Personalized Transformer for Explainable Recommendation

Lei Li\textsuperscript{1}, Yongfeng Zhang\textsuperscript{2}, Li Chen\textsuperscript{1}

\textsuperscript{1}Hong Kong Baptist University, \textsuperscript{2}Rutgers University

csleili@comp.hkbu.edu.hk

August 4, 2021

The 59th Annual Meeting of the Association for Computational Linguistics (ACL'21)
Transformer (Vaswani et al., NIPS’17)

BERT (Devlin et al., NAACL’19)

Bidirectional model for classification

OpenAI GPT (Radford et al., 18)

Unidirectional model for generation

Courtesy images from UNILM (Dong et al., NeurIPS’19)
Autoregressive Natural Language Generation

- Predict future tokens based on past tokens
  - Generate an output sequence, based on the given input sequence
Explainable Recommendation

• Given a user-item pair, provide an explanation to justify why the item is recommended to the user
  • Pre-defined template (Zhang et al., SIGIR’14)
  • Image visualization (Chen et al., SIGIR’19)
  • Natural language sentence in this work
    • E.g., “the style of the jacket is fashionable”
  • ......

You might be interested in [feature], on which this product performs well.
You might be interested in [feature], on which this product performs poorly.
Transformer for Explanation Generation

• Consider the user-item pair as an input sequence
  • Regard IDs as tokens, similar to words

The food is good
Problem Identification

• Identical generated explanations for almost every user-item pair
  • Adam Main Canteen
  • Beth Renfrew Cafeteria
  • Carol Bistro Bon
  • David Harmony Cafeteria
  • ......

• Less useful, if unable to explain the key specialty of each recommendation
• Could cause negative effects on users (Tintarev and Mashoff, 15)
Attention Visualization

• The generation relies heavily on <bos>
  • The reason why all explanations are identical
• Attention weights on userID and itemID approach 0
  • Model insensitive to IDs

Why insensitive?
Problem Analysis

- Frequency mismatch between IDs and words
  - One user/item ID vs. hundreds of words in a review
  - An ID appears in only a few reviews
- IDs being regarded as uncommon words (OOV tokens)

Restaurant review (yelp.com)
Solution: Context Prediction

• Bridge IDs and words, and give the former linguistic meanings
Our model PETER can utilize IDs for generation

PETER: PErsomalized Transformer for Explainable Recommendation
Attention Masking

- Revise Left-to-Right attention masking matrix (call it PETER masking)
  - Allow the interaction between user and item for context prediction and recommendation

\[
A_{t,h} = \text{softmax} \left( \frac{Q_{t,h} K_{t,h}^T}{\sqrt{d}} + M \right) V_{t,h}
\]

\[
Q_{t,h} = S_{t-1} W_{l,h}^Q, \quad K_{t,h} = S_{t-1} W_{l,h}^K,
\]

\[
V_{t,h} = S_{t-1} W_{l,h}^V
\]

\[
M = \begin{cases} 
0, & \text{Allow to attend} \\
-\infty, & \text{Prevent from attending}
\end{cases}
\]
Context Prediction & Explanation Generation

- Context prediction: predict explanation words in one step
  - With the item representation
- Explanation generation: generate them one by one
  - Linear layer
    \[ c_t = \text{softmax}(W^v s_{L,t} + b^v) \]
  - NLL loss for context prediction
    \[ L_c = \frac{1}{|\mathcal{T}|} \sum_{(u,i) \in \mathcal{T}} \frac{1}{|E_{u,i}|} \sum_{t=1}^{|E_{u,i}|} - \log c_{2t}^e \]
  - NLL loss for explanation generation
    \[ L_e = \frac{1}{|\mathcal{T}|} \sum_{(u,i) \in \mathcal{T}} \frac{1}{|E_{u,i}|} \sum_{t=1}^{|E_{u,i}|} - \log c_{2+|F_{u,i}|+t}^e \]
Recommendation & Targeted Explanation

• Predict a rating score for the user-item pair
  - MLP with one hidden layer
    \[ \hat{r}_{u,i} = w^f \sigma(W^f s_{L,1} + b^f) + b^f \]
  - MSE loss for rating prediction
    \[ \mathcal{L}_r = \frac{1}{|T|} \sum_{(u,i) \in T} (r_{u,i} - \hat{r}_{u,i})^2 \]

• Incorporate features for targeted explanation generation
  - E.g., conversational recommendation (Chen et al., IJCAI’20)
  - Denoted as PETER+
Multi-task Learning

- Three tasks trained in an end-to-end manner
  - Explanation generation
  - Context prediction
  - Rating prediction

\[ J = \min_{\Theta} (\lambda_e \mathcal{L}_e + \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r) \]
Datasets (Li et al., CIKM’20)

- Yelp
  - Restaurant
- Amazon
  - Movies & TV
- TripAdvisor
  - Hotel
- The explanation is a review sentence containing features

<table>
<thead>
<tr>
<th></th>
<th>Yelp</th>
<th>Amazon</th>
<th>TripAdvisor</th>
</tr>
</thead>
<tbody>
<tr>
<td>#users</td>
<td>27,147</td>
<td>7,506</td>
<td>9,765</td>
</tr>
<tr>
<td>#items</td>
<td>20,266</td>
<td>7,360</td>
<td>6,280</td>
</tr>
<tr>
<td>#records</td>
<td>1,293,247</td>
<td>441,783</td>
<td>320,023</td>
</tr>
<tr>
<td>#features</td>
<td>7,340</td>
<td>5,399</td>
<td>5,069</td>
</tr>
<tr>
<td>#records / user</td>
<td>47.64</td>
<td>58.86</td>
<td>32.77</td>
</tr>
<tr>
<td>#records / item</td>
<td>63.81</td>
<td>60.02</td>
<td>50.96</td>
</tr>
<tr>
<td>#words / exp</td>
<td>12.32</td>
<td>14.14</td>
<td>13.01</td>
</tr>
</tbody>
</table>

* exp denotes explanation.
Evaluation Metrics

• Recommendation
  • RMSE & MAE

• Explanation
  • Text quality: BLEU (Papineni et al., ACL’02) & ROUGE (Lin, ACL’04 Workshop)
    • Not equal to explainability (Chen et al., SIGIR’19 Workshop; Li et al., CIKM’20)
  • Explainability from the angle of item features (Li et al., CIKM’20)
    • Unique Sentence Ratio (USR)
    • Feature Matching Ratio (FMR)
    • Feature Coverage Ratio (FCR)
    • Feature Diversity (DIV)
## Quantitative Analysis on Explanations

- Ours the best or comparable

<table>
<thead>
<tr>
<th>Explainability</th>
<th>Text Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>FMR↑</td>
<td>FCR↑</td>
</tr>
<tr>
<td>TripAdvisor</td>
<td>Amazon</td>
</tr>
</tbody>
</table>

| Transformer | 0.06 | 0.06 | 2.46 | 0.01 | 7.39 | 0.42 | 19.18 | 10.29 | 12.56 | 1.71 | 0.92 | 1.09 |
| NRT | 0.07 | 0.11 | 2.37 | 0.12 | 11.66 | 0.65 | 17.69 | 12.11 | 13.55 | 1.76 | 1.22 | 1.33 |
| Att2Seq | 0.07 | 0.12 | 2.41 | 0.13 | 10.29 | 0.58 | 18.73 | 11.28 | 13.29 | 1.85 | 1.14 | 1.31 |
| PETER | 0.08** | 0.19** | 1.54** | 0.13 | 10.77 | 0.73** | 18.54 | 12.20 | 13.77** | 2.02** | 1.38** | 1.49** |
| ACMLM | 0.05 | 0.31 | **0.95** | **0.95** | 7.01 | 0.24 | 7.89 | 7.54 | 6.82 | 0.44 | 0.48 | 0.39 |
| NETE | 0.80 | 0.27 | 1.48 | 0.52 | 19.31 | 2.69 | 33.98 | 22.51 | 25.56 | 8.93 | 5.54 | 6.33 |
| PETER+ | **0.86**** | **0.38**** | **1.08** | **0.34** | **20.80**** | **3.43**** | **35.44**** | **26.12**** | **27.95**** | **10.65**** | **7.44**** | **7.94**** |

- **IDs only**
- **Metric problem**

**With features**: Less useful, if unable to guarantee text quality
## Qualitative Case Study on Explanations

- Context prediction task can indeed give IDs linguistic meanings
- Two tasks resemble one’s drafting-polishing process
- Incorporated features further improve text quality

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Top-15 Context Words</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PETER</td>
<td>&lt;eos&gt; the and a pool was with nice is very were to good in of and is the rooms are spacious and the bathroom has a large tub</td>
<td></td>
</tr>
<tr>
<td>PETER+</td>
<td>&lt;eos&gt; the and a was pool with to nice good very were is of in and is the pool area is nice and the gym is very well equipped &lt;eos&gt;</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Top-15 Context Words</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PETER</td>
<td>&lt;eos&gt; the and a was were separate bathroom with shower large very had in is and is the bathroom was large and the shower was great &lt;eos&gt;</td>
<td></td>
</tr>
<tr>
<td>PETER+</td>
<td>&lt;eos&gt; the and a was bathroom shower with large in separate were room very is and is the lobby was very nice and the rooms were very comfortable &lt;eos&gt;</td>
<td></td>
</tr>
</tbody>
</table>
Efficiency Comparison

- Training minutes comparison with BERT-based model under the same settings
  - PETER+ is small (only 2 attention layers), so it takes much less training time
  - PETER+ is unpretrained, and thus requires more training epochs

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Epochs</th>
<th>Time/Epoch</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACMLM</td>
<td>97.0</td>
<td>3</td>
<td>32.3</td>
</tr>
<tr>
<td>PETER+</td>
<td>57.7</td>
<td>25</td>
<td>2.3</td>
</tr>
</tbody>
</table>
**Recommendation Performance**

- Ours the best on the largest dataset with over 1 million records
- Ours comparable on small datasets
  - Not a problem to real applications, e.g., billion-scale users in Amazon

<table>
<thead>
<tr>
<th></th>
<th>Yelp</th>
<th>Amazon</th>
<th>TripAdvisor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R↓</td>
<td>M↓</td>
<td>R↓</td>
</tr>
<tr>
<td>PMF</td>
<td>1.09</td>
<td>0.88</td>
<td>1.03</td>
</tr>
<tr>
<td>SVD++</td>
<td><strong>1.01</strong></td>
<td><strong>0.78</strong></td>
<td>0.96</td>
</tr>
<tr>
<td>NRT</td>
<td><strong>1.01</strong></td>
<td><strong>0.78</strong></td>
<td><strong>0.95</strong></td>
</tr>
<tr>
<td>NETE</td>
<td><strong>1.01</strong></td>
<td>0.79</td>
<td>0.96</td>
</tr>
<tr>
<td>PETER</td>
<td><strong>1.01</strong></td>
<td><strong>0.78</strong></td>
<td><strong>0.95</strong></td>
</tr>
</tbody>
</table>
## Ablation Study

- Prove the rationale of each component

**Reduce to standard Transformer**

<table>
<thead>
<tr>
<th></th>
<th>Explainability</th>
<th>Text Quality</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FMR</td>
<td>FCR</td>
<td>DIV</td>
</tr>
<tr>
<td>Disable $\mathcal{L}_c$</td>
<td>0.06 ↓</td>
<td>0.03 ↓</td>
<td>5.75 ↓</td>
</tr>
<tr>
<td>Disable $\mathcal{L}_r$</td>
<td>0.07</td>
<td>0.14 ↑</td>
<td>2.90 ↑</td>
</tr>
<tr>
<td>Left-to-Right Masking</td>
<td>0.07</td>
<td>0.15 ↑</td>
<td>2.68 ↑</td>
</tr>
<tr>
<td>PETER</td>
<td>0.07</td>
<td>0.13</td>
<td>2.95</td>
</tr>
</tbody>
</table>

Recommendation and context prediction highly correlated

Block item information
Conclusion

• The 1\textsuperscript{st} to enable Transformer with personalized natural language generation
  • Shed light on a broader scope of fields that also need personalization
    • E.g., personalized conversational systems

• Model small and unpretrained, but effective and efficient
  • Open up a new way of exploiting Transformer by designing good tasks instead of scaling up model size

• Design a task to connect IDs and words
  • Point out a way for Transformer to deal with heterogeneous inputs
    • E.g., \textit{image generation} based on text in multimodal AI
References (1)

• Vaswani, Ashish, et al. "Attention is all you need." NIPS’17.


References (2)

Thank you!

csleili@comp.hkbu.edu.hk