멀티 뷰 특징들을 활용한 수어 번역

Exploring Sign Language Translation Using Multi-View Features

Hyunjoo Jang, Jungeun Kim, Won-Yong Shin, Ha Young Kim*
Yonsei University

{hinjuuu, jekim5418, wy.shin, hayoung.kim}@yonsei.ac.kr

Abstract

There is a growing social interest in automatic sign language translation technology for deaf individuals who primarily communicate through sign language. Translation methods using deep learning models is being developed. To recognize a sign language, it is necessary to observe the face, hand shape, and hand movements, and even for the same sign word, the degree of movement varies depending on the individual. For this reason, not only a diverse and large amount of sign language image data is required, but understanding sign language videos is also challenging. Therefore, in this study, we examine whether features from various perspectives can enhance the performance of sign language translation. We use three types of input representation, which are global image, global keypoints, and hand keypoints, to extract muti-view features. Through experiments on PHOENIX-2014T, we demonstrate that using global image along with hand keypoints is effective in sign language translation. We expect that this study will be helpful in learning various multi-modal data and features from different perspectives.

I. Introduction

Sign language data is available in German, Chinese, and English, and artificial intelligence research on sign language translation using these videos is in progress. Sign language translation is a very challenging task as it requires not only the recognition of hand gestures but also facial expressions from the entire image, making the appropriate representation of input images crucial. Therefore, in this study, we explore the performance based on the combination of multi-view features. We used three types of features, the global image, global keypoints, and hand keypoints. These were combined using element-wise summation, self-attention, and cross-attention methods to diversify the embedding data to be learned. We conduct experiments on a sign language video dataset PHOENIX-2014T [1]. Through these experiments, we show that using global images and hand keypoints is effective in sign language translation. We believe this is because it allows for the emphasis of important hand gestures in sign language translation through cross-attention mechanism [2].

II. Method

This study is a gloss free translation that uses sign language videos and directly translates them into sentences without using gloss, which is a unit for expressing sign language. We extract various feature data from sign language images through element-wise summation, self-attention, and cross-attention to strengthen the representation of sign language video. Our baseline model is GFSLT [4] and is composed of an encoder and a decoder. GFSLT only uses global images as input data. The encoder is a module that embeds sign language images, and the decoder is a module that converts them into sentences. We modify the encoder of GFSLT to use two combinations of global image, global keypoints, and hand keypoints as inputs. The global image feature is representation of one image frame and the global image keypoint is a human pos estimation information. The hand keypoint is a information of the hand where sign language movements are concentrated.

We calculated the sign language translation performance of the four cases in Table1 by utilizing these encoding input and element-wise summation, self-attention, and cross-attention mechanisms. The first case is that the global image feature and the global keypoint feature were element-wise summed, and the second case is that we use self-attention mechanism with global image feature and then it was element-wise summed with the global image feature. The third case is that we use cross-attention mechanism with the global image keypoint and the global image feature and then it was element-wise summed with global image feature. The fourth case is that use cross-attention mechanism with the global image hand keypoint and then it was element-wise summed with global image feature. Lastly, we also calculated the performance of a case using global image features as input to the encoder to perform a baseline for comparison in the same experimental environment. We consist of cross-attention mechanism that value/key is global/hand keypoint and query is global image feature, respectively. Translation results were measured by calculating the Bilingual Evaluation Understudy (BLEU) score.

III. Experimental Results

We use PHOENIX-2014T dataset including sign language videos of daily news and weather forecasts from German public broadcaster PHOENIX. The total number of videos is 8,257, divided into 7,096 Train, 519 Dev, and 642 Test videos. Our experimental results showed that the last case in Table 1 was most effective. The cross-attention mechanism by using hand keypoints is effective rather than using single global image data.

Table 1. Experimental result

Features	Attention Mechanism	Bleu4 (Test)
Global Image Feature (Baseline)	-	**20.41**
Global Image Feature + Global Keypoint Feature	-	19.30
Global Image Feature + Global Image Attention Feature	Self-Attention	19.71
Global Image Feature + Global Keypoint Attention Feature	Cross-Attention	19.85
Global Image Feature + Hand Keypoint Attention Feature	Cross-Attention	**20.50**

Table 1 shows that the fourth case is more effective than the others. Additionally, it is superior to the baseline. This case utilizes cross-attention mechanism with the global image feature and hand keypoints feature, and then it was element-wise summed with global image feature. This is because, in addition to the hand motion information in the global image, it is possible to understand hand movements from a different perspective through hand keypoints.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2023R1A2C200337911 and No. RS-2023-00220762).

REFERENCES

[1] CAMGOZ, Necati Cihan, et al. Neural sign language translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 7784-7793.

[2] VASWANI, Ashish, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[3] PAPINENI, Kishore, et al. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002. p. 311-318.

[4] ZHOU, Benjia, et al. Gloss-free Sign Language Translation: Improving from Visual-Language Pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. p. 20871-20881.