Segments, Residuals and Embeddings for Few-Example Video Event Detection

Dennis Koelma and Cees Snoek
University of Amsterdam
The Netherlands
Pipeline 10Ex 2016

- **Videos**
  - Sample 2/sec

- **Frames**
  - CNN Inception ImageNet Shuffle
  - Pool5

- **pool5**
  - SVM 10Ex M1
  - Video Story embedding
  - SVM 10Ex M2
  - SVM 10Ex M3

- **prob**
  - SVM 10Ex M2

- **dense trajectories**
  - Fisher vector

- **mfcc0**
  - Fisher vector

- **mfcc1**

- **mfcc2**

- **SVM 10Ex M4**

- **SVM 10Ex M5**
Pipeline 10Ex 2017

- Videos
  - sample 2 / sec
  - ResNet + ResNeXt
  - ImageNet
  - Shuffle
- Frames
- pool5
- difference coding
- SVM 10Ex M1
- Video Story embedding
- SVM 10Ex M5
- sliding window
- avg pool
- SVM 10Ex M2
- dense trajectories
- Fisher vector
- SVM 10Ex M3
- mfcc0, mfcc1, mfcc2
- Fisher vector
- SVM 10Ex M4
CNN Features from 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
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

Feature Difference Coding

• K-means clustering ($k = 5$) on last fully connected layer before probability layers (called flatten)

• Fisher like encoding but sigma is based on distance of points assigned to a cluster to its center
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: $L(S,W)$

VideoStory Embedding as a Feature

MAP 2014 Test Set

flatten-avg

video story

ResNet  |  ResNeXt  |  Fusion

0.300  |  0.305  |  0.310
Video Story for 0Ex

Attempting a bike trick

0.45 bike
0.30 man

Cosine similarity

1.0 attempt
1.0 bike
1.0 trick

Embedding

$X_i \xrightarrow{W} S_i$
Finding Segments to Expand Training Material

Example1

Cosine similarity
Window based Features

MAP 2014 Test Set

- ResNet
- ResNeXt
- Fusion
Result Individual Modalities on 2014 Test Set

- flatten-avg
- softmax
- trajectories
- mfcc
- video story
- flatten-dc
- flatten-window

- DC is best
- R < Rx < F
- overfit?
- VS > flatten
- window > avg

- ResNet
- ResNeXt
- Fusion
Fusion Visual Modalities on 2014 Test Set

ResNet + ResNeXt

VS
DC
Win
DC-VS
DC-Win
VS-DC-Win
Fusion on 2014 Test Set

- last year new features
- single mod
- visual fusion
- MM fusion
- + avg

Bar chart showing fusion results with different models and features:

- AVG2-DT-MFCC-VS
- DC
- VS-DC-Win
- VS-DC-Win-DT-MFCC
- VS-DC-Win-DT-MFCC-AVG2

Models compared:
- ResNeXt
- ResNet+ResNeXt
Computational Efficiency

**MAP**
- p-visualFusionTwoCNN
- c-mmFusionTwoCNN
- c-visualFusionOneCNN
- c-mmFusionOneCNN
- c-visualSingle

**Feature Extraction**
- p-visualFusionTwoCNN
- c-mmFusionTwoCNN
- c-visualFusionOneCNN
- c-mmFusionOneCNN
- c-visualSingle

**Classification**
- p-visualFusionTwoCNN
- c-mmFusionTwoCNN
- c-visualFusionOneCNN
- c-mmFusionOneCNN
- c-visualSingle
Our MED Submission

Test 2014

PS

AH

- p-visualFusionTwoCNN
- c-mmFusionTwoCNN
- c-visualFusionOneCNN
- c-mmFusionOneCNN
- c-visualSingle
Conclusions

• Visual features are still improving
• Fusion still works but other modalities need work
• 0ex helps to get more out of your examples
Thank You