Video Search when examples are scarce

Dennis Koelma and Cees Snoek
University of Amsterdam
The Netherlands
Overview

• Pipeline 10ex
  • ImageNet shuffle
  • Video Story
  • Results

• Pipeline 0ex
  • Video Story
  • Concepts
  • Results

• Conclusions
Pipeline 10Ex 2015

- **Videos**: sample 2/sec
- **Frames**: CNN ImageNet Shuffle
- **pool5**: SVM 10Ex M1
- **avg pool**: SVM 10Ex M2
- **prob**: SVM 10Ex M3
- **Fisher vector**: SVM 10Ex M4
- **dense trajectories**: SVM 10Ex M2
- **mfcc0 mfcc1 mfcc2**: SVM 10Ex M4
22k ImageNet classes

- Use as many classes as possible
- Find a balance between level of abstraction of classes and number of images in a class

Example imbalance

296 classes with 1 image

Irrelevant classes

Siderocyte

Gametophyte
CNN training on selection out of 22k ImageNet classes

• Idea
  • Increase level of abstraction of classes
  • Incorporate classes with less than 200 samples

• Heuristics
  • Roll, Bind, Promote, Subsample

• Result
  • 12,988 classes
  • 13.6M images

Video Story: Embed the story of a video

Joint optimization of $W$ and $A$ to preserve

**Descriptiveness**: preserve video descriptions: $L(A,S)$

**Predictability**: recognize terms from video content: $L(S,W)$

Results 10Ex Individual Modalities on 2014 Test Set

MAP

Regular contribution

E023-Dog_show
E025-Marriage_proposal
E027-Rock_climbing
E028-Town_hall_meeting
E031-Beekeeping
E036-felling_a_tree
E038-Playing_fetch
E040-Tuning_musical_instrument
E041-
Results 10Ex Individual Modalities on 2014 Test Set

No way

- marriage proposal
- playing fetch

- pool5
- prob
- mbh
- mfcc
- vs
Results 10Ex Individual Modalities on 2014 Test Set

Good MBH / MFCC

- Good MBH / MFCC
- Felling tree
- Tuning musical instrument

Legend:
- pool5
- prob
- mbh
- mfcc
- vs
10Ex Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test</td>
<td>EvalFull</td>
<td>EvalFull</td>
<td>EvalFull</td>
<td>EvalFull</td>
</tr>
<tr>
<td></td>
<td>Progress</td>
<td>Progress</td>
<td>Progress</td>
<td>Progress</td>
<td>Progress</td>
</tr>
<tr>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP</td>
<td>0.254</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>prob</td>
<td>0.256</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mbh</td>
<td>0.127</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mfcc</td>
<td>0.069</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vs</td>
<td>0.258</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pool5 + prob</td>
<td>0.272</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pool5 + prob + mbh + mfcc (2015)</td>
<td>0.283</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pool5 + prob + vs</td>
<td>0.279</td>
<td>0.283</td>
<td>0.368</td>
<td>0.179</td>
<td>0.445</td>
</tr>
<tr>
<td>pool5 + prob + mbh + mfcc + vs</td>
<td>0.290</td>
<td>0.290</td>
<td>0.394</td>
<td>0.187</td>
<td>0.463</td>
</tr>
</tbody>
</table>

- Top performance in 2015 and 2016
- Some progress but not a lot
- We shifted focus to 0Ex
Video Story for 0Ex

A crazy guy doing insane stunts on bike

Original Bike Tricks from Tim Knoll

0.45 bike
0.30 man

Cosine similarity

Attempting a bike trick

Embedding

W

A

A

w2v

Stunt

Bike

Motorcycle

A

A

w2v

Stunt

Bike

Motorcycle

A

A

w2v
Pipeline 0Ex

- Videos
- Frames
- pool5
- Video Story embedding
- CNN ImageNet Shuffle
- avg pool
- term vector
- cosine similarity 0Ex M1
- word2vec expanded event text
- CNN FCVID
- UCF101
- TV13
- prob
- filtered prob
- cosine similarity 0Ex M2
- top 3 concepts closest to event title in word2vec space
- sample 2 / sec
Video Story Training Sets

- Amir’s YouTube46k - [www.mediamill.com](http://www.mediamill.com)
  - 45826 videos from YouTube based on 2013 MED research set terms
- FCVID: Fudan Columbia Video Dataset
  - 87609 videos
- Merged

- Video Story dictionary: Terms that occur more than 10 times in the dataset
  - Merged: 5587 terms
- Using vocabulary of stemmed terms that occur more than 100 times in Wikipedia dump
  - With stemming: Respect the Video Story dictionary
    - 267,836 terms
- Use word2vec to expand them per video
Results 0Ex Video Story on 2014 Test Set

• Fails on 5 events (AP < 0.01), unusable on 2 events (AP < 0.05)
Concept Bank

• Datasets
  • FCVID
    • 233 concepts
    • Shot segmentation -> max 5 keyframes / video -> max 3000 keyframes / concept (expand within shot if less than 3000)
  • UCF101
    • 101 concepts
    • Shot segmentation -> max 5 keyframes / video -> max 3000 keyframes / concept (expand within shot if less than 3000)
  • TV13 SIN concepts
    • 346 concepts
    • max 3000 keyframes / concept (expand within shot if less than 3000)

• CNN is finetune on ImageNet Shuffle network trained on 13k classes

• Two CNN’s:
  • FCVID
  • FCVID + UCF101 + TV13
Results 0Ex Concept Bank on 2014 Test Set

• Fails on 9 events (AP < 0.01), unusable on 1 event (AP < 0.05)
• FCVID CNN is main contributor
• FcvidUcfTv CNN is worse but fusion makes it a bit better overall
Results 0Ex VideoStory Concept Fusion on 2014 Test Set

• Fails on 3 events (AP < 0.01), unusable on 7 events (AP < 0.05)
• Concepts are slightly better than Video Story but fusion is best
## OEx Results

<table>
<thead>
<tr>
<th>Test</th>
<th>2014</th>
<th>2016</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>EvalFull</td>
<td>MAP</td>
<td>Progress</td>
<td>InfMAP</td>
</tr>
<tr>
<td>Video Story Amir</td>
<td>0.060</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video Story FCVID</td>
<td>0.118</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video Story Merged</td>
<td>0.133</td>
<td></td>
<td></td>
</tr>
<tr>
<td>concepts FCVID</td>
<td>0.143</td>
<td></td>
<td></td>
</tr>
<tr>
<td>concepts FCVID + UCF + TV</td>
<td>0.140</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video Story</td>
<td>0.146</td>
<td></td>
<td></td>
</tr>
<tr>
<td>concepts</td>
<td>0.150</td>
<td>0.171</td>
<td>0.135</td>
</tr>
<tr>
<td>concepts + Video Story</td>
<td>0.167 (0.175)</td>
<td>0.181</td>
<td>0.149</td>
</tr>
</tbody>
</table>

- The “trend” is the same
- Top performance with fully automatic search
Conclusions

• Video Story for 0Ex benefits from “carefully” selected training material

• Concepts produce higher MAP than Video Story but Video Story is applicable to more events

• Fully automatic video search with just a few examples is becoming feasible
  • 0ex is doable when relevant concepts are present
    • You just have to find them
  • 10ex still makes a big difference
Thank You

• Video Story - http://www.mediamill.nl

• ImageNet Shuffle CNN’s - http://tinyurl.com/imagenetshufflle