Multimedia Event Detection using Deep CNNs and Zero-Shot Classifiers

Nakamasa Inoue[1], Rryosuke Yamamoto[1], Na Rong[1], Satoshi Kanai[1], Junsuke Masada[1], Chihiro Shiraishi[1], Shi-wook Lee[2], and Koichi Shinoda[1]

Tokyo Institute of Technology[1],

National Institute of Advanced Industrial Science and Technology[2]

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)

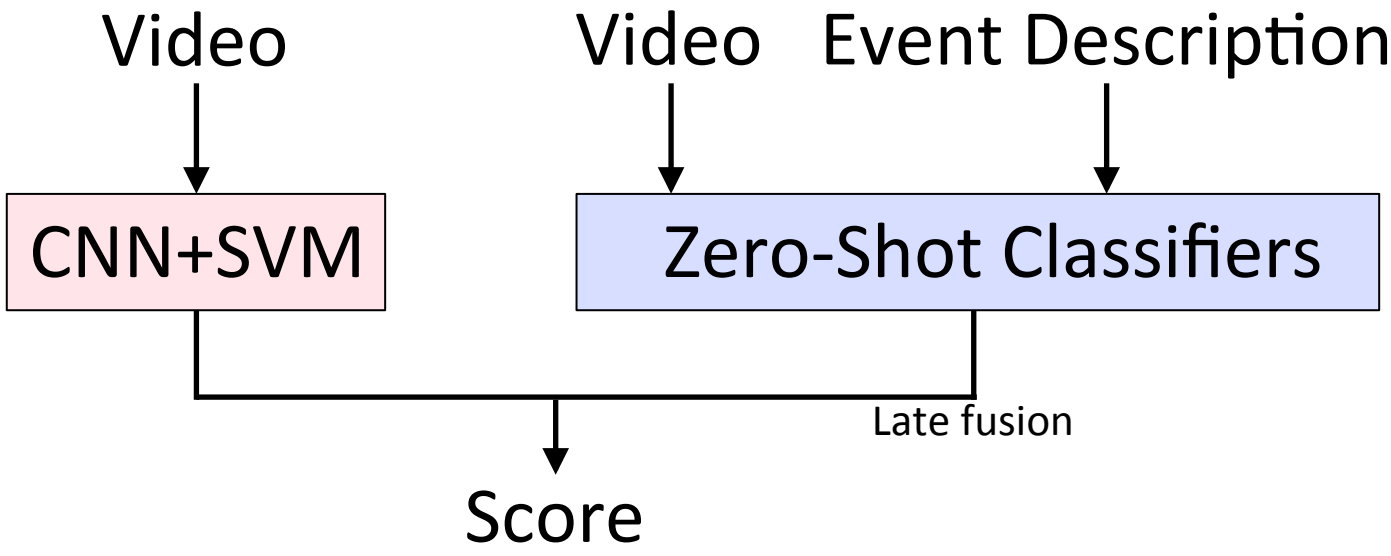
- Conclusion

  Supervised and zero-shot classifiers are complementary

  YFCC-Verb did not improve the performance

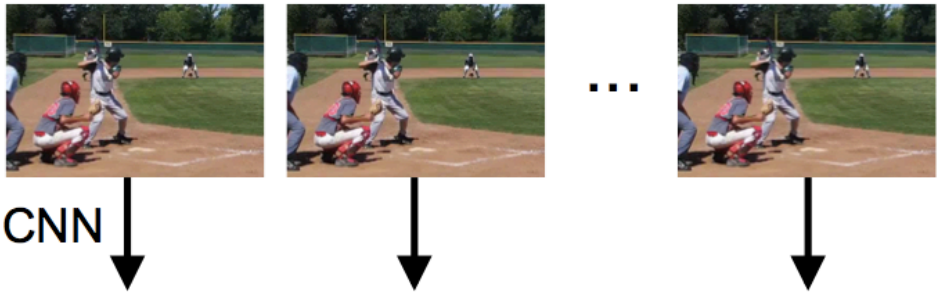
Method

A hybrid of supervised and zero-shot classifiers



Supervised Classifiers

Convolutional neural network (CNN)



every 2 seconds

Deep Features*

Model: GoogLeNet

Pooling

Visual Feature

SVM

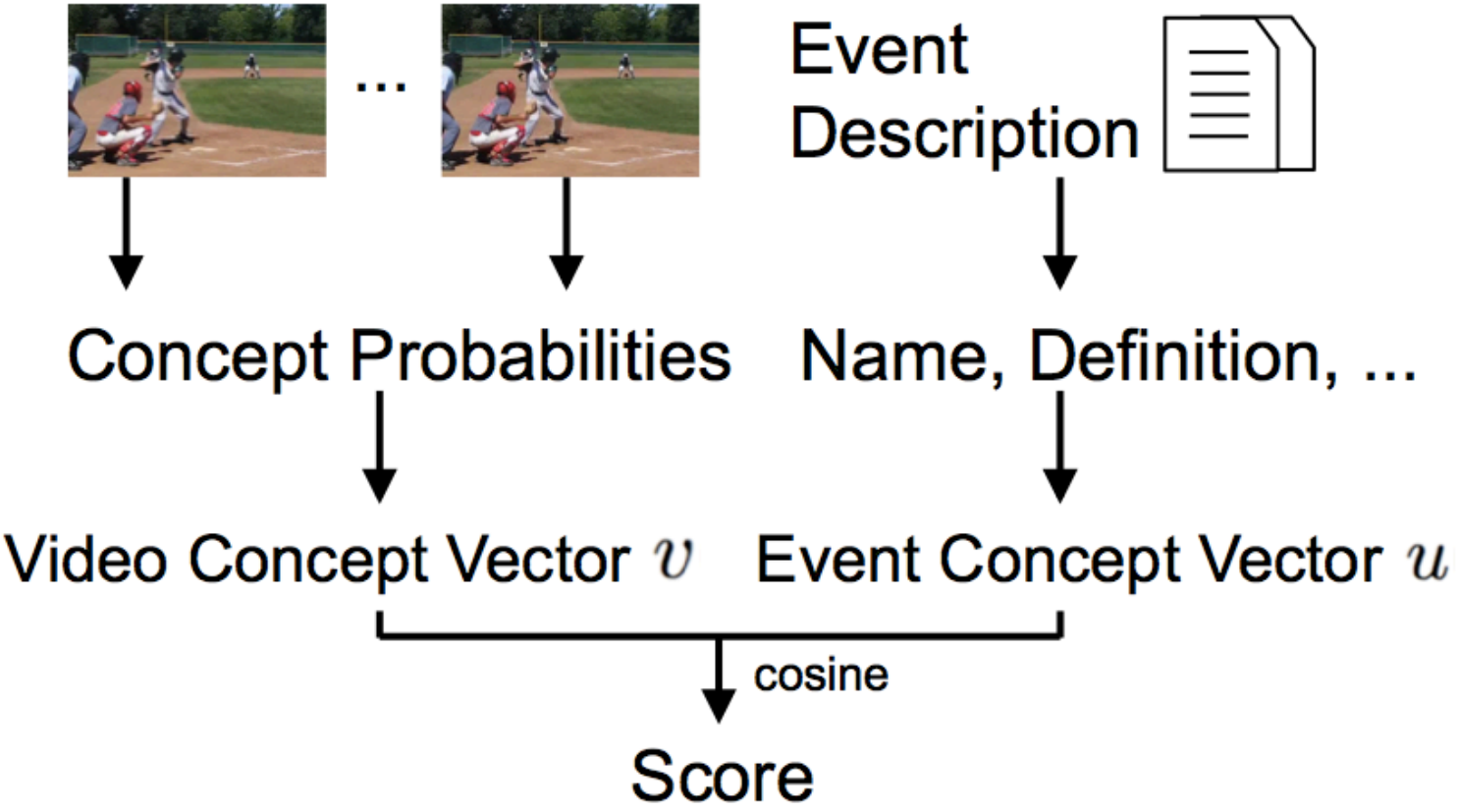
Score

SVMs are trained by 10 example videos for each event

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

Zero-Shot Classifiers

Extract video vectors and event vectors



Concept Vectors

- A video concept vector for a video clip *V*

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

Word vector

Frame index Concept name

- An event concept vector for an event *E*

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

Weight

Word vector

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

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

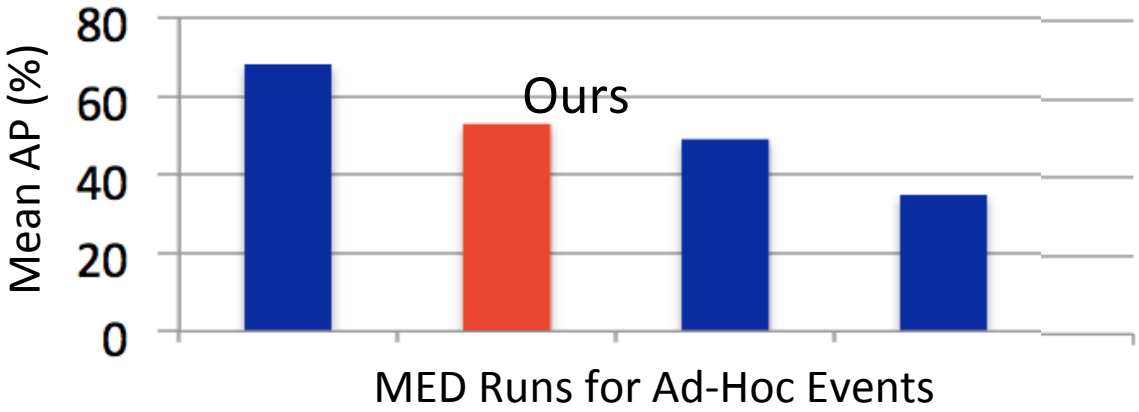
Results

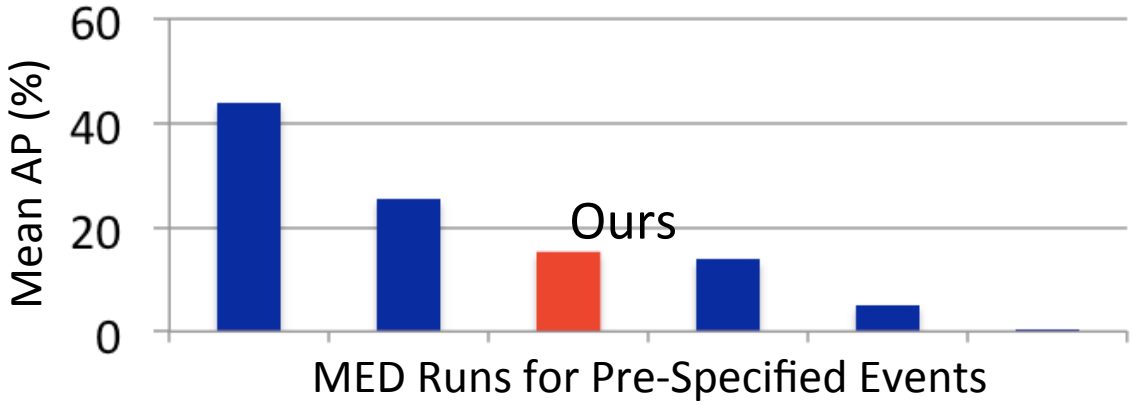
Mean Average Precision for 4 submitted runs

Method (Dataset)	MED-14 Kindred	MED-17 PS Events	MED-17 AH Events
SVM (ImageNet)	34.0	14.7	52.1
SVM (ImageNet+YFCC-Verb)	28.4	9.1	-
SVM+Zero-Shot (ImageNet)	36.4	**15.3**	-
SVM+Zero-Shot (ImageNet+Places)	38.1	15.1	**52.9**

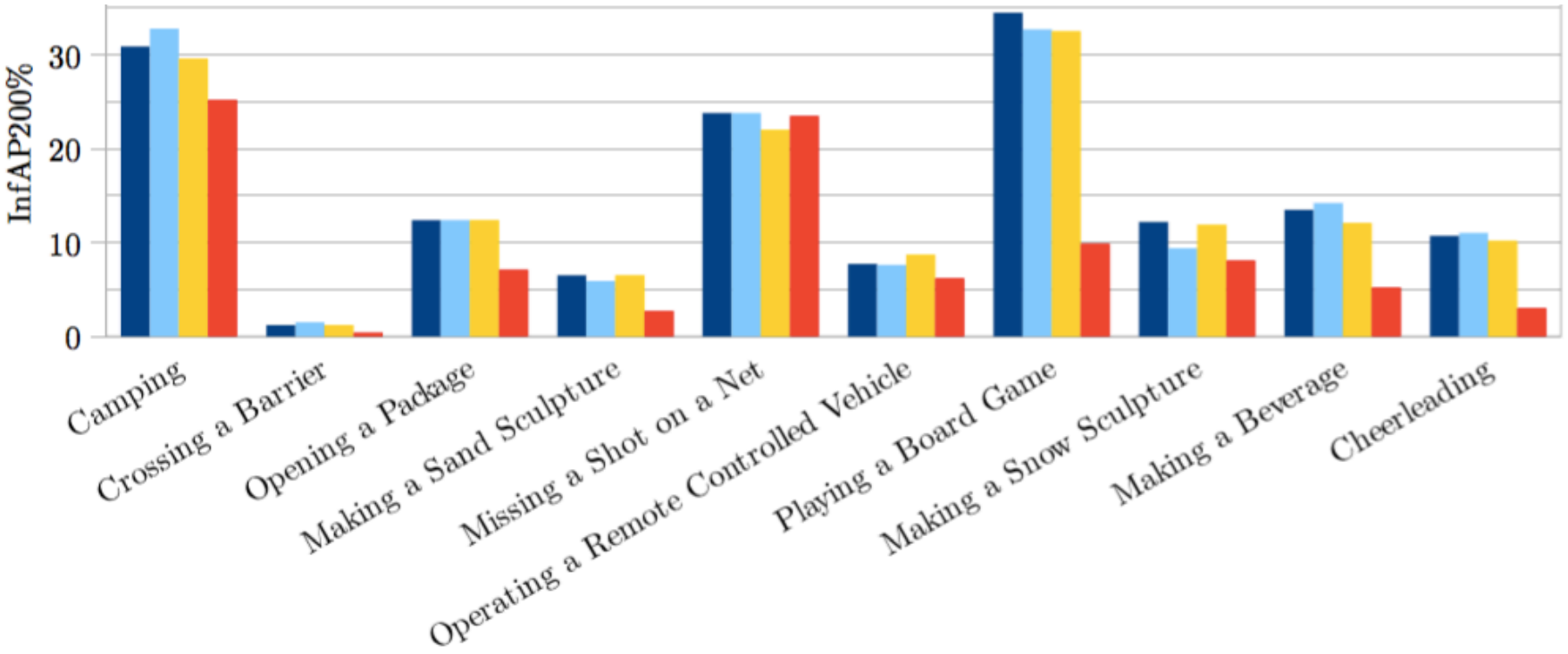
Comparison with the Other Teams

- Mean AP by teams





AP by Events



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