

MASTER'S THESIS

Deep networks for sign language video caption

Zhou, Mingjie

Date of Award:
2020

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

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

Declaration	i
Abstract	ii
Acknowledgements	iv
Table of Contents	v
List of Tables	vii
List of Figures	viii
Chapter 1 Introduction	1
1.1 Sign Language Recognition	2
1.2 Sign Language Translation	4
1.3 Thesis Outline	6
Chapter 2 Related Work	8
2.1 Data Access	8
2.2 Sign Language Recognition	9
2.3 Sign Language Translation	11
Chapter 3 Preliminary	14
3.1 Convolutional Neural Networks	14
3.1.1 The Convolution Operation	15
3.1.2 Networks' Layers	15
3.2 Recurrent Neural Networks	17

3.2.1	The Gradients in Vanilla RNN	19
3.2.2	Long Short-term Memory and Gated Recurrent unit	21
3.3	Sequence-to-Sequence Learning	24
3.3.1	Sequence-to-Sequence Learning with RNNs	25
3.3.2	Attention-based Sequence-to-sequence Model with RNN	27
3.3.3	Self-attention-based Sequence-to-sequence Model	29
Chapter 4	Sign Language Recognition	32
4.1	Overview	32
4.2	The Proposed Method	33
4.2.1	Network Architecture	33
4.2.2	Clip-level Feature Learning	35
4.2.3	Global sequential feature learning	36
4.3	Experiments	37
4.3.1	Experimental Setup	38
4.3.2	Overfitting Reduction	40
4.3.3	Experimental Results	42
4.4	Conclusion	47
Chapter 5	Sign Language Translation	48
5.1	Overview	48
5.2	From Clip Order Prediction to Direct Sign-to-text Translation	48
5.3	Experiments	50
5.3.1	Experimental Setup	50
5.3.2	Experimental Results	52
5.4	Conclusion	53
Chapter 6	Future Outlook and Conclusion	54
6.1	Future Outlook	54
6.2	Summary	55
	Bibliography	57
	Curriculum Vitae	66