Automatic Generation of Video Descriptions Using DenseNet

Toshichika Mashimo
mashimo@nut-kslab.net

Seiji Nagumo
s181059@stn.nagaokaut.ac.jp

Mutsuki Ishii
ishii@nut-kslab.net

Takashi Yukawa
yukawa@vos.nagaokaut.ac.jp

Nagaoka University of Technology, Niigata, Japan

Abstract: The kslab team participated in TRECVID 2021 on the Video to Text (VTT) task. In the previous research, ResNet has been used for feature extraction. However, misrecognition of subjects can cause a decline in the precision of output sentences. This study aims to improve the recognition accuracy by using DenseNet for feature extraction instead. Although the accuracy improved, the output of the sentences was too similar to captions in the training data. Further improvements to this model in terms of the training dataset and the sentence aggregation method are pursued by the authors to solve this problem.

1. Introduction

The TRECVID Video to Text (VTT) task involves the automatic generation of explanatory text from videos, which has been studied extensively in the past.

Recently, with the development of deep learning, various text generation methods including NIC (neural image captioning) models such as the encoder-decoder model and LSTM (long short term memory), have been proposed. However, these methods require long processing time for videos with a large number of frames.

In previous work, Shibata and Yukawa[1] have succeeded in generating highly accurate captions while significantly reducing the number of frames used for processing by using only keyframes of the video.

2. Proposal of a New Image Feature Extraction Model in The Key-Frame Based Scheme

Our model consists of three steps: frames extraction from the video, text conversion for each frame, and aggregation of all the text. The model structure is shown in Figure 1.

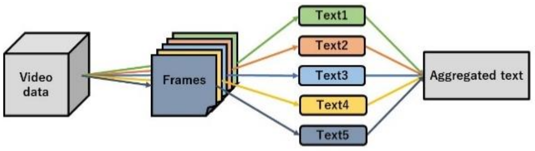

Figure 1. Model structure

According to past studies, Hoshino and Yukawa[2] found that text conversion for each frame (the second step in our model) had a significant effect on the scored quality of the final output. The score is lowered when the output contains words that are unrelated to the contents of the frame. Therefore, in the present work, the authors focused on improving the image feature extraction method, and for this reason changed the machine learning model from ResNet[3] to DenseNet[4].

ResNet and DenseNet are both convolutional neural network models for feature extraction. They have various computation layers, including convolution, pooling, and activation functions, and calculate residuals between input values through or not the layers. In Figure 2, ResNet architecture shows that each layer is connected in the next block, and in Figure 3, DenseNet architecture shows that it has connections to all layers after itself. This allows the features obtained in front of the layers to be propagated to the back of them, and this information in a partial pixel is reflected in the overall pixel.

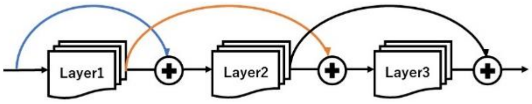

Figure2. ResNet architecture

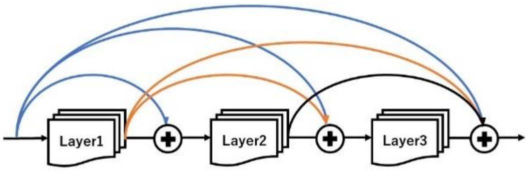

Figure3. DenseNet architecture

3. Evaluation Results

The authors tested two models, ResNet and DenseNet, for feature extraction using VTT2021 test datasets to calculate scores. In addition, MS COCO[5] datasets were used for model training, and BURTSUM[6] for text aggregation. Table 1 summarizes the METEOR, BLEU and CIDEr scores for each model.

Table 1. Test scores (VTT2021 data)

Model	METEOR	BLEU	CIDEr
ResNet	0.199	0.052	0.092
DenseNet	0.200	0.054	0.103

4. Discussion

The encoder used in the caption generation step of the NIC model was changed from ResNet to DenseNet. In ResNet, the subject of the sentence changed for the similar frames. This is because our model misinterprets a moving subject as a change to another subject. In contrast, the subject is more consistent in the sentences generated for each frame in DenseNet. This feature was particularly noticeable in videos with a small number of subjects. In addition, the accuracy was found to increase for all scores when using DenseNet.

However, even after changing the encoder from ResNet to DenseNet, the number of words in the caption may not be sufficient for a correct answer. Since the generated sentences were too short when MS COCO was used as a training dataset, it is expected that the accuracy could be improved by using a dataset that contains more detailed sentences or by applying a summarization method.

5. Conclusion

Our DenseNet VTT model was able to learn from the training data correctly as indicated by comparing the final sentences with the ground truth. However, this model does not expect to generate sentences with a sufficient number of words for each keyframe. These short sentences cause a lack of keywords in the final sentences because extraction method is used in the text aggregation step. It is necessary to generate new sentences by collecting high-priority words from captions for all keyframes. The authors will attempt to use a generative summarization method instead of an extractive summarization method for text aggregation as the next step.

References

[1] A.Shibata and T.Yukawa. An Automatic Text Generation System for Video Clips Using Machine Learning Technique. In TRECVID 2018 VTT Task paper.

[2] M.Hoshino and T.Yukawa. Automatic Caption Generation for Video Clips Using Keyframe and Document Summarization Techniques. In TRECVID 2020 VTT Task paper.

[3] K.He, X.Zhang, S.Ren, and J.Sun. Deep Residual Learning for Image Recognition. IEEE Conference on Pattern Recognition and Computer Vision (CVPR), 1026, pp. 770-778.

[4] G.Huang, Z.Liu, L.van and Maaten, and K.Q.Weinberger. Densely Connected Convolutional Networks. IEEE Conference on Pattern Recognition and Computer Vision (CVPR), 2017, pp. 4700-4708.

[5] T.-Y. Lin, M. Marie, S.Belongie, J.Hays, P.Perona, D.Ramanan, P.Dollar, and C.L.Zitnick. Microsoft COCO: Common Objects in Context. In European Conference on Computer Vision (ECCV), 2014, Lecture Notes in Computer Science, vol 8693. Springer, pp. 740-755.

[6] Y.Lie. Fine-tune BERT for Extractive Summarization. Computing Research Repository, arXiv:1903.10318, 2019