KU-ISPL TRECVID 2020 VTT Model

Junyeop Lee[1], Youngsaeng Jin[1], Gwantae Kim[1], and Hanseok Ko[1]

Intelligent Signal Processing Laboratory, Korea University

Abstract. KU-ISPL model for TRECVID 2020 Video-to-Text (VTT) is presented in this paper. In this 2020 trecvid[1], we focused on the video captioning method in the end-to-end manner.We proceed by making 3 runs for our model by combining 2 types of models to explore how the each model impacts the performance of sentence generation and the basic operation. In run 1, we deal with the baseline represented by SA-LSTM, and in run 2 and run 3, we conduct by connecting transformer and lstm.Attention mechanism is exploited for best use of contextually pertinent frames in input video. The model pays attention to the hidden states of transformer in the encoder to obtain efficient hidden states in lstm as decoder. We only used vtt 2020 data as training data for sentence generation.Experimental results show that the combined model has the potential to achieve some performance in an end-to-end manner.

Keywords: Video captioning · Transformer · LSTM

1 Model

This section presents the baseline model that we explored. As mentioned above we use as a baseline the sa-lstm . run1 uses sa-lstm to extract features and concatenates each frame into IRv2 network when sequence video input comes in, and a sentence is generated by passing through the sa-lstm network individually.Run2 extracts features using a transformer as an encoder, passes through lstm with hidden state, and finally passes through the transformer acting as a decoder and outputs. Run 3 was used by attaching a transformer as an encoder and lstm as a decoder to verify the function of the transformer as a decoder.

2 Dataset

We included only the trecvid main dataset as training data to test the model's efficiency.

References

[1] George Awad et al. "TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains". In: *Proceedings of TRECVID 2020*. NIST, USA. 2020.