

Singapore Management University

## Institutional Knowledge at Singapore Management University

---

Research Collection School Of Computing and  
Information Systems

School of Computing and Information Systems

---

11-2017

### Vireo @ TRecViD 2017: Video-to-text, ad-hoc video search and video hyperlinking

Phuong Anh NGUYEN

Qing LI

Zhi-Qi CHENG

Yi-Jie LU

Hao ZHANG

*See next page for additional authors*

Follow this and additional works at: [https://ink.library.smu.edu.sg/sis\\_research](https://ink.library.smu.edu.sg/sis_research)



Part of the [Databases and Information Systems Commons](#), and the [Graphics and Human Computer Interfaces Commons](#)

---

#### Citation

NGUYEN, Phuong Anh; LI, Qing; CHENG, Zhi-Qi; LU, Yi-Jie; ZHANG, Hao; WU, Xiao; and NGO, Chong-wah. Vireo @ TRecViD 2017: Video-to-text, ad-hoc video search and video hyperlinking. (2017). *2017 TREC Video Retrieval Evaluation, TRECVID 2017, Gaithersburg, November 13-15*. Available at: [https://ink.library.smu.edu.sg/sis\\_research/6599](https://ink.library.smu.edu.sg/sis_research/6599)

This Conference Paper is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email [cherylds@smu.edu.sg](mailto:cherylds@smu.edu.sg).

---

**Author**

Phuong Anh NGUYEN, Qing LI, Zhi-Qi CHENG, Yi-Jie LU, Hao ZHANG, Xiao WU, and Chong-wah NGO

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/341097893>

# VIREO@ TRECVID 2017: Video-to-Text, Ad-hoc Video Search and Video hyperlinking

Conference Paper · January 2017

CITATIONS

14

READS

40

7 authors, including:



[Zhi-Qi Cheng](#)

Carnegie Mellon University

24 PUBLICATIONS 466 CITATIONS

[SEE PROFILE](#)



[Yijie Lu](#)

City University of Hong Kong

9 PUBLICATIONS 112 CITATIONS

[SEE PROFILE](#)



[Hao Zhang](#)

City University of Hong Kong

12 PUBLICATIONS 105 CITATIONS

[SEE PROFILE](#)



[Xiao Wu](#)

Southwest Jiaotong University

111 PUBLICATIONS 2,692 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Multimedia Event Detection and Recounting [View project](#)

# VIREO @ TRECVID 2017: Video-to-Text, Ad-hoc Video Search and Video Hyperlinking

Phuong Anh Nguyen<sup>†</sup>, Qing Li<sup>†</sup>, Zhi-Qi Cheng<sup>\*†</sup>, Yi-Jie Lu<sup>†</sup>,  
Hao Zhang<sup>†</sup>, Xiao Wu<sup>\*</sup>, Chong-Wah Ngo<sup>†</sup>

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

<sup>\*</sup>*The School of Information Science and Technology, Southwest Jiaotong University*  
<http://www.swjtu.edu.cn>

## Abstract

In this paper, we describe the systems developed for Video-to-Text (VTT), Ad-hoc Video Search (AVS) and Video Hyper-linking (LNK) tasks at TRECVID 2017 [1] and the achieved results.

**Video-to-Text Description (VTT):** We participate in the TRECVID 2017 pilot Task of Video-to-Text Description, which consist of two subtasks, *.i.e.* Matching and Ranking, and Description Generation.

*Matching and Ranking task:* To compare the effectiveness of spatial and temporal attentions, we experiment:

- *No attention* model: Each video is represented by average pooling over both the spatial and temporal dimensions of the ResNet-152 features extracted from frames and text description is encoded by LSTM. Then we learn an embedding space to minimize the distance between the corresponding video and text description in the format of triple loss. Furthermore, C3D is utilized to extract the motion features of videos. Similarity scores from two kinds of features are averagely fused for the final ranking.
- *Spatial attention* model: Average pooling is only used in the temporal dimension and the spatial dimension is kept. Then we train attention model on the spatial feature map of video to compute the similarity score, which is used for final ranking.
- *Temporal attention* model: Different from the spatial attention model, we train attention model on frame-level, and perform average pooling over the spatial dimension.
- *No-spatial-temporal attention* model: similarity scores from the above three models are averagely fused for the final ranking.

*Description Generation task:* We adopt the similar approach as the matching and ranking task and the difference is that LSTM is used to generate the sentence word by word. More details about the model can be seen in [2, 3]. Our submission can be summarized as:

- *No attention* model: Each video is represented by average pooling over both the spatial and temporal dimension of the ResNet-152 features extracted from frames and LSTM is used to generate the sentence word-by-word. Furthermore, we concatenate the features of ResNet-152 and C3D to feed into the LSTM to generate descriptions for videos.

- *Spatial attention* model: For video features, average pooling is only used in temporal dimension and the spatial dimension is kept. Then we train attention model to access different feature when generating different words in the sentence.
- *Temporal attention* model: For video features, this model does the average pooling over the spatial dimension, and learn attention model over the temporal dimension.

**Ad-Hoc Video Search (AVS)** We merged three search systems for AVS. One is our concept-based, zero-example video search system which has been proved useful in previous years [4], one is a video captioning system which is individually trained in VTT task, the other is a text-based search system which computes similarities between query and videos using the metadata extracted from the videos. In this study, we intend to find whether the combination of the concept-based system, captioning system and text-based search system would do any help to improve search performance. We submit 5 fully automatic runs and 3 manually-assisted runs. Our runs are listed as follows:

- *F\_D\_VIREO.17\_1*: An automatic run with infAP=0.093 uses concept-based video search system only. The number of concepts in the concept bank is about 15K, which includes a collection of concepts from ImageNet Shuffle [5], FCVID [6], Sports-1M [7], SIN [8], Places and Research Set [9, 10]. Most of the concept detectors are trained or fine-tuned with ResNet-50 [11].
- *F\_D\_VIREO.17\_2*: An automatic run combines the results of *F\_D\_VIREO.17\_1* and the video captioning system. In the captioning system, both ResNet and C3D features are used. The ratio between ResNet, C3D and concept results in the weight fusion is 2:1:3. The performance stays at infAP=0.120.
- *F\_D\_VIREO.17\_3*: The results of concept-based system and the video captioning system are combined with the same approach in *F\_D\_VIREO.17\_2* but the weight fusion is the average of ResNet, C3D and concept results. This run gets infAP=0.116.
- *F\_D\_VIREO.17\_4*: The results of *F\_D\_VIREO.17\_3* and the text-based search system are combined in this run. The meta-data, the on-screen text and the speech are extracted from the videos and fed into Lucene to build the text-based search system. The ratio between *F\_D\_VIREO.17\_3* and text-based search results in the weight fusion is 10:1. The performance stays at infAP=0.116.
- *M\_D\_VIREO.17\_1*: This manual run is based on *F\_D\_VIREO.17\_1* using concept-based video search system. Human efforts are involved in two steps. (1) A user corrects the mistakes in queries after automatic NLP parsing. (2) Automatically proposed semantic concepts are screened by the user by deleting the unrelated, non-discriminative concepts. The run gets infAP=0.124.
- *M\_D\_VIREO.17\_2*: In this run, the result of *M\_D\_VIREO.17\_1* has been fused with the result of the captioning system used in *F\_D\_VIREO.17\_2*. The ratio between ResNet, C3D and concept results in the weight fusion is manually selected based on user's experiences. The run ends up with infAP=0.164.
- *M\_D\_VIREO.17\_3*: This run is an auto run which combines of the result from *F\_D\_VIREO.17\_2* with the result of the text-based search system in *F\_D\_VIREO.17\_4*. The ratio of the weight fusion between *F\_D\_VIREO.17\_2* and text-based search results is 10:1. This run achieves the best performance among our automatic runs with infAP=0.120.
- *M\_D\_VIREO.17\_4*: This run gets the best performance among manual runs submitted with infAP = 0.164. In the run, the results of *M\_D\_VIREO.17\_2* and the text-based search system are combined with the ratio of 10: $x$  in the weight fusion, where  $x$  is defined by analyzing the popularity of query terms using Google Books Ngram.

**Video Hyperlinking (LNK):** We introduce two novelties: (a) development of new semantic representation network (SRN) for evaluation of cross-modal similarities; (b) re-ranking of search result by considering data risk based on the statistical properties of hubness, local intrinsic dimension (LID) and diversity [12].

- *Run-1*: Visual baseline. The visual run relies on large concept banks including more than 14K of concept classifiers [13]. The relatedness between anchors and targets is evaluated based on the average fusion of SRN and cosine similarity.
- *Run-2*: Rerun of *Run-1* using the LID-first algorithm proposed in [12]. The goal is to promote the ranks of targets with “lower data risk”, specifically, in lower local dimension, being hubs of data, and sufficiently diverse from neighboring region.
- *Run-3*: Multimodal baseline. This run combines visual *Run-1* and the text features extracted from ASR. Using SRN, we evaluate four different kinds of similarities between anchors and targets: visual-visual, visual-text, text-visual and text-text. These similarities are averagely fused to quantify the relatedness between anchor and targets. Finally, we further fuse the three kinds of relatedness: SRN, visual cosine similarity, textual cosine similarity with the weights of 0.5, 0.3 and 0.2 respectively.
- *Run-4*: Rerun of *Run-3* using the LID-first algorithm proposed in [12].

## 1 Video-to-Text Description (VTT)

The TRECVID 2017 pilot Task of Video-to-Text Description is challenging since it involves detailed understanding of the video content including many concepts, such as objects, actions, scenes, person-object relations, temporal order of events and so on. Moreover, the inter-modality correspondence between video content and natural language sentences is also nontrivial for this task.

In this task, a set of 1,880 Vine videos are randomly selected from more than 50,000 Twitter Vine videos. Each video has a duration of around 6 seconds and is annotated multiple times by different annotators. When describing each video, annotators are asked to write a sentence include, if appropriate and applicable, four facets of the video, *.i.e.*, who, what, where and when.

### 1.1 Matching and Ranking

#### 1.1.1 Task Description

In this subtask, participants are asked to rank a set of text descriptions in terms of relevance to a given video. Different from that of last year, the matching and ranking subtask of this year split the whole video set into four testing subsets (2, 3, 4, 5) of varying size, in order to measure the impact of the set size on the performance. Concretely, subset 2 includes 1,613 videos, subset 3 includes 795 videos, subset 4 includes 388 videos, and subset 5 includes 159 videos. For subset 2, participants are asked to rank two independent description sets: A and B. Subset 3 has three description sets: A, B, and C. Subset 4 has four description sets: A, B, C, and D. Subset 5 has five description sets: A, B, C, D, and E.

#### 1.1.2 Approach

For this task, we adopt spatio-temporal attention network (STN) to directly learn the inter-modality correspondence without explicit concept detectors. Specifically, the STN model is designed to select the most salient parts of the video in both spatial and temporal dimensions for inter-modality matching to improve the accuracy.

**Video Embedding Features:** To better represent the video content, we adopt two kinds of visual features for each video: ResNet-152 [14] features, which can extract the object information from video frames, and C3D features [15], which can extract motion information from video clip (consecutive 16 frames). The input visual feature is denoted as  $f_I$ , consisting of parts  $f_i$ . Each part  $f_i$  is embedded into a new space as follows:

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

where  $\mathbf{v}_I \in \mathbb{R}^{d \times m}$  is the transformed feature matrix, with  $d$  as the dimension of embedding space and  $m$  as the number of grids or frames.  $W_I$  is the transformation matrix and  $b_I$  is the bias term.

**Text Description Embedding Features:** To represent the text description, we adopt LSTM [16], a special form of RNN, to encode the whole sentence into one fixed-length vector. Supposing we have a caption  $\mathcal{S} = \{s_1, s_2, \dots, s_L\}, s_t \in \mathcal{V}$ , where  $\mathcal{V}$  is the vocabulary,  $L$  is the caption length. First, each word of the caption  $\mathcal{S}$  can be represented as a sequence of word vector  $\mathbf{S} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_L\}$ . Then we input the sequence of word vector to a LSTM layer one by one. For timestep  $t$ ,  $\mathbf{x}_t$  and  $\mathbf{h}_t$  are the input and output vector respectively,  $\mathbf{T}$  are input weights matrices,  $\mathbf{R}$  are recurrent weight matrices and  $\mathbf{b}$  are bias vectors. *Logic sigmoid*  $\sigma(x) = \frac{1}{1+e^{-x}}$  and *hyperbolic tangent*  $\phi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$  are element-wise non-linear activation functions, mapping real values to  $(0, 1)$  and  $(-1, 1)$  separately. The dot product and sum of two vectors are denoted with  $\odot$  and  $\oplus$  respectively. Given inputs  $\mathbf{x}_t$ ,  $\mathbf{h}_{t-1}$  and  $\mathbf{c}_{t-1}$ , the LSTM unit updates for timestep  $t$  are:

$$\begin{aligned} \mathbf{g}_t &= \phi(\mathbf{T}_g \mathbf{x}_t + \mathbf{R}_g \mathbf{h}_{t-1} + \mathbf{b}_g) && \text{cell input} \\ \mathbf{i}_t &= \sigma(\mathbf{T}_i \mathbf{x}_t + \mathbf{R}_i \mathbf{h}_{t-1} + \mathbf{b}_i) && \text{input gate} \\ \mathbf{f}_t &= \sigma(\mathbf{T}_f \mathbf{x}_t + \mathbf{R}_f \mathbf{h}_{t-1} + \mathbf{b}_f) && \text{forget gate} \\ \mathbf{c}_t &= \mathbf{g}_t \odot \mathbf{i}_t + \mathbf{c}_{t-1} \odot \mathbf{f}_t && \text{cell state} \\ \mathbf{o}_t &= \sigma(\mathbf{T}_o \mathbf{x}_t + \mathbf{R}_o \mathbf{h}_{t-1} + \mathbf{b}_o) && \text{output gate} \\ \mathbf{h}_t &= \phi(\mathbf{c}_t) \odot \mathbf{o}_t && \text{cell output} \end{aligned}$$

After feeding the whole sentence into LSTM, we use the output of last step  $\mathbf{h}_L$  as the representation of the caption, i.e.,  $\mathbf{v}_T = \mathbf{h}_L$ .

**Spatio-Temporal Attention (STA):** The attention layer is to find the most relevant part from the video to the text description. Given the video representations  $\mathbf{v}_I$  and the caption representation  $\mathbf{v}_T$ , the attention mechanism is formulated as:

$$\begin{aligned} \mathbf{h}_A &= \tanh(W_{I,A} \mathbf{v}_I \odot W_{T,A} \mathbf{v}_T + \mathbf{b}_A) \\ \mathbf{p}_I &= \text{softmax}(W_P \mathbf{h}_A + \mathbf{b}_P) \\ \tilde{\mathbf{v}}_I &= \sum_{i=1}^m p_i v_i \end{aligned}$$

where  $W_{I,A}, W_{T,A}$  are the transformation matrix for video and caption, respectively.  $\mathbf{p}_I$  denotes the attention weights for different parts of the video, which represent the relevance of the part to the text description. Therefore,  $\tilde{\mathbf{v}}_I$  is the aggregated representation for the video content, which is modulated by the text description.

**Objective Function:** To this end, the similarity between video and text description is computed as follow:

$$\begin{aligned} \mathbf{u} &= W_I \tilde{\mathbf{v}}_I \odot W_T \mathbf{v}_T + \mathbf{b} \\ S < \mathbf{v}_I, \mathbf{v}_T > &= \tanh(W_{u,s} \mathbf{u} + \mathbf{b}_s) \end{aligned}$$

where  $W_{u,s} \in \mathbb{R}^d$  and  $b_s \in \mathbb{R}$  is bias.  $S \langle \mathbf{v}_I, \mathbf{v}_T \rangle$  represents the similarity score between video content and text description, which can be used to rank the set of text descriptions. The model is optimized by minimizing the following rank-based loss function with margin as the loss function:

$$\mathcal{L}(W, D_{trn}) = \sum_{(\mathbf{v}_I, \mathbf{v}_T^+, \mathbf{v}_T^-) \in D_{trn}} \max(0, \delta + S \langle \mathbf{v}_I, \mathbf{v}_T^- \rangle - S \langle \mathbf{v}_I, \mathbf{v}_T^+ \rangle) \quad (2)$$

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 determined by cross validation.

### 1.1.3 Experiments and Submissions

**Settings and Datasets:** Here we detail the parameter setting of spatio-temporal attention network (STA). The dimension of attention layer  $h_A$  is set as 512 and the hyper parameter  $\delta$  is set as 0.2 through cross validation. STA is trained using stochastic gradient descent with momentum set as 0.9 and the initial learning rate as 0.1. The size of mini-batch is 128 and the training stops after 30 epochs. To prevent overfitting, dropout [17] is used. Since the provided training dataset by TRECVID is too small, we also train our model on MSR-VTT [18].

**Submissions:** To compare the effectiveness of spatial and temporal attention, we submitted 4 runs listed in the Abstract. For subset 2, we submit four runs of four models for description set A and B, respectively. For each other subset, we only submit one run of no-spatial-temporal attention model.

### 1.1.4 Results Analysis

The testing subset 2 is composed of 1,613 videos and two description sets (A and B, each composed of 1,613 sentences). The evaluation metric is mean inverted rank. In addition to the submitted four runs, we also calculate the performance of other settings, which make use of either only ResNet-152 features or only C3D features for above four models. Figure 1 shows the results of our runs on subset 2 and description sets A and B, respectively. From the figure, we can see that the ResNet-152 features consistently outperform C3D features in no, spatial, and temporal attention mechanism. Since the ResNet-152 is much deeper than C3D (152 v.s. 19), the ResNet-152 features are more powerful than C3D features. Besides, the fusion of ResNet-152 and C3D features can further improve the performance. It indicates that motion information (C3D) can be complementary to the static appearance (ResNet-152) in video content understanding, which is due to the fact that the text descriptions include many motion verbs, such as “dancing” or “running”. Another observation from Figure 1 is that although the performance of spatial/temporal attention models is lower than no attention model, the fusion of no, spatial, and temporal attention models can boost the final performance. This shows that the attention mechanism can improve the matching performance by putting more weights on the most relevant parts of the video. Figure 2 presents the performance of different subsets of varying size. We can see that when the size of description set reduces, the mean inverted rank will increase. This is reasonable because fewer caption candidates means the bigger possibility of finding the correct correspondence between video and description.

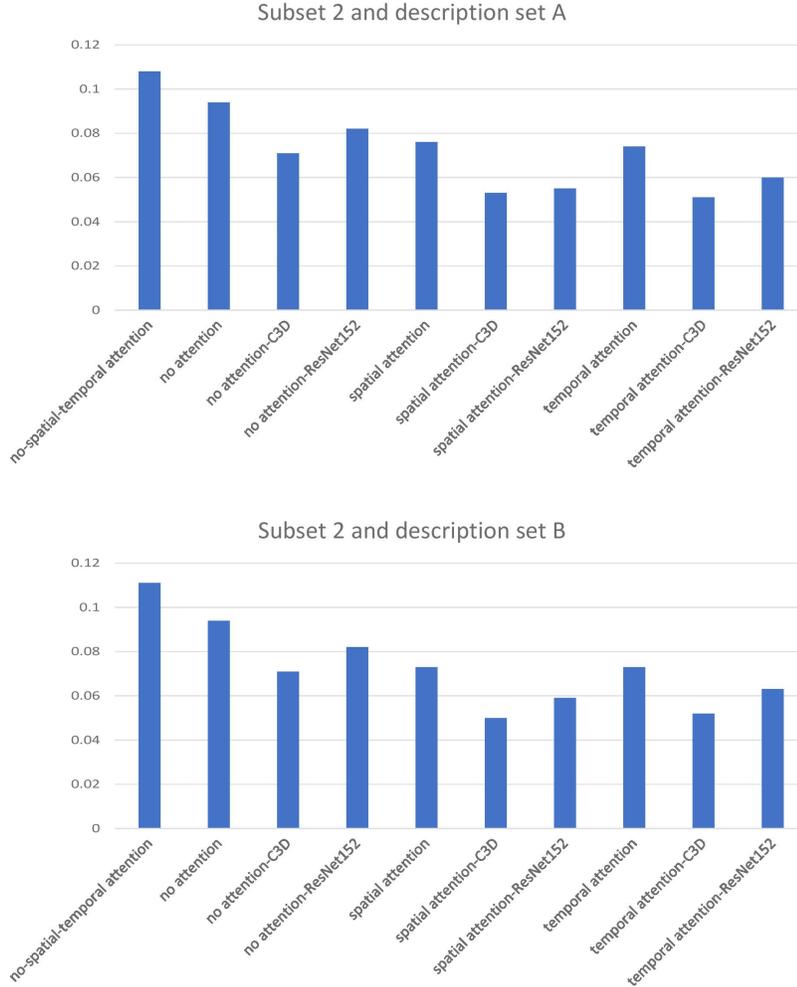


Figure 1: Mean inverted rank of no, spatial and temporal attention models.

## 1.2 Description Generation

### 1.2.1 Task Description and Submissions

In this subtask, participants were asked to generate a sentence to describe a specific test video, without taking into consideration the existence of the sentence sets in the matching and ranking subtask. This subtask includes 1,880 videos and we first train our model on MSR-VTT dataset and then make predictions on the given videos. Our submissions have been listed in the Abstract.

### 1.2.2 Results Analysis

Figure 3 presents the performance of our submissions on three evaluation metrics: meteor, BLEU and STS (Semantic Similarity). From the metrics of meteor and STS, we can see that both the spatial and temporal attention mechanism can clearly improve the quality of generated sentences. And in no attention model, the combination of C3D features and ResNet-152 features can further improve the performance. It suggests that both the motion information and the static appearance are important when generating the description for the video.

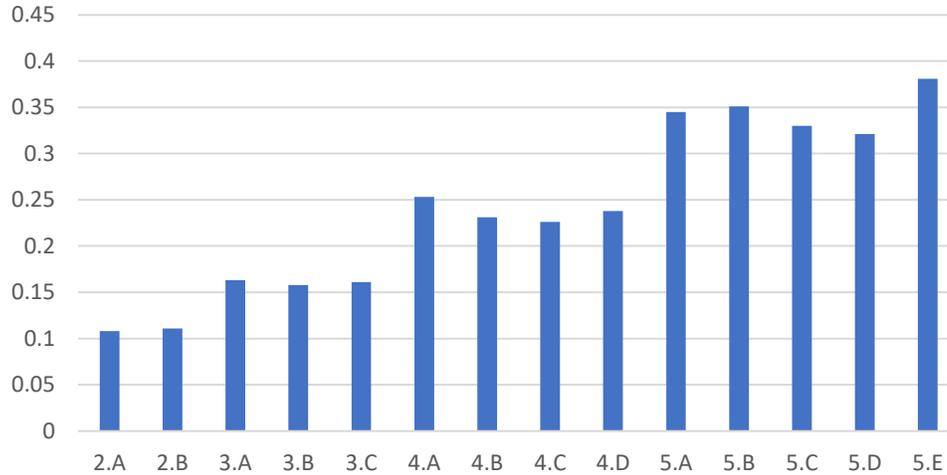


Figure 2: The impact of the set size on the mean inverted rank

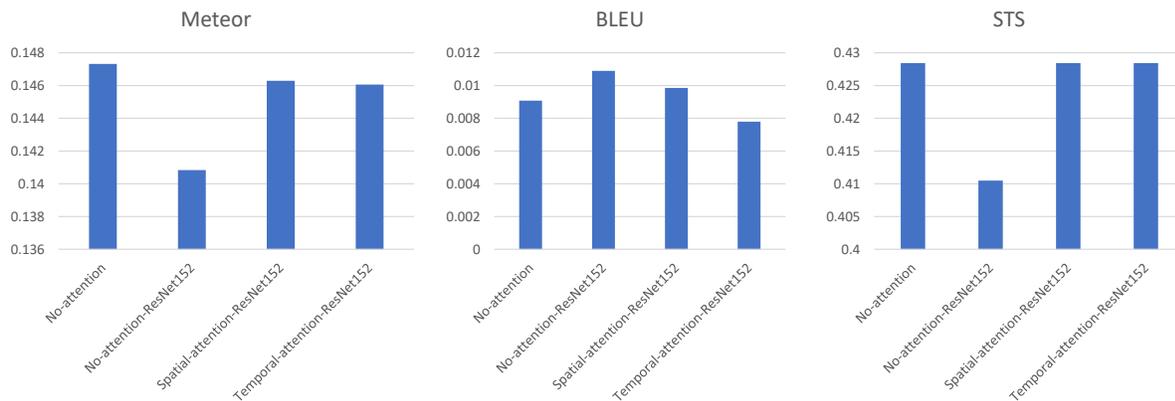


Figure 3: Meteor, BLEU, and STS of no, spatial and temporal attention models on description generation.

### 1.3 Summary

In the video-to-text description task, we have tried the spatial and temporal attention network for both the matching and ranking subtask and description generation subtask. We also use different features, i.e., ResNet-152 and C3D, to represent different aspects of the video content. From the results, we can conclude that the object information from the ResNet-152 features and the motion information from C3D features are complementary and can be combined to boost the performance. Another conclusion is that the spatial/temporal attention mechanism can help improve the performance for both the matching and generation subtasks. In the future, we are targeting at investigating how to combine different features (ResNet-152 and C3D) into one framework of the spatial-temporal attention network. In addition, we will also consider other modalities for the task, like audio information.

## 2 Ad-Hoc Video Search (AVS)

### 2.1 Results Analysis

Our experimental results show that (1) the concept-based search system still stands for the best individual run; (2) the combination of concepts and captions effectively improves the performance by about 30% compared to concept-only search system; (3) the metadata has little help in improving the performance in AVS task 2017.

Type	Method	Abbv.	Mean xInfAP
Auto	Visual concepts only	Concepts	0.093
Auto	Visual concepts + Captions	Concepts+Captions	0.120
Auto	Visual concepts + Captions + Metadata	Concepts+Captions+Metadata	0.120
Manual	Visual concepts only	Concepts	0.124
Manual	Visual concepts + Captions	Concepts+Captions	0.164
Manual	Visual concepts + Captions + Metadata	Concepts+Captions+Metadata	0.164

Table 1: The performance of official runs on TRECVID 2017 AVS task

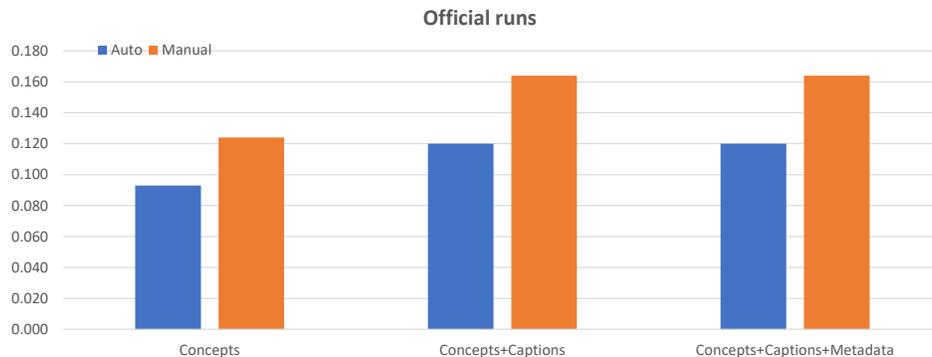


Figure 4: Performance comparison of official auto and manual runs.

Table 2 presents the mean xInfAP for the official runs on TRECVID 2017 AVS task. For the auto runs, we can see that only using visual concepts already gets a good enough performance. And combining the concepts with captioning (simply average fusion) can boost the mean xInfAP to 0.12, which is nearly 30% performance gain. This indicates that the captioning system is complementary to the visual concepts, which only consider the concepts in the video and dismiss the relationship between these concepts. Another observation for the auto runs is that the metadata has little help on improving the performance. A possible reason is that no named-entities are mentioned for this year AVS queries.

Figure 4 presents the performance comparison between the auto and manual runs. For the concept-based searching system, the manual run is done by manually selecting the good concepts from the generated visual concepts. The performance gain indicates that the manual selection of visual concepts can help filter the noisy visual concepts and then lead to the improvement on the searching results. For the combination of concepts and captioning, the manual process is to manually assign fusion weights to three different components: concepts, captioning-ResNet152, captioning-C3D. This manual assignment of fusion weights is based on our experience on the results of TRECVID 2016 AVS task. From the figure, we can see that this manual process further boosts the mean xInfAP from 0.12 to 0.164, a more than 30%

performance gain. This again indicates that the concept-based system and the caption-based system are very complementary. It is quite worth investigating in the future how to automatically assign adapted fusion weights to these two systems.

## 2.2 Interactive Video Search

We also enhanced the video interactive search system [13] showcased in Video Browser Showdown (VBS) [19] of MMM 2017. The system is enhanced in three main modules, including: a simplified color sketch modal with recommendation system, a manual concept selection modal allows defining AND, OR, NOT logical combination of concepts, and a new browsing interface with compact video frames representation [20].

In VBS, the video retrieval tasks are divided into 3 sub-tasks: visual known-item search (visual KIS), textual known-item search (textual KIS) and ad-hoc video search (AVS). In visual KIS, the participants see a 20 seconds video segment and try to find the correct segment from the video corpus. In textual KIS, a detailed-textual-description is given out and the participants try to find the described video segment. Different from textual KIS, a general textual query is given out in AVS task, then the participants need to find all the video segments described. At VBS 2018, there are 5 queries for each sub-tasks. Also, two sessions are provided with one for evaluating the effectiveness of the tool using by expert users and the other one for novice users.

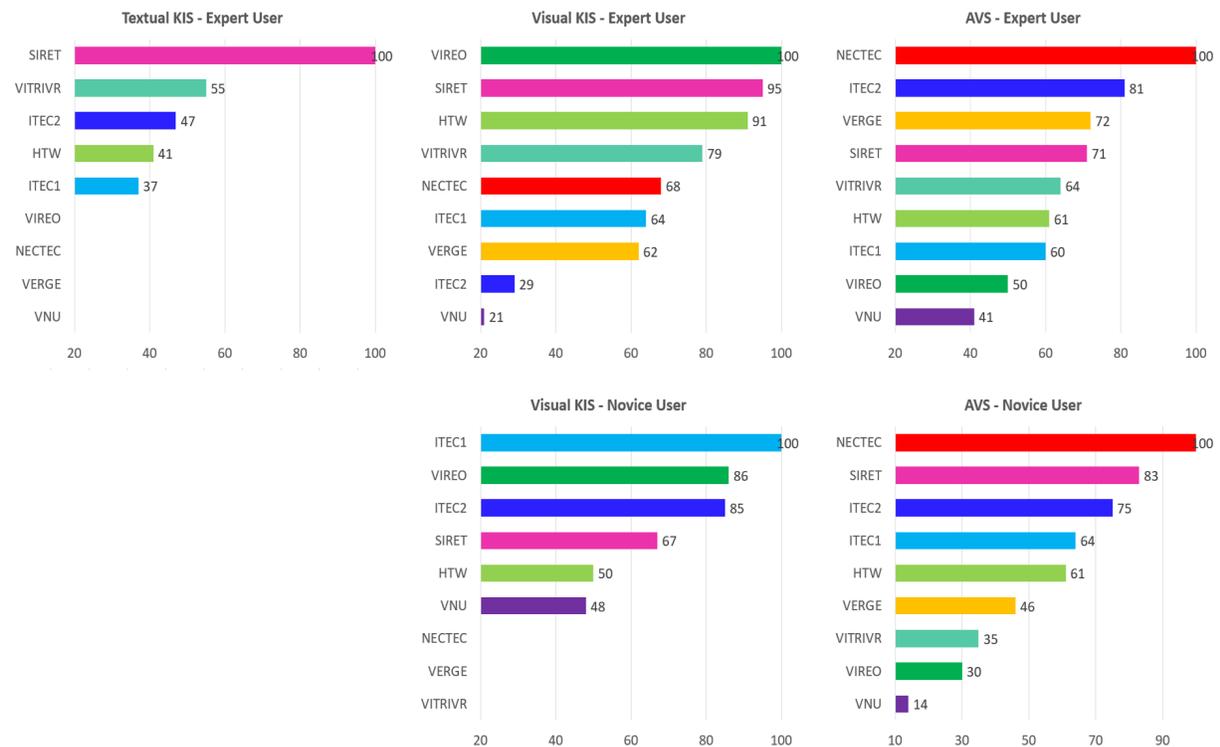


Figure 5: The result of all sub-tasks in VBS 2018. The first row shows the tasks for expert users, the second row shows the tasks for novice users. In each chart, the vertical axis shows the teams participated, the horizontal axis shows the score achieved by each team scaled to 100.

In 2018, there are 9 teams participate in the VBS task. The detail results are shown in Figure 5. Using the simplified color sketch modal, we take advantages in visual KIS tasks and finish as the top in

expert session and the second in novice session. As the winner in AVS tasks of VBS 2017, we also have advantages with a large concepts bank which is filled with 13K concepts but we just finish at the second rank from bottom up. This can be explained by the current scoring scheme of VBS which does not take into account the number of members in each team i.e. the number of users engage in searching process. In actually, VIREO team has one member joined the challenge meanwhile the other teams have more than two members. In textual KIS tasks, we miss all the queries and stay behind the other teams. When we review the textual KIS queries, we figure out that we are able to retrieve the correct video segments of some queries in the returned rank list but we miss them because of using the master-shot key-frame to represent the video segment which does not provide enough confident for judging.

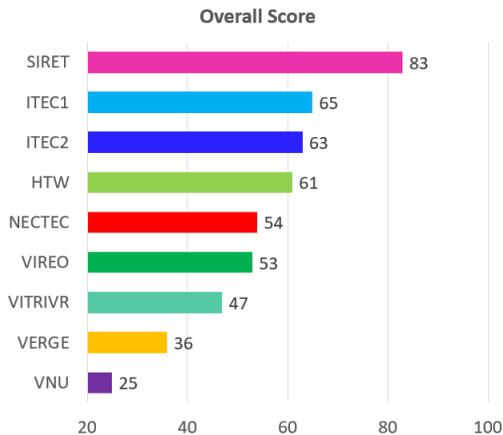


Figure 6: The overall result of VBS 2018.

The overall result of VBS 2018 is shown in Figure 6. We finished at the 6<sup>th</sup> position over 9 teams. We conclude that our proposed color-sketch modal is robust for visual KIS. Besides, our system has some factors that needs to improve including: adapting the system for textual KIS queries, providing an effective browsing interface for result browsing and video context understanding. These directions are our interested topics for the coming VBS in 2019.

### 3 Video Hyperlinking (LNK)

#### 3.1 Semantic Representation Network

Inspired by [21], SRN is a deep model proposed to assess the relatedness between visual and text words, as depicted in 7. The model consists of two LSTM networks, which share weights with each other, for inputs of anchors and targets respectively. The network encodes both target and anchor into the same feature space, before further transforming them into holographic and softmax layers. The holographic layer evaluates the relatedness between anchor and target using circular correlation [22], which directly measures vector correlation in frequency domain using FFT (fast Fourier transform). Finally, the softmax layers output the probabilities of similarity and dissimilarity between anchors and targets.

In our implementation, SRN is trained with LNK development set on top of the pre-trained model in [22]. We feed four different kinds of inputs (visual-visual, visual-text, text-visual, text-text) for each pair of anchor and target during training. The input dimension is set to be 50, i.e., top-50 text words with the highest tf-idf scores and top-50 visual words with the highest probability scores. We zero-pad the vectors if the number of words is less than the predefined dimension. The order of text words follow

the temporal sequence where they appear in ASR, and the order of visual words is based on their index values. The learning is based on cross-entropy loss function.

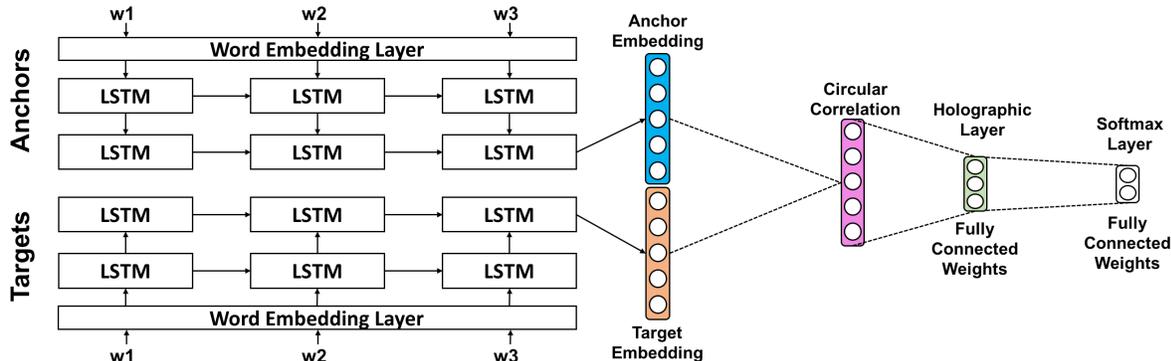


Figure 7: The structure of the semantic representation network.

### 3.2 Data Risk

Different from retrieval task, LNK should minimize the risk of linking to false positives, non-representative and redundant video fragments. We quantify the data risk using hubness, local intrinsic dimension (LID) and diversity. Hubness measures the popularity of points based on reverse nearest neighbors, specifically, the number of points that regards a point as its  $k$ -nearest neighbors. LID characterizes the intrinsic dimension of a point in local context based on its neighborhood structure. In principle, a point with high LID value refers to a “difficult” point that has a higher probability to be false than true positive. Diversity is calculated as the average pairwise distance between a point and its  $k$ -nearest neighbors. Basically, the key idea here is to promote candidate targets with low LID, high hub and diversity to higher rank. We formulate the problem as an optimization algorithm and adopt LID-first algorithm [12] for reranking of baseline results. LID-first algorithm is shown to introduce consistent improvement based on our experiment conducted on LNK development set [12].

### 3.3 Implementation Details

Due to the use of SRN and consideration of data risk in multi-modal setting, we discard the testing samples without speech from experiment. In total, there are 2,719 videos being excluded. The visual features are based on large concept banks used in AVS task [13]. The textual features are extracted from ASR provided by LIMSI [23]. We use the technique similar to [24] for video fragmentation. Basically LDA topic model is used to represent a paragraph, while logical regression is used to identify the fragment boundaries. Meanwhile, same as [25], we ensure that the fragment boundaries will not happen in the middle of a sentence.

### 3.4 Result

**Multi-modality.** Text words do not introduce improvement for  $P@5$  and  $P@10$ , however, successfully recall more true positives in other three measures  $P@20$ , MAP and MAiSP.

**Data risk.** Similarly, LID-first algorithm does not introduce improvement for  $P@5$ , but generally brings larger improvement especially for multimodal run as observed in  $P@20$  and MAP.

	P@5	P@10	P@20	MAP	MAiSP
Run-1	0.864	0.852	0.502	0.1848	0.1113
Run-2	0.864	0.860	0.530	0.1849	0.1128
Run-3	0.856	0.852	0.582	0.1951	0.1199
Run-4	0.856	0.852	0.710	0.2392	0.1473

Table 2: The performance of submitted runs on TRECVID 2017 LNK task

## 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 11250716, 11210514).

## References

- [1] G. Awad, A. Butt, J. Fiscus, D. Joy, A. Delgado, M. Michel, A. F. Smeaton, Y. Graham, W. Kraaij, G. Quénot, M. Eskevich, R. Ordelman, G. J. F. Jones, and B. Huet, “Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking,” in *Proceedings of TRECVID 2017*, 2017.
- [2] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, “Sequence to Sequence - Video to Text,” *ICCV*, pp. 4534–4542, 2015.
- [3] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, “Describing videos by exploiting temporal structure,” *ICCV*, vol. 11-18-Dece, pp. 4507–4515, feb 2015. [Online]. Available: <http://arxiv.org/abs/1502.08029>
- [4] 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*. ACM, 2016, pp. 127–134.
- [5] P. Mettes, D. C. Koelma, and C. G. Snoek, “The imagenet shuffle: Reorganized pre-training for video event detection,” in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 2016, pp. 175–182.
- [6] 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.
- [7] 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.
- [8] W. Zhang, H. Zhang, T. Yao, Y. Lu, J. Chen, and C.-W. Ngo, “Vireo@trecvid 2014: instance search and semantic indexing,” in *In NIST TRECVID Workshop*, 2014.
- [9] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” in *NIPS*, 2014.
- [10] 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)*, 2012.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *arXiv preprint arXiv:1512.03385*, 2015.
- [12] Z. Q. Cheng, H. Zhang, X. Wu, and C. W. Ngo, “On the selection of anchors and targets for video hyper-linking,” in *ICMR*, 2017, pp. 287–293.

- [13] 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.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.
- [15] 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.
- [16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [17] 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.
- [18] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5288–5296.
- [19] C. Cobârzan, K. Schoeffmann, W. Bailer, W. Hürst, A. Blažek, J. Lokoč, S. Vrochidis, K. U. Barthel, and L. Rossetto, "Interactive video search tools: a detailed analysis of the video browser showdown 2015," vol. 76, pp. 5539–5571, 2016.
- [20] P. A. Nguyen, Y.-J. Lu, H. Zhang, and C.-W. Ngo, "Enhanced video search at vbs 2018," in *MultiMedia Modeling*. Cham: Springer International Publishing, 2018, pp. 407–412.
- [21] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *AAAI*, 2016, pp. 2786–2792.
- [22] Y. Tay, M. C. Phan, L. A. Tuan, and S. C. Hui, "Learning to rank question answer pairs with holographic dual lstm architecture," in *SIGIR*, 2017, pp. 695–704.
- [23] L. Lamel, "Multilingual speech processing activities in quero: Application to multimedia search in unstructured data," in *HLT*, 2012.
- [24] H. Amoualian, L. Wei, E. Gaussier, G. Balikas, M. R. Amini, and M. Clausel, "Topical coherence in lda-based models through induced segmentation," in *ACL*, 2017, pp. 1799–1809.
- [25] R. Bois, V. Vukotić, A.-R. Simon, R. Sicre, C. Raymond, P. Sébillot, and G. Gravier, "Exploiting multi-modality in video hyperlinking to improve target diversity," in *MMM*, 2017, pp. 185–197.