A Computational Approach to Sign Language Understanding

Tejaswini Ananthanarayana
Rochester Institute of Technology
Rochester, New York 14623
ta2184@rit.edu

Abstract

Sign language is the primary mode of communication in the Deaf and Hard-of-Hearing (DHH) communities. Unfortunately, sign language is not as well understood among the non-signing hearing population leading to limited access and services to the DHH community, and also acts as a barrier between non-signing and DHH people.

In my Ph.D. thesis, I am working on improving the sequence modeling for sign language translation and understanding by considering different types of sequence models, various input features, and by understanding the semantic relation between the words and the signs. Currently, my research focuses on a popular publicly available German Sign Language dataset.

Introduction

About 5% of the world population or 466 million people suffer from hearing loss (2018). The popular mode of communication between Deaf and Hard-of-Hearing (DHH) people is using sign language. Sign language is not popularly known across hearing people leading to loss of learning, employment, and other opportunities for the DHH community. My work mainly focuses on bridging this gap by using artificial intelligence (AI) to facilitate better communication between DHH and hearing people with the development of automatic sign language translation. The use of human interpreters for sign language translation can be inconvenient, difficult to schedule, or costly, hence the use of machine learning to develop automatic translators can be very beneficial to society.

Why is Sign Language Understanding Hard?

The most popular way to estimate the performance of a sign language understanding model is by comparing the ground truth and predicted words. But in real-time, this score will not be appropriate as from one signer to another the context of the sign may change and the words or sentences may be described in different ways. To understand better, we asked two American Sign Language (ASL) signers to annotate a few videos and compared all the results together.

No.	Ground truth caption	ASL Signer 1	ASL Signer 2
1.	He called and asked if they had the right kind of dog, and the answer was yes!	You can have	they shouted ”you have it?”, ”yes”.
2.	He keeps drinking coffee	Still making out	coffee continued still.
3.	Did the teacher buy a house yesterday?	Did the teacher buy a house yesterday?	Did the teacher buy a house yesterday?

Table 1: Comparison between ground truth and predicted captions between two human annotators. The green cells show agreement between any two interpretations while yellow shows disagreement (Best viewed in color).

Sign language translation (SLT) is hard because it is very context-based and even when two signers are shown the same signs they might not necessarily interpret them in the same way, potentially due to missing context. Table 1 demonstrates this discrepancy.

My thesis looks at different aspects of SLT (1) model development - focusing on sign video to spoken language text translation; (2) feature selection - focusing on finding the best features for unconstrained datasets; and (3) model interpretation - focusing on understanding the relationship between signs and specific features.

We are also interested in exploring which features contribute the most to SLT. For example, in ASL, fingerspelling can be heavily correlated with hand joints, whereas question-asking can heavily rely on facial expressions. My work focuses on understanding the connection between different types of signs and features in order to improve model learning.

Current Work

My current research work explores sequence-to-sequence models and attention-rich transformer model. In my thesis, I study how the model's performance is affected by using different types of input features. We extract OpenPose (2017) features obtaining body, hand and face joint locations, CNN features pretrained on ImageNet (2012) to extract a multi-dimensional feature vector from the visual

frames, and Means cluster IDs (1982) based on OpenPose joint locations. To date, our research work is based on German Sign Language (GSL), a popularly available dataset collected from weather forecast airings from the RWTH-PHOENIX-Weather dataset (2018). We will extend it to ASL in the future. Many other SLT models work on finger spelling, single sign recognition, digit recognition etc. This work focuses on continuous SLT, a more comprehensive problem.

We use OpenPose (2017) as features to the SLT model and not for its original designed task of keypoints detection. It provides 25 body points, 21 points for each hand, and 70 face points. We perform frame-to-frame smoothing of the OpenPose points. Depending upon the dataset, it can be seen that sometimes the person signing can be close to the camera, away from the camera, or shifted from the center. This variation is taken care of by generating a canonical form of the frames. The canonical form scales all the points to the same size and also centers the points at the origin as shown in Fig. 1.

We perform ablations using a combination of body, hands, and face joints to understand which features contribute the most towards effective sign language translation. These ablations as shown in Table 2 is performed using a basic sequence-to-sequence model without attention on the GSL dataset. From the results in Table 2 it is clear that for the GSL dataset, hands, body, and face points together contribute to the best performance in general, but we are also interested in understanding which of the different features perform well for SLT. We are interested in the mappings between visual embeddings of the different features with the language embeddings of the translations.

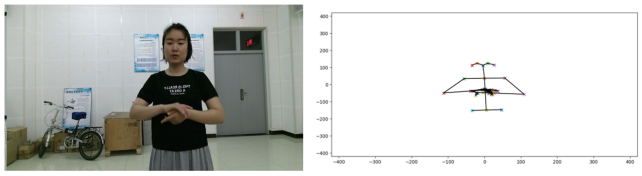

Figure 1: OpenPose canonical representation on a proprietary dataset. (P.S: Please zoom in for better visualization)

OpenPose Features	Set	BLEU 1	BLEU 4
Hands	Validation	7.31	0.07
	Test	12.14	0.12
Body	Validation	4.63	0.01
	Test	6.29	0.01
Face	Validation	2.54	0.01
	Test	2.23	0.01
Hands + Body	Validation	18.11	4.06
	Test	20.68	4.33
Hands + Face	Validation	18.13	4.06
	Test	17.68	4.04
Hands + Body + Face	Validation	**23.31**	**5.68**
	Test	**23.49**	**5.55**

Table 2: OpenPose ablation results on sequence-to-sequence model without attention on the GSL dataset.

Additionally, we experimented with a human as an oracle to compare how a human can predict the captions on a sign language video versus OpenPose joints visualization for each frame rendered as a video. The OpenPose joints are mainly used for this experiment because the OpenPose annotated body, hands, and face joints are highly interpretable by humans. For this experiment we use the American Sign Language (ASL) dataset (2012) as the signers available did not understand GSL.

Video type	BLEU 1	BLEU 2	BLEU 3	BLEU 4
RGB videos	24.86	12.17	6.87	3.81
OP videos	6.32	1.790	0.59	0.14

Table 3: Human as an oracle experiment on ASL dataset for RGB and OpenPose (OP) generated videos.

As a part of a human as an oracle experiment, we asked ASL signers to interpret regular RGB sign videos as well as skeleton videos created with OpenPose points from the RGB videos. Table 3 shows the BiLingual Evaluation Understudy (BLEU) (2002) scores obtained. The low BLEU scores for the RGB videos can be attributed to the context and to different ways of predicting the spoken sentence for a sign. The BLEU score comparison is between the human prediction and the ground truth.

Future Work

My future research work is mainly focused on: expanding the sign language dataset to enable training on a larger dataset; exploring optical flow like features; understanding and finding the relationship between visual features and word embeddings; understanding under what circumstances facial, hand joints, body features are activated; exploring the visual embeddings and word embeddings to see how they correlate while performing sign language translation.

References

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28.

Necati Cihan Camgöz, Simon Hadfield, O. K. H. N. R. B. 2018. RWTH-PHOENIX-Weather 2014 T: Parallel Corpus of Sign Language Video, Gloss and Translation.

Neidle, C.; and Vogler, C. 2012. American Sign Language Dataset. http://www.bu.edu/asllrp/. [Online; accessed 9-May-2020].

Organization, W. H. 2018. https://www.who.int/news-room/facts-in-pictures/detail/deafness/. [Online; accessed 04-Jan-2020].

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL*.