Temporal U-Nets for Video Summarization with Scene and Action Recognition

Heeseung Kwon *
Dept. of Creative IT Engineering
POSTECH
Pohang, Korea
aruno@postech.ac.kr

Woohyun Shim *
Dept. of Creative IT Engineering
POSTECH
Pohang, Korea
wh.shim@postech.ac.kr

Minsu Cho
Dept. of Computer Science and Engineering
POSTECH
Pohang, Korea
mscho@postech.ac.kr

Abstract

While videos contain long-term temporal information with diverse contents, existing approaches to video understanding usually focus on a short trimmed video clip with a specific content such as a particular action or object. For comprehensive understanding of untrimmed videos, we address an integrated video task of video summarization with scene and action recognition. We propose a novel convolutional neural network architecture for handling untrimmed videos with multiple contents. The proposed architecture is an encoder-decoder structure where the encoder captures long-term temporal dynamics from an entire video and the decoder predicts detailed temporal information of multiple contents of the video. Two-stream processing is adopted for obtaining feature representations, one for focusing on the spatial information and the other for the temporal information. We evaluate the proposed method on the benchmark of the Challenge on Comprehensive Video Understanding in the Wild (CoVieW 2019), and the experimental results demonstrate that our method achieves outstanding performance.

1. Introduction

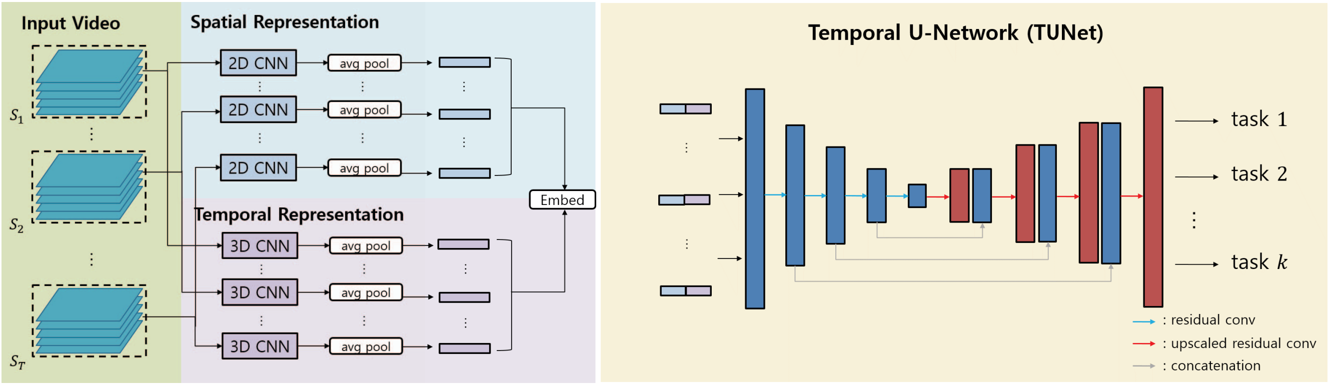
Video understanding is a rapidly growing research area of computer vision because of its wide availability for numerous practical applications. According to the advent of the deep learning era, research on the video domain have been progressed with successful deep learning methods. However, existing video understanding approaches usually consider a short video clip with a highly specific task of the video domain such as categorizing human actions or tracking particular objects. These kinds of approaches are not appropriate for understanding videos deeply because the contents of videos have strong relationships to each other and the relationships change with time. To understand video data in depth, the inherent contents need to be jointly analyzed in the dynamic scenes.

To overcome the issue that we mentioned above, this paper deals with the comprehensive task named video summarization with temporal scene and action recognition in untrimmed videos, which is the task of CoVieW 2019. The goal of the task is constructing a short video clip (30 seconds) from an untrimmed video (5 to 10 minutes) using given importance scores of video frames. At the same time, human actions and scenes are jointly classified to a set of predefined classes according to the content of the summarized video clip. For the task, we propose a new Convolutional Neural Network (CNN) architecture, which is designed to handle untrimmed videos and multiple tasks effectively. The proposed architecture consists of the encoder-decoder structure with residual modules and receives a whole untrimmed video as input. The architecture encodes long-term temporal dynamics through temporal convolution layers and outputs for every segment the multiple contents through temporal up-scaling at the decoder phase. Additionally, to enable the network to learn more effectively, we design the task-specific losses for each content and utilize the two distinctive models to obtain abundant feature representations from the input video frames.

2. Proposed Approach

This section presents the main methodology that we proposed, called Temporal U-Network (TUNet). We first introduce the feature extraction pipeline to construct the input for TUNet, then describe the structure of TUNet and how it works.

[*] indicates equal contribution.



(a) Feature Representation

(b) Temporal U-Network

Figure 1: The illustration of (a) shows the process of generating feature representations. Input video is divided into T segments, which of each consists of multiple video frames, and they become the spatial and the temporal representations obtained through the pre-trained 2D & 3D CNNs. The illustration of (b) shows the proposed architecture, TUNet. TUNet is an encoder-decoder style structure, and the architecture captures temporal dynamics through residual convolutions over temporal axis at the encoder phase, and conducts segment-wise predictions with upscaled residual convolutions over temporal axis at the decoder phase.

2.1. Feature Representation

To provide abundant feature representations to our model, we adopt the two-stream setting that consists of a spatial stream network and a temporal stream network. We exploit a ResNet-50 [4] trained on ImageNet [3] for the spatial stream network and TSM ResNet-50 [7] trained on Kinetics [2] for the temporal stream network. Figure 1a illustrates the procedure of generating feature representations. Input video is divided into N video segments with a same duration, and each video segment is converted to a spatial stream representation and a temporal stream representation after passing through the two networks and average pooling. Finally, the representations of both streams are fused to have a single representation by an embedding layer which consists of two 1 by 1 convolution layers and a concatenation layer.

2.2. Temporal U-Network

Now, we describe the TUNet architecture, and the overall architecture is illustrated in Figure 1b. TUNet is an encoder-decoder style architecture which is inspired from [8] and receives an entire video as network input. We expect that the architecture is effective for long-term untrimmed videos because the architecture can capture long-term temporal dynamics at the encoder phase while it can predict detailed temporal information at the decoder phase. In addition, the proposed architecture consists of fully convolutional networks, so the architecture is not restricted to the video length and it could be linked to diverse tasks easily.

We employ residual convolution modules from [4] and modify it to 1-d temporal version and use temporal max pooling operations to encode the interaction over video segments. The residual modules and max pooling operations are stacked in 4-level to cover large temporal receptive fields. After the encoding phase, the latent features are passed through the up-scaled residual convolution modules, which of each is composed of up-scaling bi-linear interpolation and the residual convolution module. We also concatenate the encoded feature at the earlier layer to up-scaled feature over the channel axis as common practice in encoder-decoder architecture to incorporate low-level semantics. To make the predictions on each video segment, the features are up-scaled at the decoder phase until those have the same length as the given input. Finally, the last feature representation is connected to each task with a 1 by 1 convolution layer to learn the task-specific representation.

Since we focus on the comprehensive task, video summarization with scene and action recognition, we set the loss which is composed of multiple task-specific losses as

$$\mathcal{L} = \mathcal{L}_{scene} + \mathcal{L}_{action} + \mathcal{L}_{im-score}. \qquad (1)$$

L_{scene} and L_{action} denote losses for scene and action recognition, respectively. We use segment-wise cross-entropy (CE) loss for classifying scenes and actions, which are simple extensions of naive CE losses along the temporal axis. $L_{im-score}$ denote a loss for video summarization. We adopt mean squared error (MSE) between the estimated importance scores and ground truth importance scores. Additionally, we found the CE loss also could be an another option for $L_{im-score}$ by quantizing the importance score. How these losses affect models will be discussed in experimental section.

3. Experiment

3.1. Dataset

We conduct experiments on CoVieW 2019 challenge dataset. The dataset consists of 1,500 untrimmed videos sampled from multiple video datasets [1, 6, 9], and the average duration of the videos is approximately 5 minutes. The distribution between train set and test set is 1,200 and 300, respectively. Since the label of test set is unknown, we randomly split the 1,200 training videos to 1,000 train set and 200 validation set, and we evaluate our models on the validation set. Each video is divided into a set of 5-second long segments and each segment is annotated with a scene, an action and an importance score. The number of scene and action categories are 78 and 99, and the importance score is distributed from 0 to 2 at the interval of 0.1. For the evaluation, we follow the proposed evaluation metric of CoVieW 2019.

3.2. Implementation Details

Training: The procedure of video pre-processing is similar to that of [1]. Technically, we convert each video to a set of frames at 1 frame-per-second. To generate the feature representations, frames are center cropped and fed into the two-stream networks, and the last convolutional features are extracted from both streams. Extracted feature representations are L2 normalized and embedded to 1,024-dimensional vectors with a fully connected (FC) layer. Lastly, feature representations of both streams are concatenated. For training TUNet, we use the minibatch SGD with Nestrov momentum for updating parameters, and batch size is set to 32. The initial learning rate is set to 0.01, and reduces its 1/10 after 10 epochs. The maximum epoch number is set to 15.

Baseline model: We design a simple baseline to verify the effectiveness of our architecture. The baseline is composed of one FC layer and multiple classifiers. A dropout layer is utilized after the FC layer and the loss function is same as the proposed architecture. Similar to TUNet, we use the minibatch SGD with Nestrov momentum for training, and batch size is set to 32. Since the baseline is much lighter than the proposed architecture, the initial learning rate is set to 0.01, and reduces its 1/10 after 10 epochs. The maximum epoch number is set to 15.

3.3. Experimental Results

We first investigate the effectiveness of combining the spatial and temporal representation. Each representation has its own characteristic in performing different tasks; Spatial representation was significant for scene recognition and temporal representation seems to provide useful cues in predicting the action categories (Table 1). Moreover, the performance is increased when both features are utilized together validating that those features are complementary to each other. Hence, we used the both features for all the experiments afterwards, unless explicitly stated otherwise.

Table 1: The result of the settings combined with two-stream representations. The scores are computed according to the metric stated in CoVieW 2019.

Model	Scene	Action	Summarization
Spatial	56.92 / 83.08	50.50 / 79.92	77.71
Temporal	56.00 / 80.42	56.50 / 85.08	78.02
Two-stream	56.83 / 84.58	56.67 / 85.58	77.09

Now, we compare the baseline architecture with our proposed model, TUNet. TUNet has several advantages over the baseline model: 1) capturing long-term temporal dynamics, 2) rich semantics from the hierarchical encoding and decoding and 3) information flows from low-level features through skip-connections. With all these forementioned components, the improvement on Top-1 accuracy at scene and action recognition are 4.09% and 4.16% point, while the performance on summarization task has not improved much. (Table 2). This is because MSE loss itself does not significantly improve the performance of regressing the importance score, as it will be seen later.

Table 2: The results of the baseline model and TUNet.

Model	Scene	Action	Summarization
Baseline	56.83 / 84.58	56.67 / 85.58	77.09
TUNet	60.92 / 83.08	60.83 / 86.67	77.34

Table 3: Comparative study of different loss function for importance score regression. CI denotes the class interval.

Model	Scene	Action	Summarization
CI: 0.1	61.08 / 83.58	53.42 / 83.92	79.82
CI: 0.3	59.42 / 83.67	51.83 / 85.33	79.41
CI: 0.7	59.75 / 83.42	52.25 / 85.50	79.65
MSE	60.92 / 83.08	60.83 / 86.67	77.34

At last, we analyze the use of CE loss for the importance score regression. Since the importance score ranges from 0 to 2, this can be partitioned into sub-classes with a certain interval. We split the classes with the interval of 0.1, 0.3 and 0.7 resulting in 21, 7 and 3 classes respectively. Then, we compare those with the model trained using MSE loss (Table 3). For scene and action recognition, the model with MSE loss attains similar or better accuracy, but it demonstrates worse scores on the summarization task. The prior work for the video summarization [5] also shows the models with MSE loss perform worse than the models with CE loss, indicating that CE loss is stronger than MSE loss for making the model to focus on the corresponding task. Among the variants of CE loss, we have not found the any significant difference with regard to the number of classes.

Unfortunately, CE loss for the video summarization task degrades the performance of other tasks, especially action recognition. However, if we measure the performance on all segments in the validation set, the average accuracy for CE loss model having 21 classes are 58% and 57% for scene and action recognition, respectively. These scores are almost identical to those of the MSE model; This implies that the action recognition task are less relevant to summarization task. To fully utilize the benefits of each task, loss function should be applied adaptively according to the task relationships, but we leave it to the future work.

4. Conclusion

We have presented a novel video understanding approach that addresses the comprehensive task, video summarization with scene and action recognition in untrimmed videos. We propose a new CNN architecture named TUNet to handle untrimmed videos and multiple tasks together. TUNet captures long-term temporal dynamics of untrimmed videos through temporal convolutions and obtains detailed temporal information of multiple contents by segment-wise predictions. The proposed architecture reports great performance on the challenge dataset and demonstrates the effectiveness of the TUNet architecture. We expect that our approach could be extended to other comprehensive video tasks in the future.

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 preprint arXiv:1609.08675*, 2016.

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] H.-I. Ho, W.-C. Chiu, and Y.-C. Frank Wang. Summarizing first-person videos from third persons' points of view. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–85, 2018.

[6] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[7] J. Lin, C. Gan, and S. Han. Temporal shift module for efficient video understanding. *arXiv preprint arXiv:1811.08383*, 2018.

[8] M. Rochan, L. Ye, and Y. Wang. Video summarization using fully convolutional sequence networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 347–363, 2018.

[9] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015.