17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Video Summarization With Frame Index Vision Transformer

Tzu-Chun Hsu, Yi-Sheng Liao, Chun-Rong Huang
Department of Computer Science and Engineering, National Chung Hsing University
Taichung, Taiwan, 402
{g109056029, g107056049, crhuang}@nchu.edu.tw

Abstract

In this paper, we propose a novel frame index vision transformer for video summarization. Given training frames, we linearly project the content of the frames to obtain frame embedding. By incorporating the frame embedding with the index embedding and class embedding, the proposed frame index vision transformer can be efficiently and effectively applied to learn the importance of the input frames. As shown in the experimental results, the proposed method outperforms the state-of-the-art deep learning methods including recurrent neural network (RNN) and convolutional neural network (CNN) based methods in both of the SumMe and TVSum datasets. In addition, our method can achieve real-time computational efficiency during testing.

1 Introduction

With the increasing number of videos in the Internet, efficiently browsing videos becomes one of the most important issues in the multimedia domain. To reduce the browsing time, video summarization [1, 2, 3] is proposed to extract important video content and generates a compact summary for efficient browsing. It can also be applied to video saliency analysis [4, 5] and video surveillance [6, 7].

Video summarization methods can be mainly divided to unsupervised methods [8, 9] and supervised methods [10, 11]. Although unsupervised methods can extract keyframes based on similarity and representative properties of frames, they are hard to grab the important semantic concept which is meaningful for humans. In addition, human-created summaries are hard to be learned by unsupervised methods. To solve the aforementioned problems, supervised methods are proposed which can learn the frame importance from human-created summaries. Nevertheless, the summaries labeled by different people will be variant because of individual differences. How to explicitly learn the frame importance from human-created summaries becomes a novel issue in video summarization.

Recently, deep learning based methods especially recurrent neural network (RNN) based methods [11, 12] have been shown their effectiveness for supervised video summarization. RNN based methods formulate the video summarization problem as a sequence-to-sequence problem. Thus, these methods can well learn the relationship between continuous frames. Most RNN based methods consider continuous neighbor frames when learning the importance of the human-created summaries. Nevertheless, the important content of the video may not continuously appear due to the shot changes [13] of the video content.

To address aforementioned issues and effectively represent the frame information for video summarization, we propose a novel frame index vision transformer (FIVT) which is inspired by the vision transformer [14]. Different from [14] which only considers split words of a single image, we consider the continuous frames as the words to describe the importance of the video for video summarization. To apply the training frames to the transformer encoders, we linearly project the continuous frames to obtain the frame embedding. Besides the frame embedding, we also add the index embedding to maintain the temporal information of the continuous frames. Finally, the class embedding is added to indicate if the continuous frames belong to important video content or not for video summarization. By consisting of these three embeddings, embedded frames are constructed and serve as the input of the transformer encoder.

Because the transformer encoder has been shown its performance in learning long-term relationship [15], we use the transformer encoder to learn the important video content based on the embedded frames. Each transformer encoder contains a self-attention module which aims to learn representative features of the video content, and a fully connected module which feeds forward the features to the next layer. We have evaluated the proposed method in two popular video summarization datasets, SumMe [16] and TVSum [3]. As shown in the experiments, the proposed method can achieve better results compared with the state-of-the-art methods. The paper is organized as follows. Sec. 2 gives the related work. The proposed frame index vision transformer is presented in Sec. 3. Experimental results and comparisons are shown in Sec. 4. Finally, the conclusions are given in Sec. 5.

2 Related Work

Unsupervised video summarization methods can be mainly divided into clustering based methods [8, 17] and dictionary based methods [9, 18]. Although un-

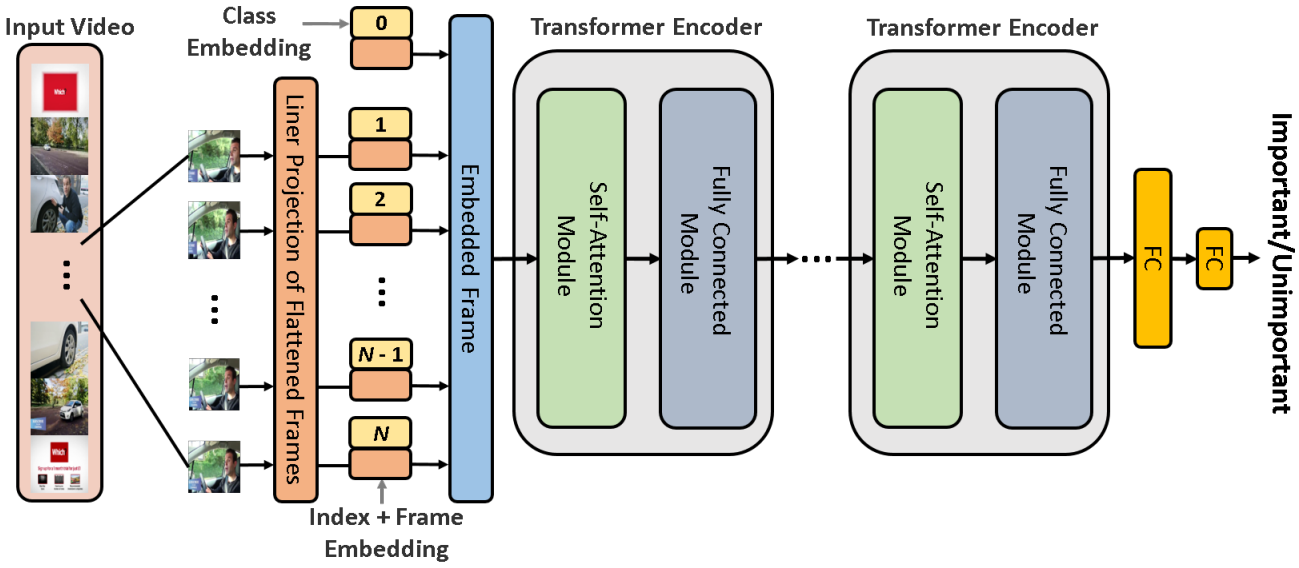


Figure 1. The proposed frame index vision transformer. The frame embedding, index embedding and class embedding are used to generate the embedded frame for the transformer encoder.

supervised methods can extract keyframes from target videos, the results may not fit the semantic concept of human beings.

To solve this problem, supervised methods have been proposed to learn what is the important content of the video based on human-created summaries. For example, Gong et al. [10] propose the sequential determinantal point process (seqDPP) to model the important video content for video summarization. In [19], sequential and hierarchical DPP is further proposed to improve the performance. Potapov et al. [20] perform temporal segmentation to obtain semantically-consistent segments and apply the support vector machine [21] to assign importance scores to each segment. Gygli et al. [22] consider jointly optimization of multiple objective functions to obtain video summarization.

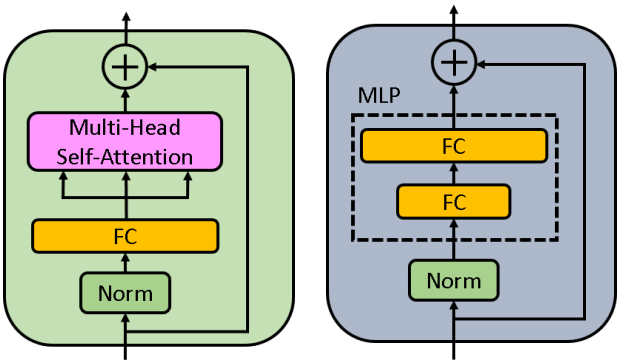
More recently, deep learning based video summarization methods are proposed. Zhang et al. [11] propose using long short-term memory (LSTM) of RNN to solve the video summarization problem. Mahasseni et al. [23] propose an adversarial LSTM network where the discriminator LSTM aims to distinguish the original video and its reconstruction from the LSTM summarizer. Both supervised and unsupervised versions are provided in [23]. Zhao et al. [24] propose a hierarchical structure-adaptive RNN (HAS-RNN) to solve the shot segmentation and video summarization problems. Huang et al. [25] sequentially combine 2-D CNNs, 1-D CNNs, and LSTM to learn the frame-level importance scores. To select semantic keyshots, Ji et al. [12] propose an attentive encoder–decoder network which contains bidirectional LSTM encoders and attention based LSTM decoders. Zhu et al. [26] propose a detect-to-summarize network (DSNet) to obtain temporal interest proposals and directly predict the importance of video segments. Instead of imposing RNN, Rochan et al. [27] propose a fully convolutional sequence network (FCSN) to model the complex dependency among input frames. Fajtl et al. [28] propose a self-attention based sequence to sequence network which consists of the attention network and the regressor network for video summarization. Compared with deep learning based methods, the proposed frame index vision transformer can learn the importance from frames based on the transformer encoders and achieve comparable results.

3 Method

The overview of the proposed frame index vision transformer is shown in Figure 1. The input frames are linearly projected to flattened frames as the frame embedding, and then combined with the index embedding and the class embedding to generate the embedded frame for the input of the transformer encoder. Each transformer encoder contains a self-attention module and a fully connected module to learn representative features. Finally, two fully connected layers are implemented to generate video summarization results.

3.1 Frame Index Vision Transformer

The proposed frame index vision transformer is inspired by the vision transformer [14]. In the vision



(a) (b)

Figure 2. (a) The self-attention module and (b) The fully connected module. The black dot rectangle represents the MLP.

transformer, each image is split into a sequence of 2-D image patches. These patches provide the words of the image and serve as the input of the vision transformer. The vision transformer is then used to learn the correlations between the image patches and the target class. Different from the vision transformer, we would like to learn if the frames contain important information for video summarization. Given a training video of F frames, we sequentially and equally partition the video to subsets for training. Each subset contains N continuous frames which are denoted as $f = \{f_1, f_2, ...f_N\}$ and the classification label of the subset is given by the human-created ground truth. f in each subset serves as the input of the frame index vision transformer and we impose the indices of the frames for the temporal representation.

Let the dimension of the frame be $\mathcal{R}^{H\times W\times C}$, where W, H and C are the width, the height and the number of channels of the frame. Similar to previous deep learning approaches, the resolution of all of the frames is resized to 224×224 and the number of channels is 3. We generate a sequence $f \in \mathcal{R}^{N\times M}$, where $N = 9$ is the number of input frames and $M = H\times W\times C$. For each sequence, a learnable frame embedding is prepended to indicate if the input sequence contains important scenes. The index embedding is added to the frame embedding to present temporal information for the input frames. The frame embedding, index embedding and class embedding of the training frames are combined to an embedded frame which serves as the input of the transformer encoder. In this way, it is not necessary to modify the structure of the transformer encoder.

The transformer encoder is composed of a self-attention module, and a fully connected module. The initial input (the embedded frame) of the transformer encoder is defined as follows:

$$\boldsymbol{z}_0 = [x_{class}; x_f^1\boldsymbol{E}; x_f^2\boldsymbol{E}; ...; x_f^N\boldsymbol{E}] + \boldsymbol{E}_{idx}, \quad (1)$$

where x_{class} is the class embedding, $\boldsymbol{E}$ is the embedding projection of all frames, x_f^n is the nth frame, $\boldsymbol{E}_{idx}$ is the index embedding, and the dimensions of $\boldsymbol{E}$ and $\boldsymbol{E}_{idx}$ are $\mathcal{R}^{(W\cdot H\cdot C)\times D}$ and $\mathcal{R}^{(N+1)\times D}$, respectively. $D = 768$ is the dimension of the linear projection suggested in [14]. Layer normalization (LN) is applied to each module for feature normalization and a residual connection [29] is applied to connect in each module.

The self-attention module contains the multi-head attention [15] and is shown in Figure 2(a). Here, the fully connected layer is used to expand the normalized features and generate new representation subspaces. The output of the multi-head attention is based on weighted sums of the values of different representation subspaces at different positions. Due to limited space, please refer to [15] for details. The weights of the values are computed from the pairwise similarity of their corresponding respective query and key representations.

Given the input vector $\boldsymbol{z}_0$ of the first transformer encoder, assume that the number of the self-attention operation is k. The multi-head self-attention of the ℓth self-attention module of the transformer encoder is defined as follows:

$$MSA(\hat{\boldsymbol{z}}_{\ell-1}) = [SA_1(\hat{\boldsymbol{z}}_{\ell-1}), ..., SA_k(\hat{\boldsymbol{z}}_{\ell-1})]\boldsymbol{U}_{msa}, \quad (2)$$

where $\boldsymbol{z}_{\ell-1}$ is the output of the $(\ell-1)$th transformer encoder, $\hat{\boldsymbol{z}}_{\ell-1} = LN(\boldsymbol{z}_{\ell-1})$ is the layer normalization result of $\boldsymbol{z}_{\ell-1}$, $LN(\cdot)$ is the layer normalization function, SA_k is the self-attention [15] and $\boldsymbol{U}_{msa} \in \mathcal{R}^{k\cdot D_h\times D}$. As suggested in [14], we set $k = 12$ and $D_h = D/k$. The output $\boldsymbol{z}'_\ell$ of the ℓth self-attention module is composed of the multi-head attention and a residual connection of the input as follows:

$$\boldsymbol{z}'_\ell = MSA(\hat{\boldsymbol{z}}_{\ell-1}) + \boldsymbol{z}_{\ell-1}. \quad (3)$$

The fully connected module is applied to the output of the self-attention module. This module consists of a LN and a multilayer perceptron (MLP). The MLP contains two fully connected layers as shown in Figure 2(b) and is represented by the function $MLP(\cdot)$. The output of the ℓth fully connected module is defined as

$$\boldsymbol{z}_\ell = MLP(\hat{\boldsymbol{z}}'_\ell) + \boldsymbol{z}'_\ell, \quad (4)$$

where $\hat{\boldsymbol{z}}'_\ell = LN(\boldsymbol{z}'_\ell)$, and $\boldsymbol{z}'_\ell$ is the residual connection of the output of the self-attention module.

Finally, two fully connected layers are implemented after the last transformer encoder, and a soft-max function is applied after the final fully connected layer for the importance classification. In our implementation, the number of the transformer encoders is 12. The sizes of the last two fully connected layers are 768 and 2, respectively.

4 Experimental Results

In the experiments, two popular video summarization datasets, SumMe [16] and TVSum [3], were used

Table 1. Ablation Study in F-Score (%)

Methods	SumMe	TVSum
w/o Warmup and Index	43.5	54.1
w/o Index	45.2	58.3
w/o Warmup	48.1	61.9
Proposed	**49.0**	**62.3**

Table 2. Comparisons with State-of-the-Art Methods in F-Score (%)

Methods	SumMe	TVSum
dppLSTM [11]	38.6	54.7
SUM-GAN$_{sup}$ [23]	41.7	56.3
HAS-RNN [24]	44.1	59.8
FCSN [27]	47.5	56.8
TS-STN [25]	46.1	60.0
M-AVS [12]	44.4	61.0
Proposed	**49.0**	**62.3**

for evaluation. The SumMe dataset contains 25 videos of user events with manually labeled summaries. The TVSum dataset contains 50 videos with shot-level scores and 10 video categories. We followed the experimental settings in [11] for the evaluation in both datasets and computed the F-Score values of the proposed method for comparisons.

During training, we used a warm up strategy which applied the learning rate from 0.05 to 0.5 in the first 10 epochs and the cosine learning rate decay [30] in the following epochs. The batch size is 30 and the parameters of the FIVT is updated by using SGD with a momentum of 0.9. We performed our experiments on an Intel i7 computer with Nvidia GTX 2080Ti GPU and implemented our method by using PyTorch.

4.1 Ablation Study

In our approach, we add the index embedding to help the training of the transformer encoder. Moreover, we consider the warm up strategy to update the parameters of the proposed FIVT. To address the effectiveness of these two factors, ablation study is performed as shown in Table 1. As we expected, the proposed method without the warm up strategy and the index embedding (w/o Warmup and Index) achieves the worst results. Compared with the proposed method without the index embedding (w/o Index), the proposed method without the warm up strategy (w/o Warmup) has better results. These results imply that the importance of applying the index embedding in the proposed method for video summarization. Nevertheless, the proposed method with both factors achieves the best results in both datasets.

4.2 Quantitative Results

We compared the proposed method with several state-of-the-art video summarization methods including dppLSTM [11], SUM-GAN$_{sup}$ [23], HAS-RNN [24], FCSN [27], TS-STN [25], and M-AVS [12]. As shown in Table 2, the proposed method outperforms these state-of-the-art methods in both of the SumMe and TVSum datasets. Compared with deep learning based methods which use frame pixels as the inputs of the networks, the embedded frames provide more representative information by simultaneously consisting of the frame embedding, index embedding and class embedding for the transformer encoder learning. By using the embedded frames, our frame index vision transformer can effectively learn the correlations between the frames and human-created summaries. Thus, it can achieve better F-Score values compared with deep learning based methods.

During testing, the average frames per second (fps) of the proposed method are about 58 and 56 for the SumMe and TVSum datasets, respectively. The computational efficiency of the proposed method relies on the the design of the embedded frames and the transformer encoder. As a result, the proposed method can be applied for real-time content analysis.

5 Conclusions

In this paper, we propose a frame index vision transformer to achieve video summarization. Different from RNN based methods, the proposed method treats the frames as a sequence of words and input the frames to the transformer encoder. Because frames contain continuous information of the important content of the videos, transformer encoder can learn representative features for the video summarization. In addition, by considering the index embedding in the embedded frames, the proposed method can achieve better results compared with the state-of-the-art methods. In the future, we will focus on the challenge of improving self-supervised pre-training to further improve the performance in learning the important content of the training videos.

Acknowledgments

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST109-2221-E-005-063, MOST109-2314-B-006-024 and MOST110-2327-B-006-006.

References

[1] M. Yu-Fei, L. Lie, Z. Hong-Jiang, and L. Mingjing, "A user attention model for video summarization," in *Proc. ACM Conf. Multimedia*, 2002, pp. 533–542.

[2] J. Almeida, N. J. Leite, and R. da S. Torres, "Vison: Video summarization for online applications," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 397–409, 2012.

[3] Y. Song, J.Vallmitjana, A.Stent, and A.Jaimes, "Tvsum: Summarizing web videos using titles," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015, pp. 5179–5187.

[4] R. Cong, J. Lei, H. Fu, M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 2941–2959, 2019.

[5] C.-R. Huang, Y.-J. Chang, Z.-X. Yang, and Y.-Y. Lin, "Video saliency map detection by dominant camera motion removal," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1336–1349, 2014.

[6] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1971–1984, 2008.

[7] C.-R. Huang, P.-C. J. Chung, D.-K. Yang, H.-C. Chen, and G.-J. Huang, "Maximum a posteriori probability estimation for online surveillance video synopsis," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1417–1429, 2014.

[8] A. Aner and J. R. Kender, "Video summaries through mosaic-based shot and scene clustering," in *Proc. Euro. Conf. Computer Vision*, 2002, pp. 388–402.

[9] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1600–1607.

[10] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *Advances in Neural Information Processing Systems*, vol. 27, 2014, pp. 1–9.

[11] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Euro. Conf. Computer Vision*, 2016, pp. 766–782.

[12] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder-decoder networks," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1709–1717, 2020.

[13] C.-R. Huang, H.-P. Lee, and C.-S. Chen, "Shot change detection via local keypoint matching," *IEEE Trans. on Multimedia*, vol. 10, no. 6, pp. 1097–1108, 2008.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Intl. Conf. Learning Representations*, 2021, pp. 1–21.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 1–11.

[16] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating summaries from user videos," in *Proc. Euro. Conf. Computer Vision*, 2014, pp. 505–520.

[17] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Automatic video summarization by graph modeling," in *Proc. Int'l Conf. Computer Vision*, vol. 1, 2003, pp. 104–109.

[18] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Trans. on Multimedia*, vol. 14, no. 1, pp. 66–75, 2012.

[19] A. Sharghi, B. Gong, and M. Shah, "Query-focused extractive video summarization," in *Proc. Euro. Conf. Computer Vision*, 2016, pp. 3–19.

[20] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. Euro. Conf. Computer Vision*, 2014, pp. 540–555.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, p. 273–297, 1995.

[22] M. Gygli, H. Grabner, and L. V. Gool, "Video summarization by learning submodular mixtures of objectives," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3090–3098.

[23] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2017, pp. 2982–2991.

[24] B. Zhao, X. Li, and X. Lu, "HSA-RNN: Hierarchical structure-adaptive rnn for video summarization," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2018, pp. 7405–7414.

[25] S. Huang, X. Li, Z. Zhang, F. Wu, and J. Han, "User-ranking video summarization with multi-stage spatio-temporal representation," *IEEE Trans. on Image Processing*, vol. 28, no. 6, pp. 2654–2664, 2019.

[26] W. Zhu, J. Lu, J. Li, and J. Zhou, "Dsnet: A flexible detect-to-summarize network for video summarization," *IEEE Trans. on Image Processing*, vol. 30, pp. 948–962, 2021.

[27] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proc. Euro. Conf. Computer Vision*, 2018, pp. 358–374.

[28] J. Fajtl, H. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *Proc. Asian Conf. on Computer Vision Workshops*, 2018, pp. 39–54.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[30] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *Proc. Int'l Conf. Learning Representations*, 2017, pp. 1–16.