MASTER'S THESIS

Deep networks for sign language video caption

Zhou, Mingjie

Date of Award:
2020

Abstract

In the hearing-loss community, sign language is a primary tool to communicate with people while there is a communication gap between hearing-loss people with normal hearing people. Sign language is different from spoken language. It has its own vocabulary and grammar. Recent works concentrate on the sign language video caption which consists of sign language recognition and sign language translation.

Continuous sign language recognition, which can bridge the communication gap, is a challenging task because of the weakly supervised ordered annotations where no frame-level label is provided. To overcome this problem, connectionist temporal classification (CTC) is the most widely used method. However, CTC learning could perform badly if the extracted features are not good. For better feature extraction, this thesis presents the novel self-attention-based fully-inception (SAFI) networks for vision-based end-to-end continuous sign language recognition. Considering the length of sign words differs from each other, we introduce the fully inception network with different receptive fields to extract dynamic clip-level features. To further boost the performance, the fully inception network with an auxiliary classifier is trained with aggregation cross entropy (ACE) loss. Then the encoder of self-attention networks as the global sequential feature extractor is used to model the clip-level features with CTC. The proposed model is optimized by jointly training with ACE on clip-level feature learning and CTC on global sequential feature learning in an end-to-end fashion. The best method in the baselines achieves 35.6% WER on the validation set and 34.5% WER on the test set. It employs a better decoding algorithm for generating pseudo labels to do the EM-like optimization to fine-tune the CNN module. In contrast, our approach focuses on the better feature extraction for end-to-end learning. To alleviate the overfitting on the limited dataset, we employ temporal elastic deformation to triple the real-world dataset RWTH-PHOENIX-Weather 2014. Experimental results on the real-world dataset RWTH-PHOENIX-Weather 2014 demonstrate the effectiveness of our approach which achieves 31.7% WER on the validation set and 31.2% WER on the test set.

Even though sign language recognition can, to some extent, help bridge the com-

munication gap, it is still organized in sign language grammar which is different from spoken language. Unlike sign language recognition that recognizes sign gestures, sign language translation (SLT) converts sign language to a target spoken language text which normal hearing people commonly use in their daily life. To achieve this goal, this thesis provides an effective sign language translation approach which gains state-of-the-art performance on the largest real-life German sign language translation database, RWTH-PHOENIX-Weather 2014T. Besides, a direct end-to-end sign language translation approach gives out promising results (an impressive gain from 9.94 to 13.75 BLEU and 9.58 to 14.07 BLEU on the validation set and test set) without intermediate recognition annotations. The comparative and promising experimental results show the feasibility of the direct end-to-end SLT.

Keywords: Deep Learning; Video Caption; Sign Language; Sign Language Recognition; Sign Language Translation; Machine Translation.

Table of Contents