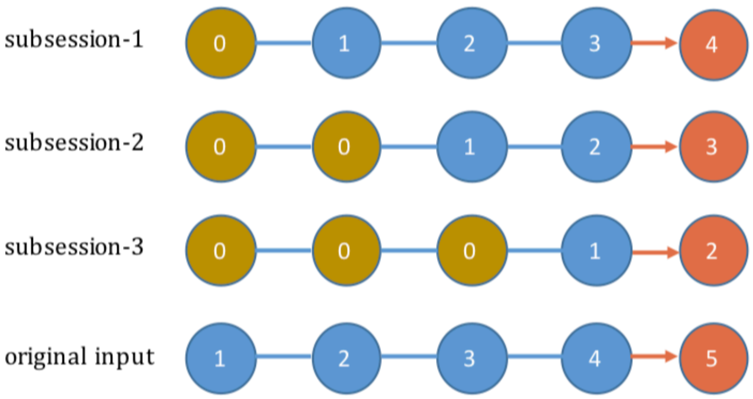


Pros: better for utilizing GPU
Cons: quadratic complexity, particularly for longer sequeces

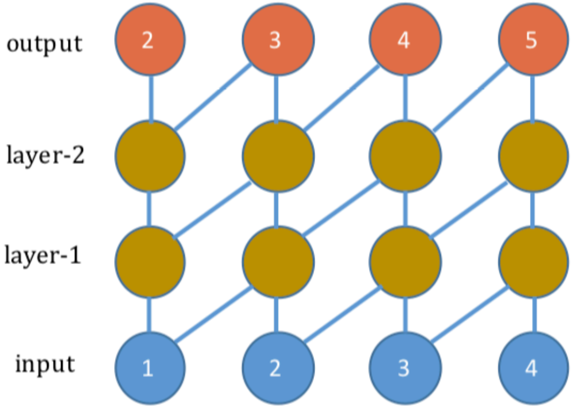
[1] Self-Attentive Sequential Recommendation. ICDM2018

Future Data Helps Training:

- Training Method (Left-to-Right-Style) for Long Sequences



(a) Data augmentation (DA).



(b) Autoregressive models (AR).

Future Data Helps Training:

Data augmentation & Autoregressive

$$Caser/GRURec : \underbrace{\{x_0, x_1, ..., x_{14}\}}_{input} \Rightarrow \underbrace{x_{15}}_{output}$$

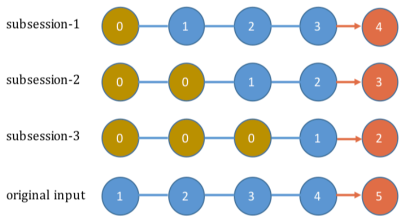
$$NextItNet : \underbrace{\{x_0, x_1, ..., x_{14}\}}_{input} \Rightarrow \underbrace{\{x_1, x_2, ..., x_{15}\}}_{output}$$

$$Caser/GRURec\ sub - session - 1 : \{x_{-1}, x_0, ..., x_{13}\} \Rightarrow x_{14}$$

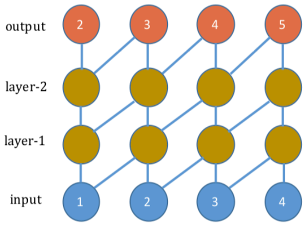
$$Caser/GRURec\ sub - session - 2 : \{x_{-1}, x_{-1}, ..., x_{12}\} \Rightarrow x_{13}$$

......

$$Caser/GRURec\ sub - session - 12 : \{x_{-1}, x_{-1}, ..., x_2\} \Rightarrow x_3$$



(a) Data augmentation (DA).



(b) Autoregressive models (AR).

No future data is used when modeling a prediction function during training!

Tencent 腾讯

Future Data Helps Training:

- Is a strict order necessary? Seems Not

My watching session in Tik Tok

Fine to me if changing the playing order

Future Data Helps Training:

- Is a strict order necessary?

My purchase session in Alibaba

phone-->phone case--> earphone--> screen protector

An alternative purchase session for me

phone--> screen protector --> earphone--> phone case

Also fine to me!

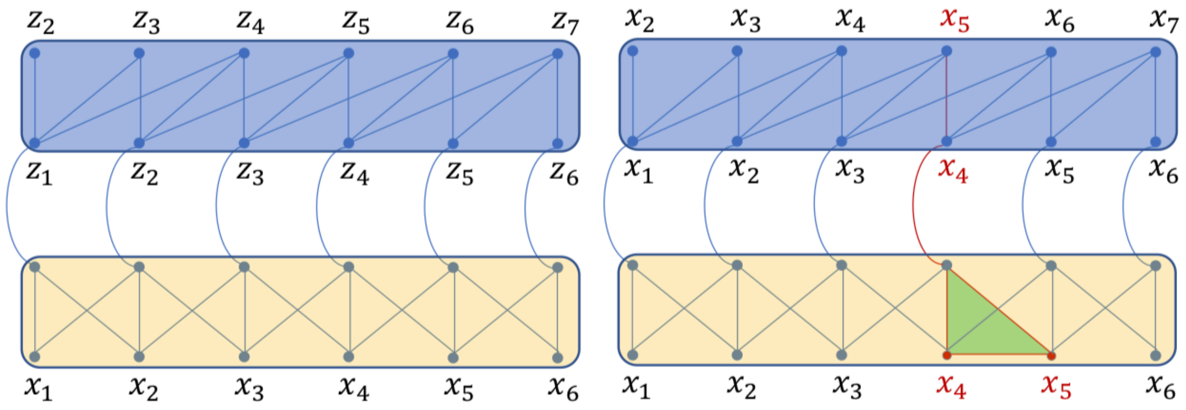
phone--> earphone--> screen protector --> phone case

Future Data Helps Training:

- Modeling two-side contexts

  Modeling Future interactions could help build better prediction function
  Allievate data sparsity

- Modeling two-side contexts straightly causes data leakage



(a) Typical seq2seq learning (√). (b) Seq2seq learning for SRS (×).

x_5 *follows* x_4
seen by the encoder

Future Data Helps Training:

- Other trivial methods:

$$NextItNet+: \underbrace{\{x_1, ..., x_{t-1}\}}_{input} \Rightarrow \underbrace{\{x_2, ..., x_t\}}_{output}$$

$$\underbrace{\{x_t, ..., x_2\}}_{input} \Rightarrow \underbrace{\{x_{t-1}, ..., x_1\}}_{output}$$

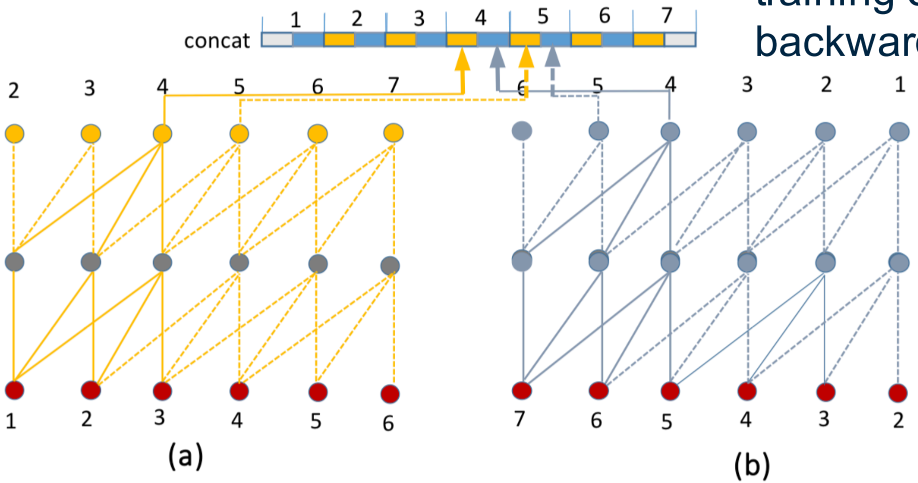
- Drawbacks

(1) Using the same set of parameters to model two side contexts is not accurate
(2) Modeling the left & right context separately is suboptimal, and has mutual interference

Future Data Helps Training:

- Other trivial methods:

two-way NextItNets (tNextItNets)



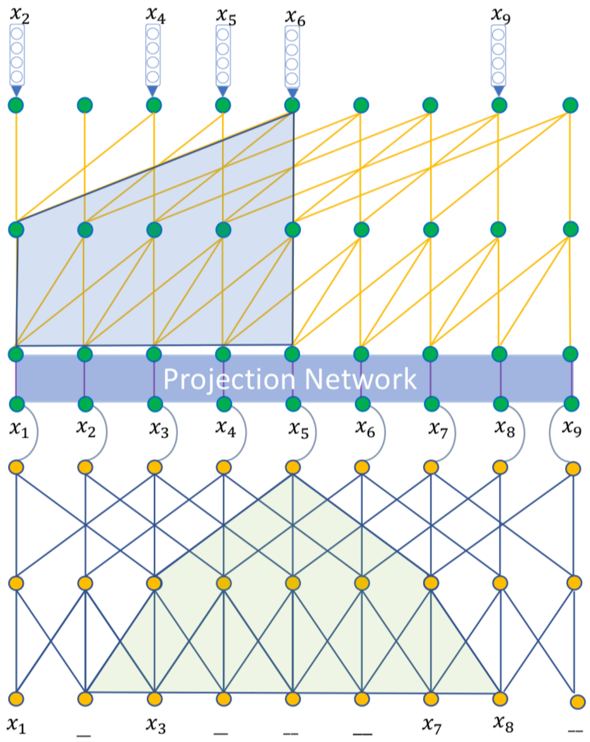
Forward NextItNet   Backward NextItNet

Drawbacks:

training & inference has discrepancies since backward network is useless during inference

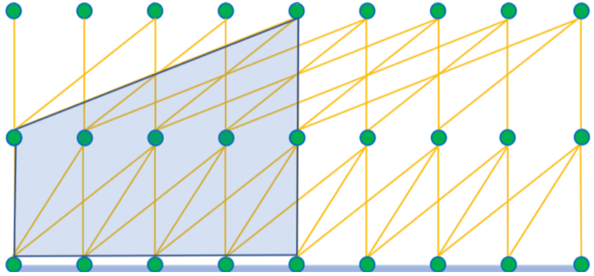
Future Data Helps Training:

- Our solution: GRec

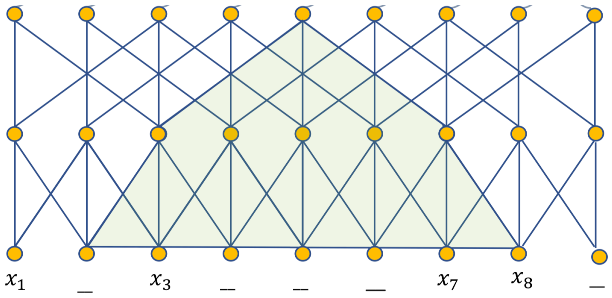


NextItNet-style Decoder

BERT-style Encoder



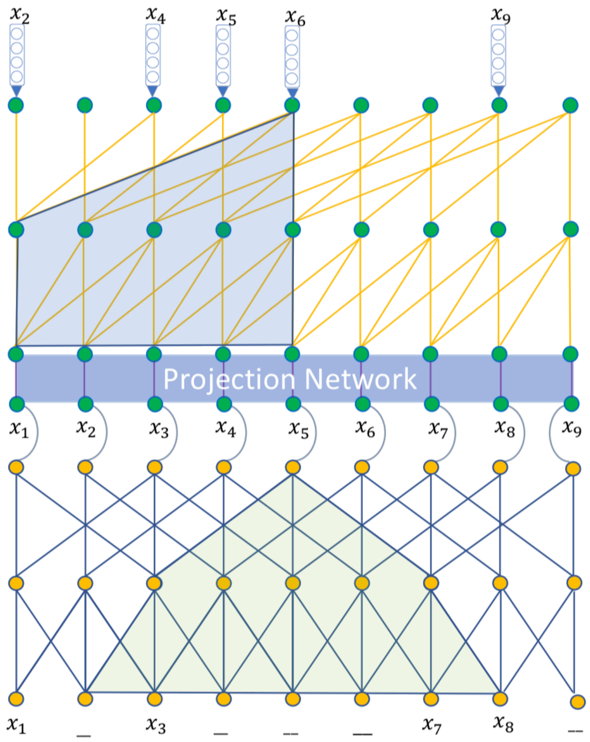
Causal CNN



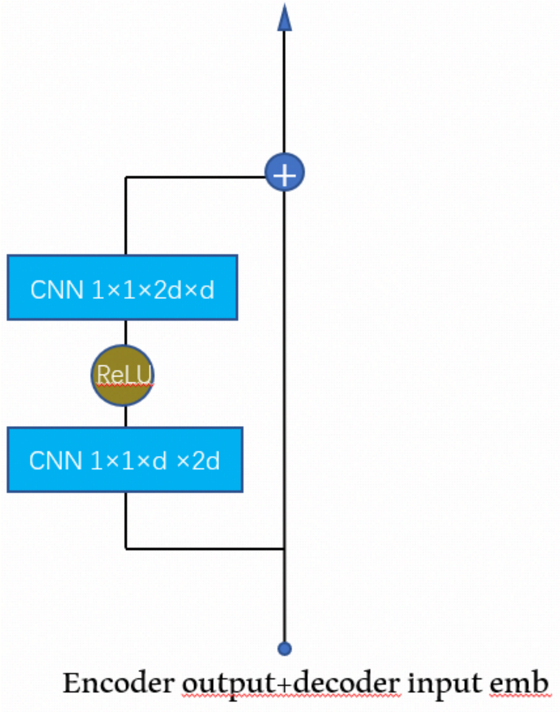
Non-causal CNN

Future Data Helps Training:

- Our solution: GRec



projector



Future Data Helps Training:

- GRec:

$$G(\mathcal{X};\Theta) = \sum_{x \in \mathcal{X}} \log p(x_{\triangle} | \tilde{x}; \Theta)$$

$$= \sum_{x \in \mathcal{X}} \log \prod_{i=1}^{m} p(\boxed{x_{\triangle_i}} | \boxed{x_{1:\triangle_{i-1}}}, \boxed{\tilde{x};} \Theta)$$

masked items

previous items

items without mask

Masked items are predicted given its previous items and other items without masking

Future Data Helps Training:

- Grec vs. NextItNet:

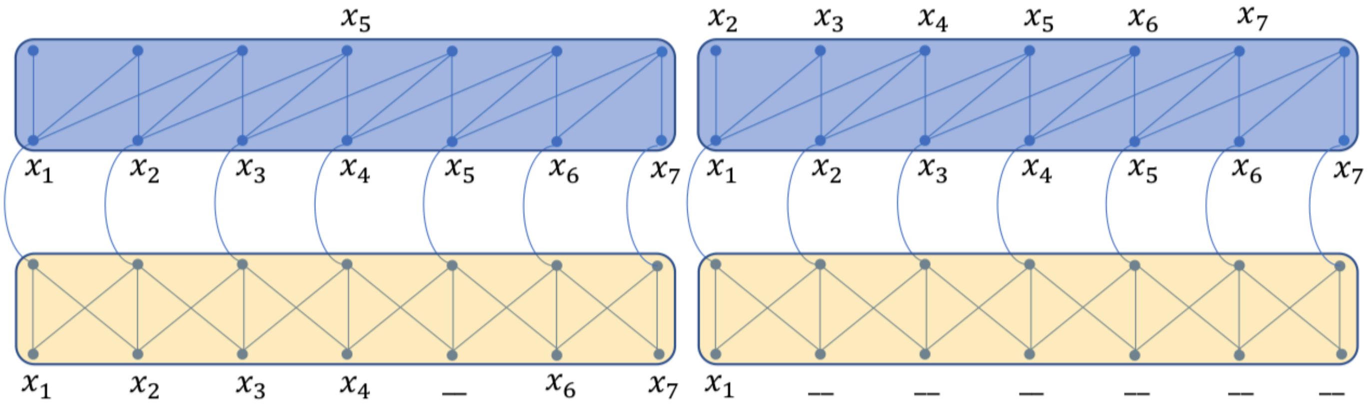
$$NextItNet : \underbrace{\{x_1, x_2, x_3, ..., x_7, x_8\}}_{decoder\ input} \Rightarrow \underbrace{\{x_2, x_3, x_4, ..., x_8, x_9\}}_{decoder\ output}$$

$$GRec : \underbrace{\{x_1, \_\_, x_3, \_\_, \_\_, \_\_, x_7, x_8, \_\_, \}}_{encoder\ input} + \underbrace{\{x_1, x_2, x_3, ..., x_9\}}_{decoder\ input}$$

$$\Rightarrow \underbrace{\{x_2, x_4, x_5, x_6, x_9\}}_{decoder\ output}$$

Future Data Helps Training:

- Connections:



(a) GRec with one masked item

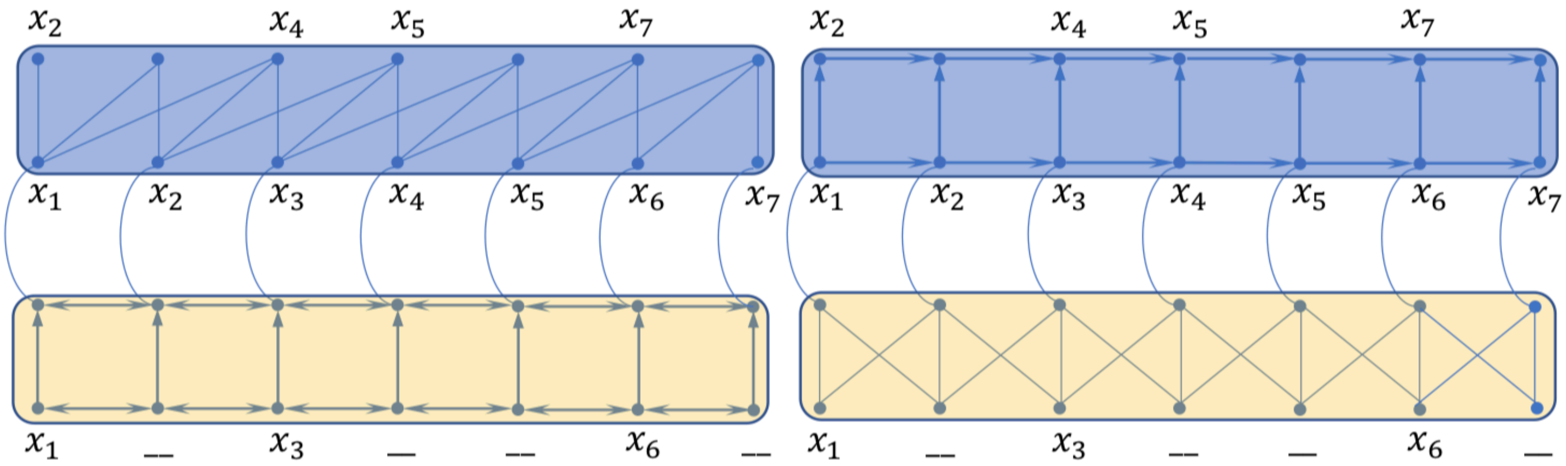
(b) GRec with t masked items.

Similar to BERT with a useless decoder

Similar to NextItNet, with a useless encoder

Future Data Helps Training:

- Generality：



(a) ReCd: BiRNN+ causal CNN. **(b) CeRd: non-causal CNN + RNN.**

Future Data Helps Training:

- Datasets:

DATA	*#actions*	*#sequences*	*#items*	*k*
TW10	9,986,953	1,048,575	65,997	10
ML30	25,368,155	858,160	18,273	30
ML100	25,240,741	300,624	18,226	100

Tencent腾讯 Future Data Helps Training:

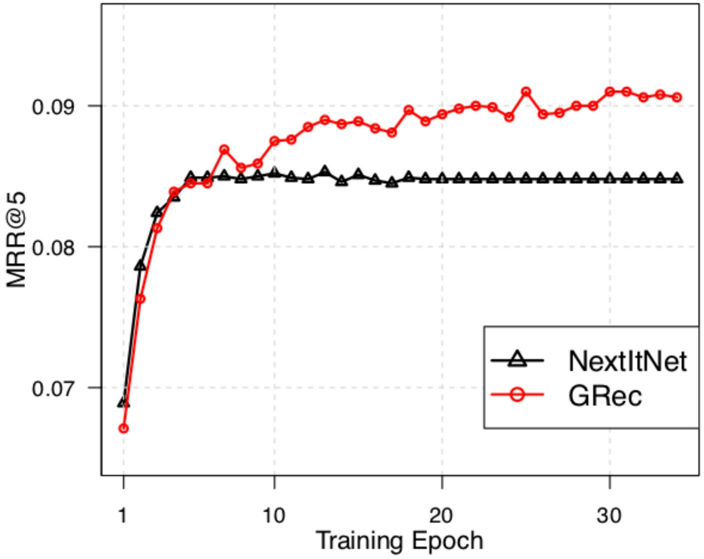
- Results compared with baselines:

Table 2: Accuracy comparison. MostPop returns item lists ranked by popularity. For each measure, the best result is indicated in bold.

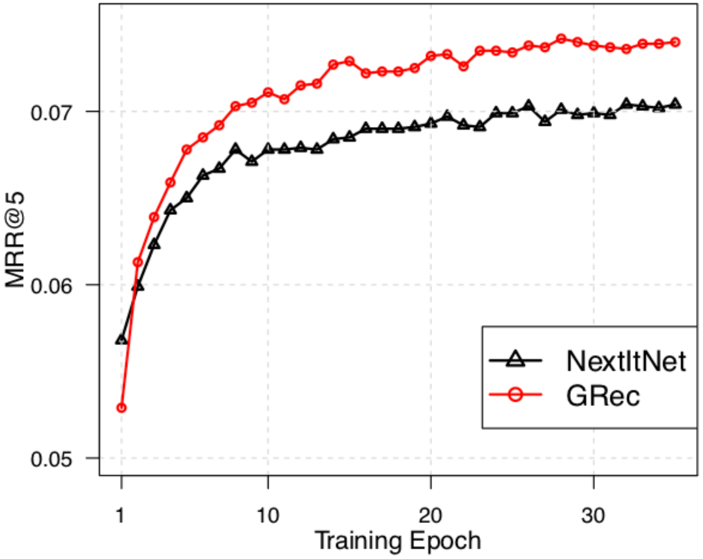
DATA	*Models*	*MRR@5*	*MRR@20*	*HR@5*	*HR@20*	*NDCG@5*	*NDCG@20*
TW10	*MostPop*	0.0055	0.0127	0.0203	0.0970	0.0091	0.0305
	Caser	0.0780	0.0916	0.1330	0.2757	0.0916	0.1317
	GRU4Rec	0.0786	0.0926	0.1325	0.2808	0.0919	0.1335
	NextItNet	0.0848	0.0992	0.1408	0.2931	0.0986	0.1414
	NextItNet+	0.0698	0.0844	0.1214	0.2775	0.0825	0.1218
	tNextItNet	0.0813	0.0958	0.1376	0.2896	0.0953	0.1380
	GRec	**0.0901**	**0.1046**	**0.1498**	**0.3021**	**0.1049**	**0.1477**
ML30	*MostPop*	0.0030	0.0058	0.0098	0.0405	0.0047	0.0132
	Caser	0.0622	0.0739	0.1074	0.2323	0.0733	0.1083
	GRU4Rec	0.0652	0.0788	0.1156	0.2589	0.0776	0.1179
	NextItNet	0.0704	0.0849	0.1242	0.2756	0.0837	0.1263
	NextItNet+	0.0564	0.0711	0.1051	0.2609	0.0685	0.1121
	tNextItNet	0.0658	0.0795	0.1164	0.2605	0.0782	0.1188
	GRec	**0.0742**	**0.0889**	**0.1300**	**0.2850**	**0.0879**	**0.1315**

Future Data Helps Training:

- Convergence Results:



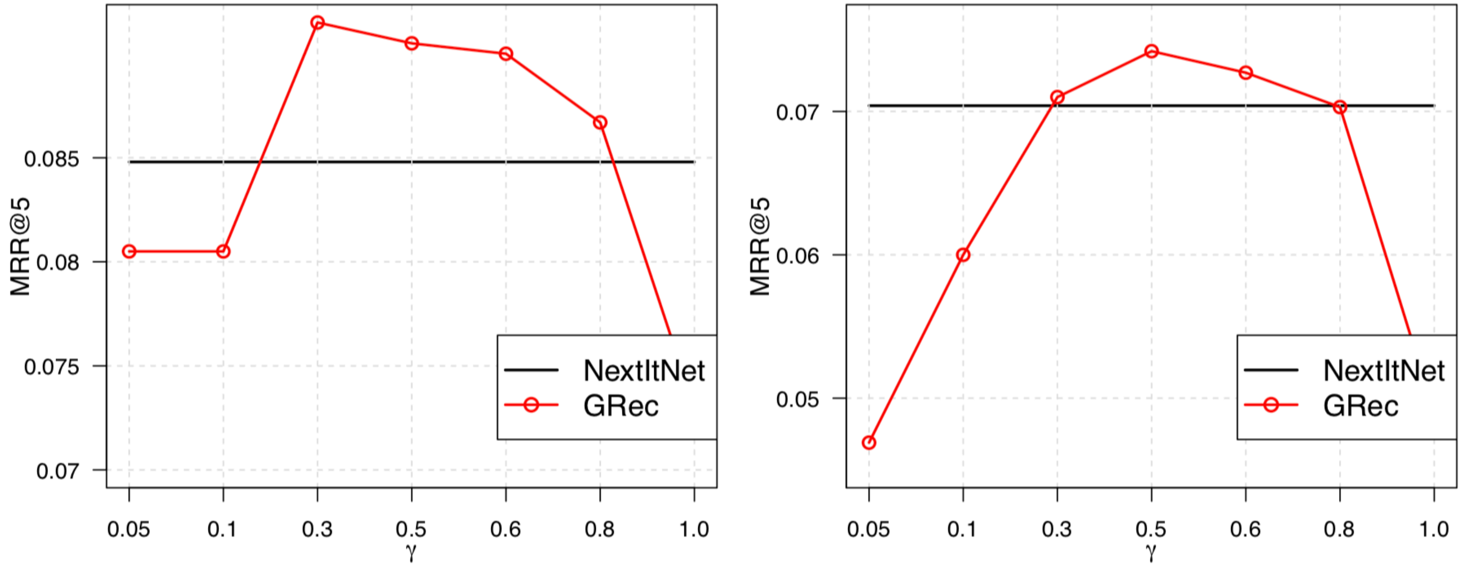
(a) TW10



(b) ML30

Future Data Helps Training:

- Results with different gap-filling percentage:



(a) TW10 (b) ML30

Tencent 腾讯

Future Data Helps Training:

- Results with/without projector:

DATA	*NextItNet*	*NextItNetP*	*GRec*	*GRecN*
TW10	0.0848	0.0843	0.0901	0.0880
ML30	0.0704	0.0702	0.0742	0.0720
ML100	0.0552	0.0558	0.0588	0.0577

- Results with different encoder or decoder networks:

DATA	*ReCd*	*NextItNet*	*CeRd*	*GRU*
TW10	0.0879	0.0843	0.0876	0.0786
ML30	0.0728	0.0704	0.0712	0.0652
ML100	0.0582	0.0552	0.0571	0.0509

Future Data Helps Training:

- Thanks!