KU-ISPL TRECVID 2017 VTT System

Daehun Kim, Joungmin Beh, Youngseng Chen, Hanseok Ko[1]

Intelligent Signal Processing Laboratory, Korea University

Abstract

KU-ISPL system for TRECVID 2017 Video to Text (VTT) is presented in this paper. The main method of the system is a stacked LSTM model for sentence generation. Input descriptors of the system consist of various deep learning-based features and multi-object detection results to obtain diversity of characteristics and key information from videos. We choose mid-level features of VGGnet and SoundNet as major features to capture multimodality about image and acoustic. Additionally, the visual attribution about objects and places is used for high-level feature. Finally, visual syntax detection is fine-tuned by sigmoid loss function for finding key words. We make 4 runs for the stacked LSTM model by combining various types of features to see how the information impacts the performance of sentence generation. Word2Vec is adopted for effective encoding of sentences. The embedded words by Word2Vec are used at state value and target of the LSTM. On the other side, the sentence matching method is based on the fusion score of Meteor, Bleu and the detection. The output of detection represents the probability that a word exists. Because the TRECVID VTT task is open domain, the sentence generation and sentence matching system is trained by various database such as MSVD, MPII-MD, MVAD, MSR-VTT, and TRECVID-VTT 2016.

Methods

1. System overview

We investigate various methods to participate in the TRECVID 2017[16] Video to Text (VTT) task [1], and through various experiments it is determined that applying sequential RNN based method is most efficient to the system. The overall system architecture for the sentence generation is shown in Figure 1. The main approach of the system is the Sequence-to-Sequence model [2] for sentence generation as shown in Figure 2. We aim at improving the training data instead of model improvement for more precise sentence generation. We use CNN feature with object detection [3] and SoundNet [4] result for this purpose. Additionally, the visual attribution about objects and places are used for high-level feature. Finally, the probability of key word occurrence is used by visual syntax detection which is fine-tuned with sigmoid loss function for multi-class detection. The corpus of database is represented using Word2Vec method [5]. These embedded words by Word2Vec are used at the state value and target of the LSTM. Through a combination of these methods, we construct four sentence generation runs and subsequently proceed our primary run by self - evaluation using Meteor [6] and Bleu [7].

---

[1] *Director of Intelligent Signal Processing Laboratory*

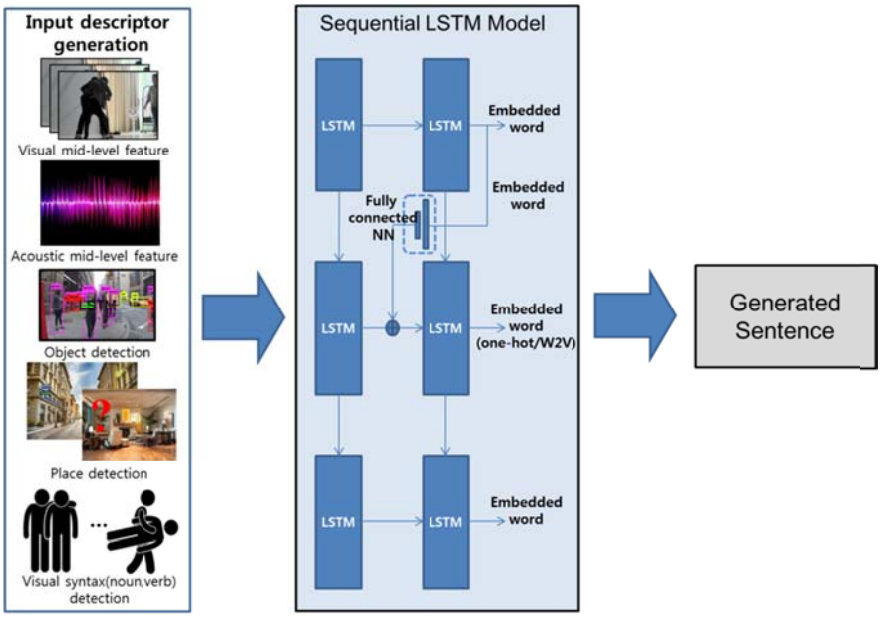

Figure 1. Overall structure of KU-ISPL TRECVID 2017 VTT system

2. Input descriptor generation

The proposed method uses multi-modal features and detection results as inputs to reflect richer information. Each run has a different combination among the five types of input. To prevent over-fitting, 60 frames are randomly selected from the whole videos in each training epoch and the features are extracted in order.

2.1 Visual mid-level feature

CNN models are proved effective for visual recognition tasks such as object detection and scene classification. It is used very widely in many image processing fields. We employ VGG.net [8] fc7 layer as midlevel feature for building primary data. It is already used in the sequence-to-sequence model [1]. We generate training data with pre-trained VGGnet to obtain the fc7 layer. We extract fc7 layer from all databases with every frame. Each frame is randomly cropped with random region. Cropping region has over 90% of full frame size. Dimension of the visual mid-level feature is 4096.

2.2 Acoustic mid-level feature

Some sounds have enough features to represent an event. So our system adopted the 1D CNN based acoustic feature extraction method [9]. The main idea in this method is to leverage the natural synchronization between vision and sound in unlabeled video in order to learn a representation for sound. Because the domain of VTT task is open, “conv5” layer feature rather than the final layer feature is chosen for the universal acoustic deep feature. Dimension of the acoustic mid-level feature is 256.

2.3 Object detection

We decided to use object detection [3] to highlight keywords. Keywords mean dictionary from corpus of each training videos. 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. Our idea is motivated by the fact that detection results can represent each image's keywords. For example, if detection results include male and dog, we can envision "petting" or "hunting" activities with these keywords. If an appropriate verb can be found in the dictionary, we think it would affect the weight during training. Objects Detection results are generated and used for training together with size n/20 frame * 100 (number of classes). The dimension of object detection results is 100.

2.4 Place detection

Powerful visual features can be obtained with CNN models and they are robust on diverse visual tasks such as scene recognition. The proposed system employs Places205-AlexNet (AN) model [10]. The model contains 5 convolutional layers and 3 fully-connected layers. The convolutional filters have various sizes from 3x3 to 11x11. The output layers represent the probabilities of 205 different places.

2.5 Visual syntax detection

The results of object detection and place detection cannot cover all words in training sentences. Hence, our system includes self-trained visual syntax detection. First, Syntaxnet [11] in Tensorflow-models is used for analysis of sentences. Therefore, the words are classified according to the part of speech tags and pick out what happens frequently in noun and verb. Next, VGGnet is fine-tuned for these words by sigmoid loss function because it is a multi-class problem. The number of trained words is 649.

3. Sequential LSTM model

3.1 Sequence to Sequence model

As shown in Figure 2, while first stage encodes input sequence to a fixed length vector using one LSTM, and 2nd stage uses another LSTM to map the vector to a sequence of outputs. We can use one sequential LSTM (thicker gray) for both the encoding and decoding stage. This means that we can share parameters on encoding and decoding stage. This model uses a stack of two LSTMs with 700 hidden units each. We also apply drop-out layers to this model to prevent overfitting. Basically it has same structure of S2VT model [2] except drop-out layers and putting embedded words if Word2Vec is used.

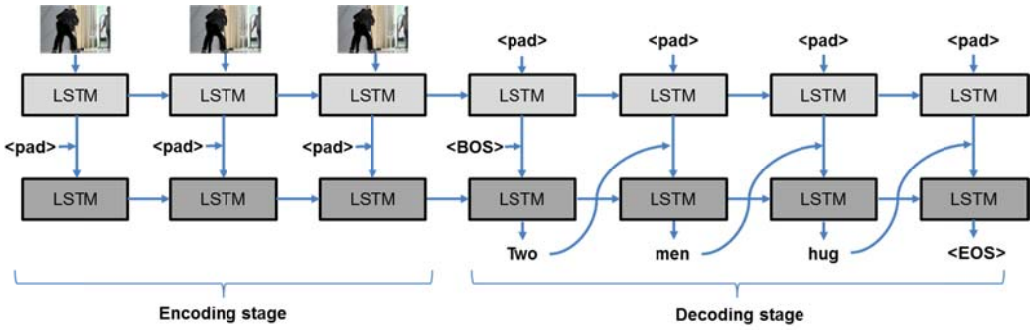

Figure 2. Fundamental network structure of LSTM network

4. Sentence Generation

4.1 Database

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

The MSVD [12] obtained from Microsoft is a collection of Youtube clips. This dataset originally consists of multi-lingual descriptions. In this work, only English descriptions are 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 MPII-MD dataset [13] from Max Planck Institute for Informatics contains 68,374 video clips extracted from movies. Each video clip has a single description which is from movie scripts and audio description (AD). The scripts and AD contribute to more accurate descriptions again diverse and complex scenes. 21,221 vocabularies are included.

The MVAD [14] collected from Montreal Institute for Learning Algorithm(MILA) is another 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 dataset are 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.

The TRECVID-VTT 2016 is a test set for TRECVID Video to Text Challenge 2016. It contains over 30k Twitter Vine videos with automatically annotated descriptions. 1,875 videos are contained with two descriptions each. 3,724 descriptions and 2,487 vocabularies are included in total.

We extract 1/4 randomly from all databases because of the problem of training time. We refined dictionary from corpus with most used words and verbs from internet.

Table 1. Statistics about 5 dataset in our task.

	MSVD	MPII-MD	MVAD	MSR-VTT	TRECVID-VTT 2016
# video	1,564	68,374	4,951	6,074	1,875
# description	67,139	68,374	4,951	121,021	3,724
# avg description	40	1	1	20	2
# vocab	12,316	22,221	10,984	22,451	2,487

4.2 Word2Vec

The word or sentence-based task with uncountable corpora represented as one-hot vectors is too complex to deal with due to length of the vectors. Word2vec can alleviate the problem. Word2vec embeds words to a lower dimensional space. Moreover, semantically similar words are embedded nearby, so it contributes to high-quality word representations. The proposed system employs Inspect_Word2Vec model from Google [15]. It was trained on almost 100 billion words from a Google news dataset and contains exactly 3 million words in dictionary. Each word is mapped as 300-dimensional feature. In addition, six key words (“a”, “and”, “the”, “of, “also”, “should”) are used as stop words and excluded from the dictionary. Additional 6 zeros are suffixed at the end of preexisting words, and the above six words represented as concatenating 300 zero values with 6-dimensional one-hot vectors. Finally, each word is represented as 306-dimensional vectors.

4.3 Sentence Generation Run

Our goal is to improve the performance of sentence generation through a combination of various features. For this goal, we organized four runs through training for some combinations. We tried a combination as shown in Table2. Run2 is the result using only the CNN mid-level features to compare the results with each run. We make runs with a combination of CNN mid-level feature and object detection, all features, all features and using word2vec.representation.

Table 2. Combined method for each run.

	VGGNet mid-feat	Object Detection	SoundNet mid-feat	Place Detection	Visual Syntax	Word2vec
Run1	O	O	X	X	X	X
Run2	O	X	X	X	X	X
Run3	O	O	O	O	O	X
Run4	O	O	O	O	O	O

4. Sentence Matching

The second subtask in TRECVID VTT task is sentence matching and ranking. The sentence matching method is based on the fusion score of Meteor, Bleu and the detection. The output of detection represents the probability that a word exists. The score of detection defined as the ratio of the number of words whose probabilities exceed a certain value and the word number of the entire generated sentence. Our system finds all the scores between each example and the generated sentences, and determines the order between them.

5. Conclusion

Through various experiments we determined that applying sequential RNN based method is most efficient to the system. We aimed at finding appropriate feature combinations for more precise sentence generation. We used CNN mid-level feature with object detection and SoundNet for this purpose. Additionally, the visual attribution about objects and places are used for high-level feature. Finally, the probabilities of occurrence of key words are used by the visual syntax detection. We also tried to represent the corpus of the database using the Word2Vec method. The average performance index of our team's results is shown in Figure 3. We expected the acoustic features to be effective, but the performance was not good. Our team tried to train with too many DBs for about 3 weeks, but the loss did not decrease enough. Therefore, the learning performance is not excellent somewhat.

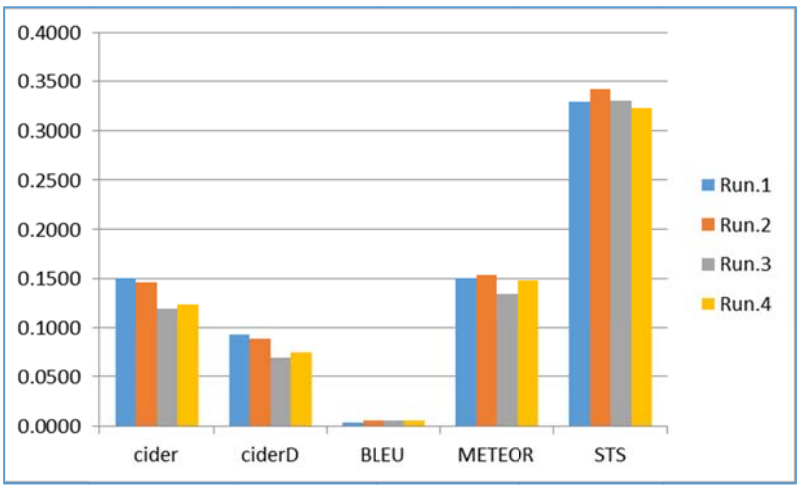

Figure 3. Overall structure of KU-ISPL TRECVID 2017 VTT system

References

1. Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006).

2. Venugopalan, Subhashini and Rohrbach, Marcus and Donahue, Jeff and Mooney, Raymond and Darrell, Trevor and Saenko, KateSmeaton, Sequence to Sequence - Video to Text, Proceedings of the IEEE International Conference on Computer Vision, 2015
3. Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z, Song Y, Guadarrama S, Murphy K, Speed/accuracy trade-offs for modern convolutional object detectors, CVPR, 2017
4. Yusuf Aytar, Carl Vondrick, Antonio Torralba , SoundNet: Learning Sound Representations from Unlabeled Video, NIPS, 2016
5. T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems, 3111-3119, 2013
6. Michael Denkowski and Alon Lavie, Meteor Universal: Language Specific Translation Evaluation for Any Target Language, Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, 2014
7. Kishore Papineni, Salim Roukos, ToddWard, andWei-Jing Zhu, BLEU: A method for automatic evalution of machine translation., In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July. 2002
8. K. Simonyan and A. Zisserman., Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.
9. Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." Advances in Neural Information Processing Systems. 2016.
10. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014) Learning deep features for scene recognition using places database. Advances in Neural Information Processing Systems
11. Andor, Daniel, et al. "Globally normalized transition-based neural networks." arXiv preprint arXiv:1603.06042 (2016).
12. Chen D, L., and Dolan W, B. (2011) Collecting highly parallel data for paraphrase evaluation. In ACL.
13. Rohrbach, A., Rohrbach, M., Tandon, N., and Schiele, B. (2015) A Dataset for Movie Description. In CVPR.
14. Torabi, A., Pal, C., Larochelle, H., and Courville, A. (2015) Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research. In CVPR.
15. Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR.
16. George Awad, Asad Butt, Jonathan Fiscus, David Joy, Andrew Delgado, Martial Michel, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot, Maria Eskevich, Roeland Ordelman, Gareth J. F. Jones and Benoit Huet (2017) TRECVID 2017: Evaluating Ad-hoc and Instance Video Search, Events Detection, Video Captioning and Hyperlinking. In Proceedings of TRECVID 2017