

Automatic Caption Generation for Video Clips Using Keyframe and Document Summarization Techniques

Masaki Hoshino

Takashi Yukawa

Team KsLab_NUT, Nagaoka University of Technology

1. Introduction

The VTT task requires to generate a single sentence that describes the content from a video.

We aim to generate caption with high precision and at the same time significantly reduce the number of frames used for the processing.

We propose a method that combines the generation of captions from keyframes extracted from a video with a technique to summarize them as a document.

2. Approach

System consists of three steps: keyframe extraction, caption generation, and caption aggregation.

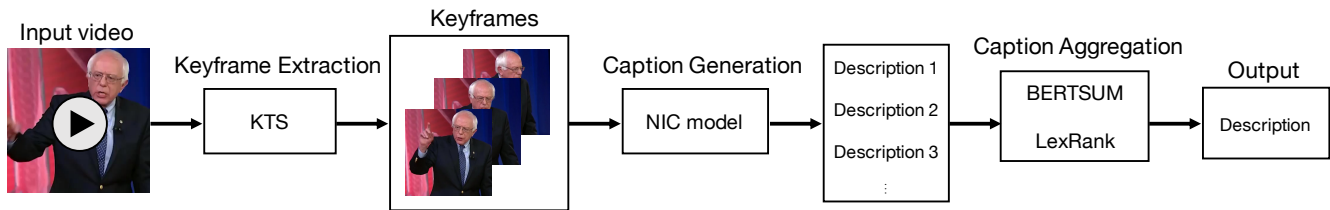


Figure 1. Our approach overview

1. Keyframe extraction step

The keyframes are extracted using Kernel Temporal Segmentation (KTS).

2. Caption generation step

Captions are generated for each key frame using the NIC model.

3. Caption aggregation step

Output a single caption using the extractive method used in video summarization tasks.

We compared the performance of two extractive methods, BERTSUM and LexRank.

3. Results

Table 1. Our Method Scores (VTT 2020 data)

Run	METEOR	CIDEr
run1.bsum.primary	0.195	0.137
run2.lex065	0.210	0.137

Table 2. Average frames per video (VTT2020 data)

	Use whole frame	Our run
Number of frames	147	5

Table 3. Scores by VTT2020 participating teams

Team	METEOR	CIDEr
RUC AIM3	0.310	0.538
PicSOM	0.262	0.319
MMCUniAugsburg	0.202	0.140
KsLab_NUT	0.195	0.137
IMFD_IMPREEE	0.194	0.087
KU_ISPL	0.191	0.074

4. Conclusion

From Table 1, proposed method achieves similar scores to many of the other teams' methods.

Also, there was no significant difference in the scores between BERTSUM and LexRank.

Table 2 shows that the proposed method reduces the processing frames by **96.6%** on average compared with the method using all video frames.

For further improvements in accuracy, possible approaches include revising KTS parameters, using abstractive methods to aggregate captions into a single caption rather than extracting a single sentence, and modifying the dataset used to train the NIC model.