Nagaoka University of Technology Submissions to TRECVID 2022 Video to Text Task

Toshichika Mashimo, Mutsuki Ishii, Takashi Yukawa

Nagaoka University of Technology, Niigata, Japan

Abstract

The Kslab team participated in the TRECVID 2022 video to text (VTT) task and submitted four runs with different captioning methods and aggregation methods. Our system consists of three phases: frame extraction from the video, captioning for each frame, and aggregation of the captions. This year, we adopted the NIC and OFA models for the still image captioning phase, and sentence-based and word-by-word Lexrank methods for the sentence aggregation phase.

In the captioning phase, the still image captioning model using OFA provides drastic improvement in vocabulary size and sentence length compared with the NIC-based model. In the aggregation phase, much better results were obtained with sentence-based Lexrank. This is attributed to the high relevance of words within a single sentence.

Additionally, we observed that low-scoring sentences have the following features. The first is that inaccurate captions are sometimes generated in the captioning phase due to inappropriate frames being extracted. Second, sometimes the correct words are included in the captioning phase, but they are not in the final sentence. Based on these results, future improvements should focus on frame extraction and the aggregation phases.

1. Introduction

For the TRECVID 2022 video to text (VTT) task, the Kslab team proposes to use only a part of the frames extracted from the video. This is because it is possible to generate a caption of sufficient quality only from frames with significant change, instead of using all frames extracted from the video for motion prediction in time-series analysis.

According to research by Shibata et al., the system [1] that uses only a part of the frame generates a caption using the first frame, last frame, and keyframes [2], which are scene-changing frames in the video.

The system consists of three phases: frame extraction from the video, captioning for each frame, and aggregation of the captions. Figure 1 shows a flowchart of these processes.

In previous years, based on this system, we used the NIC model [3] for the captioning phase and Lexrank [4] for the aggregation phase. This year, an OFA model was newly added to the captioning phase. We also used two different Lexrank methods in the aggregation phase: sentence-based and word-by-word. The following section provides a comparison of the new methods with previous methods.

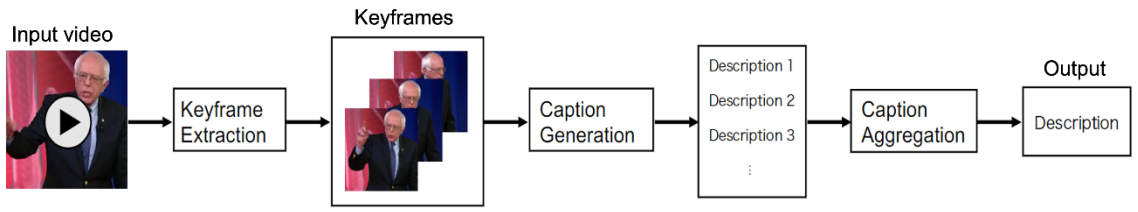

Figure 1. Overview of the keyframe method

2. New Methods for the Captioning and Aggregation

2.1 Sentence generation with OFA

The captioning phase uses an image captioning task from the OFA framework, which is capable of generating highly accurate descriptive text from images.

OFA [5] is a framework that uses an encoder and a decoder based on the Seq2Seq model [6] combined with the transformer [7]. It is capable of image generation from descriptions, image classification, language modeling, and question answering, in addition to description generation from images.

The encoder consists of a self-attention mechanism and a feed-forward network (FFN), and the decoder also uses these, plus cross-attention, to build a connection between the output of the decoder and the encoder. Also, a normalization layer is added after the first layer of attention and the FFN for training stability and speed.

For every task that can be performed using OFA, the output is obtained by taking an image and an imperative sentence as input. The imperative sentence is text that expresses what the output should be regarding the target image. In our system, the sentence, "What does the image describe?" is used for this input.

One of the advantages of this technique is that the Seq2Seq model used in OFA is robust to training time-series and sequential data since the model uses recurrent neural networks (RNNs). The model can generate captions even with only one frame, because it takes the previous word as an input when generating sentences. Therefore, when keyframes are used as input, separate captions can be generated for each frame.

2.2 Word-by-word Lexrank

In the aggregation phase, the goal is to extract phrases from the captions output in the generation phase and form captions from the aggregated results using Lexrank for each phrase.

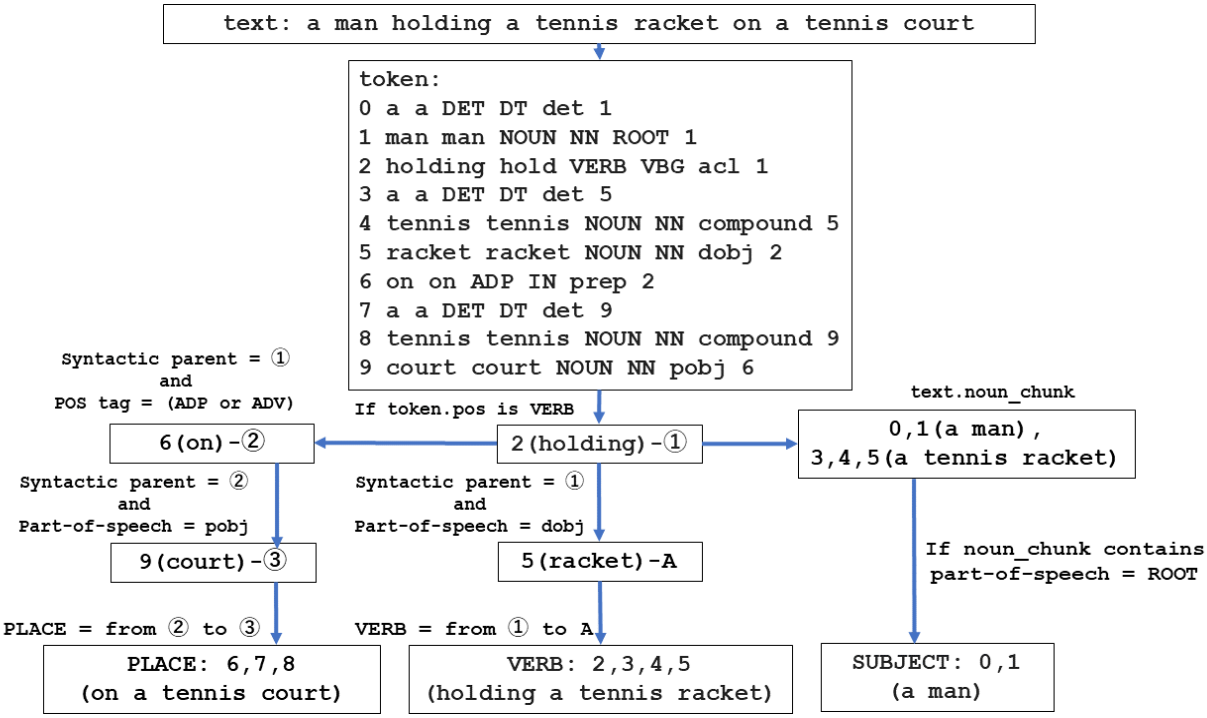

Figure 2. Flow of phrase extraction in the sentence of video ID 1711 generated using OFA

First, from the seven sentences generated in the generation phase, word groups related to the subject, verb, and location are extracted using spaCy. The extraction method is shown in Figure 2.

Tokens including index, verbatim text content, base form of the token, part-of-speech (POS) tag, dependency relation, and index of syntactic parent is obtained from the input document. Based on these tokens, the VERB, PLACE, and SUBJECT phrases are extracted from the text.

In the case of VERB phrases, the words whose POS tag is VERB are determined to be automatic or transitive based on their dependency relations. If it is an intransitive verb, only the word is used as the VERB phrase, and if it is a transitive verb, the subsequent object is used as part of the VERB phrase.

For PLACE phrases, the first step is to find a preposition or adverb in a dependency relationship with a word whose POS tag is VERB. If it is a preposition, the PLACE phrase is defined up to the following prepositional object (pobj). If it is an adverb, only “there” and “here” are treated as PLACE phrases.

Regarding SUBJECT phrases, the words defined in a noun chunk phrase that have ROOT in the dependency relationship are used as SUBJECT phrases.

This flow is run once for each of the seven sentences. For each phrase, aggregation is performed using Lexrank, and the three extracted phrases are combined to form a single sentence.

2.3 Submitted Runs

Our team submitted four different runs of the system. Each system is a combination of two different generation phases and two different aggregation phases. For the sentence generation method, we used the NIC method, which was used in previous years, and the OFA method proposed here. For the sentence aggregation method, we used the sentence-based Lexrank [4], which was also used in previous years, and word-by-word-based Lexrank. Table 1 lists the names of the submitted runs and the methods used.

Table 1. Names and methods of runs

Name	Caption generation	Caption aggregation
kslab_NUT_1	OFA	Sentence-Based
kslab_NUT_2	NIC	Sentence-Based
kslab_NUT_3	OFA	Word-By-Word
kslab_NUT_4	NIC	Word-By-Word

3. Results and Discussion

Table 2 summarizes the evaluation metric scores for each run submitted to the VTT task.

Table 2. Scores for each run

	BLEU	CIDEr	CIDEr-D	METEOR	spice
kslab_NUT_1	0.1142	0.619	0.194	0.2764	0.097
kslab_NUT_2	0.0749	0.163	0.048	0.1971	0.049
kslab_NUT_3	0.0928	0.51	0.11	0.2217	0.071
kslab_NUT_4	0.0692	0.141	0.027	0.1642	0.036

As can be seen in Table 2, kslab_NUT_1 achieved higher scores in all metrics than kslab_NUT_2, and that similarly, kslab_NUT_3 achieved higher scores than kslab_NUT_4, confirming the effectiveness of the method of using OFA for sentence generation. It was also found that the more correct sentences were generated during the generation phase, the more strongly they affected the output. The effectiveness of the word-by-word Lexrank method was not confirmed, as kslab_NUT_3 scored lower than kslab_NUT_1, and kslab_NUT_4 scored lower than kslab_NUT_2 as well.

Reviewing the generated text that scored low in kslab_NUT_1, the generated text often lacked words for time and place. In addition, blurred frames were sometimes selected as keyframes, and the correct words or phrases were not selected during the aggregation phase.

We also found that the outputs of kslab_NUT_3 and kslab_NUT_4 tended to fail to correctly recognize the sentence structure—for example, the object of transitive verbs sometimes could not be obtained.

Therefore, the accuracy can be expected to improve if frames can be extracted in which the subject is clearly visible in the keyframe extraction phase, and if noise can be removed from blurred frames. In addition, the score can likely be improved by making it possible to recognize both time and place in the captioning phase, and by making it possible to distinguish more complex sentences by increasing the number of cases in the sentence structure in the word-by-word aggregation phase.

4. Conclusion

The proposed system showed improved accuracy when introducing OFA in the captioning phase while following the framework for generating explanatory text using keyframes. In the aggregation phase, we experimented with a word-by-word based Lexrank, but it could not achieve much more accuracy than the sentence-based Lexrank.

To further improve the accuracy in the future, using a method that can select clear keyframes in the keyframe generation phase, enabling the output of time and location in the captioning phase, and recognizing sentence structure in sentence-based Lexrank may prove effective.

References

[1] A. Shibata and T. Yukawa. An automatic text generation system for video clips using machine learning technique. In TRECVID 2017 VTT Task paper. Nagaoka University of Technology, 2018.

[2] G. F. Woodman and M. M. Chun. The role of working memory and long-term memory in visual search. Visual Cognition, Vol. 14, No. 4–8, pp. 808–830, 2006.

[3] O.l Vinyals et al. Show and tell: A neural image caption generator. Computing Research Repository, arXiv:1411.4555, 2015.

[4] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, Vol. 22, No. 1, pp. 457–479, 2004.

[5] I. Sutskever, O. Vinyals, and Q. Le. Sequence to sequence learning with neural

networks. Advances in Neural Information Processing Systems, Vol. 27, pp.3104–3112, 2014.
[6] P. Wang et al. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. Computing Research Repository, arXiv2202.03052, 2022.
[7] A. Vaswani et al. Attention is all you need. Advances in Neural Information Processing Systems, Vol. 30, pp. 5998–6008, 2017.