University of Amsterdam’s Deep Net for Video Event Detection

Pascal Mettes, Spencer Cappallo, Dennis Koelma, Cees G. M. Snoek

University of Amsterdam
Summary

Top performance for example-based event detection tasks.
This talk

Organizing ImageNet Hierarchy

Training Deep Network

Sampling frames

Extracting features

Pooling to video representation

Train videos

Training SVM

Learning the frame representation.

Pooling frames to video representation.
This talk

Organizing ImageNet Hierarchy → Training Deep Network → Sampling frames → Extracting features → Pooling to video representation → Training SVM

Learning the frame representation.
Google’s Inception Network [Szegedy et al. CVPR 2015].

- Very deep network with inception modules.
- Trained with standard ImageNet setup.
- 1.2 million images from 1,000 classes.
Observation

Not all 1,000 classes are equally relevant for event detection.

Only 8% of complete ImageNet hierarchy is used.

- Full ImageNet hierarchy contains 14 million images from 21,841 classes.

We leverage the complete ImageNet hierarchy for training.
Problems with the complete hierarchy

Imbalance in image distribution.
- ‘Yorkshire terrier’ has 3047 examples.
- 296 classes have 1 example.

Over-specific classes for event detection.
- ‘siderocyte’ and ‘gametophyte’ not likely to be relevant for event detection.
Four proposals for reorganizing ImageNet
Four proposals for reorganizing ImageNet

Proposal 1: **Roll up** all classes with only 1 child.
Proposal 2: **Bind** all subtrees with less than 3000 examples.
Proposal 3: **Promote** all classes with less than 200 examples.
Proposal 4: Sample for classes with more than 2000 examples.
Advantages of our proposal

1. All images in the ImageNet hierarchy are used.

2. Over-specific and small classes are merged with their parents.

3. Compact semantic frame representations (12,988 classes).
Pooling frames to video representation.
Pooling: Main idea

An event video is an interplay of sub-events.

We aim to pool over individual sub-events, not average over all.
Algorithm overview

Find the most discriminative fragments from training videos.

Encode a video using a score for each discriminative fragment.

Step 1: Propose

Step 2: Select

Step 3: Encode

[Mettes et al. ICMR 2015]
Algorithm overview

Find the most discriminative fragments from training videos.

Encode a video using a score for each discriminative fragment.

Step 1: Propose

Step 2: Select

Step 3: Encode

[Mettes et al. ICMR 2015]
Algorithm overview

[Mettes et al. ICMR 2015]

Find the most discriminative fragments from training videos.

Encode a video using a score for each discriminative fragment.

**Step 1: Propose**

Training video

**Step 2: Select**

**Step 3: Encode**

Video

Encoding
Experiments

- Organizing ImageNet Hierarchy
- Training Deep Network
  - Sampling frames
    - Extracting features
      - Pooling to video representation
        - Training SVM
Experiment 1: AlexNet vs. GoogleNet

GoogleNet outperforms AlexNet.
Experiment 2: 1,000 vs. all ImageNet classes

GoogleNet outperforms AlexNet.
Using all ImageNet classes helps.
Experiment 3: Our ImageNet reorganization

GoogleNet outperforms AlexNet.

Using all ImageNet classes helps.

We do better than directly using all classes.

Our feature vector is twice as small.
Experiment 4: 100 Example results

GoogleNet outperforms AlexNet.

Using all ImageNet classes helps.

We do better than directly using all classes.

Our feature vector is twice as small.

Idem for 100 Examples.
Experiment 5: Average pooling vs. Bag-of-Fragments

MED 2014 100 Examples:

<table>
<thead>
<tr>
<th>Method</th>
<th>AlexNet [ICMR results]</th>
<th>GoogleNet [new results]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averaging</td>
<td>0.232</td>
<td>0.351</td>
</tr>
<tr>
<td>Bag-of-Fragments</td>
<td>0.276</td>
<td>0.317</td>
</tr>
<tr>
<td>Combination</td>
<td>0.373</td>
<td>0.381</td>
</tr>
</tbody>
</table>

Bag-of-Fragments is both competitive and complementary to average pooling.
Fusion:
- Deep Net with averaging.
- Motion (MBH with Fisher Vectors).
- Audio (MFCC with Fisher Vectors).

Results:
- Our fusion yields top result.
- ‘Deep Net only’ already near top.
Fusion:
- Deep Net with averaging.
- Deep Net with Bag-of-Fragments.
- Motion (MBH with Fisher Vectors).
- Audio (MFCC with Fisher Vectors).

Results:
- Our fusion yields top result.
- ‘Deep Net only’ second place.
Conclusions

Training on organized ImageNet hierarchy helps event detection.

Bag-of-Fragments yields complementary video representations.
Pascal Mettes

- mail: P.S.M.Mettes@uva.nl
- address: Science Park 904, Amsterdam