Multimedia Event Detection using Deep CNNs and Zero-Shot Classifiers

Nakamasa Inoue¹, Ryosuke Yamamoto¹, Na Rong¹, Satoshi Kanai¹, Junsuke Masada¹, Chihiro Shiraishi¹, Shi-wook Lee², and Koichi Shinoda¹
Tokyo Institute of Technology¹,
National Institute of Advanced Industrial Science and Technology²
Overview

- Method
  Supervised Classifiers + Zero-shot Classifiers
- Datasets for training
  ImageNet, Places, YFCC-Verb
- Results
  Mean AP: 52.9% (Ad-Hoc), 15.3% (Pre-Specified)
- Conclusion
  Supervised and zero-shot classifiers are complementary
  YFCC-Verb did not improve the performance
Method

A hybrid of supervised and zero-shot classifiers

Video → CNN+SVM

Video → Event Description → Zero-Shot Classifiers

Score

Late fusion
Supervised Classifiers

Convolutional neural network (CNN)

Deep Features*

Pooling

Visual Feature

SVM

Score

*1024 dimensional features are extracted from the pool5/7x7 layer3

Model: GoogLeNet

SVMs are trained by 10 example videos for each event

every 2 seconds
Zero-Shot Classifiers

Extract video vectors and event vectors

Concept Probabilities

Name, Definition, ...

Video Concept Vector \( \mathbf{v} \)

Event Concept Vector \( \mathbf{u} \)

\[ \text{cosine} \]

Score
Concept Vectors

• A video concept vector for a video clip $V$

$$v(V) = \sum_{i,c} p_{i,c} \phi(c)$$

Frame index Concept name

Word vector

• An event concept vector for an event $E$

$$u(E) = \sum_d \sum_{w \in W_d} \frac{\alpha_d}{|W_d|} \phi(w)$$

Set of words for description type $d$ (Name, Definition, etc.)

Weight

Word vector
Datasets for Training

• ImageNet for objects
  - ImageNet Shuffle [Mettes 2016]
  - 12,988 objects
• Places for scenes
  - 365 scenes [Zhou 2015]
• YFCC-Verb for actions
  - 4,126 verbs
  - 18,839 video clips
  - labels are generated from metadata
Verb Labels for YFCC

• 4,126 verb labels, 18,839 videos
• A subset of YLI-MED dataset [Bernd 2015]
• Labels are extracted from tags and video descriptions made by users
## Results

### Mean Average Precision for 4 submitted runs

<table>
<thead>
<tr>
<th>Method (Dataset)</th>
<th>MED-14 Kindred</th>
<th>MED-17 PS Events</th>
<th>MED-17 AH Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (ImageNet)</td>
<td>34.0</td>
<td>14.7</td>
<td>52.1</td>
</tr>
<tr>
<td>SVM (ImageNet+YFCC-Verb)</td>
<td>28.4</td>
<td>9.1</td>
<td>-</td>
</tr>
<tr>
<td>SVM+Zero-Shot (ImageNet)</td>
<td>36.4</td>
<td><strong>15.3</strong></td>
<td>-</td>
</tr>
<tr>
<td>SVM+Zero-Shot (ImageNet+Places)</td>
<td>38.1</td>
<td>15.1</td>
<td><strong>52.9</strong></td>
</tr>
</tbody>
</table>
Comparison with the Other Teams

- Mean AP by teams
AP by Events

![Graph showing AP by Events with different methods: SVM, SVM+Zero-Shot(ImageNet), SVM+Zero-Shot(ImageNet+Places), SVM(ImageNet), SVM(ImageNet+YFCC-Verb).]
Conclusion and Future Work

- Method: A hybrid system of supervised classifiers and zero-shot classifiers
- Mean AP: 52.9% (Ad-Hoc), 15.3% (Pre-Specified)
  - Supervised and zero-shot classifiers are complementary
  - YFCC-Verb did not improve the performance
- Future Work
  - action recognition, audio analysis