Transforming Multi-Concept Attention into Video Summarization

Yen-Ting Liu[1], Yu-Jhe Li[2], Yu-Chiang Frank Wang[1]

[1]National Taiwan University, Taipei, Taiwan
[2]Carnegie Mellon University, Pittsburgh, PA, U.S.A.
{r06942114,ycwang}@ntu.edu.tw, yujheli@cs.cmu.edu

Abstract. Video summarization is among the challenging tasks in computer vision, which aims at identifying highlight frames or shots over a lengthy video input. In this paper, we propose a novel attention-based framework for video summarization with complex video data. Unlike previous works which only apply attention mechanism on the correspondence between frames, our *multi-concept video self-attention (MC-VSA)* model is presented to identify informative regions across temporal and concept video features, which jointly exploit context diversity over time and space for summarization purposes. Together with consistency between video and summary enforced in our framework, our model can be applied to both labeled and unlabeled data, making our method preferable to real-world applications. Extensive and complete experiments on two benchmarks demonstrate the effectiveness of our model both quantitatively and qualitatively, and confirms its superiority over the state-of-the-arts.

1 Introduction

Video summarization [1–4] aims at identifying highlighted video frames or shots, which is among the challenging tasks in computer vision and machine learning. Real-world applications such as video surveillance, video understanding and retrieval would benefit from successful video summarization outputs. To address this challenging task, several deep learning-based models [5–9] employing long short-term memory (LSTM) cells [10] have been recently proposed. However, the use of such recurrent neural network (RNN) based techniques might fail if the length of the input video is long [11]. Therefore, even the training video data are with ground-truth labels, there is no guarantee that RNN-based models would achieve satisfactory results using the last output state. To address the aforementioned issue, several approaches (also based on deep learning) are proposed [12–14]. For example, [12, 13] advances hierarchical structure LSTMs to capture longer video, which is shown to be able to handle video with longer lengths. [14] proposes SUM-FCN which considers a CNN-based semantic segmentation model to deal with videos while alleviating the above concern. Yet, these existing techniques might not exhibit capabilities in modeling the relationship between video frames, since they generally treat each frame equally important. Thus, their summarization performance might be limited.

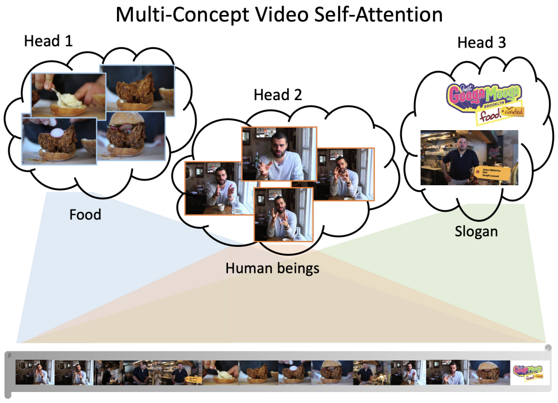

Fig. 1: Illustration of advancing self-attention while preserving visual diversity for video summarization. Noted that Head 1 to 3 denote attention performed in a subspace, which describes proper visual concept information across frames. We proposed a multi-concept video self-attention (MC-VSA) framework for solving this task.

To advance the attention mechanism for video summarization, a number of methods are recently proposed [15–17]. With the ability of learning importance weights across all video frames, attention-based models are expected to be more robust while it is still able to tackle lengthy video as well as the video semantic meanings. For example, [16] firstly proposes an attentive video summarization model (AVS) for improved attention on summarization tasks. [17] also employs attention models for properly identifying video shot boundaries. Nevertheless, these attention-based methods might not generalize to general videos with complex content information, since they typically perform attention on pre-selected feature representations or latent spaces. To make a summarization model more robust to real-world video, one needs to better observe and relate the temporal-concept information within and across video frames, rather than exclusively attend correlation between frames in the video.

In this paper, we propose a novel attention-based deep learning framework for video summarization. With the goal to attend temporally and concept informative features for summarization in the sense, we present a *multi-concept video self-attention (MC-VSA)* model in a discriminative learning mechanism. Based on the idea of [18], we add the multi-head attention mechanism in our model to transform input video frames features into different subspaces. Different from the previous attention model [15–17], this allows us to exploit a variety of visual appearances during the attention process, and thus identify visual concept informative regions across frames for both *summarization* and *video semantic consistency* purposes, which we call multi-concept attention cross whole video time step (temporal and concept attention).

Take an example as illustrated in Fig. 1, it would be desirable to be able to extract different visual concepts corresponding to different semantics or objects with the highlight guidance, so that the joint attention across these concepts would allow satisfactory video summarization outputs. More importantly, our learning framework can be generalized well in a semi-supervised setting, i.e., only a number of training data are with ground-truth labels. More details of our proposed framework will be presented in Sec. 3. In addition, we found that the current evaluation protocol using pre-defined procedure has some problems [19, 3] (e.p., random summaries outperform even the human-generated summaries in leave-one-out experiments), which are mentioned in [20]. Therefore, our quantitative experiment and ablation study are based on both the current [19, 3] and the new [20] evaluation protocol.

The contributions of this paper are highlighted as follows:

- We present a multi-concept video self-attention model (MC-VSA) that aims at attending temporally and concept informative features via transforming input video frames in different subspaces, which is beneficial to video summarization purposes.
- We are among the first to propose the attention-based framework that observes semantic consistency between input videos and learned summarization outputs, which allows the video summarization model can be generalized in semi-supervised settings.
- Experimental results on benchmark datasets confirm that our proposed method achieves favorable performances against the state-of-the-art approaches in two evaluation protocols.

2 Related Work

Video summarization. Video summarization is among the active research topics in computer vision. Several deep methods [5–9] developed for video summarization choose to employ long short-term memory (LSTM) cells [10]. For instance, [5] consider video summarization as a key-frame/shot selection task, and propose an LSTM-based model for addressing this task. Since most of the videos contain hundreds even thousands of frames, it might not be easy for LSTMs to handle such long-term temporal dependency of videos. Hence, some existing approaches [12–14] are further developed to deal with long videos. [12, 13] propose a hierarchical structure of RNN to exploit intra and inter-shot temporal dependency via two LSTM layers, respectively. Such a hierarchical structure is considered to be more preferable for handling video data with hundreds of frames. On the other hand, [14] develop fully convolutional networks for video summarization which requires less training time due to the use of parallel computation techniques. Nevertheless, solving video summarization problems typically requires one to consider the importance across video frames. Existing models generally view the contributions of each frame are equally important during their training stage, which might limit the summarization performance.

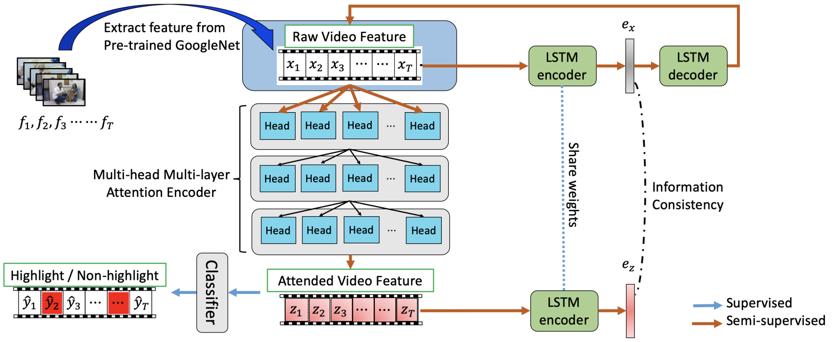

Fig. 2: Overview of our Multi-Concept Video Self-Attention (MC-VSA) for video summarization. Our MC-VSA is composed of three modules: the multi-head multi-layer attention encoder, classifier, and the LSTM-based auto-encoder module. Note that the attention encoder takes input videos $X = \{x_i\}_{i=1}^{T}$ and transforms them into attended features $Z = \{z_i\}_{i=1}^{T}$, followed by the classifier to output the associated highlight scores $\hat{y}_i$. The LSTM-based auto-encoder module preserves data representation ability while enforcing visual concept similarity, allowing guidance of self-attention for summarization purposes.

Attention based summarization. Attention-based models [17, 16, 15, 21] have been proposed to video summarization tasks to alleviate the aforementioned problems. For example, [17] introduces an attention mechanism for detecting the shot boundaries, aiming at improving the summarization performances. An attentive encoder-decoder framework is presented in [16], with the models AVS to video summarization via matrix addition and multiplication techniques. [15] utilizes adversarial learning for visual attention with models of [22], which aims at learning a discriminator to detect highlighted fragments as a summary in the input video. Yet, these attention-based methods typically require ground-truth highlighted supervision, and thus it might not be easy to extend to the cases when such labels are not available.

Semi-supervised and unsupervised summarization. In order to overcome the above concerns, unsupervised [7, 21] and semi-supervised [8] method have been proposed. [7] is the first unsupervised deep learning paper for video summarization, which uses GAN$_{dpp}$ [7] with different LSTM modules to select key-frames from the input video via adversarial learning. Similarly, [21] uses transformer [18] in conditional GAN and achieve improved performance in unsupervised setting. On the other hand, [8] uses an encoder-decoder mechanism aim at enforcing similarity between the input and the summarized outputs. Nevertheless, the above models [7, 8] take two different LSTM module to maintain the information consistency between raw video and the summary, which cannot ensure the video information is in the embedding from the LSTM module. Also, even though [21] improves the unsupervised model using an additional attentive module with GAN, there is still a large gap compared with SOTAs in a supervised setting, and their model lacks discussion for the attentive module. To overcome the above limitations, our MC-VSA model uses a share-weighted LSTM module to encode the video feature and attended feature (resulting in the summary later) enforcing the embedding from the LSTM encoder represents the semantic meaning of the video, and focus on the benefit the attention module brings.

3 Proposed Method

3.1 Problem Formulation and Notation

Given an input video with a total of T frames, our goal is to select the most important key-frames, about 15% of the whole video, as the summarization output. We now define the notations to be used in this paper. Assume that we have frame set $F = \{f_i\}_{f=1}^{T}$ with the associated label set $Y = \{y_i\}_{i=1}^{T}$, where $f_i \in \mathbb{R}^{H \times W \times 3}$ and $y_i \in \mathbb{R}$ represent the i^{th} frame in the target video. To extract the visual features from the frame set F, we apply a CNN (pre-trained on ImageNet) and obtain the video feature set $X = \{x_i\}_{i=1}^{T}$, where $x_i \in \mathbb{R}^d$ (d denotes the dimension of the visual feature).

3.2 Overview of MC-VSA

As shown in Figure 2, we propose a Multi-Concept Video Self-Attention model (MC-VSA) to identify the most representative frames for summarization purposes. Our Multi-Concept Video Self-Attention model is composed of three modules: the multi-head multi-layer attention encoder, classifier, and the LSTM auto encoder decoder module. First, the model takes $X = \{x_i\}_{i=1}^{T}$ with T sequential features as input of its attention encoder. Our attention encoder then transforms input features x_i in X into different subspaces where attention can be performed accordingly. As stated earlier, the attention encoder allows one to exploit various visual appearances during the attention process, and thus identify concept informative regions across frames.

We note that, while the learning of MC-VSA can be simply trained using unlabeled data, we further introduce the visual concept loss, which would guide the MC-VSA if the input video data is with ground-truth highlighted labels. To be more precise, we encourage the learned attention weighted features $Z = \{x_i\}_{i=1}^{T}$ to preserve the same video information as the original one ($X = \{x_i\}_{i=1}^{T}$). To achieve this, the shared LSTM encoder in our framework is designed to match latent vectors e_z and e_x from the same video, thus implying visual concept similarity. If ground-truth highlighted labels are available, the final classifier thus takes the attended features $Z = \{z_i\}_{i=1}^{T}$ to produce the final highlighted scores $\hat{y}_i$ for each z_i. With label supervision available, we encourage the output labels $\hat{Y} = \{\hat{y}_i\}_{i=1}^{T}$ to match the corresponding ground truths $Y = \{y_i\}_{i=1}^{T}$. More details about our MC-VSA model are elaborated in the following.

As for testing, we take the input video with T' frames and produce the final summarization outputs from $\hat{Y} = \{\hat{y}_i\}_{i=1}^{T}$. We note that our MC-VSA is able to handle a different number of frames in a video as input. The experimental results will be presented in the next section.

3.3 Video Self-Attention for Summarization

The multi-head multi-layer attention encoder in our MC-VSA is inspired by the Transformer [18]. To perform concept-temporal attention from the input video, we project the "temporal" video features across video frames onto different subspaces. Each sub-space aims at observing distinct visual concepts as verified later. To be more specific, this attention encoder module is developed to transform input video frames into N subspaces by the introduced N self-attention heads, with the goal of observing and distilling potentially representative information across video frames. It then aggregates the attended results from each subspace and produces the final attended features, which can be seen as jointly incorporating the temporal and concept information. In addition, we also perform such multi-head attention across image layers to exhibit robustness in identifying representative visual concepts.

Standard Self-Attention. For the sake of completeness, we briefly review the self-attention module [23]. Typical self-attention mechanisms transform the input features into three inputs: query Q, key K, and value V by matrix multiplication with transforming matrix. The softmax layer will take the result of the multiplication of Q and K, and produce the attention weights. Hence, the target attention result is produced from the result of the final matrix multiplication of softmax and the V.

Multi-Concept Visual Self-Attention for Video Summarization. To observe both temporal and concept information from the input video frames $F = \{f_i\}_{i=1}^{T}$, we advance the idea of multi-head multi-layer self-attention as described below. As depicted in Fig 3, we have the attention encoder comprise of N self-attention modules (i.e., the head number equals N), and each of them is developed to derive the attended feature each of N subspaces. We firstly transform the input X into N subspace by the N projection layers P_n ($\mathbb{R}^{d^n} \leftarrow \mathbb{R}^d$) where n denotes the projection layer number ($n = 1 \sim N$) and d^n denotes the subspace dimension.

To produce the finalized attended results from all of the N subspaces, we introduce the linear projection layer M^R to derive the final attended features $R = \{r_i\}_{i=1}^{T}$, where $r_i \in \mathbb{R}^d$ (same dimention as X_i), for the original input features $X = \{x_i\}_{i=1}^{T}$, which can be formulated as:

$$R = M^R \cdot \mathrm{concat}(O_{1:N}), \tag{1}$$

where concat means we concatenate the outputs $O_{1:N}$ from all of the N self-attention blocks in the subspace.

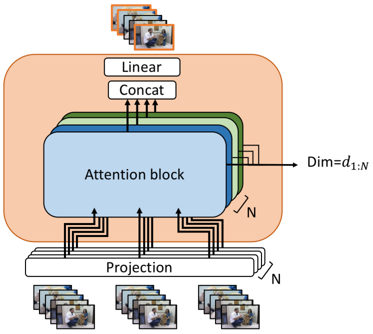

Fig. 3: Illustration of multi-head multi-layer self-attention module of our attention encoder (note that only a single layer is shown for simplicity). With N different single-head attention blocks (each with a projection matrix layer), self-attention can be performed in different subspaces (dimension d_n for each) for capturing diverse visual concepts. We concatenate the outputs $O_{1:N}$ from all attention blocks and obtain the joint attention result R at the output of the final linear transform layer.

To extract rich information from video, we employ L layers in the attention encoder as shown in Figure 2. Namely, the output of R' at the first layer will be passed to the second one to produce the fine-grained output R''. Hence, the finalized attention features $Z = \{z_i\}_{i=1}^{T}$ is denoted as $Z = R^{(L)}$.

Later in our experiments, we will present example visual self-attention results produced by different heads e.g., Figure 4 confirming that the attention encoder exhibits sufficient capability in exploiting visual appearance variations across video frames for attention.

3.4 Self-Learning of Video Semantic Consistency for Video Summarization

The aforementioned attention mechanism can be viewed as a self-summarization process, but lack the ability to ensure that the attended outputs produced by the attention modules would preserve the information in the input video.

To alleviate this limitation, we apply a Siamese network based on a shared LSTM encoder and a single decoder as illustrated in Figure 2. This shared LSTM encoder aims at deriving the compressed latent vectors e_z and e_x for the attended feature set Z and the original feature set X, while the LSTM decoder is to recover the encoded representation for reconstruction purposes. Thus, we have the reconstruction loss $\mathcal{L}_{\text{rec}}$ observe the output of this auto-encoder module:

$$\mathcal{L}_{\text{rec}} = \sum_{i=1}^{T} \|\hat{x}_i - x_i\|^2, \tag{2}$$

where $\hat{X} = \{\hat{x}_i\}_{i=1}^{T}$ denotes the reconstructed feature set and $\hat{x}_i$ indicates the ith recovered video frame feature.

More importantly, to preserve visual concept consistency, we require the encoded vectors e_z and e_x to be close if they are from the same video input. As a result, we enforce the visual concept consistency loss $\mathcal{L}_{\text{con}}$ as follows:

$$\mathcal{L}_{\text{con}} = \|e_x - e_z\|^2. \tag{3}$$

It is worth noting that, our reconstruction loss $\mathcal{L}_{\text{rec}}$ and consistency loss $\mathcal{L}_{\text{con}}$ are both computed without observing any ground-truth label. That is, the introduction of this module allows training using unlabeled video. Together with the labeled ones, our proposed framework can be learned in a semi-supervised fashion. As later verified in our experiments, this would result in promising video summarization performances when comparing against the state of the arts.

3.5 Full Objectives

As depicted in Figure 2, our proposed framework can be learned with fully labeled video. That is, the classification layer takes the resulted attended feature set $Z = \{z_i\}_{i=1}^{T}$ to produce the final highlight potential score $\hat{y}_i$ for each attended feature z_i. More precisely, we encourage the output highlight labels $\hat{Y} = \{\hat{y}_i\}_{i=1}^{T}$ produced by our method can be closed to the ground truth $Y = \{y_i\}_{i=1}^{T}$ and the binary cross-entropy classification loss $\mathcal{L}_{\text{cls}}$ is formulated as below:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{T}\sum_{t=1}^{T} y_t \;\log(\hat{y}_t) + (1 - y_t)\;\log(1 - \hat{y}_t). \tag{4}$$

Thus, the total loss $\mathcal{L}$ is summarized as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{con}}, \tag{5}$$

where $\mathcal{L}_{\text{cls}}$ is calculated by labeled data, while the $\mathcal{L}_{rec}$ and $\mathcal{L}_{\text{con}}$ are derived by both the labeled and unlabeled ones.

We note that, if there is no labeled video data is available during training, $\mathcal{L}_{cls}$ in (5) cannot be computed. Following [9, 14], we train our MC-VSA in such *unsupervised* setting and introduce a diversity loss $\mathcal{L}_{\text{div}}$ (6) to (5). This modification would encourage MC-VSA to select informative yet distinct frames with representative information in an unsupervised learning scenario.

$$\mathcal{L}_{\text{div}} = \sum_{s=1}^{S} \sum_{x^s \in c^s} \sum_{\substack{x^{s\prime} \in c^s \\ x^s \neq x^{s\prime}}} d(x^s, x^{s\prime}). \tag{6}$$

Table 1: Comparisons with existing supervised summarization methods on SumMe and TvSum in differnt experimental settings. The numbers in bold and under line indicate the best and the second result.

Method	SumMe			TvSum		
	C	A	T	C	A	T
Bi-LSTM [5]	37.6	41.6	40.7	54.2	57.9	56.9
DPP-LSTM [5]	38.6	42.9	41.8	54.7	59.6	58.7
GAN$_{sup}$ [7]	41.7	43.6	-	56.3	61.2	-
DR-DSN$_{sup}$ [9]	42.1	43.9	42.6	58.1	59.8	58.9
SUM-FCN [14]	47.5	51.1	44.1	56.8	59.2	58.2
re-SEQ2SEQ [8]	44.9	-	-	**63.9**	-	-
UnpairedVSN [24]	47.5	-	41.6	55.6	-	55.7
H-RNN [12]	44.3	-	-	62.1	-	-
HSA-RNN [13]	44.1	-	-	59.8	-	-
M-AVS [16]	44.4	46.1	-	61.0	61.8	-
VASNet [17]	49.7	51.1	-	61.4	62.4	-
MC-VSA (Ours)	**51.6**	**53.0**	**48.1**	63.7	**64.0**	**59.5**

Table 2: Comparisons with recent unsupervised approaches for video summarization using SumMe and TvSum. Note that * indicates the non deep-learning based methods. The number in bold indicates the best performance.

DATASET	[19]*	[7]	[14]	[9]	[24]	MC-VSA (Ours)
SumMe	26.6	39.1	41.5	41.4	**47.5**	44.6
TvSum	50.0	51.7	52.7	57.6	55.6	**58.0**

4 Experiment

In this section, we first describe the datasets in Sec. 4.1, followed by the experimental protocols and implementation details in Sec. 4.2. For evaluating our MC-VSA, we present quantitative results in Sec. 4.3 and Sec. 4.4. We also provide ablation studies in Sec. 4.6. Finally, we provide qualitative results and visual analysis in Sec. 4.5.

4.1 Datasets

We evaluate our method on two public benchmark datasets SumMe [2] and TvSum [19], and use the additional dataset: OVP and YouTube [9] in the Augmented and Transfer settings:

SumMe. SumMe consists of 25 videos with several different topics such as holidays and sports. Each video ranges from 1 to 6 minutes and annotated by 15 to 18 persons. Thus, there are multiple ground truth summaries for each video.

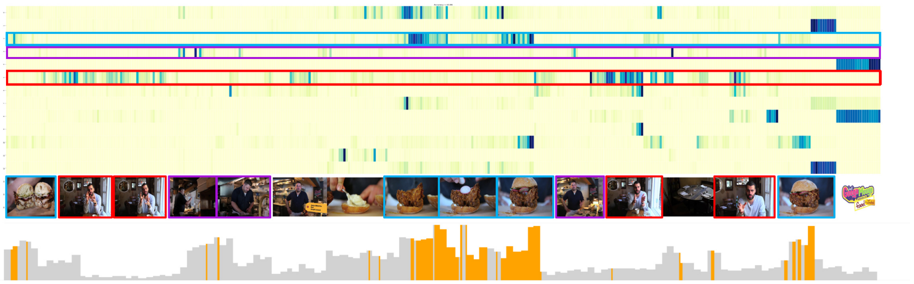

Fig. 4: Visualization example of our MC-VSA for video summarization on TvSum. We visualize selected attention maps generated by the second layer in the attention encoder, with ground truth (grey) and predicted outputs (orange) summarization shown below. Note that the attention outputs bounded in different colors (blue, red and purple) correspond to different multi visual concepts (e.g., burger, commentator, and chef) in this video.

TvSum. TvSum is a larger dataset with 50 videos and includes topics like news and documentaries. The duration of each video is from 2 to 10 minutes. Same as SumMe, TvSum dataset has 20 annotators providing frame-level importance scores. Following [5] and [9], we convert important scores to shot-based summaries for evaluation.

OVP and YouTube. Followed by [5] and [9], we consider two additional challenging datasets released by [9]: OVP and YouTube, which contain 50 videos and 39 videos in the augmented and transfer settings.

4.2 Protocols and Implementation Details

Evaluation protocols We follow the three settings adopted in [5, 9, 14] to evaluate our methods:

- Canonical: we use the standard supervised learning on the dataset, i.e., 80% for training and the rest for testing following previous work [5, 7, 16, 9, 14].
- Augmented: we use the standard supervised training as the canonical setting but augment the training data with OVP and YouTube datasets.
- Transfer: We use three datasets as the training data and a target dataset (e.g. SumMe or TvSum) as the testing data to evaluate the transfer ability of our model.

For a fair comparison, we follow the commonly adopted metric in previous works [5, 9, 14], and computed F-score to assess the similarity between automatic and ground-truth summaries. As for the training/testing data, we apply

Table 3: Ablation studies and performance comparisons on TvSum dataset. We take dppLSTM [5] and [9] for comparisons in Kendall's τ, and Spearman's ρ evaluation protocol. The number in bold denotes the best performance.

Method	w/ knapsack algo.	w/o knapsack algo.	
	F1 score	Kendall's τ	Spearman's ρ
dppLSTM	60.0	0.042	0.055
DR-DSN$_{dpp}$	58.0	0.020	0.026
[26]	58.4	0.078	0.116
VASNet	61.4	-	-
Ours (w/o attention)	59.7	0.005	0.006
Ours (1layer-1head)	60.1	0.065	0.079
Ours (3layers-24heads)	**63.7**	**0.116**	**0.142**

the same standard supervised learning setting as [5, 7, 16, 9, 14] where the training and testing are from the disjoint part of the same dataset. We report the results at F-score in all of the settings. To avoid the shortage with F-score, which is mentioned by [20], we additionally conduct the experiments using new protocols [20] and make comparisons with state-of-the-arts as well in Table. 3.

Implementation details We downsample the video data into frame sequences in 2 fps as previous work [5, 9]. For fair comparisons with [5, 14, 9, 16], we also employ GoogleNet [25] pre-trained on ImageNet as backbone as our CNN for extracting the video features while the output dimension d is 1024 (output of pool5 layer of the GoogleNet). All of the attention modules are composed of linear projection matrices as mentioned in Section 3. We set the number of heads N as 24 while the dimension d_n of each subspace features are set as $\{64 \ |n = 1 \sim 12\}$ and $\{128 \ |n = 13 \sim 24\}$. Our MC-VSA comprises of 3 multi-head attention layer, i.e., we set L as 3. The classifier is composed of a fully connected layer followed by a sigmoid activation. The LSTM encoder and decoder in our model contain 512 units. Besides, we set the learning rate as $1e^{-4}$ for all of our components. We use Adam optimizer to train the MC-VSA by optimizing the objective loss. We produce the summary outputs by KNAPSACK algorithm following [5, 14]

4.3 Comparison with Supervised Approaches

We compare our proposed MC-VSA with state-of-the-art methods on two benchmark datasets and summarize the results in Table 1. In canonical setting, we see that our MC-VSA performed favorably against recent LSTM based approaches (e.g., Bi-LSTM [5], DPP-LSTM [5], GAN$_s up$ [7], and DR-DSN$_s up$ [9]), and the CNN-based model (SUM-FCN [14]). Our model also achieves the improvement over LSTM module Bi-LSTM [5], DPP-LSTM [5], GAN$_s up$ [7], and DR-DSN$_s up$ [9]) by a large margin. For both augment and transfer settings, we also observe similar trends and achieve improved performances against state-of-the-art methods. It is worth noting that, though our model exhibit inferior to

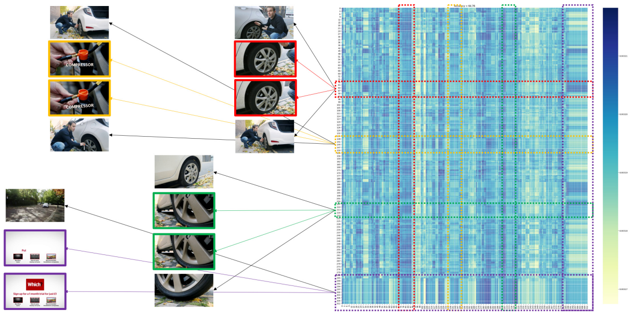

Fig. 5: Visualization for multi-head self-attention at the first layer of MC-VSA. We show that this low-level attention layer not only implies shot-level attention, visual concepts associated with similar objects are properly attended across video frames (e.g., attention outputs bounded in difference colors).

re-SEQ2SEQ [8] by 0.2% at F-score on the TvSum dataset, our approach and several competitors outperform [8] by a large margin on the SumMe dataset.

4.4 Comparisons with Unsupervised Approaches

We report our unsupervised learning results and comparisons in Table 2. With training strategies presented in Sect. 3, we evaluate the effectiveness of our MC-VAS in the unsupervised setting by comparing it with five existing unsupervised summarization methods [7, 9, 14, 6]. As shown in Table 2, our MC-VAS was able to achieve comparable results with the state-of-the-art did on both SumMe and TvSum. Thus, even without any supervision, we can confirm that our model takes advantage of multi-concept video self-attention with visual concept consistency for video recovery and summarization.

4.5 Qualitative Results

To analyze the effectiveness of the self-attention module in MC-VSA, we present visualization results in Fig. 4, in which the attention outputs were observed from the second (high-level) layer in our model. In Fig. 4, the upper half part illustrates frame-level attention weights for the selected 13 heads in our model. Note that each row in the upper part of this figure represents a single head, in which the darkness of the color indicates the importance of the associated frame. From this example result, we see that the attention weights for different heads are quite different, which confirms that our multi-head self-attention mechanism leads to visual concept diversity. For example, by comparing the learned attention weights and the corresponding frames (i.e., upper vs. lower parts of Fig. 4), we see that one head in the blue rectangular box exhibits the semantic meaning of hamburger, while the red one indicates the appearance of the food critic. And, as confirmed by earlier quantitative experiments, these resulting attention weights across different heads are indeed correlated with the summarization outputs.

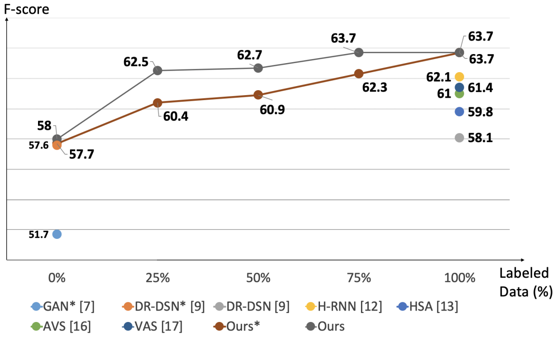

Fig. 6: Performance analysis of our model in semi-supervised settings on TvSum. The x/y-axis indicate the percentage of labels in the training set and the F-score, respectively. Results of recent supervised and unsupervised approaches are depicted for comparison purposes. Note that Ours* denotes our MC-VSA excluding the LSTM auto-encoder module while Ours represents the full model.

On the other hand, Fig. 5 illustrates the attention observed in the first layer of our MC-VSA, which can be viewed as a low-level self-attention of multiple heads from the input video. Take the entry of the ith column at the jth row, its value reflects the attention for the corresponding frame pair. From this figure, we see that the attention boundaries were visible and generally matched the shot boundaries. In addition, we see that visual concepts with similar visual appearances (e.g., wheel, car, etc.) were properly identified in particular video segments, which reflect the concept-specific video shot information of this input video. Guided by the classification and data recovery losses, this explains why our proposed model is able to capture multiple representative visual information, achieving satisfactory summarization outputs.

4.6 Ablation Studies

Semi-supervised settings. We first conduct a semi-supervised learning analysis of our proposed MC-VSA on the TvSum dataset. As illustrated in Figure 6, the vertical axis indicates the F-score, and the horizontal axis represents the percentage of the labeled data observed during training. For the completeness of analysis, we compare our approach with 5 supervised or unsupervised summarization methods in the same figure. From the results presented in the figure, we see that our MC-VSA achieved improved performances over others. Moreover, we see that our method was able to perform favorably against existing supervised approaches by a large margin, even when only 25% labels were observed by our model during training. Furthermore, Figure 6 compares our model with its variants in semi-supervised settings. Note that Ours* denotes our model excluding both reconstruction loss and visual concept consistency loss. Refer to the semi-supervised analysis, the performance drop between Ours and Ours* confirms that such two loss terms are crucial when observing unlabeled data. We note that in figure 6 for cases 100%, the performances achieved by ours and ours* respectively are the same. This is because when we use the entire label set in 100 %, the LSTM module only serves to train our model more stable instead of achieving improved performance.

Network architecture design. We now further discuss the design of our model architecture. In Figure 7, we show the performance of our multi-head multi-layer attention model with varying numbers of layers and heads (x-axis). From this figure, we see that while such hyperparameters need to be determined in advance, the results were not sensitive to their choices. In other words, with a sufficient number of heads and layers, multiple visual concepts can be extracted for summarization purposes as shown in our supplementary video. As shown in Table 3, we apply three evaluation protocols, including F1, Kendall's τ, and Spearman's ρ, to evaluate our MC-VSA model. Kendall's τ, and Spearman's ρ are proposed by [20] for impartial comparison. We compare our full model (3layers-24heads) with other baseline models. To be more specific, we take the VASNet [17] as the naive self-attention model baseline. Ours (w/o attention) represents the MC-VSA model consisting of the only classifier while ours (1layer-1head) indicates a only single layer and head in the attention encoder. The performance drop is observed when comparing ours with the above-mentioned baseline models. We additionally report the performance provided by [5] and [9] in [20] in Kendall's τ and Spearman's ρ evaluation protocol for benchmark comparison.

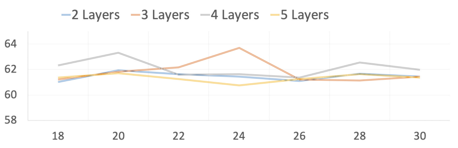

Fig. 7: Performance of our attention model with varyingnumbers of layers (L) and heads (N) (i.e., x-axis). We fix L= 3 and N = 24 for our model in all experiments.

5 Conclusion

We presented a novel deep learning framework *multi-concept video self-attention (MC-VSA)* and consistency constraint between the input video and the output summary for video summarization. The core technical novelty lies in the unique design of multi-concept visual self-attention model, which jointly exploits concept and temporal attention diversity in the input videos, while enforcing the summarized outputs to have consistency with original video. Our proposed framework not only generalized in supervised, semi-supervised and unsupervised settings but also in both evaluation protocols. Also, our experiments and qualitative results confirmed the effectiveness of our proposed model and its ability to identify certain informative visual concepts.

References

1. Chao, W.L., Gong, B., Grauman, K., Sha, F.: Large-margin determinantal point processes. In: UAI. (2015) 191–200
2. Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: European conference on computer vision, Springer (2014) 505–520
3. Gygli, M., Grabner, H., Van Gool, L.: Video summarization by learning submodular mixtures of objectives. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3090–3098
4. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Summary transfer: Exemplar-based subset selection for video summarization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 1059–1067
5. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: Proceedings of the European Conference on Computer Vision (ECCV). (2016)
6. Jung, Y., Cho, D., Kim, D., Woo, S., Kweon, I.S.: Discriminative feature learning for unsupervised video summarization. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). (2018)
7. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial lstm networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
8. Zhang, K., Grauman, K., Sha, F.: Retrospective encoders for video summarization. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018)
9. Zhou, K., Qiao, Y., Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). (2018)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9** (1997) 1735–1780
11. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 4694–4702
12. Zhao, B., Li, X., Lu, X.: Hierarchical recurrent neural network for video summarization. In: Proceedings of the 25th ACM international conference on Multimedia, ACM (2017) 863–871
13. Zhao, B., Li, X., Lu, X.: Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7405–7414
14. Rochan, M., Ye, L., Wang, Y.: Video summarization using fully convolutional sequence networks. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 347–363
15. Fu, T.J., Tai, S.H., Chen, H.T.: Attentive and adversarial learning for video summarization. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2019) 1579–1587
16. Ji, Z., Xiong, K., Pang, Y., Li, X.: Video summarization with attention-based encoder-decoder networks. arXiv preprint arXiv:1708.09545 (2017)
17. Ji, Z., Hajar Sadeghi, S., Vasileios, A., Dorothy, M., Paolo, R.: Summarizing videos with attention. Proceedings of the AAAI Conference on Artificial Intelligence Workshops (AAAI workshops) (2018)

18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. (2017) 5998–6008
19. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 5179–5187
20. Otani, M., Nakashima, Y., Rahtu, E., Heikkila, J.: Rethinking the evaluation of video summaries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 7596–7604
21. He, X., Hua, Y., Song, T., Zhang, Z., Xue, Z., Ma, R., Robertson, N., Guan, H.: Unsupervised video summarization with attentive conditional generative adversarial networks. In: Proceedings of the ACM Conference on Multimedia (MM). (2019)
22. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. In: Advances in Neural Information Processing Systems. (2015) 2692–2700
23. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. (2015) 2048–2057
24. Rochan, M., Wang, Y.: Video summarization by learning from unpaired data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 7902–7911
25. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: the IEEE conference on computer vision and pattern recognition. (2015) 1–9
26. Chen, Y., Tao, L., Wang, X., Yamasaki, T.: Weakly supervised video summarization by hierarchical reinforcement learning. In: Proceedings of the ACM Multimedia Asia. (2019) 1–6