VIREO@TRECVID 2016: Multimedia Event Detection, Ad-hoc Video Search, Video to Text Description

Hao Zhang, Lei Pang, Yi-Jie Lu, Chong-Wah Ngo

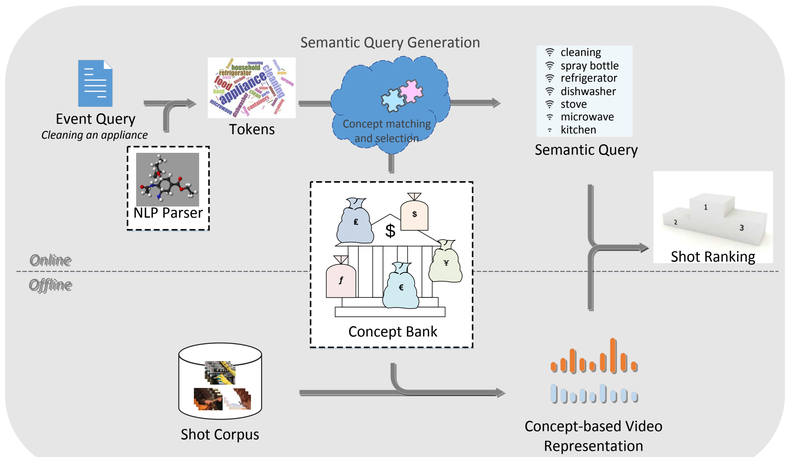
VIREO Video Retrieval Group

Video Retrieval Group (VIREO), City University of Hong Kong

http://vireo.cs.cityu.edu.hk

Ad-hoc Video Search (AVS)

Framework



Highlighted Concept Reranking (HCR)

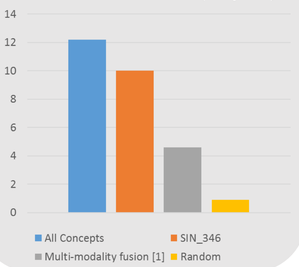
- After ***Semantic Query Generation***, human experts can highlight a few important concepts .
- The videos on the top which contain the highlighted concepts are given high weights and their ranking are largely boosted.

Vocabulary

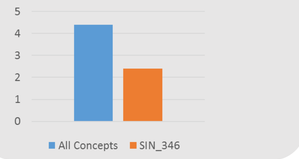
ImageNetShuffle-13K
ImageNet-1000
Places-205
SIN-346
Research Collection-497

Experiments

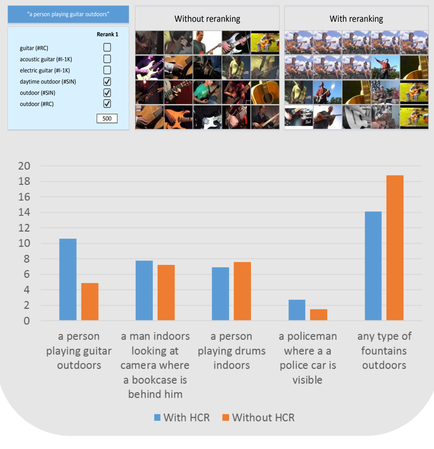
Performance for different concept settings in Video-Search 2008 dataset (MinfAP%)



Performance for different concept settings in AVS 2016 evaluation dataset (MinfAP%)

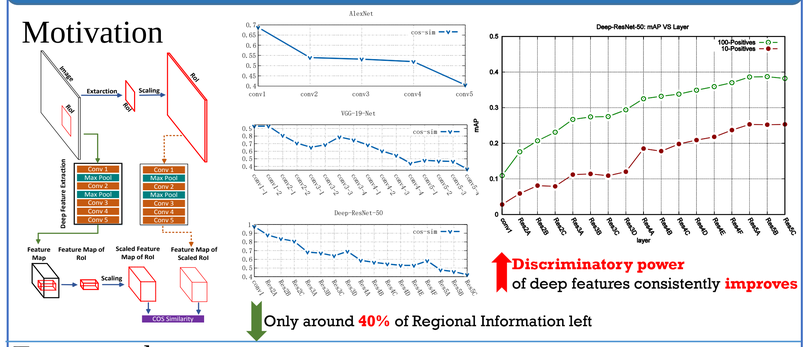


Performance with and without highlighted concept ranking (HCR)

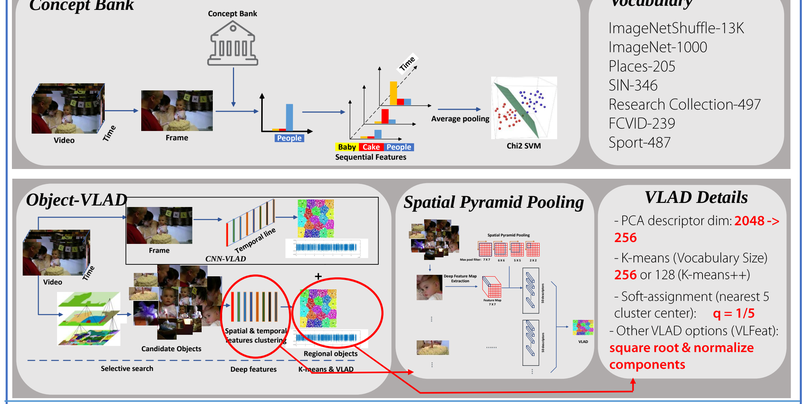


Multimedia Event Detection (MED)

Motivation



Framework



Concept Bank

Vocabulary

ImageNetShuffle-13K
ImageNet-1000
Places-205
SIN-346
Research Collection-497
FCVID-239
Sport-487

Object-VLAD

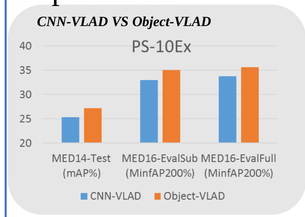
Spatial Pyramid Pooling

VLAD Details

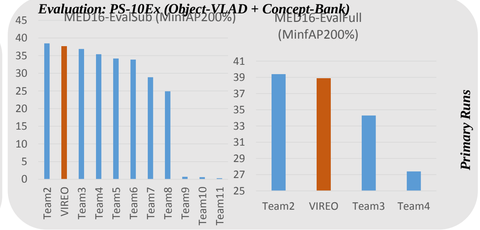
- PCA descriptor dim: **2048 -> 256**
- K-means (Vocabulary Size) **256** or 128 (K-means++)
- Soft-assignment (nearest 5 cluster center): **q = 1/5**
- Other VLAD options (VLFeat): **square root & normalize components**

Experiments

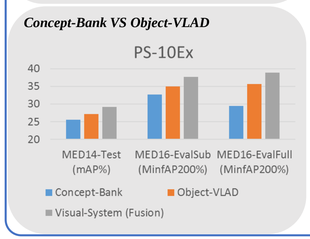
CNN-VLAD VS Object-VLAD



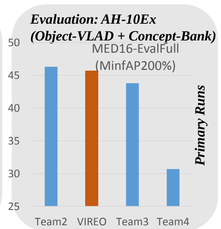
Evaluation: PS-10Ex (Object-VLAD + Concept-Bank)



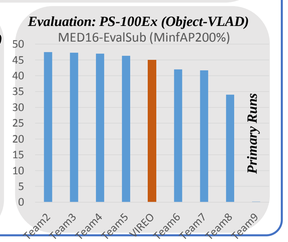
Concept-Bank VS Object-VLAD



Evaluation: AH-10Ex (Object-VLAD + Concept-Bank)



Evaluation: PS-100Ex (Object-VLAD)



Video to Text Description (VTT)

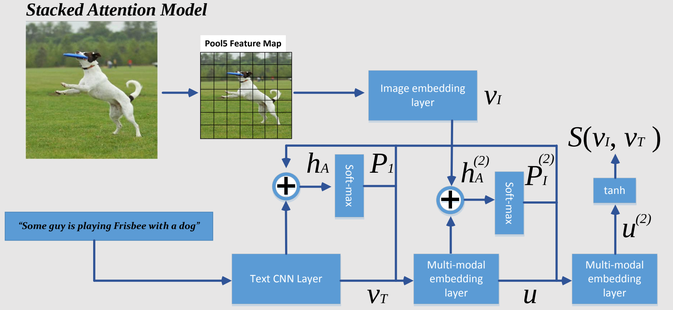
Framework

Concept-based Text Description Matching

Zero-example MED system in reversed pipeline for concept-based video to text retrieval.

- Each video is represented as concept vector (max pooling in temporal domain)
- Query sentences are feed in SQG module for generating the concept-based sentence representation (WordNet to find synonyms).
- Similarity calculation between video and query representation

Stacked Attention Model



f_I **: feature for each region**

1st Attention Layer

$$v_I = \tanh(W_I f_I + b_I)$$
$$h_A = \tanh(W_{I,A} v_I \oplus (W_{R,A} v_T + b_A))$$
$$P_I = \mathrm{softmax}(W_P h_A + b_P)$$
$$\widetilde{v_I} = \sum_1^m p_i v_i$$
$$u = \widetilde{v_I} + v_T$$

2nd Attention Layer

$$h_A^{(2)} = \tanh(W_{I,A}{}^{(2)} v_I \oplus (W_{R,A}{}^{(2)} u + b_A{}^{(2)}))$$
$$P_I^{(2)} = \mathrm{softmax}(W_P{}^{(2)} h_A{}^{(2)} + b_P{}^{(2)}))$$
$$\widetilde{v_I}^{(2)} = \sum_1^m p_i^{(2)} v_i$$
$$u^{(2)} = \widetilde{v_I}^{(2)} + u$$

Objective-Function

$$S < v_I, v_T > = \tanh(W_{u,s} u^{(2)} + b_s)$$
$$l(W, D_{trn}) = \sum_{(v_I, v_T^-, v_T^+) \in D_{trn}} \max(0, \delta + S < v_I, v_T^- > \; - S < v_I, v_T^+ >)$$

Training Dataset

Flickr30K
Coco
TGIF

Deep Features

- Visual Content: VGG Feature
  Motion: C3D Feature
- Sentence: Textual CNN

Experiments