KU-ISPL TRECVID 2018 VTT Model

Youngsaeng Jin, Junggi Kwak, Younglo Lee, Jeongseop Yun, Hanseok Ko[1]

Intelligent Signal Processing Laboratory, Korea University

Abstract

KU-ISPL model for TRECVID 2018 Video-to-Text (VTT) is presented in this paper. A stack of two LSTM with attention mechanism is structured in the VTT architecture. We employ a sequence-to-sequence model to deal with sequential input and output. The encoder in our model encodes video frames into visual representations and the decoder decodes the visual representations into textual words. Attention mechanism is exploited for best use of contextually pertinent frames in input video. The model pays attention to the hidden states of 2nd LSTM in the encoder to obtain efficient hidden states in the decoder. Visual feature, acoustic feature and detection result of videos are extracted from deep learning models and the resulting features are subsequently concatenated into one. It is used for an input descriptor of the model. The stacked LSTM and attention weights are jointly trained and the whole model is an end-to-end trainable network. We proceed by making 4 runs for our model by combining various types of features to explore how the information impacts the performance of sentence generation. The sentence matching method is based on the fusion score of Meteor and Bleu. Because the TRECVID VTT task is open domain, the sentence generation and sentence matching system are trained by various database such as MSVD, MVAD, and MSR-VTT. Experimental results show that the proposed model performs better than the model without attention mechanism.

Methods

1. System overview

Video-to-Text (VTT) [1, 2] is a process of generating text that describes the content of a video. With rapid advancement of deep learning techniques, VTT is widely researched but still faces limitations because it is hard for a machine to understand what event occurs in a video clip. Since both video and corresponding textural description are sequential data, we employed Sequence-to-Sequence – Video-to-Text (S2VT) model [3] as shown in Figure 2 as baseline. We explored this model under TRECVID 2017 VTT (KU-ISPL) [4] by combining various types of input features. One of the shortcomings of this model, however, was determined later that due to encoding an input video into a single fixed-length vector, it was unable to compress all information. This led to inadequate descriptions generated. To alleviate this problem, we employ attention mechanism [5], which has recently made a big impact in Neural Machine Translation (NMT). Attention mechanism allows the decoder in the network to pay more attention on important frames and obtain efficient decoded information. The model is learned to generate a context vector which is computed as the weighted sum of the input sequence. In this paper, an attention-based S2VT model is proposed to

[1] *Director of Intelligent Signal Processing Laboratory*

mitigate the aforementioned problem.

For the matching and ranking subtask, each generated description is scored by weighted sum of METEOR [9] and BLEU [10]. The scores are obtained by comparing with the reference descriptions given by TRECVID.

Visual feature, acoustic feature and detection result of videos are used for input features. They are extracted from a deep learning model, VGG-Net [6], VGGish-Net [7] and mask-RCNN [8] respectively. Four sentence generation runs are implemented wherein each run combines different type of those features. The main run is ***run2*** which is attention-based S2VT with the visual feature only. The model is trained by various database such as MSVD [11], MVAD [12], MSR-VTT. The overall system architecture for the sentence generation is shown in Figure 1.

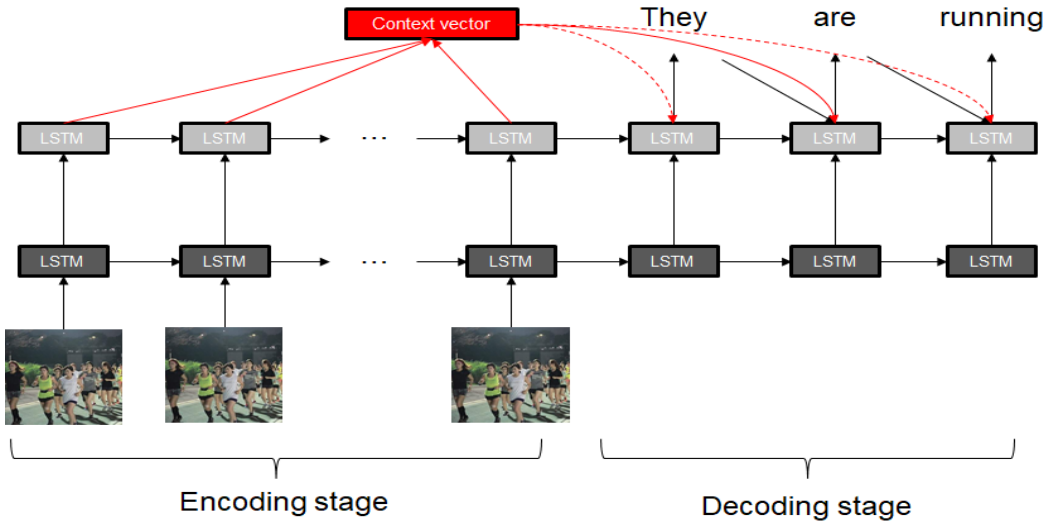

Figure 1. Overall structure of KU-ISPL TRECVID 2018 VTT model

2. Input descriptor generation

We employ visual features, acoustic features and detection results as input descriptors to capture diverse information contained in the video clip. To prevent over-fitting, 50 frames are randomly selected from the whole videos in each training epoch. In this section, the process of various feature extraction scheme is described in detail.

2.1 Visual mid-level feature

CNN models are proved effective for visual recognition tasks such as object detection and scene classification. It is widely used for feature extractors. We employ VGG-Net [6] fc7 layer as midlevel feature for building primary data. It is trained by large scale dataset. The visual features are extracted from pre-trained VGG-Net. Dimension of the visual mid-level feature is 4096 at each timestep.

2.2 Acoustic mid-level feature

Sound is a good information to represent an event. We adopt VGGish [7] model which is a pre-trained VGG-like audio classification model. It is trained using a dataset of over 2 million human-

labeled 10-second YouTube video soundtracks, with labels taken from an ontology of more than 600 audio event classes. First sound wave is extracted from input videos and the wave is resampled to 16 kHz. Then it is preprocessed to get 64-dim log-mel spectrogram (0.025 window length, 0.01 hop length). Concatenate consecutive 96 log-mel spectrograms (96×64, about 1s) and feed it into pretrained VGGish model to obtain final embedding audio feature vector at last. Dimension of acoustic feature is 128 at each timestep.

2.3 Object detection

Object detection result is used to capture objects appeared in videos. With detection results, we hoped to determine the subject of the scene for better generating description. While all the words used in training are in the corpus, there are cases where it is difficult to ascertain whether the corpus correctly express the training videos. If an appropriate word can be chosen, we think it would affect the weight during training. We exploit Mask-RCNN [8] model to obtain precise detection results. Objects detection results are generated and used for training. The dimension of object detection results is 81 at each timestep.

3. Architecture

3.1 Sequence to Sequence model

Models for video to text contain encoder and decoder. Encoder works for transforming frames into visual representation and decoder does for generating textual descriptions using the visual representation. Since both videos and descriptions are sequential data, it is best to use sequence to sequence model. Therefore, S2VT [3] is employed for generating descriptions, which is structured with a stack of two LSTM for better encoding and decoding. As shown in Figure 2, while encoding stage encodes input sequence to a fixed-length vector, decoding stage maps the vector to a sequence of output sequence. Dimensions of the hidden state of both LSTM is 1024. Dropout is applied on both LSTM layers to prevent overfitting. Basically, it has same structure of S2VT model [3] except the dropout.

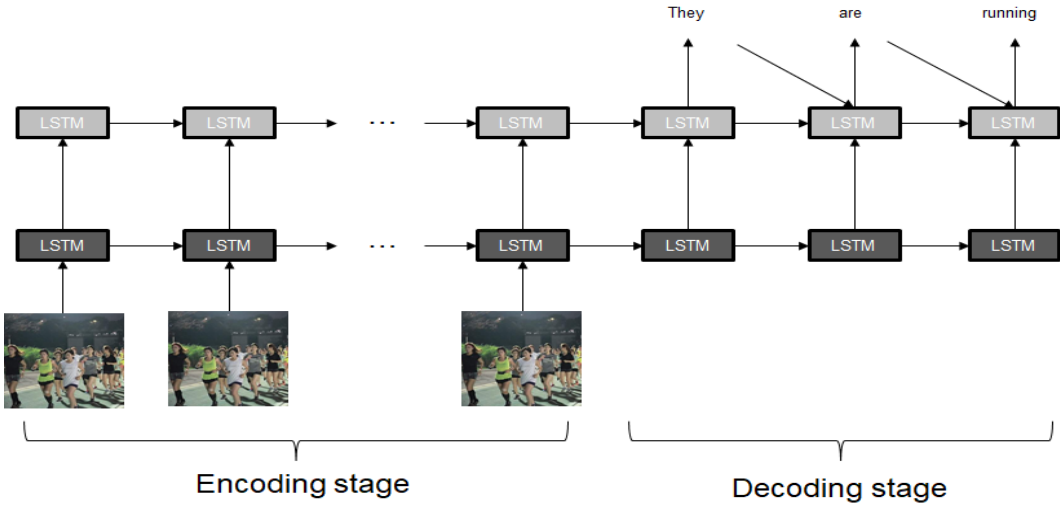

Figure 2. Fundamental structure of S2VT model

3.2 Attention mechanism

As an output from encoding stage in a general sequence to sequence model, such as S2VT, is a fixed-length vector, it leads to a limitation on representing all information in an input video. Owing to information loss, an inadequate description can be generated. To mitigate this problem, an attention mechanism is emplaced to look over all the information included in the input video. It allows the decoder in the network to pay more attention to contextually pertinent frames and generate the proper word through a context vector as shown in Figure 3. The context vector is generated for each output time step. Essentially, the model is learned to obtain the context vector. The model pays attention on hidden states of 2nd LSTM in the encoder to obtain efficient hidden states in the decoder so that words at each timestep are generated more properly.

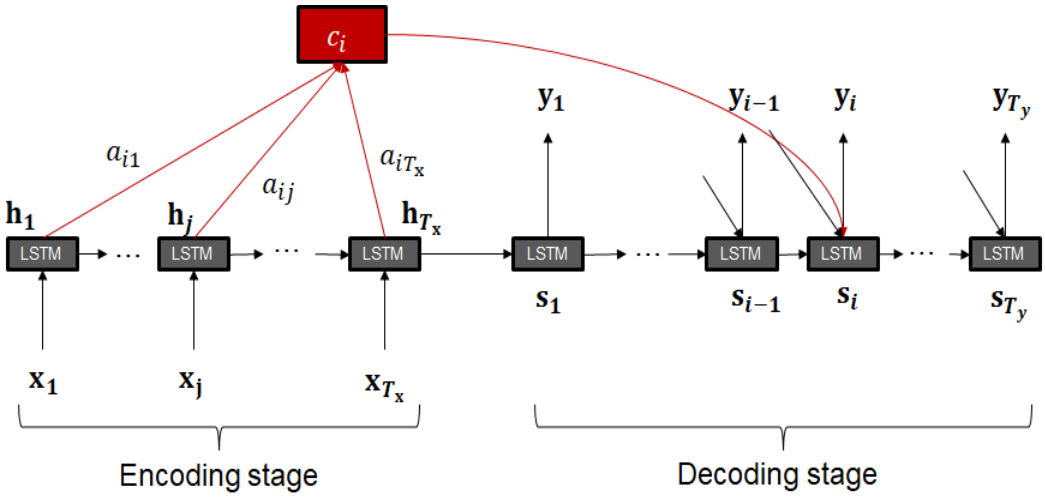

Figure 3. Schema of attention mechanism

With the context vector c_i, the hidden state s_i in the decoding stage is computed as

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

where y_i is an output at timestep i. With the above hidden state in decoding stage, the model generates more natural description.

$$y_i = g(s_i)$$

4. Sentence Generation

4.1 Dataset

We collected three public datasets from online for training and testing our model. The datasets will be described in detail below.

The MSVD [11] obtained from Microsoft is a collection of Youtube clips. This dataset originally consists of multi-lingual descriptions. In this work, only English description is used and its amount

is 1,564. There are roughly 40 human-annotated descriptions for each video and 67,139 descriptions overall with 12,316 vocabularies. The dataset has most descriptions in comparison with other existing datasets.

The MVAD [12] collected from Montreal Institute for Learning Algorithm (MILA) is another set of video clips from 92 Hollywood movies with a total of 46,589 clips. Each clip is accompanied with a single automatically annotated description. In this work, 4,951 of the data set is used. Video clips are 7 seconds on average and 10,984 vocabularies are included in total.

The MSR-VTT dataset is used in MSR Video to Language Challenge organized by ACM. It contains a total of 46,589 clips with 20 categories. Each clip is with about 20 human-annotated descriptions. In our work, we use 6,074 videos with 121,021 descriptions. Video clips are 20 seconds on average and 22,451 vocabularies are contained.

Table 1. Statistics about 5 dataset in our task.

	MSVD	MVAD	MSR-VTT
# video	1,564	4,951	6,074
# description	67,139	4,951	121,021
# avg description	40	1	20
# vocab	12,316	10,984	22,451

4.2 Sentence Generation Run

Our goal is to improve the performance of sentence generation through S2VT model with attention mechanism. Moreover, various combinations of features are exploited to verify how the information impacts the performance of sentence generation. Experiment configuration for each run is shown in Table2. ***Run2*** is only using the visual feature to compare the results with each run, which is our primary run for TRECVID 2018. ***Run3*** and ***run4*** based on ***run2*** input the additional features. Note that ***Run1*** is the model we submitted last year (KU-ISPL TRECVID 2017 VTT System [4]).

Table 2. Combined method for each run.

	Attention Mechanism	VGGNet mid-feat	VGGish mid-feat	Object Detection
Run1	X	O	X	X
Run2	O	O	X	X
Run3	O	O	O	X
Run4	O	O	O	O

5. Sentence Matching

The second subtask in TRECVID VTT task is sentence matching and ranking. The sentence matching method is based on the total score of METEOR [9] and BLEU [10]. They are evaluation metrics for sentence comparison. The score is measured by comparison of our generated descriptions and reference descriptions given by TRECVID. Due to the difference scale between METEOR and BLEU, the total score for rank is 1.2 * METEOR + 1 * BLEU. The reference descriptions for each dataset (A, B, C, D, E) are ranked in descending order of the total score.

6. Conclusion

Through various experiments we determined that applying attention-based sequence-to-sequence model was effective to accurate description generation of video clips. With the attention mechanism, generated descriptions were found more natural and accurate. Additionally, attention mechanism allowed the decoder in the model to pay more attention to contextually important frames. Visual feature, acoustic feature and detection result of videos were respectively extracted from a deep learning model. The performance measurements of our team's results are shown in Figure 4 and an example of generated descriptions for a given video is shown in Figure 5. ***Run2*** (Attention, Visual) and ***run4*** (Attention, Visual + Acoustic + Detection) are better than the others. We expected the acoustic features to be effective, but the performance was not good because many videos in test dataset contain background music which is unrelated to the event in the videos. Therefore, the learning performance with the acoustic feature naively employed by inadvertently including background music was determined not effective.

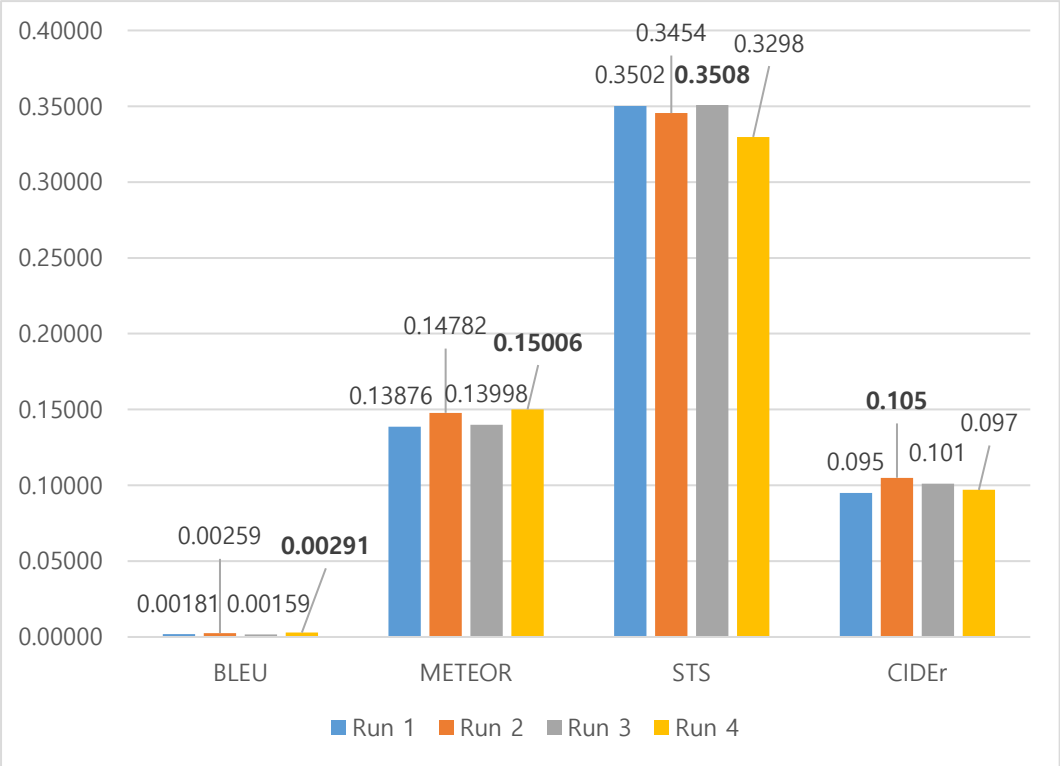

Figure 4. Quantitative results of KU-ISPL TRECVID 2018 VTT model

GT	Running from the kitchen into the living room, a very tall man in a striped black shirt and shorts, with the head of a baby, dribbles an orange ball.
S2VT (V)	A man is a
+Attention (V)	A woman is putting some items in the oven
+Attention (V+A)	She leaves someone at the mirror
+Attention (V+A+O)	Someone and someone watch from the second floor as someone steps into the .

Figure 5. Qualitative results of KU-ISPL TRECVID 2018 VTT model

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2017R1A2B4012720).

References

1. S. Alan F., O. Paul and K. Wessel. Evaluation campaigns and TRECVid. In proceedings of the 8th ACM international workshop on Multimedia information retrieval. ACM, 2006.
2. G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. F. Smeaton, Y. Graham and W. Kraaij. TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search. Proceedings of TRECVID 2018. NIST, USA, 2018
3. S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko. Sequence to sequence-video to text. In proceedings of the IEEE international conference on computer vision, 2015.
4. D. Kim, J. Beh, Y. Chen, and H. Ko. KU-ISPL TRECVID 2017 VTT System. In TRECVID 2017.
5. D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
6. K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

7. JF. Gemmeke et al. "Audio set: An ontology and human-labeled dataset for audio events." Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.
8. K. He, G. Gkioxari, P. Dollár and R. Girshick. "Mask R-CNN. Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017.
9. M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In EACL, 2014.
10. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In ACL, 2002.
11. A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In CVPR, 2015.
12. A. Torabi, C. Pal, H. Larochelle, and A. Courville. Using descriptive video services to create a large data source for video annotation research. arXiv:1503.01070v1, 2015.