Video Summarization by Learning Relationships between Action and Scene

Jungin Park, Jiyoung Lee, Sangryul Jeon, and Kwanghoon Sohn*
School of Electrical and Electronic Engineering
Yonsei University, Seoul, Korea
{newrun, easy00, cheonjsr, khsohn}@yonsei.ac.kr

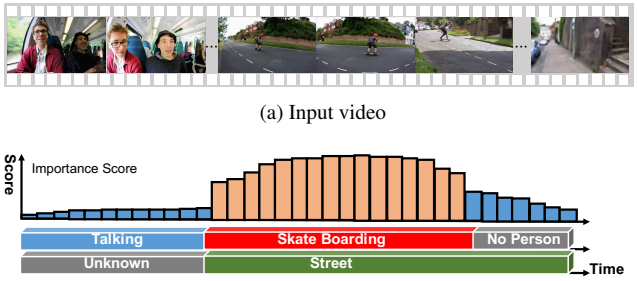
Abstract

We propose a novel deep architecture for video summarization in untrimmed videos that simultaneously recognizes action and scene classes for every video segments. Our networks accomplish this through a multi-task fusion approach based on two types of attention modules to explore semantic correlations between action and scene in the videos. The proposed networks consist of the feature embedding networks and attention inference networks to stochastically leverage the inferred action and scene feature representations. Additionally, we design a new center loss function that learns the feature representations by enforcing to minimize the intra-class variations and to maximize the inter-class variations. Our model achieves a score of 0.8409 *for summarization and accuracy of* 0.7294 *for action and scene recognition on test set of CoVieW'19 dataset, which is ranked 3rd.*

1. Introduction

Multimedia on the Internet is growing rapidly with a development of online video service platforms such as YouTube and Flickr. With this remarkable growth, current efforts on understanding untrimmed videos have been focused on tackling video summarization task that produces a shorter video to convey the important and relevant content of the input video.

Video summarization has achieved great success in recent years by leveraging deep convolutional neural networks (CNNs) with their high invariance to semantic variations, being considered as a structured prediction problem [40, 19, 41, 4, 37, 38, 12, 20, 27]. Most existing summarization methods, however, ignore semantic context priors (*e.g.* human activities and surrounding scenes), even though they have been shown to be effective in comprehensive understanding of videos [30]. Our key observation is that creating a brief yet informative synopsis of a long video is highly correlated with certain events that consist of human action, and usually constrained by particular scenes. For example, generating the summarization of a video with 'Go Skateboarding Day' is highly related to action (*e.g.*, skate boarding) and scene class (*e.g.*, street) as illustrated in Fig. 1. Therefore, to extract representative frames from videos, it is desirable to understand both the action and scene of the video at the same time.



(a) Input video

(b) Video Understanding

Figure 1. Illustration of our video summarization with temporal action and scene recognition. Given an input video, we aim to predict importance scores for summarizing the video while simultaneously recognizing its highly relevant action and scene classes for every clips of the video.

With this motivation, we present a novel network architecture that incorporates action and scene information of an untrimmed video to determine the continuous importance scores along the temporal axis. We studies the discovery of highlights in a video considering the context based on action and scene recognition. The key idea of the proposed networks is to weave the advantages of action recognition, scene recognition and video summarization in a joint and boosting manner. Our networks accomplish this through an unified network by learning three objectives from a shared representation to improve learning efficiency and prediction accuracy in an end-to-end manner.

Moreover, we investigate an aspect of self-attention module in feature embedding networks inspired by CBAM [36]. Self-attention module focuses on important features and suppresses unnecessary ones at the channel and

*Corresponding author

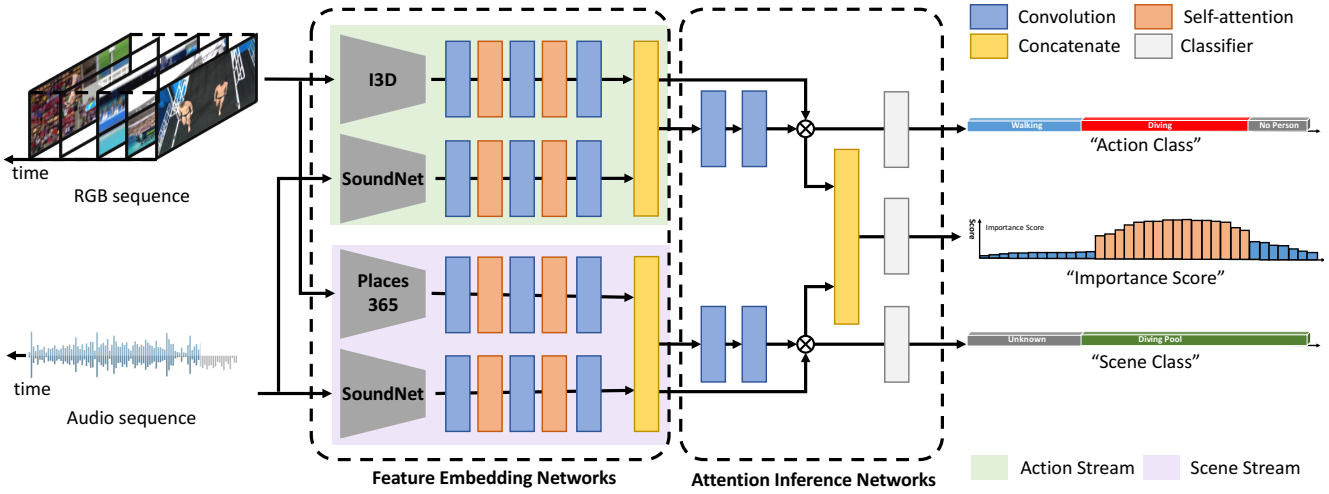


Figure 2. An overall architecture of proposed model for video summarization with action and scene recognition. We first extract visual and audio feature representations for a set of uniformly sampled video clips using pre-trained networks. At each stream, visual and audio features are fused to predict each action and scene classes for every clips. And action and scene features are fused to predict importance scores by exploiting the semantic correlations between action and scene.

temporal axis. In attention inference networks, we obtain actionness and sceneness probabilities that indicate 'what (*i.e.* action)' and 'where (*i.e.* scene)' to attend for three different video understanding tasks, respectively. We carefully present a modified center loss [35] for simultaneously minimizing intra-class variations and maximizing inter-class variations. Experimental results on the CoView'19 dataset demonstrate the effectiveness of the proposed networks in video summarization with temporal action and scene recognition.

2. Related Work

2.1. Action Recognition

Action recognition is the one of the most important task in video understanding. Many studies have been extensively studied by leveraging the recent advances of CNNs [29, 33, 3, 10, 39, 34] to encode spatio-temporal information. One natural way is to leverage CNNs with various formulation such as devising two-stream architecture on visual frames and stacked optical flows [29], extension of convolution kernels in CNNs from 2D to 3D [33], or combining two-stream processing and 3D convolutions [3]. Another alternative solution is to utilize recurrent neural networks (RNNs) over the activation of the last fully-connected layer in a 2D CNNs [10, 39]. Although those methods have shown improved performance in action recognition, it can be observed that most aforementioned methods mainly focuses on improving action classification performance only.

2.2. Scene Recognition

Scene recognition in images gives helpful context information for object recognition. Since objects are main components of scenes in images, accurate recognition of scenes requires knowledge about both scene and objects [9]. Likewise, scene recognition in untrimmed videos helps recognizing action. Recently, some methods [8] for action recognition have investigated this aspect to comprehensively video understanding. Marszaek *et al.* [21] show the relevance of the cooccurrence between action and scene for retrieving actions in movie. Heilbron *et al.* [8] described semantic context, *i.e.* action-object and action-scene relationships, in the detecting action process. Coview'18 challenge [30] has been held for studying strong mutual relationships among action and scene. However, relevance of action and scene have not been studied yet for video summarization.

2.3. Video Summarization

Several approaches have studied video summarization with various formulations including video synopsis Pritch08, time-lapses [11, 16, 26], montages [13, 32] and storyboards [5, 7, 6, 18, 20, 37, 40]. Our problem statement is related to storyboards which select representative video frames to summarize key events present in the entire video. Recent works on video summarization have learned how to select informative video subsets closed to human-created summaries [18, 5, 7, 6, 28]. In the last few years, several deep learning based approaches are presented [40, 19, 41, 4, 37, 38, 12, 20, 27] by learning with labelled videos. Rochan *et al.* [27] have proposed fully con-

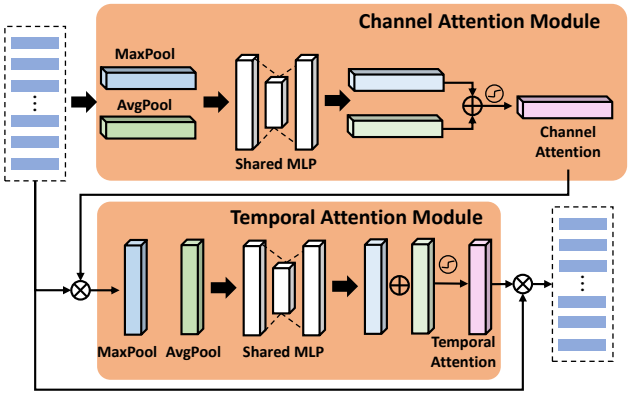


Figure 3. Illustration of self-attention module. The inter-channel relationship and the inter-temporal relationship of features are extracted in the channel attention module and the temporal attention module, respectively. We can extract the coarse- and finer-level attention by using two pooling layers.

volutional sequence model to summarize long, egocentric videos into short representative videos.

More related to our work is highlight detection that predicts highlight score to each segment [38]. Yao *et al*. [38] retrieved highlights from first-person videos with pairwise deep ranking model to learn the relationship between highlight and non-highlight segments. However, there is lack of attention inference considering action and scene context to determine the importance of each segment in the same video.

3. Proposed Method

3.1. Problem Formulation

Given an input video composed of T clips v_i for $i \in \{1, \ldots, T\}$, we aim to determine the importance score of each video clip $\mathbf{y}_i^{\text{sum}}$, and simultaneously recognize action and scene classes, $\mathbf{y}_i^{\text{act}}$ and $\mathbf{y}_i^{\text{scn}}$, in an end-to-end manner. To this end, we formulate *feature embedding networks* to extract shared action and scene feature representations, and *attention inference networks* to infer the each class and score with shared representations. Concretely, we first extract feature representations and then automatically detect the semantically meaningful clips in temporal detection networks. Finally, classification networks are composed of three classifiers estimate action class $\mathbf{y}_i^{\text{act}}$, scene class $\mathbf{y}_i^{\text{scn}}$, and importance score $\mathbf{y}_i^{\text{sum}}$ of the clips. The importance scores are continuous scores ranged from 0 (*i.e*. not important) to 2 (*i.e*. important). The overall architecture of our approach is illustrated in Fig. 2.

3.2. Network Architecture

Feature embedding networks. As shown in Fig. 2, feature embedding networks are formulated in two-stream configuration, *i.e*., action stream and scene stream. Each stream has same structure that incorporates both visual and audio information to leverage complementary modalities from the raw video.

Formally, let $\mathbf{x}^{\text{act}} \in \mathbb{R}^{D\times T}$, $\mathbf{x}^{\text{scn}} \in \mathbb{R}^{D\times T}$ and $\mathbf{x}^{\text{aud}} \in \mathbb{R}^{D\times T}$ be the video level visual feature map from pre-trained action networks [14], visual feature map from pre-trained scene networks [42] and audio feature representation from pre-trained networks [2], respectively. Each stream in the feature embedding networks is designed to extract action and scene features, denoted as $\mathbf{f}^{\text{act}}$ and $\mathbf{f}^{\text{scn}}$ by passing through a feed-forward and concatenating visual and audio features such that,

$$\begin{aligned} \mathbf{f}^{\text{act}} &= \mathcal{F}(\mathbf{x}^{\text{act}}; \mathbf{W}_{\text{v}}^{\text{act}}) || \mathcal{F}(\mathbf{x}^{\text{aud}}; \mathbf{W}_{\text{a}}^{\text{act}}), \\ \mathbf{f}^{\text{scn}} &= \mathcal{F}(\mathbf{x}^{\text{scn}}; \mathbf{W}_{\text{v}}^{\text{scn}}) || \mathcal{F}(\mathbf{x}^{\text{aud}}; \mathbf{W}_{\text{a}}^{\text{scn}}), \end{aligned} \tag{1}$$

where $||$ represents the concatenation operator, $\mathbf{W}^{\text{act}}$ and $\mathbf{W}^{\text{scn}}$ are the network parameters for each action and scene stream, respectively. The feature embedding networks are composed of three convolution layers with rectified linear unit (ReLU), and two self-attention modules [36] between every convolution layers. We refine the features by sequentially applying self-attention modules that emphasize meaningful features along two principal dimensions respectively: channel and temporal axes. The goal of self-attention module is to increase representation power by focusing on important features and suppressing unnecessary ones based on attention mechanism. To this end, we exploit the inter-channel relationship of features in the channel attention module, and utilize the inter-temporal relationship of features in the temporal attention module. As shown in Fig. 3, each module is consisted of two pooling layers and fully-connected (FC) layers with sigmoid activation function. We adopt both average-pooling and max-pooling operations to features aggregating general attention and finer attention as in [36].

Given intermediate average-pooled features and max-pooled features $\mathbf{h} \in \mathbb{R}^{D\times T}$ as input, the self-attention module infers a 1D channel attention $\mathbf{A}_c \in \mathbb{R}^{D\times 1}$ and a temporal attention $\mathbf{A}_t \in \mathbb{R}^{1\times T}$. The overall self-attention process can be formulated as,

$$\begin{aligned} \mathbf{h}' &= \mathbf{h} + \mathbf{h} \otimes \mathbf{A}_c(\mathbf{h}), \\ \mathbf{h}'' &= \mathbf{h}' + \mathbf{h}' \otimes \mathbf{A}_t(\mathbf{h}'), \end{aligned} \tag{2}$$

where $\otimes$ represents element-wise multiplication. Motivated by BAM [23], residual learning scheme is adopted along with the attention mechanism to facilitate the gradient.

Attention inference networks Although the embedded features from action and scene stream characterize the video, a direct fusion (*e.g*., concatenation) of these inputs does not present optimal performance for video summarization. Thus, we suggest attention inference networks for

adaptive fusion of action and scene context. While the self-attention module detects discriminative parts in the inter-channel and the inter-temporal relationship, attention inference networks extract *actionness* and *sceneness* by exploiting correlations between the visual and the audio features.

More precisely, the temporal detection network is composed of two convolution layers with batch normalization and ReLU function between two layers. The second fully-convolution layer determines the weights for each clip in the form of a probability distribution. The output of the second convolution layer is element-wisely multiplied to the embedded features similar to a soft attention mechanism [22] such that,

$$\begin{aligned} \mathbf{z}^{\text{act}} &= \mathbf{w}^{\text{act}} \otimes \mathbf{f}^{\text{act}}, \\ \mathbf{z}^{\text{scn}} &= \mathbf{w}^{\text{scn}} \otimes \mathbf{f}^{\text{scn}}, \end{aligned} \tag{3}$$

where $\mathbf{w}^{\text{act}}$ and $\mathbf{w}^{\text{scn}}$ are the attention weight vectors for action and scene feature representations.

While the attention weighted features from attention inference networks are fed into action and scene classifiers to predict action and scene class for each clip, a concatenated feature of $\mathbf{z}^{\text{act}}$ and $\mathbf{z}^{\text{scn}}$ utilizes to estimate importance score for video summarization. The action and scene classification module is consisted of a convolution layer with batch normalization and a FC layer. The importance score classifier has same structure with the other two classifiers but the number of parameters are twice since the dimension of the input. The output of the importance score classifier is passed through the sigmoid function to make the value of the output from 0 to 1. Since groundtruth scores are ranged from 0 to 2, we scale prediction of importance score by multiplying the scaling factor 2.

3.3. Loss Functions

To optimize the proposed network, we define the loss function as the sum of three loss functions:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{act}} + \alpha\mathcal{L}_{\text{scn}} + \beta\mathcal{L}_{\text{sum}}, \tag{4}$$

where $\mathcal{L}_{\text{act}}$ is the action loss, $\mathcal{L}_{\text{scn}}$ is the scene loss, $\mathcal{L}_{\text{sum}}$ is the summarization loss, and α, β are the hyper parameters to balance three loss functions.

The action and scene losses are composed of the classification loss and the center loss [35] as follows:

$$\mathcal{L}_{\text{act,scn}} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{cent}}, \tag{5}$$

where $\mathcal{L}_{\text{cls}}$ is a standard multi-label cross-entropy loss between groundtruth labels and predicted classes, $\mathcal{L}_{\text{cent}}$ is the center loss which minimizes intra-class variations and maximizes inter-class variations, and λ is the hyper parameter to balance two loss functions.

For the center loss, we first define a class center feature $\mathbf{c}_k \in \mathbb{R}^d$ for k-th class, where d is the dimension of the weighted feature $\mathbf{f}^*$. The class center features are randomly generated according to each class except for the background class (*no person* for the action and *unknown* for the scene). The background class center features are set to 0 to satisfy the assumption that the importance score in the background scene should be close to zero. With the weighted feature and the class center feature, the center loss can be formulated as,

$$\mathcal{L}_{\text{cent}} = \frac{1}{2}\sum_{i=1}^{T}[||\mathbf{z}_i - \mathbf{c}_{y_i}||_2^2 - \frac{1}{N-1}\sum_{y_j \neq y_i} ||\mathbf{z}_i - \mathbf{c}_{y_j}||_2^2], \tag{6}$$

where T is the total number of clips, $\mathbf{z}_i$ is the weighted feature of i-th clip, $\mathbf{c}_{y_i}$ is the class center feature for i-th clip, and N is the total number of classes. We can minimize intra-class variations by minimizing the first term of the center loss and maximize inter-class variations by minimizing the second term of the center loss. Unlike [35] which focuses on minimizing intra-class variations, our loss function allows for more discriminative feature by considering inter-class variations.

In our case, a importance score of each clip reflects its degree of interest within a video. To this end, we use mean-squared error (MSE) as the summarization loss to optimize the importance scores. Note that the whole networks are learn parameters with action labels, scene labels, and importance score labels for every clips in the end-to-end manner.

4. Experiments

4.1. Implementation Details

For RGB sequences, we use I3D networks [14] trained on the Kinetics dataset [15] to extract action features and ResNet50 trained on the Places365 dataset [42] to extract scene features for video clips. We rescale the smallest dimension of a frame to 240 and perform the cropping of size 224×224. In cropping process, we perform different 5 crops (left-upper, right-upper, center, left-bottom, and right-bottom) and randomly used during training for data augmentation. The inputs to the I3D models are stacks of 16 frames sampled at 16 frames per second, and the input to the ResNet50 is 1 frame sampled at 1 frame per second. The action and the scene features are averaged at intervals of every 5 seconds to obtain clip-level features. For audio sequences, we use SoundNet [2] trained on the ESC50 dataset [25] to extract audio features. We convert the video to sound MP3s and reduce the sampling rate to 22kHz, and convert to single channel audio. We also scale the waveform to be in the range of $[-256, 256]$. The window size of the model set to 1 second and audio features are also averaged at every 5 seconds.

We sample 160 clips at uniform interval from each video in both training and testing. The networks are trained us-

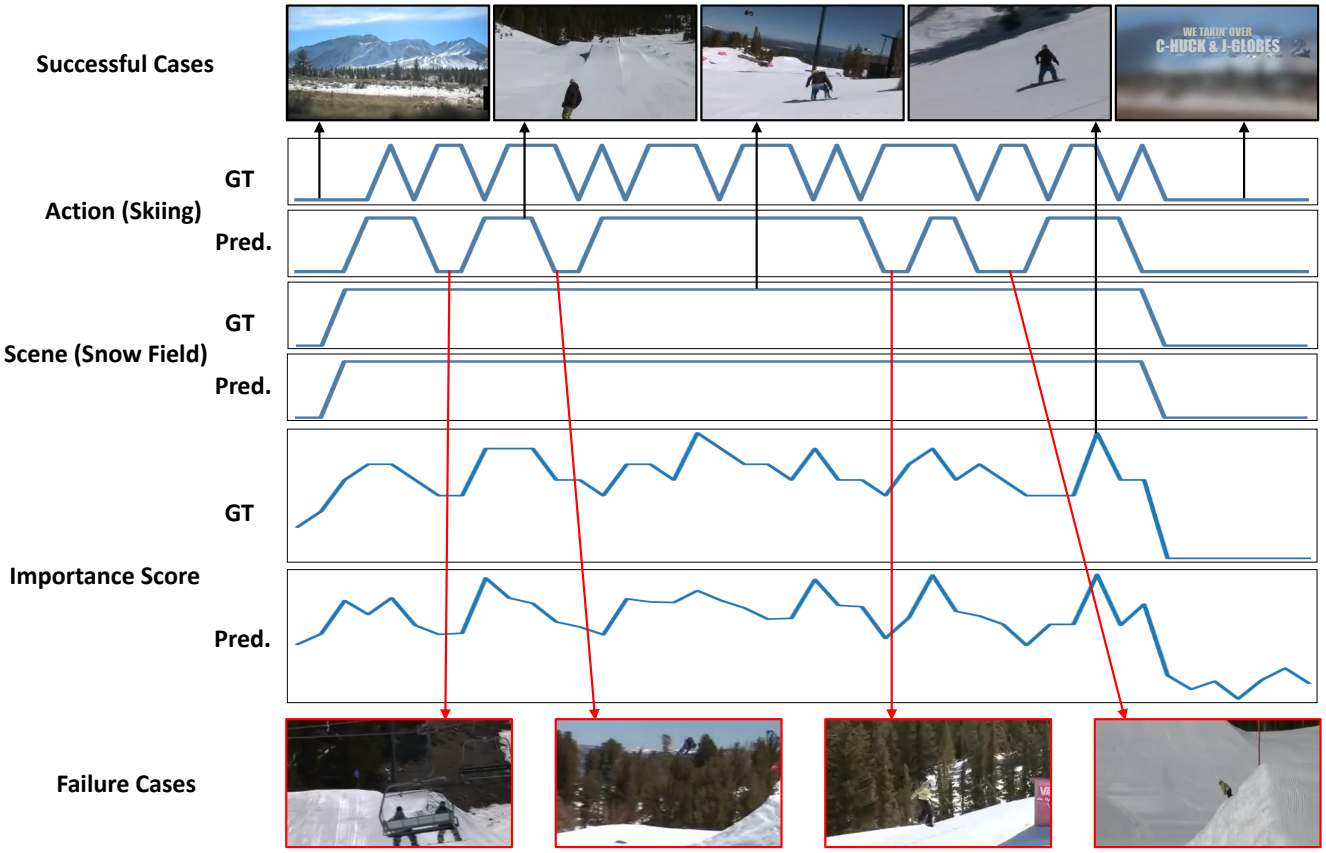


Figure 4. Video summarization with action and scene recognition result for qualitative analysis on CoVieW'19 dataset. We represent only the class of "skiing" for action and "snow field" for scene although there are few other classes.

ing stochastic gradient descent (SGD) algorithm with learning rate 10^{-4} for classification losses and 10^{-2} for center losses. The hyper parameters are set to $\alpha = 1, \beta = 10$ in (4), and $\lambda = 10^{-1}$ in (5). Our algorithm is implemented in PyTorch [24].

4.2. Experimental Settings

Dataset. Our method is evaluated on the CoVieW'19 challenge dataset for video summarization with action and scene recognition task. The CoVieW'19 dataset consists of untrimmed videos sampled from the Youtube-8M [1] dataset, Dense-captioning dataset [17], and TVSum dataset [31] with annotated importance score, action and scene class labels for each video. Each label is annotated with every 5 second long segments. The importance score indicates how important each segment is compared to other segments in the video and its value is scaled from 0 (not important) to 2 (most important). The number of action and scene classes are 99 and 78, respectively.

Dataset is consisted of 1,500 videos which are splitted into 1,200 videos for training and 300 videos for testing. In training phase, we randomly split training set to 1,080 and 120 videos for training and validation, respectively. For the challenge, we evaluate the performance on test set which composed of 300 videos.

Evaluation metric. For the quantitative evaluation, importance score for video summarization, and temporal action and scene recognition performance are measured separately. The video summarization metric is defined as

$$\frac{1}{N}\sum_{k=1}^{N}\frac{\sum_{i=1}^{N_s} I(k, pred_i)}{\sum_{i=1}^{N_s} I(k, GT_i)},$$

where N is the total number of videos, N_s is the number of selected clips, $I(k, pred_i)$ is ground truth importance score of the clip predicted to have top-i score, and $I(k, GT_i)$ are ground truth importance score of the clip have top-i score for k-th video, respectively. In CoVieW'19 challenge, N_s is set to 6. The action and scene recognition performances are measured differently in the validation set and the test set. For the validation set, we evaluate separately the action and scene recognition performances for whole clips by top-1 and top-5 accuracy as used in action recognition [34]. For

Validation set	Summarization score
Validation set 1	0.9093
Validation set 2	0.8768
Validation set 3	0.8971
Average	**0.8944**

Table 1. Performance evaluation with the proposed model on the randomly divided validation set of CoVieW'19 dataset. Accuracies are measured using summarization metric.

Task	scores	rank
Video summarization	0.8409	2
Action & Scene recognition	0.7294	4

Table 2. Performance evaluation with the proposed model on the test set of CoVieW'19 dataset. The scores are measured using summarization metric and recognition challenge metric in (7). Rank represents the ranking for each task.

Audio	Action@1	Action@5	Scene@1	Scene@5
✗	58.95	82.76	59.56	84.92
✓	**60.10**	**85.46**	**63.31**	**86.94**

Table 3. Performance comparison of ablation study for audio stream on validation set of CoVieW'19 dataset. Accuracies are measured using the classification accuracy at top1 and top5 predictions.

the test set, we evaluate action and scene recognition performance using top-k hamming scores for selected clips in video summarization, such that,

$$H(K) = \frac{1}{N}\sum_{n=1}^{N}\sum_{label=1}^{L}\sum_{k=1}^{K}\frac{AND(k-\text{th Pred}_{label}, GT_{label})}{L}, \tag{7}$$

where $AND(a, b) = 1$ only if a and b has exactly same label index on action or scene. K is set to 5 in this challenge.

4.3. Results

In this section, we analyze our proposed network with the qualitative and the quantitative evaluations. We investigate the contribution of components proposed in our architecture with respect to 1) the effects on combination of different modalities such as visual and audio, and 2) the effectiveness of the loss functions.

Fig. 4 shows the qualitative result for video summarization with action and scene recognition. In fact, there are few other classes (*e.g.* talking for action, mountain for scene), we only represent "skiing" class for action and "snow field" class for scene that are dominant classes of the video. The examples in top row are the successful cases in our tasks and failure cases are shown in bottom row. While we can obtain satisfactory results in the scene recognition, several failure cases are shown in the action recognition. We observe that most failure cases are caused by the background clutter.

Loss	Action@1	Action@5
$\mathcal{L}_{\text{cls}}$	52.19	80.72
$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{cent}}$	**60.10**	**85.46**

Table 4. Action recognition performance comparison of different loss functions on validation set of CoVieW'19 dataset. Accuracies are measured using top1 and top5 predictions at clip level.

Loss	Scene@1	Scene@5
$\mathcal{L}_{\text{cls}}$	54.07	81.25
$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{cent}}$	**63.31**	**86.94**

Table 5. Scene recognition performance comparison of different loss functions on validation set of CoVieW'19 dataset. Accuracies are measured using top1 and top5 predictions at clip level.

Table 1 shows the results on three validation set of CoVieW'19 dataset for video summarization. Three validation set were randomly selected for each training phase, and the performance is measured by the metric in Sec. 4.2. Our network provides score of 0.8944 in the video summarization results. The results on test set of CoVieW'19 dataset are shown in Table 2. The video summarization and action scene recognition performances are evaluated and ranked separately. For the video summarization, our model shows the score of 0.8409, which placed 2nd among participants. Also, action scene recognition performance evaluated using (7) shows the score of 0.7294, which ranked 4th.

The effectiveness of multi-modalities. As mentioned in Sec. 3.2, we use both visual and audio sequences for detecting action and scene recognition. Table 3 shows the effectiveness of a single visual modality and the combination of visual and audio modalities. Comparing the performance of the visual modality only, the combination of two modalities provides 2.7% improvement at results of top5 action classification and 1.8% improvement at results of top5 scene classification, which show the importance to consider both modalities for action and scene classification.

The effectiveness of loss functions. Our premise is that discriminative features corresponding to classes can boost the performance of recognition tasks. When we learn networks for action and scene classification, two loss terms are employed, *i.e.*, the classification loss and the center loss. Table 4 and Table 5 summarize the results of action recognition and scene recognition according to the different loss functions. All accuracies are measured using classification accuracy at top1 and top5. The first row of each table shows the result with only classification loss, and the second row represents the result with the center loss. As our baseline, the loss function without the center loss provides the 52.19% accuracy. We observe that the performance is considerably enhanced by using the center loss.

5. Conclusion

We presented a novel deep architecture for comprehensive video understanding that performed video summarization, and action and scene recognition tasks. The classification and summarization are performed by attention weighted features, where two types of attention inference module are proposed to refine features. Furthermore, we proposed a novel loss function, called center loss, to minimize intra-class variations and to maximize inter-class variations to learn discriminative feature representations. We hope that the results of this study will be able to further advances in comprehensive video understanding area.

6. Acknowledgement

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science and ICT (NRF-2017M3C4A7069370).

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv*, 2016.

[2] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. *In: NIPS*, 2016.

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *In: CVPR*, 2017.

[4] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. *In: CVPR*, 2015.

[5] B. Gong, W.-L. Chao, K. Grauman, and F. Sha. Diverse sequential subset selection for supervised video summarization. *In: NeurIPS*, 2014.

[6] M. Gygli, H. Grabner, and L. V. Gool. Video summarization by learning submodular mixtures of objectives. *In: CVPR*, 2015.

[7] M. Gygli, H. Grabner, H. Reimenschneider, and L. V. Gool. Creating summaries from user videos. *In: ECCV*, 2014.

[8] F. C. Heilbron, W. Barrios, V. Escorcia, and B. Ghanem. Scc: Semantic context cascade for efficient action detection. *In: CVPR*, 2017.

[9] L. Herranz, S. Jiang, and X. Li. Scene recognition with cnns: objects, scales and dataset bias. *In: CVPR*, 2016.

[10] 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. *IEEE Trans. on PAMI*, 40(2):352–364, 2017.

[11] N. Joshi, W. Kienzle, M. Toelle, M. Uyttendaele, and M. F. Cohen. Real-time hyperlapse creation via optimal frame selection. *ACM Trans. on Graph.*, 2015.

[12] A. Kanehira, L. V. Gool, Y. Ushiku, and T. Harada. Viewpoint-aware video summarization. *In: CVPR*, 2018.

[13] H.-W. Kang, Y. Matsushita, X. Tang, and X.-Q. Chen. Space-time video montage. *In: CVPR*, 2006.

[14] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hiller, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset, 2017.

[15] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *arXiv*, 2017.

[16] J. Kopf, M. F. Cohen, and R. Szeliski. First-person hyperlapse videos. *ACM Trans. on Graph.*, 2014.

[17] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. *In: ICCV*, 2017.

[18] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. *In: CVPR*, 2012.

[19] Y. Li, L. Wang, T. Yang, and B. Gong. How local is the local diversity? reinforcing sequential determinantal point processes with dynamic ground sets for supervised video summarization. *In: ECCV*, 2018.

[20] B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial lstm networks. *In: CVPR*, 2017.

[21] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *In: CVPR*, 2009.

[22] J. Park, S. Jeon, S. Kim, J. Lee, S. Kim, and K. Sohn. Learning to detect, associate, and recognize human actions and surrounding scenes in untrimmed videos. *In: ACMMM Workshop*, 2018.

[23] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon. Bam: Bottleneck attention module. *In: BMVC*, 2018.

[24] Paszke, Adam, Gross, Sam, Chintala, Soumith, Chanan, Gregory, Yang, Edward, DeVito, Zachary, Lin, Zeming, Desmaison, Alban, Antiga, Luca, Lerer, and Adam. Automatic differentiation in pytorch. *In: NIPS Workshop*, 2017.

[25] K. J. Piczak. Esc: Dataset for environmental sound classification.

[26] Y. Poleg, T. Halperin, C. Arora, and S. Peleg. Egosampling: Fast-forward and stereo for egocentric videos. *In: CVPR*, 2015.

[27] M. Rochan, L. Ye, and Y. Wang. Video summarization using fully convolutional sequence networks. *In: ECCV*, 2018.

[28] A. Sharghi, J. S. Laurel, and B. Gong. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. *In: CVPR*, 2017.

[29] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *In: NIPS*, 2014.

[30] K. Sohn, J. Lim, M. H. Yang, J. Hsu, H. Byun, S. Lin, E. Kim, and S. Kim. The 1st workshop and challenge on comprehensive video understanding in the wild. *In: ACMMM Workshop*, 2018.

[31] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. *In: CVPR*, 2015.

[32] M. Sun, A. Farhadi, B. Taskar, and S. Seitz. Salient montages from unconstrained videos. *In: ECCV*, 2014.

[33] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *In: ICCV*, 2015.

[34] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. *In: CVPR*, 2018.

[35] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. *In: ECCV*, 2016.

[36] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon. Cbam: Convolutional block attention module. *In: ECCV*, 2018.

[37] H. Yang, B. Wang, S. Lin, D. Wipf, and M. Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. *In: ICCV*, 2015.

[38] T. Yao, T. Mei, and Y. Rui. Highlight detection with pairwise deep ranking for first-person video summarization. *In: CVPR*, 2016.

[39] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. *In: CVPR*, 2015.

[40] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. *In: ECCV*, 2016.

[41] B. Zhao, X. Li, and X. Lu. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. *In: CVPR*, 2018.

[42] B. Zhou, A. Lapedriza, A. Khosal, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. on PAMI*, 2017.