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

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

[†] *Video Retrieval Group (VIREO), City University of Hong Kong*
http://vireo.cs.cityu.edu.hk

Abstract

The vireo group participates in 3 tasks: multimedia event detection, ad-hoc video search, video-to-text description. In this paper, we will separately present frameworks for these tasks and discuss experimental results.

Multimedia Event Detection (MED): We will present an overview and comparative analysis of our system designed for the TRECVID 2016 [1] multimedia event detection (MED) task. We submit 12 runs, of which 3 runs for the pre-specified zero example (PS-0Ex), 4 runs for the pre-specified ten examples (PS-10Ex), 3 runs for the pre-specified hundred examples (PS-100Ex) and 2 runs for the ad-hoc ten examples (AH-10Ex). We do not participate in the interactive run. This year we focus on two aspects: 1) Refining concept bank by extending its vocabulary size and replacing its classifiers with the state-of-the-art deep architecture; 2) Capturing visual scenario of video with a finer granularity object-level representation. Among our 12 submitted runs, the following runs achieved the top-2 performances:

- VIREO_MED16_MED16EvalFull_PS_0Ex_MED_p-VisualTextualRernk_4: Zero-example system with manually refined semantic query, highlighted concept reranking and late fusion with OCR (Best 0Ex performance).
- VIREO_MED16_MED16EvalFull_PS_10Ex_MED_p-Fus_4: 10-example system using Concept-Bank feature and Object-VLAD video representation (Ranked 2nd on EvalSub/Full set in terms of MinfAP200).
- VIREO_MED16_MED16EvalFull_AH_10Ex_MED_p-fus_4: 10-example system using Concept-Bank feature and Object-VLAD video representation (Ranked 2nd on EvalFull set and 1st on EvalSub set in terms of MinfAP200).

Ad-hoc Video Search (AVS): For AVS, we adopt our successful zero-example MED system developed in the last year. In addition, we add a new *highlighted concept reranking* feature to further enhance human-in-the-loop. This technique allows a user to emphasize a group of concepts highlighting a specific characteristic. For example, "outdoors" and "at night." We submit a total of 6 runs, with variations include (1) fully automatic or manually assisted query processing, (2) with or without highlighted concept reranking, and (3) using concepts learned only from IACC data or not. The best runs in the respective two training set types are:

- M_D_VIREO.16_5: Manually assisted query processing. Concept detectors of more than 10K which are collected from several off-the-shelf public datasets.
- M_A_VIREO.16_3: Manually assisted query processing. 346 concept detectors are only learned from IACC.1 and IACC.2 data themselves.

Video-to-Text Description (VTT): We only participate in the matching and ranking subtask. The task here is to return for each video URL a ranked list of the most likely text description that correspond to the video. This year we focus on the concept-based retrieval of sentence and inter-modality correspondence learning by salient region selection. Our four runs are summarised below:

- Vireo-Average-VGG-C3D: Each video is represented by average pooling over the VGG [2] features extracted from frames and text description is represented by the TCNN [3] feature. Stacked attention network (SAN) [4] is adopted to learn the implicit alignment between visual and textual features. Furthermore, C3D [5] is also employed for the extraction of motion features. Similarity scores from the two SANs trained on VGG-TCNN and C3D-TCNN are averagely fused for the final ranking.
- Vireo-Flat-VGG-C3D: Rather than average pooling, frame features are concatenated into one feature vector to preserve the unique information of different frames. Then, two SANs are trained on the concatenated features, whose similarity scores are averagely fused for final ranking.
- Vireo-Fusion: To observe whether Vireo-Average-VGG-C3D and Vireo-Flat-VGG-C3D are complementary to each other, we further averagely fuse the scores from these two runs for ranking.
- Vireo-Concept: Each video is represented by our Concept-Bank feature and the detected concepts are matched to the keywords in text descriptions based on a flexible phrase matching method, which is combined with word similarities of WordNet [6] and weighting strategies, such as TF-IDF.

1 Multimedia Event Detection (MED)

Compared with our multimedia event detection system proposed in TRECVID 2015 benchmark, we modify visual systems for PS-10Ex, AH-10Ex and PS-100Ex subtasks, whereas, keep the subsystem for PS-0Ex. Specifically, the differences of visual systems are summarized as follows: 1) Hand-crafted features (e.g., the improved dense trajectory [7]) are no longer used in visual system; 2) By including ImageNet-Shuffel_13K [8], the size of concept bank is increased to 15,762 and parts of concept classifiers are re-trained with the architecture of deep residual neural networks [9]; 3) A finer granularity object-level video representation named Object-VLAD is introduced in our visual systems. Experiments conducted on MED14-Test and MED16-EvalFull datasets verify the effectiveness of our modification. The next section explains how to generate different modalities for our system. Section 1.1 and Section 1.2 describe our systems and runs for the different amount of training examples, i.e., 0Ex, 10Ex, or 100Ex. Section 1.3 describes results and analysis. Section 1.4 concludes the MED section.

1.1 Modalities

In our system we use three types of modalities: visual, texutal and speech.

1.1.1 Visual

For our visual system, we first decompose each video into two-granularity levels – keyframe level and shot level. The keyframe sampling rate is set to be one frame per two seconds and the time duration of shot is set to be five seconds. For each video, we generate 7 kinds of high-level concept feature and 1 kind of object-level deep feature. All the features used in our visual system are summarized as below:

- *ImageNet_1000*
  We use the architecture of deep residual networks (i.e., ResNet-50) proposed by K. He [9] to train ImageNet-1000 classifiers. The parameters of ResNet-50 are learned on ILSVRC dataset [10], which is a subset of ImageNet dataset with 1.26 million training images for 1,000 categories. The neural responses of the output layer (soft-max) for each keyframe are averagely pooled to form the video-level feature vector.

- *SIN_346*

  A set of 346 concept detectors fine-tuned with the ResNet-50 architecture on TRECVID-SIN 2014 dataset [11] is applied on all keyframes and averagely pooled to form a video-level representation.

- *RC_497*

  Similar to [12], we select 497 concepts from the MED14 Research Collection dataset [13]. We manually annotate at most 200 positive keyframes for each concept and fine-tuned 497 concept detectors using the RestNet-50 architecture. Same as the prior methods, we concatenate the responses of the concept classifiers on each keyframe and averagely pool the resulting feature vectors to form a video feature representation.

- *Places_205*

  To capture the scenery information of multimedia events, we fine-tune 205 scene categories on MIT places dataset [14] with the RestNet-50 architecture. The scenery concept responses are extracted from each keyframe and further averagely pooled to generate the video-level representation.

- *FCVID_239*

  To capture the action/motion information of multimedia events, 239 concept detectors are trained with SVM on FCVID [15] dataset, which contains 91,223 web videos. Since concepts from this dataset are mainly action/motion concepts annotated at video-level, we extract neural responses from fc7 layers of AlexNet [16] for each frame and pool them across the temporal domain to generate training features and train video-level concept detectors. We concatenate the responses of 205 concept detectors of an Image.

- *Sports_487*

  487 concept detectors are trained with 3D CNN structure [5] on the Sport-1M dataset [17] containing 1 million videos. Similar to FCVID, Sports-1M are mainly action/motion concepts annotated at video level.

- *ImageNet-Shuffle_12988*

  In [8], P. Mettes et. al proposed to reorganize 21,814 classes of full ImageNet [10] dataset by merging them into 12,988 categories for training a large-scale concept bank with GoogLeNet architecture [18]. We extract neural responses from output layer of the published ImageNet-Shuffle_13K-GoogLeNet for each frame and averagely pool them across the temporal domain to generate video-level representation.

  To capture semantical meaning of multimedia event, we collect all the above concept features and form a large ***Concept-Bank*** with 15,762 semantical concepts related to visual objects, background scenery and actions.

- *Object-VLAD*

  Due to the employment of max-pooling and softmax layer in DCNN structure, concept detectors tend to overlook activations of small size objects, emphasize those of primary objects and produce a very sparse semantical vector. To equally accumulate responses of multiple objects within a frame, we propose a finer granularity object-level representation named Object-VLAD for videos. Specifically, for each video, we separately encode frame-level features and object-level features with VLAD. For the prior, we directly adopt CNN-VLAD [19], which encodes frame-level features with VLAD, whereas for the latter, we utilize selective search [20] for proposing candidate objects, describe objects with deep features and encode them with VLAD. We adopt neural responses from Res5A layer of the RestNet-50 architecture (trained on ImageNet 1000 categories) [9] as feature descriptors for each frame and regional object.

1.1.2 Textual

Tesseract OCR [21] is used to extract text from video frames. The engine is applied in a brute-force manner with post processing on the resulting extracted texts. From a video segment, every key video frame (roughly one per second) is decoded using the FFmpeg open-source library. All key frames are then fed to the Tesseract engine for OCR. The resulting texts are analyzed to check whether meaningful text is extracted from the video frame. This post-processing checks the words in the extracted text per frame using the following rules:

- every word has at least 3 characters,
- every word should have at least one vowel,
- every word should match with US-English dictionary (using Python Enchant spell-checking library).

Words that abide to all conditions are kept.

The results from Tesseract per video are fed into Lucene to index and search in the text. A manually defined Boolean query based on the event description and Wikipedia is used in combination with the term frequency to retrieve the positive videos.

1.2 System per Subtask

In this section, we will elaborate the system and submitted runs for each subtask.

1.2.1 Zero Example

For the 20 pre-specified events, three runs are submitted for 0-Ex subtask:

- VIREO_MED16_MED16EvalFull_PS_0Ex_MED_p-VisualTextualRernk_4: Zero-example system with manually refined semantic query, highlighted concept reranking and late fusion with OCR (Best 0Ex performance).
- VIREO_MED16_MED16EvalFull_PS_0Ex_MED_c-VisualRernk_4: The system of primary run, but with visual features only.
- VIREO_MED16_MED16EvalFull_PS_0Ex_MED_c-VisualTextualNorernk_4: The system of primary run, but without highlighted concept reranking.

Our zero-example system aims to pick up and score the query-relevant concepts given a concept bank. This year we basically follow the successful zero-example pipeline last year [22]. Minor changes are made to refine the system. First, the deep features are mostly replaced by deep residual network's output. Second, a new ImageNet-13K dataset [8] is used to replace ImageNet-17K used in the last year. Last but not least, we implement a new user interaction *highlighted concept reranking* also for 0Ex in MED. This technique render a user a way to *emphasize* a group of concepts that share a specific characteristic worth to be strengthened. The idea is borrowed from our implementation for AVS task. Other components in the pipeline, such as automatic semantic query generation, manual concept screening, and OCR late fusion, are all kept the same as in [22].

1.2.2 Ten Example

We submitted 4 runs for the PS-10Ex subtask and 2 runs for the AH-10Ex subtask:

- VIREO_MED16_MED16EvalFull_PS_10Ex_MED_p-Fus_4: 10-example system based on ConceptBank feature fused with the Object-VLAD representation of video.
- VIREO_MED16_MED16EvalFull_PS_10Ex_MED_c-CNNVLAD_4: 10-example system based on the CNN-VLAD representation of video [19].

- VIREO_MED16_MED16EvalFull_PS_10Ex_MED_c-ConceptBank_4: 10-example system based on Concept-Bank feature.

- VIREO_MED16_MED16EvalFull_PS_10Ex_MED_c-ObjectVLAD_4: 10-example system based on the Object-VLAD representation of video.

Visual Classifiers: The Concept-Bank features, i.e., ImageNet_1000, SIN_346, RC_497, Places_205, FCVID_239, Sports_487 and ImageNet-Shuffle_12988 are first concatenated to one feature vector and then used to train an event classifier using the Chi-Square SVM. For the event classifiers based on the Object-VLAD representation, we train the classifier with linear SVM.
Fusion (Concept-Bank + Object-VLAD): For 10Ex system, average fusion is used to directly combine scores of Concept-Bank based SVM and Object-VLAD SVM. We name the fusion system as Visual-System.

1.2.3 Hundred Example

In the PS-100Ex subtask we submit 3 runs:

- VIREO_MED16_MED16EvalFull_PS_100Ex_MED_p-ObjectVLAD_4: 100-example system based on the Object-VLAD representation of video.

- VIREO_MED16_MED16EvalFull_PS_100Ex_MED_c-CNNVLAD_4: 100-example system based on the CNN-VLAD representation of video [19].

- VIREO_MED16_MED16EvalFull_PS_100Ex_MED_c-ConceptBank_4: 100-example system based on Concept-Bank feature.

The training methods with the concept bank feature and the Object-VLAD representation are the same as explained in the previous section (section 1.2.2).

1.3 MED Results and Analysis

In this section we will explain the results per subtask.

1.3.1 Zero Example

Table 1: Zero-example performance comparison on MED16EvalFull reported by NIST.

Method	**MInfAP200%**
Primary run	21.3
Primary run without OCR late fusion	19.3
Primary run without highlighted concept reranking	19.4

Table 2: InfAP200% comparison with and without highlighted concept reranking (HCR).

Event name	**Without HCR**	**With HCR**
Dog show	59.7	81.3
Rock climbing	48.1	49.1
Beekeeping	64.4	71.8
Fixing musical instrument	7.2	14.9
Horse riding competition	23.2	20.5
Parking a vehicle	1.1	3.2
Playing fetch	5.3	6.5

Table 1 lists the comparative runs we submitted for evaluation. As we see from the table, late fusion with OCR used in previous years has equal importance with highlighted concept reranking introduced in this year. It is worth noting that, due to the preference of a human evaluator, highlighted concept reranking is not applicable for all the 20 events. Table 2 shows the 7 events having highlighted concepts assigned by a human evaluator. Most events show a promising improvement by this technique. The event "horse riding competition" is an exception. The overall MInfAP is improved by 2% by these mere 7 events applied highlighted concept reranking.

1.3.2 Ten/Hundred Examples

Table 3: MED PS-10/100Ex: Performances of concept feature on MED14-Test (Mean AP%)

Feature	**PS-10Ex** mAP%	**PS-100Ex** mAP%
Concept-Bank (2015)	21.7	31.1
Concept-Bank_N1 (2016) exclude ImageNet-Shuffle	23.9	33.9
Concept-Bank_N2 (2016) include ImageNet-Shuffle	**25.5**	**36.0**

In this section, we first discuss performances of the new concept bank, followed by analyzing performances of object-level video representation. Finally, we introduce performance of visual system which fuses results of different features.

Improvement by the New Concept Bank: Table 3 compares performance of Concept-Bank used in our MED-2015 system with those of new Concept-Banks implemented this year. Specifically, Concpet-Bank (2015) [23] and Concept-Bank_N1 (2016) are trained with the same training data but by different types of classifiers, i.e., the prior adopts the AlexNet structure and the latter adopts the RestNet-50 architecture. Compared with Concept-Bank (2015), we can easily observe that Concept-Bank_N1 boosts mAP of MED from 21.7 to 23.9 by a relative improvement of 10.1% under PS-10Ex condition and from 31.1 to 33.9 by a relative improvement of 9.0% under PS-100Ex condition on MED14-Test dataset. Additionally, by combining Concept-Bank_N1 with ImageNet-Shuffle_12988, we create Concept-Bank_N2 with a large vocabulary. With more concepts involved, mAP is further boosted from 23.9 to 25.5 under PS-10Ex condition and from 33.9 to 36.0 under PS-100Ex condition.

In sum, we conclude that:

- Video representation relies on the architecture of concept classifiers, i.e., the more accurate, the better.
- Multimedia event detection depends on the vocabulary size of concept bank, i.e., the larger, the better.

Table 4: MED PS-10Ex: Improvements by encoding regional objects on MED14-Test and MED16-EvalSub/Full datasets.

PS-10Ex System	**MED14-Test** mAP%	**MED16-EvalSub** MinfAP200%	**MED16-EvalFull** MinfAP200%
CNN-VLAD	25.3	33.0	33.8
Object-VLAD	**27.2**	**35.0**	**35.6**

Table 5: MED PS-100Ex: Improvements by encoding regional objects on MED14-Test and MED16-EvalSub/Full datasets.

PS-100Ex System	**MED14-Test** mAP%	**MED16-EvalSub** MinfAP200%	**MED16-EvalFull** MinfAP200%
CNN-VLAD	38.6	42.9	34.4
Object-VLAD	**40.0**	**45.0**	**36.1**

Imporvement by Object-VLAD: We compare Object-VLAD with a strong baseline named CNN-VLAD [19]. Both methods adopt neural responses from Res5A layer of the ResNet-50 architecture as feature descriptor for image/object. Except for selective search [20] which is launched for candidate objects proposal, Object-VLAD shares the same operations, such as spatial pyramid pooling (SPP) [24], VLAD encoding and etc, with CNN-VLAD. As shown in Table 4 and Table 5, we observe that:

- By adding the phase of regional objects encoding, Object-VLAD consistently outperforms CNN-VLAD under PS-10Ex and PS-100Ex conditions on MED14-Test and MED16-EvalSub/EvalFull datasets.
- For PS-10Ex, Object-VLAD boosts performance of CNN-VLAD by a relative improvement of 6.3% on average with different testing datasets; whereas, for PS-100Ex, the relative improvement is only 4.3%, indicating that Object-VLAD brings larger improvement especially where there are less training positives.

Table 6: MED PS-10Ex: Performances of single/multiple features on MED14-Test and MED16-EvalSub/Full datasets

PS-10Ex System	**MED14-Test** mAP%	**MED16-EvalSub** MinfAP200%	**MED16-EvalFull** MinfAP200%
Concept-Bank_N2	25.5	32.7	29.4
Object-VLAD	27.2	35.0	35.6
Visual-System (Concept-Bank + Object-VLAD)	**29.2**	**37.7**	**38.9**

Improvement by Fusion: For our MED-2016 system with very few positive examples, we mainly explore capacities of deep feature-based representations (i.e., Concept-Bank_N2, Object-VLAD) without considering OCR/ASR as last year. We use average strategy to fuse results of Concept-Bank_N2 and Object-VLAD under PS-10Ex conditions and have the following observations:

- Even though Res5A feature descriptor is trained with only 1,000 ImageNet categories, with the Object-VLAD representation, we still obtain much higher performances (i.e., mAP: 25.5 vs 27.2 on MED14-Test, MinfAP200: 32.7 vs 35.0 on MED16-EvalSub, MinfAP200: 29.4 vs 35.6 on MED16-EvalFull) than those of a much larger scale Concept-Bank, which indicates that regional objects are also essential factors for video representations.
- Object-VLAD is a complement to Concept-Bank_N2, and average fusion of the two models brings a large improvement to overall performances (i.e., mAP: 25.5 vs 29.2 on MED14-Test, MinfAP200: 32.7 vs 37.7 on MED16-EvalSub, MinfAP200: 29.4 vs 38.9 on MED16-EvalFull).

We also compared our primary runs with other teams in TRECVID 2016 benchmark and summarized them as below:

For PS-10Ex, our primary run (Visual-System) achieves MinfAP200 of 37.7 on MED16-EvalSub dataset ranking 2nd among 11 teams), and MinfAP200 of 38.9 on MED16-EvalFull dataset ranking 2nd among 4 teams.

For AH-10Ex, our primary run (Visual-System) ranks 1st (MinfAP200: 42.4) on MED16-EvalSub dataset and 2nd (MinfAP200: 45.7) on MED16-EvalFull dataset among 4 teams.

For PS-100Ex, with single feature, our primary run (Object-VLAD) achieves MinfAP200 of 45.0 on MED16-EvalSub dataset ranking 5th among 9 teams and MinfAP200 of 36.1 on MED16-EvalFull dataset, ranking 2nd among 4 teams.

1.4 Conclusion and Discussion

Based on the results attained on TRECVID Multimedia Event Detection task of 2016, we can conclude that our large-scale Concept-Bank and Object-level video representation bring a larger improvement compared to our visual systems in 2015 [23] and achieve the state-of-the-art performances. Particularly, for MED with positive examples, increasing the vocabulary size of Concept-Bank and fine-tuning concept classifiers with the ResNet-50 architecture are proved to be beneficial. Additionally, the strategies of accumulating object-level features (e.g., Object-VLAD) surprisingly complements to Concept-Bank based model, which inspires us that collection of regional object information is worth of paying attention.

2 Ad-Hoc Video Search (AVS)

2.1 System

The implementation of ad-hoc video search is mainly based on our zero-example MED framework [22]. Specifically, the most important change is that we add a component called *highlighted concept reranking* in the step when a human evaluator is involved to screen the automatically proposed concept candidates. Other minor refinements include (1) replacing the AlexNet features by deep residual network's features, (2) the use of a more refined ImageNet-13K concepts [8] instead of ImageNet-17K concepts, and (3) a simplified term weighting mechanism to process the query sentence.

For highlighted concept reranking, the idea is inspired by a scenario in the AVS queries. For instance, "one or more people sitting outdoors" and "a street scene at night." In these query cases, a user may want to emphasize characteristics such as "outdoors" and "at night" to tighten the search results. In addition to concept screening, we hereby allow a user to highlight a group of concepts sharing a certain characteristic and boost the top videos containing these concepts. This is a reranking technique in which only the ranking of the top videos are refined. The videos on the top which contain the highlighted concepts are given high weights and their ranking are largely boosted. As this reranking technique is only confined in the top videos of the first search result, the videos containing the highlighted concepts but unrelated to the query can be largely excluded.

2.2 AVS Results

Table 7: MInfAP performance for different concept settings in Video Search 2008 dataset.

Concept Bank	All concepts	SIN_346 concepts	Multi-modality fusion in 2008 [25]	Random
MInfAP	0.122	0.100	0.046	0.009

By adapting our zero-example MED system to AVS, we test the objective performance for automatic runs using the dataset for Video Search task in 2008. The dataset contains 48 queries with ground truth available for testing. As shown in table 7, we see the performance using only SIN_346 concepts nowadays is more than doubled compared to the best results in 2009 [25]. With all the available concept detectors,

the performance improves further by 2%. However, for the everyday topics in AVS 2016 queries, SIN_346 concepts are much less prominent due to their limited coverage. The large gap is shown in table 8.

Table 8: Performance difference between using all the concept detectors and only SIN_346 detectors in AVS 2016 evaluation.

Concept Bank	All concepts	SIN_346 concepts
MInfAP	0.044	0.014

On the other hand, we would like to analyze how much improvement is brought about by highlighted concept reranking. Table 9 shows all the queries having highlighted concepts specified by a human evaluator. As we see from this result, the improvement is less prominent compared to the same technique applied in 0Ex MED. Moreover, as there are only 5 events in total adopting this technique, the overall performance of 30 queries stay relatively the same in around 0.044.

Table 9: Performance with and without highlighted concept reranking (HCR).

Topic ID	**Query**	**With HCR**	**Without HCR**
501	a person playing guitar outdoors	0.106	0.049
502	a man indoors looking at camera where a bookcase is behind him	0.078	0.072
503	a person playing drums indoors	0.069	0.076
517	a policeman where a police car is visible	0.027	0.015
520	any type of fountains outdoors	0.141	0.188

2.3 Interactive Video Search

We also extend the system for interactive search [26]. The system is showcased in Video Browser Showdown (VBS) of MMM 2017 [27] achieves the top performance in the subtask of ad-hoc video search.

3 Video-to-Text Description (VTT)

3.1 System Overview

The task is challenging since it requires detailed understanding of the video content, natural language sentences as well as inter-modality correspondence. A straightforward way is by training explicit concepts detectors (such as objects, actions and scenes) to find the key concepts in a video and matching the concepts to the keywords in text description based on word ontologies, such as WordNet. The main drawback of this method lies in the concept detector training. First, the number of concepts are limited due to the lack of training data. Second, the performance of concept detectors are not guaranteed and the error will be accumulated when matching sentence. To tackle these problems, we adopt stacked attention network (SAN) [4] to directly learn the inter-modality correspondence without explicit concept detectors. Specifically, the SAN model is designed to filter the noisy regions in a frame and only keep the salient regions for inter-modality matching to improve the accuracy.

3.1.1 Concept-based Text Description Matching

We adapt our successful zero-example MED system [22] in a reversed pipeline for concept-based video-to-text retrieval. As we do in the zero-example MED system, we prepared a concept bank of more than 2,000 pre-trained concept detectors. The concepts were collected from the four off-the-shelf datasets [22]. The concept features of a query video were extracted by max pooling over all the detector responses of the keyframes. On the other hand, all the available sentences were input to the system to

generate the concept-based sentence representations. This generation process is analogous to the semantic query generation in our zero-example MED system. The difference is in the use of WordNet ontology. Previously in a zero-example MED system, we only use WordNet to find synonyms for a matched synset in the query. But in video-to-text task, by extending the noun synsets in the query to their hyponyms of is-a relationship, it helps to improve the performance. Finally, the internal ranking process regards the concept features of the query video as a semantic query, then performs ranking for all the sentences by comparing the concept-based sentence representations to the semantic query.

3.1.2 Stacked Attention Model

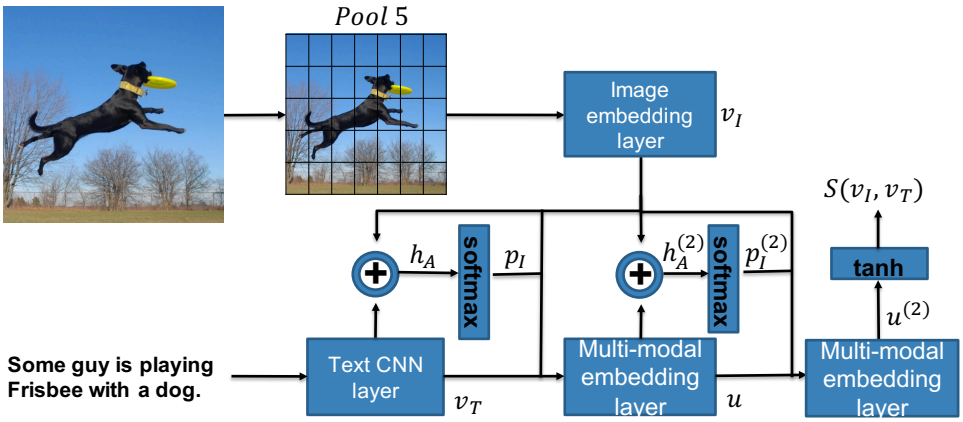

Figure 1: Framework of stacked attention network.

Figure 1 illustrates the SAN model, with visual and text features respectively extracted from videos and text description as input. The model learns a joint space that boosts the similarity between videos and their corresponding text descriptions. Different from [4], where the output layer is for classification, we modify SAN so as to maximize the similarity for video-text pairs.

Image Embedding Feature: We have two kinds of input visual features, which are the last pooling layer of VGG and C3D that retains the spatial information of the frames in a video. For each video, four frames are uniformly selected based on the video duration and the video is represented by average pooling over the frame features or concatenating into one spatial vector. Denote f_I as the input visual feature and is composed of regions f_i. Each region f_i is transformed to a new vector or embedding feature as following:

$$\mathbf{v}_I = tanh(W_I f_I + b_I) \tag{1}$$

where $\mathbf{v}_I \in \mathbb{R}^{d \times m}$ is the transformed feature matrix, with d as the dimension of new vector and m is the number of grids or regions. The embedding feature of f_i is indexed with i-th column of $\mathbf{v}_I$, denoted as $\mathbf{v}_i$. The transformation is performed region-wise, W_I is the transformation matrix and b_I is the bias term.

Text Description Embedding Feature: We explore to use a CNN similar to [3] for text representation. First, each word is embedded by word vector $\mathbf{x}_t$, provided by [28] and get the sentence vector by concatenating the word vectors:

$$\mathbf{x}_{1:T} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T] \tag{2}$$

Then we apply convolution operation on the word embedding vectors. We use three convolution filters, which have the size of three, four and five respectively and the output dimension for these three filters

are all set to 256. The t-th convolution output using window size c is given by:

$$\mathbf{h}_{c,t} = tanh(W_c \mathbf{x}_{t:t+c-1} + b_c) \tag{3}$$

The filter is applied only to window $t : t + c - 1$ of size c. W_c is the convolution weight and b_c is the bias. The feature map of the filter with convlution size c is given by:

$$\mathbf{h}_c = [\mathbf{h}_{c,1}, \mathbf{h}_{c,2}, \ldots, \mathbf{h}_{c,T-c+1}] \tag{4}$$

Then we apply max-pooling over the feature maps of the convolution size c and denote it as

$$\widetilde{\mathbf{h}}_c = max_t[\mathbf{h}_{c,1}, \mathbf{h}_{c,2}, \ldots, \mathbf{h}_{c,T-c+1}] \tag{5}$$

The max-pooling over these vectors is a coordinate-wise max operation. For convolution feature maps of different sizes $c = 3, 4, 5$, we concatenate them to form the feature representation vector of the whole question sentences:

$$\mathbf{h} = [\widetilde{\mathbf{h}}_3, \widetilde{\mathbf{h}}_4, \widetilde{\mathbf{h}}_5] \tag{6}$$

hence $\mathbf{v}_T = \mathbf{h}$ is the CNN based text description vector.

Joint Embedding Feature: The attention layer is to learn the joint feature by trying to locate the salient visual regions that correspond to the text descriptions. There are two transformation matrices, $W_{I,A} \in \mathbb{R}^{k \times d}$ for video I and $W_{T,A}$ for text description T, mimicking the attention localization, formulated as following:

$$h_A = tanh(W_{I,A}\mathbf{v}_I \oplus (W_{R,A}\mathbf{v}_T + b_A)) \tag{7}$$

$$p_I = softmax(W_P h_A + b_P) \tag{8}$$

where $h_A \in \mathbb{R}^{k \times m}$, $p_I \in \mathbb{R}^m$, $W_P \in \mathbb{R}^{1 \times k}$. Note that p_I aims to capture the attention, or more precisely relevance, of frame regions to text description. The significance of a region f_i is indicated by the value in the corresponding element $p_i \in p_I$.

The joint visual-text feature is basically generated by adding the embedding features $\mathbf{v}_I$ and $\mathbf{v}_T$. To incorporate attention value, regions $\mathbf{v}_i$ are linearly weighted and summed before the addition operation with $\mathbf{v}_T$, as following:

$$\widetilde{\mathbf{v}}_I = \sum_{i=1}^{m} p_i v_i \tag{9}$$

$$\mathbf{u} = \widetilde{\mathbf{v}}_I + \mathbf{v}_T \tag{10}$$

where $\widetilde{\mathbf{v}}_I \in \mathbb{R}^d$, and $\mathbf{u} \in \mathbb{R}^d$ represents the joint embedding feature.

As suggested in [4], progressive learning by stacking multiple attention layers can boost the performance, but will heavily increase the training cost. We consider two layer SAN, by feeding the output of first attention layer, $\mathbf{u}(1)$, into the second layer to generate new joint embedding feature $\mathbf{u}(2)$ as following:

$$\mathbf{h}_A^{(2)} = tanh(W_{I,A}^{(2)}\mathbf{v}_I \oplus (W_{R,A}^{(2)}\mathbf{u} + b_A^{(2)})) \tag{11}$$

$$p_I^{(2)} = softmax(W_P^{(2)} h_A^{(2)} + b_P^{(2)}) \tag{12}$$

$$\widetilde{\mathbf{v}}_I^{(2)} = \sum_i p_i^{(2)} \mathbf{v}_i \tag{13}$$

$$\mathbf{u}^{(2)} = \widetilde{\mathbf{v}}_I^{(2)} + \mathbf{u}^{(1)} \tag{14}$$

Objective Function: To this end, the similarity between video and text description is generated as following:

$$S < \mathbf{v}_I, \mathbf{v}_T >= tanh(W_{u,s}\mathbf{u}^{(2)} + b_s) \tag{15}$$

where $W_{u,s} \in \mathbb{R}^d$ and $b_s \in \mathbb{R}$ is bias. $S < \mathbf{v}_I, \mathbf{v}_T >$ outputs a score indicating the association between the embedding features of video and text description. The learning is based on the following rank-based loss function with a large margin form as the objective function:

$$\mathcal{L}(W, D_{trn}) = \sum_{(\mathbf{v}_I, \mathbf{v}_T^+, \mathbf{v}_T^-) \in D_{trn}} max(0, \delta + S < \mathbf{v}_I, \mathbf{v}_T^- > - S < \mathbf{v}_I, \mathbf{v}_T^+ >) \tag{16}$$

The training set, D_{trn}, consists of triples in the form of $(\mathbf{v}_I, \mathbf{v}_T^+, \mathbf{v}_T^-)$, where $\mathbf{v}_R^+(\mathbf{v}_R^-)$ is true (false) text description for video v_I. The matrix W represents the network parameters, and $\delta \in (0, 1)$ controls the margin in training and is cross-validated.

3.2 Experiments

3.2.1 Settings and Datasets

Here we detail the parameter setting of SAN. The dimension of attention layer h_A is set as 1,024 and the hyper parameter δ is set as 0.35 through cross-validation. SAN is trained using stochastic gradient descent with momentum set as 0.9 and the initial learning rate as 1. The size of mini-batch is 50 and the training stops after 30 epochs. To prevent overfitting, dropout [29] is used. Since the provided training dataset by TRECVID is too small, we also train our model on Flickr30k [30], Coco [31] and TGIF [32] datasets. Here, we aggregate the sentences from these three datasets and the size of vocabulary is 12,270.

3.2.2 Result Analysis

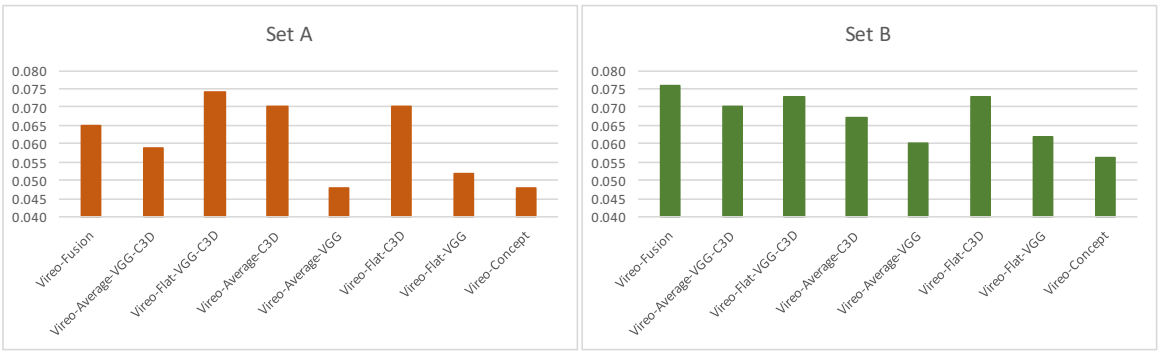

Figure 2: Mean inverted rank of SAN and concept-based ranking methods.

The testing dataset is composed of 1,915 videos and two sets (a and b) of text descriptions (each composed of 1,915 sentences). In addition to the submitted four runs, we also calculate the performance of other four runs, which only make use of the VGG feature by average pooling (Vireo-Average-VGG) over frames or concatenating frame features (Vireo-Flat-VGG). Similarly, two runs (Vireo-Average-C3D and Vireo-Flat-VGG) are evaluated on C3D features. Here, the mean inverted rank is adopted for evaluation. Table 2 presents the results of our eight runs. From the table, we can see that the C3D features consistently achieve better performance than VGG features in both average pooling and frame feature concatenation. It indicates that motion (C3D) features is much more important than static (VGG) features in video content understanding, which is due to the fact that there are many verbs in the text description, such as "dancing outdoor" or "itching herself indoor" etc. Another observation is that concatenating frame features achieves better performance than simple average pooling. This is also reasonable since the concatenation not only keep the salient information in different frames but also preserves the sequential information between different frames, which is very important when only using static (VGG) features, as

indicated by the performance of Vireo-Average-VGG and Vireo-Flat-VGG. However, we also find that the shallow fusion between static (VGG) and motion (C3D) features may even harm the performance, which is attributed to the weak action recognition ability of static (VGG) features. Finally, it is also easy to observe that SAN models consistently perform better than concept-based methods, which means that SAN model gracefully solves the problems in concept-based methods as mentioned in Section 3.1.

3.3 Summary

We have tested the traditional concept-based approach and the attention-based deep models for VTT matching task. Deep models perform much better than concept-based matching. We also observe that motion (C3D) features dominate the performance and the feature concatenation also performs better than simple average pooling, which indicates that the temporal context information is important for video content understanding. In the future, we are targeting at making use of temporal context information to select more representative frames for video representation. In addition, we will also consider auditory modalities for matching.

Acknowledgment

The work described in this paper was supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 120213, CityU 11210514).

References

[1] G. Awad, J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Qunot, M. Eskevich, R. Aly, and R. Ordelman, "Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking," in *Proceedings of TRECVID 2016*. NIST, USA, 2016.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[3] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014.

[4] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," *CoRR*, vol. abs/1511.02274, 2015.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[6] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[7] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013. [Online]. Available: http://hal.inria.fr/hal-00873267

[8] P. Mettes, D. C. Koelma, and C. G. Snoek, "The imagenet shuffle: Reorganized pre-training for video event detection," *arXiv preprint arXiv:1602.07119*, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[11] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quenot, "Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2014*. NIST, USA, 2014.

[12] P. Natarajan, S. Wu, F. Luisier, X. Zhuang, and M. Tickoo, "BBN VISER TRECVID 2013 multimedia event detection and multimedia event recounting systems," in *NIST TRECVID workshop*, 2013.

[13] S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel, "Creating HAVIC: Heterogeneous audio visual internet collection," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. DoAYan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[14] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014.

[15] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," in *arXiv preprint arXiv:1502.07209*, 2015.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[19] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative cnn video representation for event detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1798–1807.

[20] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[21] R. Smith, "An overview of the tesseract ocr engine," in *International Conference on Document Analysis and Recognition*, 2007.

[22] Y.-J. Lu, H. Zhang, M. de Boer, and C.-W. Ngo, "Event detection with zero example: Select the right and suppress the wrong concepts," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 2016, pp. 127–134.

[23] H. Zhang, Y.-J. Lu, M. de Boer, F. ter Haar, Z. Qiu, K. Schutte, W. Kraaij, and C.-W. Ngo, "VIREO-TNO@TRECVID 2015: Multimedia event detection," in *NIST TRECVID workshop*, 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 346–361.

[25] X. Y. Wei, Y. G. Jiang, and C. W. Ngo, "Concept-driven multi-modality fusion for video search," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 1, pp. 62–73, 2011.

[26] Y.-J. Lu, P. A. Nguyen, H. Zhang, and C.-W. Ngo, "Concept-based interactive search system," in *International Conference on Multimedia Modeling*. Springer, 2017, pp. 463–468.

[27] K. Schoeffmann, D. Ahlström, W. Bailer, C. Cobârzan, F. Hopfgartner, K. McGuinness, C. Gurrin, C. Frisson, D.-D. Le, M. Del Fabro *et al.*, "The video browser showdown: a live evaluation of interactive video search tools," *International Journal of Multimedia Information Retrieval*, vol. 3, no. 2, pp. 113–127, 2014.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[30] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft COCO: Common Objects in Context*. Springer International Publishing, 2014, pp. 740–755.

[32] Y. Li, Y. Song, L. Cao, J. R. Tetreault, L. Goldberg, A. Jaimes, and J. Luo, "TGIF: A new dataset and benchmark on animated GIF description," *CoRR*, vol. abs/1604.02748, 2016.