Fast Video Classification via Adaptive Cascading of Deep Models

Haichen Shen
Arvind Krishnamurthy
Seungyeop Han
Matthai Philipose

University of Washington
Rubrik
Microsoft
Recognizing entities in every frame of videos

- Convolutional neural networks ("Oracle" model)
  - ✔ High accuracy in recognizing thousands of classes
  - ✗ Expensive to execute

- Simpler convolutional neural networks ("Compact" model)
  - ✗ Low accuracy in recognizing thousands of classes
  - ✔ Cheap to execute

How can we reconcile this?
Object Skew in 1-minute video segments

- **DominantObjectCount**: # of objects that account for 80% of all object occurrences in 1-minute segments

- **Day-to-day video contains a tiny subset of classes in a short interval.**

70% of segments have DominantObjectCount $\leq 10$
Object Skew in 1-minute video segments

- DominantObjectCount: # of objects that account for 80% of all object occurrences in 1-minute segments

- Day-to-day video contains a tiny subset of classes in a short interval.

Can we exploit temporal skew in a video to accelerate the recognition speed?
Approach: Cascade oracle model with a less expensive "compact" model

Challenges:

• Can specialized models have accuracy comparable to oracle models?
• Can we produce specialized models fast enough during runtime?
• How to determine when to switch specialized models without any ground truth data?
Specialized models have comparable accuracy under skewed distributions

<table>
<thead>
<tr>
<th>Model</th>
<th>FLOPS</th>
<th>CPU lat.</th>
<th>GPU lat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogLeNet (oracle)</td>
<td>3.17G</td>
<td>779 ms</td>
<td>11.0 ms</td>
</tr>
<tr>
<td>Compact CNN</td>
<td>0.82G</td>
<td>218 ms</td>
<td>4.4 ms</td>
</tr>
</tbody>
</table>

Object recognition (1000 classes)
Producing specialized models can be fast

• We pre-train the compact models on the full, unskewed datasets during development time.

• At the test time, fix the lower layers and only re-train the top fully connected layer of the compact model.

• Cache feature vectors of compact models for all inputs in the training datasets.

Generate the specialized model ~10 seconds.
Bandit-style algorithm to determine when to switch specialized models

- Oracle Bandit Problem
  - Exploration: use the oracle model to estimate the distribution.
  - Exploitation: use a specialized model to accelerate the recognition.

- Windowed $\varepsilon$-Greedy (WEG) Algorithm
  - Adaptively select the windows size for sampling.
  - Produce a specialized model when a skew is detected.
  - Use heuristics to detect skew changes while “exploiting” specialized models.
## Evaluation

<table>
<thead>
<tr>
<th>video</th>
<th>length (min)</th>
<th>oracle</th>
<th>WEG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>acc. (%)</td>
<td>GPU lat. (ms)</td>
</tr>
<tr>
<td>Friends</td>
<td>24</td>
<td>93.2</td>
<td>28.97</td>
</tr>
<tr>
<td>Good Will Hunting</td>
<td>14</td>
<td>97.6</td>
<td>28.84</td>
</tr>
<tr>
<td>Ellen Show</td>
<td>11</td>
<td>98.6</td>
<td>29.26</td>
</tr>
<tr>
<td>The Departed</td>
<td>9</td>
<td>93.9</td>
<td>29.18</td>
</tr>
<tr>
<td>Ocean’s Eleven / Twelve</td>
<td>6</td>
<td>97.9</td>
<td>28.97</td>
</tr>
</tbody>
</table>