MLink: Linking Black-box Models for Collaborative Multi-model Inference

Mu Yuan, Lan Zhang, Xiang-Yang Li
University of Science and Technology of China
Introduction

Cost-effective Inference

- Multi-task learning and zipping
- Model compression
- Inference reusing
- Source filtering
- Multi-model scheduling

How to obtain as accurate inference results as possible without the exact execution of ML models?
Introduction

Linking Black-box Models

• Multi-task learning and zipping
• Model compression
• Inference reusing
• Source filtering
• Multi-model scheduling

How to obtain as accurate inference results as possible without the exact execution of ML models?
Introduction
Linking Black-box Models

- Model Linking
  - machine over-learning
  - cross-task semantic correlation

Predict un-executed models’ inference results based on executed models’?
Introduction

Linking Black-box Models

• Model Linking
  • machine over-learning
  • cross-task semantic correlation

• Target application
  • inference results of multiple models are required
  • cost budget is too limited to run them all
Black-box Model Linking

Model link architecture

- output formats determine the model link’s architecture
  - fixed-length vector & variable-length sequence
- 4 types of model link architectures
  - Vec-to-Vec
  - Seq-to-Vec
  - Vec-to-Seq
  - Seq-to-Seq
Collaborative Multi-model Inference

Algorithm

- select greedily w.r.t. activation probability under the cost budget

- activated models do exact inference while the others’ outputs will be predicted by the model link ensemble of activated sources.

- periodic re-profiling and re-selection
  
  - By reasonably setting the period length and the proportion of data used for profiling, we can amortize the overheads of loading/unloading ML models to negligible.
Evaluation

Real Systems

• Smart Building
  • two days (one weekday & one weekend) of videos (1 frame per minute) from 58 cameras
  • 3 models deployed
    • person counting, action classification, object counting
Evaluation

Real Systems

• City Traffic
  • two days (one weekday & one weekend) of videos (1 FPS) from 10 cameras at road intersections
  • 3 models deployed
    • person counting, traffic condition classification, vehicle counting
Evaluation

Baselines

- **Standalone**: selects models in ascending order of delay and runs models independently
- **MTL**: a multi-task learning approach
- **DRLS**: a deep reinforcement learning-based scheduling approach
- **Reducto**: a low-level feature difference-based frame filtering approach

**Target Application**

- inference results of multiple models are required
- cost budget is too limited to run them all
## Evaluation

### Video Analytics with Model Links

- GPU Memory as the cost budget

### Table 4: Comparisons of MLink, MTL, Reducto, DRLS, and Standalone

<table>
<thead>
<tr>
<th>Method</th>
<th>Building (5/9 GB Mem.)</th>
<th>City (5/9 GB Mem.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc. (%)</td>
<td>Time (ms)</td>
</tr>
<tr>
<td>Standalone</td>
<td>33.3/66.7</td>
<td>30/74</td>
</tr>
<tr>
<td>MTL</td>
<td>53.3</td>
<td>32.8</td>
</tr>
<tr>
<td>DRLS</td>
<td>45.7/81.3</td>
<td>58.7/107</td>
</tr>
<tr>
<td>Reducto</td>
<td>91.8/96.9</td>
<td>45.7/89</td>
</tr>
<tr>
<td><strong>MLink</strong></td>
<td><strong>94.1/97.9</strong></td>
<td><strong>39.3/84</strong></td>
</tr>
</tbody>
</table>

Accurate, lightweight, and widely applicable
MLink: Linking Black-box Models for Collaborative Multi-model Inference

Thanks for your listening.

Mu Yuan (ym0813@mail.ustc.edu.cn), Lan Zhang, Xiang-Yang Li
University of Science and Technology of China