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)
- 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

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

Future Data Helps Training:

- **CNN**: Caser[1]

[1] Personalized Top-N Sequential Recommendation via Convolutional Sequence Embedding. WSDM2018
Future Data Helps Training:

- **Dilated CNN:** NextItNet[1]

---

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

Future Data Helps Training:

- **Attention[1]**

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

[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**: \( \{x_0, x_1, ..., x_{14}\} \Rightarrow x_{15} \)

**NextItNet**: \( \{x_0, x_1, ..., x_{14}\} \Rightarrow \{x_1, x_2, ..., x_{15}\} \)

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

**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**: \( \{x_{-1}, x_{-1}, ..., x_2\} \Rightarrow x_3 \)
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
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**

  \[ x \rightarrow f \rightarrow o \rightarrow l \rightarrow o \rightarrow w \rightarrow s \rightarrow x \]

  seen by the encoder

\(x_5 \text{ follows } x_4\)
Future Data Helps Training:

- Other trivial methods:

\[
\text{NextItNet}^+: \{x_1, \ldots, x_{t-1}\} \Rightarrow \{x_2, \ldots, x_t\} \\
\begin{align*}
\text{input} & \quad \text{output} \\
\{x_t, \ldots, x_2\} & \Rightarrow \{x_{t-1}, \ldots, x_1\} \\
\text{input} & \quad \text{output}
\end{align*}
\]

- 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)**

**Drawbacks:**
training & inference has discrepancies since backward network is useless during inference
Future Data Helps Training:

- Our solution: GRec

![Diagram](image)
Future Data Helps Training:

- Our solution: GRec
Future Data Helps Training:

- **GRec:**

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

\[
= \sum_{x \in \mathcal{X}} \log \prod_{i=1}^{m} p(x_\triangle; x_1; \Delta_{i-1}; \tilde{x}; \Theta)
\]

- Masked items are predicted given its previous items and other items without masking.
Future Data Helps Training:

- Grec vs. NextItNet:

\[
\text{NextItNet} : \{x_1, x_2, x_3, \ldots, x_7, x_8\} \Rightarrow \{x_2, x_3, x_4, \ldots, x_8, x_9\}\\
\text{encoder input} \quad \text{decoder output}
\]

\[
\text{GRec} : \{x_1, __, x_3, __, __, __, x_7, x_8, __\} + \{x_1, x_2, x_3, \ldots, x_9\}\\
\text{encoder input} \quad \text{decoder input}
\]

\[
\Rightarrow \{x_2, x_4, x_5, x_6, x_9\}\\
\text{decoder output}
\]
Future Data Helps Training:

- **Connections:**

  - 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:

<table>
<thead>
<tr>
<th>DATA</th>
<th>#actions</th>
<th>#sequences</th>
<th>#items</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>TW10</td>
<td>9,986,953</td>
<td>1,048,575</td>
<td>65,997</td>
<td>10</td>
</tr>
<tr>
<td>ML30</td>
<td>25,368,155</td>
<td>858,160</td>
<td>18,273</td>
<td>30</td>
</tr>
<tr>
<td>ML100</td>
<td>25,240,741</td>
<td>300,624</td>
<td>18,226</td>
<td>100</td>
</tr>
</tbody>
</table>
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.

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

- Convergence Results:
Future Data Helps Training:

- Results with different gap-filling percentage:
Future Data Helps Training:

• Results with/without projector:

<table>
<thead>
<tr>
<th>DATA</th>
<th>NextItNet</th>
<th>NextItNetP</th>
<th>GRec</th>
<th>GRecN</th>
</tr>
</thead>
<tbody>
<tr>
<td>TW10</td>
<td>0.0848</td>
<td>0.0843</td>
<td>0.0901</td>
<td>0.0880</td>
</tr>
<tr>
<td>ML30</td>
<td>0.0704</td>
<td>0.0702</td>
<td>0.0742</td>
<td>0.0720</td>
</tr>
<tr>
<td>ML100</td>
<td>0.0552</td>
<td>0.0558</td>
<td>0.0588</td>
<td>0.0577</td>
</tr>
</tbody>
</table>

• Results with different encoder or decoder networks:

<table>
<thead>
<tr>
<th>DATA</th>
<th>ReCd</th>
<th>NextItNet</th>
<th>CeRd</th>
<th>GRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>TW10</td>
<td>0.0879</td>
<td>0.0843</td>
<td>0.0876</td>
<td>0.0786</td>
</tr>
<tr>
<td>ML30</td>
<td>0.0728</td>
<td>0.0704</td>
<td>0.0712</td>
<td>0.0652</td>
</tr>
<tr>
<td>ML100</td>
<td>0.0582</td>
<td>0.0552</td>
<td>0.0571</td>
<td>0.0509</td>
</tr>
</tbody>
</table>
Future Data Helps Training:

• Thanks!