An Automatic Text Generation System for Video Clips using Machine Learning Technique

Akira Shibata, Takashi Yukawa

Nagaoka University of Technology, Nagaoka, Japan 940-2188,   s153410@stn.nagaokaut.ac.jp

1. Introduction

The conventional method used all sampled frames which were extracted at regular time intervals from a video clip to generate an explanation text[1][2]. The method had encoder-decoder framework consisting CNN(Convolutional Neural Network) and LSTM(Long Short-Term Memory) network[2]. CNN encoder with ResNet200 model extracted the representation like objects, actions, backgrounds and times from 50 frames of each video[2]. LSTM decoder combined fragments of the representation to generate an explanation text[2]. This type of system has a problem that the processing time proportionally increases with the length of the video clip.

It is known from Neuroscience that humans use fewer frames when humans create explanation texts from consecutive images like a video clip[3]. Generating the caption with all sampling frames, the system outputs the sentence including representations for unnecessary frames. Our approach focuses on fewer frames referred as key frames instead of all sampled frames as the previous studies. The key frames are determined based on theories in Neuroscience. The system can output the explanation text including important representations which use key frames. The proposed system generates explanation text by combining representations of the key frames using LSTM network. These presentations are created by image caption method.

2. Approach

According to the findings in Neuroscience and Psychology, it is easier for humans to remember key frames in consecutive images[3]. Therefore, our method extracts the representation from only key frames that are easier to remain in human memory to generate explanation text.

There are five types of an event in a video clip. The types are “surprise like scene change”, “First and Latest frames”, “Repetition” [3], “Images that recall positive emotion” and “Images that recall negative emotion”[6]. The former three types are used in our method because they can be objectively judged. There are separate methods to detect key frames for different types of events. Figure 1 shows key frames detected by our method for three types of an event.

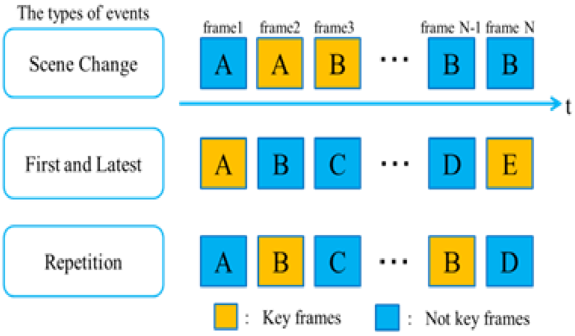


Fig.1 the example of key frames

Figure 2 shows the outline of our proposed system using the mentioned method. First, the system samples all frames from a video clip before detecting the key frames. The key frame selection process presumes types of the event using RGB difference between consecutive frames. The process outputs key frames by the method corresponding to the types of an event.

The representation extraction process creates the explanation text from each key frame based on NIC(Natural Image Caption) model[6]. The model has a CNN image encoder and a LSTM decoder. A CNN output information on key frames is transferred to a LSTM-based sentence generator[6].

Figure 3 shows the NIC model outline. The model trained by using MSCOCO and retrained the model by using Inception v3.

Finally, another LSTM network is used to generate the explanation text combining the representations of previous key frames. The later phase uses so-called single LSTM text generation method.

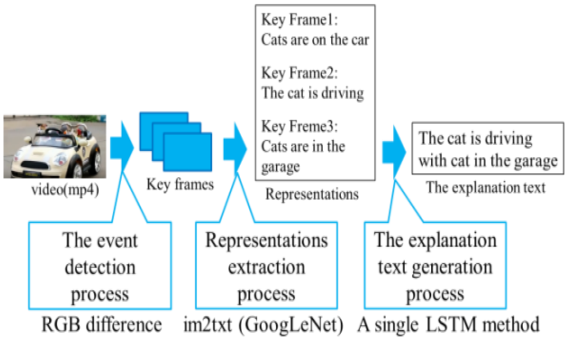


Fig.2 the suggested system

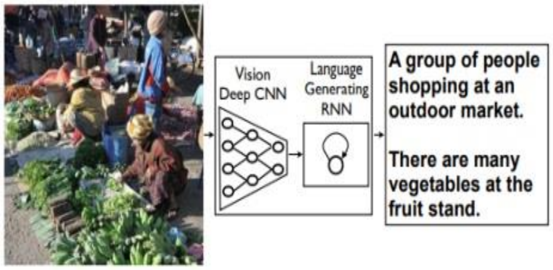

Vision Deep CNN
Language Generating RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

Fig.3 the example of NIC model[6]

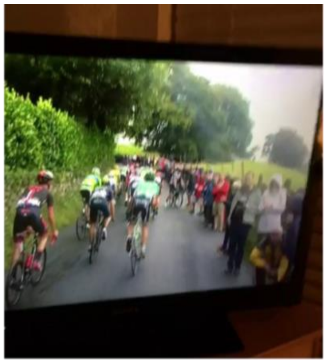
3. Results

We evaluate the systems by the average of METEOR score to compare the proposed system with other systems of the Video To Text task of trecVID 2018. Table 1 shows the scores of the participating systems released by the organizer. However, the proposed system (Kslab) included bugs when we submitted the results and the released score is 0.255. The re-evaluated score based on the debugged system is 0.270.

Tab.1 METEOR scores of teams which joined trecVID 2018

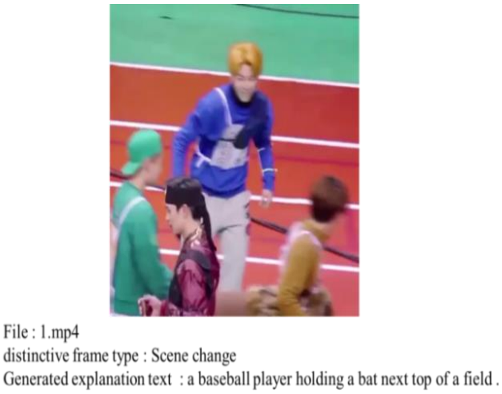
Runtype	Team name	METEOR score
N	UTS	0.306
V	INF	0.305
V	UPCer	0.276
N	Kslab	0.270 (0.255)
V	KU_ISPL	0.250
N	PicSOM	0.221
N	MMsys	0.207
V	NTU	0.206

Figure 4 shows a video clip which got a better score (0.321) from the proposed system. Figure 5 shows the video clip which got a worse score (0.185).



File : 999.mp4
distinctive frame type : Primacy and Reccency
Generated explanation text : a group of people riding bikes down a street .

Fig.4 example of good result



File : 1.mp4
distinctive frame type : Scene change
Generated explanation text : a baseball player holding a bat next top of a field .

Fig.5 example of bad result

We also compared changes in the number of processed frames in the proposed and the conventional systems. We focused on the "1811.mp4" with the largest number of frames in the test data of trecVID 2018. Table 2 shows the number of processed frames and METEOR score.

Tab.2 The number of frames and METEOR scores in 1811.mp4

	the number of processed frames	METEOR score
The proposed method	4	0.284
The conventional method	87	0.275

There were 9.2 processed frames on average in the test data of trecVID 2018 for the proposed system while the conventional system had 60.8 on average.

Finally, we evaluated the proposed system using test data in the VTT task in 2016 and 2017. Table 3 shows METEOR score when our approach used earlier data.

Tab.3 METEOR scores on past data

Test data type	METEOR score
2016	0.345
2017	0.318

4. Discussion

According to Table 1, we found that our method ranked fourth among the participant teams in the VTT task. The proposed method ranked second among the group of “N” Runtype.

When putting the Video To Text task into practical use, it is difficult for us to get the training data that match videos trend. Our approach does not use training data that match videos trend. Therefore, the system of Runtype "N" will get approximately the same score used the real video data.

From Figure 4 and Figure 5, our method need to process higher resolution images to get higher scores. There is slight difference between the proposed method and the conventional method showed in Table 2. The proposed system obtained better METEOR score for a video clip that had more frames. From Table 3, METEOR scores for past data are slightly better than that for this year. It is reasonable that the proposed system’s score ranges between 0.3 ± 0.1 in various test data.

An advantage of the proposed system is that the system could generate explanation text from fewer frames. The conventional method used over 60 frames which were made from one video. From Table 2, the proposed system used fewer frames, accounted for 4.6 percent of the total frames from the conventional system. It is evident from the average number of processed frames on all test dataset greatly decreased by using the proposed method. All things considered, the proposed approach shows effectiveness in generating explanation text with a reduction of frames. However, there are some limitations. The problems are that the system only processes short videos and mis-detects video effects. The proposed system assumes that all of key frames are memorized in “short term memory” in LSTM network. However, when processing longer videos, previous short term key frames need to disappear before the other multiple key frames come out. This process is similar to how human brain works.

The proposed system can extract key frames for the types of scene changes and image repetition. However, there is a variety of video effects such as fade-out, zoom-in and flash lights, which may cause the system to confuse to extract key frames for explanation texts. Therefore, video effects are taken into account when putting the proposed system in practical use.

In the future, the proposed method will be greatly improved for accuracy by the evolution of deep learning methods such as Picture to Text methods. This system will optimize the number of key frames. The proposed system used three types of an event, however more types should be used to detect better correct key frames from long videos and various video effects.

5. Conclusion

The final METEOR score in this study is not lower than that used by the method of the same training data type. The proposed method had great effects on the reduction of processed frames. In summary, this study needs further improvement in the extraction method of key frames and expansion of training dataset to achieve higher scores at METEOR.

6. Reference

[1] George Awad and Asad Butt and Keith Curtis and Yooyoung Lee and Jonathan Fiscus and Afzal Godil and David Joy and Andrew Delgado and Alan F. Smeaton and Yvette Graham and Wessel Kraaij and Georges Quénot and Joao Magalhaes and David Semedo and Saverio Blasi, TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search, Proceedings of TRECVID 2018, 2018, NIST, USA

[2] An-An Liu, Yurui Qiu, Yongkang Wong, Ning Xu, Yuting Su, Mohan S. Kankanhalli, “Tianjin University and National University of Shingapore”, TRECVID, Video to Text Description, 2017

[3] Geoffrey F. Woodman, Marvin M. Chun, “The role of working memory and long-term memory in visual search”, VISUAL COGNITION, 2006, 14, 808 830

[4] Craig E. Geis, “Memory: How Memories are Formed, Stored, and Retrieved with Personal Tips for Trainers”, California Training Institute, Memory

[5] Cleotilde Gonzalez, Jason Dana, Hideya Koshino, Marcel Just, “The framing effect and risky decisions:Examining cognitive functions with fMRI”, Journal of Economic Psychology 26, 2005, 1–20

[6] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, “Show and Tell: A Neural Image Caption Generator”, arXiv:1411.4555 [cs.CV]

[7] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, Kate Saenko, "Sequence to Sequence -- Video to Text", arXiv:1505.00487 [cs.CV]

[8] Benedetto De Martino,* Dharshan Kumaran, Ben Seymour, Raymond J. Dolan , “Frames, Biases, and Rational Decision-Making in the Human Brain”, SCIENCE, 2006.4.AUG, VOL.313

[9] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. “Translating videos to natural language using deep recurrent neural networks”, NAACL, 2015

[10] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, “Deep captioning with multimodal recurrent neural networks (mrnn)”, arXiv:1412.6632[cs.CV], 2014