UEC at TRECVID 2021: ActEV and VTT

Sosuke Mizuno Yang Jing Keiji Yanai
The University of Electro-Communications, Tokyo , Japan
{mizuno-s, yang-j, yanai}@mm.inf.uec.ac.jp

***Abstract*—In this paper, we report our systems and the evaluation results at TRECVID 2021. This year we participated two tasks, Activity in Extended Video (ActEV) and Video to Text (VTT).**

I. ActEV: Activity in Extended Video

ActEV is a very challenging task because it requires precise spatial and temporal localization. Our approach consists of three parts, proposal generation, action classification, and post-processing. Figure 1 shows the overall of our approach.

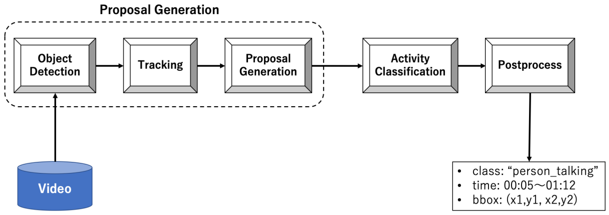


Fig. 1. The overall of our approach of the ActEV task.

A. Proposal generation

Here, we extract candidate areas for action from the input video. First, we use Faster R-CNN [1] to detect humans and cars from the input frame. We utilize Faster R-CNN with feature pyramid network [2] on ResNet-101. The model trained on the COCO dataset was fine-tuned using the VIRAT dataset. Next, we use deep SORT [3] to generate a tracking trail for each object. Finally, we generate event proposals from the trajectory of a single object, a person and a car. An event proposal can be treated as a row of bounding boxes cut out of each frame. In this study, we classify each of the proposals into one of three categories, Person, Vehicle and Person-Vehicle. "Person" category includes only events that occurred in a single person. "Vehicle" category includes only events occurring in a single vehicle. "Person-Vehicle" category proposes events in relation to a human and a vehicle. If the spatial distance between the human trajectory and the vehicle trajectory is less than the pre-defined threshold, a bounding box containing a human and a vehicle is proposed.

B. Activity Classification

1) Feature Extract: We extracted features for action classification in a 3D-ResNet [4] model. We used a 3D ResNet-101 model pre-trained with Kinetics-600.

2) Spatial-Temporal Classification: We utilize a bi-directional LSTM [5] to perform temporal classification to localize activities within spatial-temporal proposals.

C. Post-processing

Candidates after localization and classification may be spatially and temporally overlap. We employ a spatially-temporal NMS to avoid overlapping candidates.

D. Results

Table I shows the results of our systems, "UEC-1".

TABLE I
Our system results.

System	nAUDC
UEC-1	0.96405

E. Discussion

We found that our method was less accurate than other methods. This may be due to the fact that our method was hardly able to detect action classes with little training data. Another reason could be the poor recognition of action classes where person and vehicles interacted with each other. In order to improve accuracy, these problems need to be solved.

II. VTT: Video to Text

Video To Text (VTT) is a quite challenging task. Not like image to text where a single image is being conducted, we have to consider and analyze the relation among the multiple frames of each video. In addition, due to ambiguities of natural language, the candidate of corresponding texts can be in a wide range. Figure 2 shows the overall of our approach.

A. Method

In this section, we explain our method for the VTT task. Firstly, we extract several frames from a given video as inputs to our VTT model. We fine-tune the image captioning model [6] to the VTT task.

We use the pre-trained ResNet-101 model to extract the visual features. We then feed these visual features to the captioning model to obtain the caption results. As the original paper, our captioning model is implemented with attention mechanism [7]. Each of the image is encoded into a set of spatial features corresponding to each sub-region of the image. And the word prediction also makes use of this mechanism.

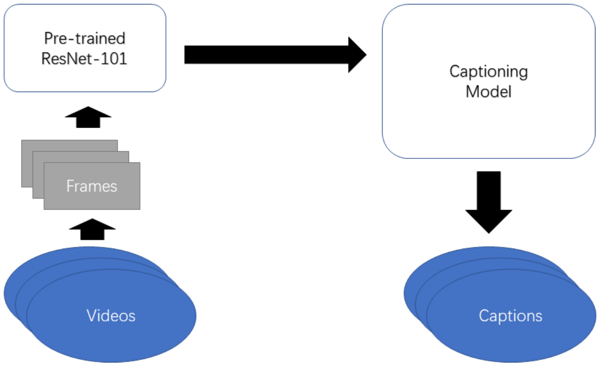


Fig. 2. The overall of our approach of the VTT task.

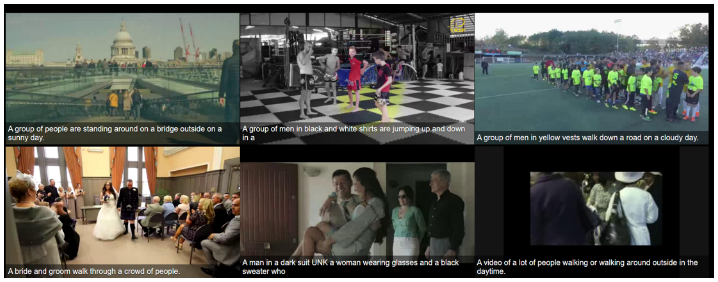
B. Dataset

Firstly, we only used the official VTT dataset which consists of about 1k videos for training the model. However, we found that our model worked poorly. So we determined to use the COCO dataset [8] by integrating it with the VTT dataset. Finally we found that the results using both the VTT dataset and the COCO dataset were improved with this expansion of dataset.

C. Results

Figure 3 shows some video captions generated by our model.

Table II shows the evaluation result from our model. Notice that we set different epochs and learning rates in each run, which are shown in Table III.


A group of people are standing around on a bridge outside on a sunny day.

A group of men in black and white shirts are jumping up and down in a

A group of men in yellow vests walk down a road on a cloudy day.

A bride and groom walk through a crowd of people.

A man in a dark suit UNK a woman wearing glasses and a black sweater who

A video of a lot of people walking or walking around outside in the daytime.

Fig. 3. Some video captions generated by our model.

TABLE II
VTT MODEL EVALUATION RESULT.

run	BLEU	METEOR	SPICE	CIDEr	CIDEr-D
UEC.run_1	0.0480	0.1553	0.031	0.036	0.018
UEC.run_2	0.0492	0.1628	0.032	0.039	0.021
UEC.run_3	0.0511	0.1723	0.034	0.047	0.024
UEC.run_4	0.0487	0.1607	0.033	0.043	0.023

TABLE III
SETTING CONDITIONS WITH EACH OF THE RUNS.

run	learning rate	epochs
UEC.run_1	3e-4	20
UEC.run_2	3e-4	30
UEC.run_3	5e-4	30
UEC.run_4	3e-4	50

D. Discussion

In this task, we extract several frames from each video while not considering their relation and adopted a naive learning strategy. Hence, We suppose we can improve our results if we introduce better learning strategy and improve our model with some methods such as optical flow.

III. CONCLUSION

This was our second time participation in the ActEV task in TRECVID [9]. This year, we experimented with a baseline method. Our results of this year were not as good as those of other teams. In addition, we participated the VTT task for the first time. We will keep improving our systems for TRECVID 2022.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc.of Neural Information Processing System*, 2015.

[2] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belong. Feature pyramid networks for object detection. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2017.

[3] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[4] K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-temporal features with 3d residual networksfor action recognition. In *Proc.of IEEE International Conference on Computer Vision*, 2017.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. In *Neural computation 9*, page 1735–1780, 1997.

[6] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. *arXiv preprint arXiv:1803.04376*, 2018.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[9] G. Awad, A. A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, J. Liu, Alan F. Smeaton, Yvette Graham, G. J. F. Jones, W. Kraaij, and G. Quénot. TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *Proc. of TRECVID 2020*, 2020.