Towards the Better Ranking Consistency: A Multi-task Learning Framework for Early Stage Ads Ranking

Xuewei Wang
Meta
AdKDD 23
Ads Ranking system

- Multi-stage ranking system: Trade-off between capacity and efficiency
  - Early stage: Simplified model with latency constraint
  - Final stage: Large capacity model with good accuracy

- Multi-objective ranking system
  - Ad Auction depend on **Total Value**
    - Bid placed by an advertiser for that ad
    - Estimated action rates (e.g. CTR, CVR)
    - Ad Quality for user’s ads experience
      - E.g. hide ads, report bad ads
02 Motivation

Ranking Consistency

- **Ideal Status**
  - Early stage and final stage have same ranking orders for ads

- **Ranking consistency issue:**
  - Top ads in the final stage are ranked low in the early stage

- **Gap between final stage and early stage:**
  - Performance gap
  - Total value definition inconsistency
    - The early stage models’ (i.e. ad quality models) development lags behind final stage models.
  - Selection bias
Why recall?
○ We only care about the top ad candidates for user impression, rather than the lower-ranked ones

Challenge:
○ The accurate recall is difficult to compute with large candidates

Solutions:
○ Offline simulated recall
  ■ We replay a small traffic with full ad requests in simulators, with relaxed timeouts between stages, to ensure that all ads from retrieval stages are ranked.
Multi-task learning for early stage ranking

- Ranking consistency improvement
  - Learn from final stage ads quality models
  - Learn from final stage CTR model
- Resource saving by model consolidation
  - Ads quality & CTR models need to predict on most ads
- Mitigation of selection bias
  - Data augmentation with non-impression data
04 Methods

Multi-task learning for early stage ranking

- New objective:
  - Consolidated Quality Score (CQS): Final stage total quality score
    
    \[
    \text{AdQuality} = f(\text{CQS})
    \]
    
    \[
    \text{CQS} = \sum_{i=1}^{N} \text{scalar}_i \times p\text{QualityEvent}_i
    \]
    
    \[
    L_{cqs} = \frac{1}{n} \sum_{i=1}^{n} (\text{CQS}_i - y_{cqs})^2
    \]

- Final stage CTR teacher distillation
  
  \[
  L_{teacher} = -[e\text{CTR} \times \log(y_{ctr}) + (1 - e\text{CTR}) \times \log(1 - y_{ctr})]
  \]

- Data Augmentation
  - Final stage eCTR as pseudo label for CTR task
  - Train on impression ads + non-impression ads
  - Help de-bias on both ads quality and CTR
05 Experiments

Consolidate Ads Quality Models

- Offline Soft Recall:
  - the sum of final stage ads total value of top $K$ ads picked by the model divided by sum of total value of the golden set.
- Total Value
  - Sum of total value for impression ads
- TVD
  - Total value divergence between final stage and early stage
- Ads quality metrics:
  - Xout rate: the ads cross-out rate

<table>
<thead>
<tr>
<th>Metric</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall (+)</td>
<td>+3.2%</td>
</tr>
<tr>
<td>Xout rate (-)</td>
<td>-1.8%</td>
</tr>
<tr>
<td>ASQ (+)</td>
<td>+0.02</td>
</tr>
<tr>
<td>TVD (-)</td>
<td>-7.9%</td>
</tr>
<tr>
<td>CTR (+)</td>
<td>+1.7%</td>
</tr>
<tr>
<td>CVR (+)</td>
<td>+2.0%</td>
</tr>
<tr>
<td>Total Value (+)</td>
<td>+1.0%</td>
</tr>
<tr>
<td>total CPU (-)</td>
<td>-0.7%</td>
</tr>
</tbody>
</table>

Table 1: The CQS model’s relative performance compared with production early stage quality models. The token (+) means better performance with higher values, and (-) means better performance with lower values.
05 Experiments

Multi-task Learning of CQS and CTR

- Significant improvement on ads recall & total value
- CTR and CVR also increased

<table>
<thead>
<tr>
<th>Metric</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall (+)</td>
<td>+12.2%</td>
</tr>
<tr>
<td>Xout rate (-)</td>
<td>-3.5%</td>
</tr>
<tr>
<td>ASQ (+)</td>
<td>+0.005</td>
</tr>
<tr>
<td>TVD (-)</td>
<td>-5.7%</td>
</tr>
<tr>
<td>CTR (+)</td>
<td>+0.4%</td>
</tr>
<tr>
<td>CVR (+)</td>
<td>+0.8%</td>
</tr>
<tr>
<td>Total Value (+)</td>
<td>+3.0%</td>
</tr>
<tr>
<td>total CPU (-)</td>
<td>-0.06%</td>
</tr>
</tbody>
</table>

Table 2: The multi-task learning framework’s relative performance compared with individual CQS model and CTR model. The token (+) means better performance with higher values, and (-) means better performance with lower values.
Ablation study

- Dedicated CTR
  - Remove CQS task in MT framework
- Dedicated CQS
  - Remove CTR & teacher task in MT framework
- MT w/o teacher:
  - Remove teacher task
- MT w/o augmented data
  - Train only on impression data

<table>
<thead>
<tr>
<th>Comparison</th>
<th>NE diff (%)</th>
<th>MSE diff (%)</th>
<th>Recall (+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedicated CTR + CQS</td>
<td>-0.04%</td>
<td>-0.6%</td>
<td>-0.6%</td>
</tr>
<tr>
<td>MT w/o Teacher task</td>
<td>+0.3%</td>
<td>-0.5%</td>
<td>-1.6%</td>
</tr>
<tr>
<td>MT w/o Augmented data</td>
<td>-</td>
<td>-</td>
<td>-11.9%</td>
</tr>
</tbody>
</table>
06 Conclusions

- Each component in our multi-task learning framework is essential to improve the performance.
- This framework can be generalized to other user cases since the CQS can be applied to any ads ranking system with the ads quality component.
- Compared with NE and MSE metrics, the offline recall evaluation metric can reflect online performance (i.e. total value) better
  - Single offline metric for an individual ranking model may not be reliable to reflect online performance.
Thank you!