A Model of Two Tales: Dual Transfer Learning Framework for Improved Long-tail Item Recommendation

Yin Zhang¹, Derek Zhiyuan Cheng², Tiansheng Yao², Xinyang Yi², Lichan Hong², Ed H. Chi²

¹Texas A&M University  ²Google Inc
Long-Tail Distributions in Large Systems

- [Zhang et al. 2017](Zhang et al. 2017)
- [Wang et al. 2017](Wang et al. 2017)
- [Olmedilla et al. 2016](Olmedilla et al. 2016)

Recommendation

- [MovienLen1M](MovienLen1M)

- [Olmedilla et al. 2016](Olmedilla et al. 2016)
Long-tail Item Distribution in Recommendation

- For items, a small fraction of items receive most of the user feedback (very popular), while most items only have few user feedback.

- Issues:
  - Performance
  - Reproducibility (e.g. tail under-trained)
  - Model Bias (e.g. rich gets richer)
  - ...

Our Goal: improve recommendation quality on tail slices of items, while keeping overall performance neutral or better.
Related Work

- Add user/item content information;
- Re-sampling:  --&gt; Shifted the original distribution of the items
  - upsampling and downsampling [TKDE'09, ECCV'18]
- Regularization:  --&gt; Hard to automatically weight different items
  - focused learning [WWW'17]:
- Correcting sampling bias:  --&gt; Mainly for in-batch sampling
  - logQ [RecSys'19]
- Share information:  --&gt; Only based on the item id
  - clustering embeddings into centroids [MGQE]

Not fully utilize the long tail distribution patterns and the semantic relations among items (item embeddings are learned independently)
We propose—Transfer learning;
- Utilize the **semantic relations** among items;
- Utilize the **frequency information** in the item long-tail distribution;

Can we **effectively transfer** information inside the item long-tail distribution for improved recommendation?
Transfer Knowledge in Long-tail Distribution: Challenges

- How to distill **knowledge** from head items that can be **well transferred** to tail items?

  Head Item Distribution

  Tail Item Distribution

- How to improve on tail items and also keep/improve the overall performance?

  Head Item Features ↔ Tail Item Features
Proposed Method: Dual Transfer Learning -- MIRec

- Transfer learning **across Models** through meta-learning: learn a meta-mapping from a few-shot model’s parameters to a many-shot model’s parameters;

- Transfer learning **across Items** through curriculum learning: connect head and tail items by their **features** to smoothly transfer the learned meta-level knowledge between head and tail items;
Two-Tower Model

MIRec Model Framework

Meta-learning for user and item towers

Curriculum learning with different data distributions
Transfer Learning Across Models -- Meta Learning

Many-shot Models Parameters (trained with rich data samples)

Few-shot Models Parameters (trained with few data samples)

\[ L(w, \theta | \Omega^*, \Omega(k)) = \| F(\theta; w) - \theta^* \|^2 + \lambda L_g(\theta | \Omega(k)) \]

Meta-mapping

Few-shot Model Learning
Transfer Learning Across Items -- Curriculum Learning

- The basic idea for curriculum learning is to simulate the human learning process, and apply it to machine learning process.

$\Omega^* := \{(u, i, r(u, i))\}$

Includes all the user feedback from both head and tail items

$\Omega(k) := \{(u, i, r(u, i))\}$

- Downsampling head items that have more than k user feedbacks
- \textit{Includes all the tail items}

- Tail items are \textbf{fully trained} in both many-shot and few-shot models;
- \textbf{Alleviate bias} among tail items that brings by the new distribution;
Experiment
Research Questions

- How well does the dual transfer learning framework MIRec perform compared to the state-of-the-art methods?
- Does MIRec learn better representations for tail items? Could we see the differences visually?
Experiment Dataset and Evaluation Criteria

**Experiment Dataset**

- **MovieLens1M**
- **Bookcrossing**

**Evaluation Criteria**

<table>
<thead>
<tr>
<th>HR@K/NDCG@K</th>
<th>Overall</th>
<th>Head Items</th>
<th>Tail Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good Results</td>
<td>–</td>
<td>↓</td>
<td>↑</td>
</tr>
<tr>
<td>Better Results</td>
<td>↑</td>
<td>–</td>
<td>↑</td>
</tr>
<tr>
<td>Great Results</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
</tr>
</tbody>
</table>
### Experiments: Recommendation Effectiveness

<table>
<thead>
<tr>
<th>Measure</th>
<th>Overall</th>
<th></th>
<th>Head</th>
<th></th>
<th>Tail</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR@10</td>
<td>NDCG@10</td>
<td>HR@10</td>
<td>NDCG@10</td>
<td>HR@10</td>
<td>NDCG@10</td>
</tr>
<tr>
<td>Two-tower</td>
<td>6.74</td>
<td>3.26</td>
<td>8.98</td>
<td>4.33</td>
<td>3.22</td>
<td>1.59</td>
</tr>
<tr>
<td>Over-sampling</td>
<td>0.23</td>
<td>0.10</td>
<td>0.14</td>
<td>0.04</td>
<td>0.37</td>
<td>0.19</td>
</tr>
<tr>
<td>Under-sampling</td>
<td>0.72</td>
<td>0.33</td>
<td>0.41</td>
<td>0.20</td>
<td>1.19</td>
<td>0.53</td>
</tr>
<tr>
<td>ClassBalance</td>
<td>6.32</td>
<td>2.98</td>
<td>9.10</td>
<td>4.43</td>
<td>1.45</td>
<td>0.71</td>
</tr>
<tr>
<td>LogQ</td>
<td>2.24</td>
<td>0.95</td>
<td>1.41</td>
<td>0.54</td>
<td>3.53</td>
<td>1.59</td>
</tr>
<tr>
<td>Head2Tail</td>
<td>1.31</td>
<td>0.66</td>
<td>1.90</td>
<td>0.97</td>
<td>0.40</td>
<td>0.18</td>
</tr>
<tr>
<td>Tail2Head</td>
<td>1.71</td>
<td>0.85</td>
<td>2.74</td>
<td>1.34</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td>MeLu</td>
<td>5.96</td>
<td>2.88</td>
<td>8.98</td>
<td>4.35</td>
<td>1.23</td>
<td>0.58</td>
</tr>
<tr>
<td>MIRec</td>
<td><strong>7.02</strong></td>
<td><strong>3.36</strong></td>
<td><strong>9.13</strong></td>
<td><strong>4.36</strong></td>
<td><strong>3.70</strong></td>
<td><strong>1.81</strong></td>
</tr>
</tbody>
</table>

MIRec **consistently** outperforms the other state-of-the-art methods.
Experiments: Visualization of Tail Item Embeddings

The movie clusters from MIRec is more **coherent** with respect to the movie genres information.
Conclusions and Future Work

● We propose a novel dual transfer learning framework MIRec to explore the item long tail distribution in recommendation:
  ○ Meta learning -- mapping relations between few shot model and many shot model;
  ○ Curriculum learning -- the correlations among items through training strategies;
● The MIRec brings improvement for tail items, and at same time, relatively keep/improve the overall performance;
● The learned item embeddings by MIRec contains more semantic information;

Future work:
● Improve the curriculum; 
● Explore other types of content information;


A Model of Two Tales: Dual Transfer Learning Framework for Improved Long-tail Item Recommendation

Yin Zhang\textsuperscript{1}, Derek Zhiyuan Cheng\textsuperscript{2}, Tiansheng Yao\textsuperscript{2}, Xinyang Yi\textsuperscript{2}, Lichan Hong\textsuperscript{2}, Ed H. Chi\textsuperscript{2}

\textsuperscript{1}Texas A&M University \hspace{1em} \textsuperscript{2}Google Inc

Thank you!
Experiments: Impact of Influence Factors

Both meta-learning and curriculum learning in MIRec play important role to help improve the recommendation performance.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Overall</th>
<th></th>
<th>Head</th>
<th></th>
<th>Tail</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR@10</td>
<td>NDCG@10</td>
<td>HR@10</td>
<td>NDCG@10</td>
<td>HR@10</td>
<td>NDCG@10</td>
</tr>
<tr>
<td>Default</td>
<td>6.74</td>
<td>3.26</td>
<td>8.98</td>
<td>4.33</td>
<td>3.22</td>
<td>1.59</td>
</tr>
<tr>
<td>MIRec</td>
<td>7.02</td>
<td>3.36</td>
<td>9.13</td>
<td>4.36</td>
<td>3.70</td>
<td>1.81</td>
</tr>
<tr>
<td>MIRec\textsubscript{M}</td>
<td>6.48</td>
<td>3.13</td>
<td>8.75</td>
<td>4.17</td>
<td>2.88</td>
<td>1.43</td>
</tr>
<tr>
<td>MIRec\textsubscript{C}</td>
<td>6.83</td>
<td>3.30</td>
<td>8.77</td>
<td>4.21</td>
<td>3.78</td>
<td>1.87</td>
</tr>
<tr>
<td>MIRec – LogQ</td>
<td>6.98</td>
<td>3.36</td>
<td>9.75</td>
<td>4.71</td>
<td>2.62</td>
<td>1.26</td>
</tr>
<tr>
<td>Two – tower\textsubscript{2}</td>
<td>6.90</td>
<td>3.22</td>
<td>9.46</td>
<td>4.37</td>
<td>2.89</td>
<td>1.41</td>
</tr>
</tbody>
</table>