2023 32nd IEEE International Conference on Robot and
Human Interactive Communication (RO-MAN)
August 28-31, 2023. Paradise Hotel, Busan, Korea

2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) | 979-8-3503-3670-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/RO-MAN57019.2023.10309531

Improving Sign Language Understanding Introducing Label Smoothing

Tan Sihan[1], Khan Nabeela Khanum[1], Itoyama Katsutoshi[1,2], Nakadai Kazuhiro[1]

Abstract— **Sign language is one of the most important communication methods when considering equality, diversity, and inclusion. Sign language understanding implies understanding sign language using machines, and it involves mainly two functions; sign language recognition and sign language translation. To improve sign language understanding performance, this paper proposes to use label smoothing with CTC (Connectionist Temporal Classification) loss as training criteria for the sign language understanding neural network. Experimental results showed the effectiveness of the proposed method in both sign language recognition and translation.**

I. Introduction

As the native language for deaf and hard-of-hearing people to communicate, Sign Language (SL) plays an indispensable role in their daily lives. However, they suffer from language barriers nowadays and are encouraged to use spoken language, i.e., text-based communication and lip-reading. The truth is that sign language has been developed independently and does not share the same grammar with their spoken counterparts [19], which ignores the interests of the deaf communities who overwhelmingly favor using signed languages for daily contact in person and online, as well as when communicating with spoken language groups.

SL understanding has been established with the aim of better communication between the deaf and hearing communities. Generally, SL Understanding requires two functions: Sign Language Recognition (SLR) and Sign Language Translation (SLT). SLR is to recognize sign actions from given videos, and SLT is to generate spoken language sentences from the information embedded in the sign videos or sign representations.

However, the development between the SLR and SLT is unbalanced. SL Understanding has been mainly focused on the visual aspect, with little Natural Language Processing (NLP) involved, that is, it just recognizes each sign action, but ignores the information of the spoken syntax behind it. For this, SLR systems cannot grasp the underlying spoken grammar and complexity of sign language on their own, and SLT faces the additional challenge of considering the unique linguistic features during translation. Although several studies [2, 3, 21, 23] have been done on the SLT, their attention is still insufficient. Thus, to achieve better and complete SL Understanding, we strongly suggest that SLT should be given more attention, and it is of great importance to integrate SLR and SLT tasks for better SL Understanding.

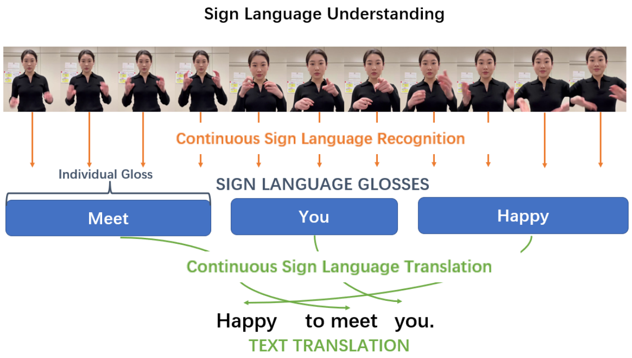
Sign Language Understanding
Continuous Sign Language Recognition
Individual Gloss
SIGN LANGUAGE GLOSSES
Meet
You
Happy
Continuous Sign Language Translation
Happy
to meet
you.
TEXT TRANSLATION

Fig. 1. Sign Language Understanding pipeline. The Sign Language Understanding system conducts CSLR task as the first step to tokenize the input video into glosses. The next step is to translate the glosses into the corresponding spoken language text.

The contributions of this paper can be summarized as follows:

- We study SL Understanding defined as the integration of SLR and SLT, providing significant and thorough insights to the related tasks at hand and pointing out the future research direction.
- We review public datasets for SL Understanding and display their features, specifying the limitations of the publicly available datasets.
- We propose to introduce label smoothing to Connectionist Temporal Classification (CTC) loss as sequence learning training criteria for SL Understanding to mitigate the overfitting problem in SL Understanding, and experimental results indicates the effectiveness of the proposed method.

The rest of this paper is organized as follows: Section II involves the preliminary SL Understanding for a better comprehension of the whole picture. Section III discusses the related work. Furthermore, Section IV provides the review of mainstreaming datasets for SL Understanding. In Section V, a detailed description of the proposed sequence learning training criteria for SL Understanding is provided. In the following Section, we report the evaluation results. Finally, the conclusion and future research direction are drawn.

II. Sign Language Understanding

Despite considerable advancements achieved in machine translation (MT) between spoken language [6, 22] and computer vision in the classification task [7], SL Understanding lags behind for many reasons. Different from spoken

This work was supported by JSPS KAKENHI Grant No. JP19KK0260, JP20H00475 and JP23K11160.

1 Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, 2-12-1-W8-30 Ookayama, Meguro-ku, Tokyo, 152-8552, JAPAN. {tansihan,nabeela,itoyama,nakadai}@ra.sc.e.titech.ac.jp

2 Honda Research Institute Japan Co., Ltd.

language and motion videos, the multidimensional feature of sign language poses additional challenges for computer vision because it relies on both manual ($i.e.$, hand shape, position, movement, orientation of the palm or fingers) and non-manual ($i.e.$, eye gaze, head-nods/shakes, shoulder orientations, various types of facial expression as mouthing and mouth gestures) signals. While spoken language follows a sequential pattern in which words are processed one at a time, these cues might happen concurrently. Moreover, signs fluctuate in space and time, and the number of video frames corresponding to a single sign is likewise not constant. A complete SL Understanding system involves the following sub-tasks:

Sign Language Glossing: Glossing is the process to transcribe sign language word-for-word by another written language. Sign language videos can be divided into different segments, each representing a gloss, a word with an independent meaning. As glosses merely indicate what part of the sign language sentence means but do not form an appropriate sentence in spoken language, they differ significantly from spoken text.

Sign Language Recognition: After sign language glossing, what the SLR system needs to do is recognizing the gloss meaning. Generally, SLR consists of isolated SLR and continuous sign language recognition (CSLR) [1]. Isolated SLR is to identify the isolated single signs in videos, and is much like action recognition. At the same time, CSLR is a more challenging task that recognizes the sequence of glosses that are present in a continuous/non-segmented video sequence. The input of the sign language model is high dimensional spatio-temporal data, and the model needs to understand what a signer looks like and what their signs mean, and then comprehend what the sign means in combination.

Sign Language Translation: Once the system has understood the meaning of the sign language video, the final step is to generate the spoken language sentence. Like any other natural language, sign languages have their own grammatical and linguistic structures that frequently do not correspond to those of their spoken language counterparts. Hence, in a real sense, this issue is a machine translation work. Figure. 1. demonstrates the SL Understanding pipeline, the function of glossing, sign language recognition, and translation.

III. Related Work

Based on the demonstration above, we overview the related works in SL Understanding in this section.

A. Isolated Sign language recognition

The objective of isolated SLR is to deal with the video segment classification (where the segment boundaries are provided), based on the fundamental assumption that a single gloss is present [4].

B. Continuous Sign Language Recognition

Glosses in sign language have shorter durations than actions ($i.e.$, they may only contain a very small number of frames), and the transitions between them are frequently very subtle, making it difficult to identify their temporal bounds accurately. Additionally, CSLR is usually characterized as a weakly supervised learning undertaking due to the lack of gloss-level annotations.

In most CSLR systems, a feature extractor is typically followed by a temporal modeling mechanism. The feature extractor is used to obtain the feature representations from the individual input frames [5] or sets of neighboring frames [16]. Meanwhile, the SL unit feature representations ($i.e.$, gloss-level and sentence-level) can be modeled thanks to temporal modeling techniques. Sequence learning for temporal modeling can be accomplished using HMMs, CTC [11], or Dynamic Time Warping (DTW) techniques. Given that CTC has consistently demonstrated superior performance than the aforementioned ones, it has been established as the principal sequence training criteria in the majority of CSLR research. However, CTC often results in overconfident peak distributions that are prone to overfitting, and provides limited contribution to the feature extractor's optimization [24].

Considering the uniqueness of sign language and limitations of available dataset and CTC, CSLR is quite challenging task. Not only the spatial information from the sign videos need to be extracted but also it is crucial to take into account the temporal relationships between different signs in the videos.

C. Sign Language Translation

Generally, sequence-to-sequence-based SLT methods can be classified into following protocols:

Gloss2Text: it is a text-to-text task on which the objective is to translate the ground truth gloss sequences to the spoken language sentences.

Sign2Gloss→Gloss2Text: Sign2Gloss model is trained first, and the Gloss2Text model is trained on ground truth glosses (independently of the Sign2Gloss model), then during inference make predictions conditioned on the output of the Sign2Gloss model. In [2] Camgoz $et\ al.$ use this protocol to do SLT, and they assert that Gloss2Text model should be the upper bound for translation performance, but in this assertion, ground truth gloss annotations are treat as the fully understanding of sign language, ignoring the information bottleneck in glosses.

Sign2Gloss2Text: This is currently the most mainstream and state-of-the-art approach in SLT. This approach use the glosses extracted from the sign videos by the CSLR model. Then, the translation task is converted into text-to-text problem, which can be solved by utilizing the Gloss2Text network trained by the CSLR predictions. However, same as the Sign2Gloss→Gloss2Text, an information bottleneck is inevitably introduced since this method uses sign glosses as intermediate supervision.

Sign2(Gloss+Text): When the original sign video is transformed into glosses, some spatio-temporal information is lost for the following SLT task. To relieve the problem above, [3] introduced a new protocol: Sign2(Gloss+Text), and this protocol follow the same naming convention. Sign2(Gloss+Text) is the joint learning of continuous sign language recognition and translation in an end-to-end manner. Camgoz $et\ al.$, introduce Transformer to SL Understaing. This model achieves encouraging results in translation, however, their CSLR performance is sub-optimal, with a higher Word Error Rate than baseline models, which suggests their model may be weaker in processing the videos [23] or there are limitations in their training criteria.

Sign2Text: It is the end goal of SLT. The objective of sign2text is to translate directly from the continuous sign videos to spoken language sentences without using any intermediary representation ($e.g.$, glosses).

IV. Datasets

The lack of sufficient annotated datasets is one of the most significant challenges that has restricted the advancement of SL Understanding research. Depending on whether annotations are provided at the gloss-level, continuous gloss-level, or in the form of spoken text, existing SL datasets can be divided into three categories: isolated SLR dataset, CSLR dataset, and SLT dataset. Besides, those datasets can also be classified as Signer Dependent (SD) or Signer Independent (SI) based on the evaluation scheme. For instance, a signer cannot be present in both the training and test sets in the SI datasets. Table I depicts the most well-known public SLR datasets, together with their essential characteristics.

Isolated SLR Dataset: The isolated SLR datasets are particularly crucial for certain scenarios ($e.g.$, creating a sign language dictionary, or for teaching purposes).

The isolated Signum dataset [20], the isolated Chinese Sign Language (CSL) Dataset [18], and the isolated Greek Sign Language (GSL) dataset [1] consist of frequent daily glosses, and are recorded in the predefined environment. Meanwhile, the American Sign Language (ASL) dataset [13] is a real-life large-scale isolated sign language dataset, they constitute challenging material with large variation in view, background, lighting and positioning, since they are not official recordings. The contents of this ASL dataset is from the ASL tutorial books.

CSLR Dataset: Most of the daily life communications require continuous sign language, and because of this, a number of CSLR datasets has been released for linguistic purposes.

Most CSLR datasets are about interactions between the deaf in daily life and recorded in prefined enviroments with stationary equipment.

SLT Dataset: PHOENIX Weather 2014 T (PHOENIX14T) dataset [2]. It is an extension of the PHOENIX14 corpus [9], originating from the weather forecast domain, focusing on sign language translation, which has recently become the primary benchmark for CSLR and CSLT. It consists of parallel sign language videos, gloss annotations and their corresponding translation. Additionally, How2sign dataset [8] provides multimodal and multiview continuous American SL for CSLR and SLT.

In summary, the current publicly available datasets are constrained by one or more of the following:

- Limited vocabulary size.
- Short video or total duration.
- Restricted domain.
- Lack of corresponding spoken language text.

V. Proposed method

The goal of SL Understanding ($i.e.$, CSLR and SLT) is to generate continuous glosses and spoken language text from sign video.

Given a sign language video $\mathcal{V} = (v_1, ..., v_T)$ with T frames, the CSLR model learns the probabilities $p(\mathcal{G}|\mathcal{V})$ of predicting a sign gloss sequence $\mathcal{G} = (g_1, ..., g_N)$ with N glosses and a spoken language sentence $\mathcal{S} = (s_1, ..., s_X)$ with x words. Modeling these conditional probabilities is a challenging undertaking as it is a sequence-to-sequence task. Also, the sequence length of the source token is much larger than that of the target, that is, $T \gg N$ and $T \gg X$. One way to train the CSLR network would be using cross-entropy loss with frame level annotations. However, there is no corpus that has frame-level annotations and constructing such precisious datasets is also a challenging task. An alternative form of weaker supervision is to use a sequence-to-sequence learning loss functions, CTC.

A. Vanilla CTC Criterion

CTC trains the neural network by computing a maximum-probability training criterion over all possible alignments. The probability of the possible label sequence is modeled as being conditionally independent by the product of each label probability. CTC is widely utilized for labelling unsegmented sequences ($e.g.$, speech recognition and optical character recognition).

In CTC-based CSLR task, CTC introduce $blank$ label, representing the slience or or transition between two consecutive gloss. The extended glosses can be defined as $\mathcal{G} = (g_1, ..., g_N) \cup \{blank\} \in R^L$, where L is the total number of labels.

The CTC is utilized to compute the $p(\mathcal{G}|\mathcal{V})$, marginalizing over all possible $\mathcal{V}$ to $\mathcal{G}$ alignment as:

$$p(\mathcal{G}|\mathcal{V}) = \sum_{\pi \in \mathcal{B}} p(\pi|\mathcal{V}), \tag{1}$$

where π is a path and $\mathcal{B}$ is the collection of all possible paths that lead to $\mathcal{G}$. The CTC loss in CSLR can be defined as:

$$\mathcal{L}_{ctc} = 1 - p(\mathcal{G}^*|\mathcal{V}), \tag{2}$$

where $\mathcal{G}^*$ is the ground truth gloss sequence.

Although CTC-based CSLR methods provide remarkable training convenience, overfitting is one of the main issues with CTC-based approaches, which results in insufficient training of the feature extractor.

In sign language, a main problem is large variance, individual differences tend to be larger than speech, and no rules in motion between glosses exists. These make sign languages characterized as data with large variance. Also, the size of training dataset is extremely smaller than datasets in other fileds. These are unique properties of sign languages, especially, the latter property leads to overfitting.

TABLE I

LARGE-SCALE PUBLICLY AVAILABLE SL UNDSTANDING DATASETS

Datasets	Language	Signers	Vocabulary size	Instances	Duration(h)	Resolution	fps	Type	Modalities
ASL 100 [13]	English	189	100	5,736	5.55	Varying	Varying	Isolated SLR	RGB
ASL 1000 [13]	English	**222**	1,000	25,513	24.65	Varying	Varying	Isolated SLR	RGB
CSL SD [12]	Chinese	50	178	25,000	**100+**	**1920x1080**	**30**	CSLR	**RGB+D**
CSL SI [12]	Chinese	50	178	25,000	**100+**	**1920x1080**	**30**	CSLR	**RGB+D**
Isolated CSL [18]	Chinese	50	500	**125,000**	67.75	**1920x1080**	**30**	Isolated SLR	**RGB+D**
GSL SD [1]	Greek	7	310	10,290	9.59	848x480	**30**	CSLR	**RGB+D**
GSL SI [1]	Greek	7	310	10,290	9.59	848x480	**30**	CSLR	**RGB+D**
Isolated GSL [1]	Greek	7	310	40,785	6.44	848x480	**30**	Isolated SLR	**RGB+D**
How2Sign [8]	English	11	**16,000**	35,000	79	1280x720	**30**	**CSLR+SLT**	**RGB+D**
PHOENIX14 SD [9]	German	9	1,231	6,841	10.71	210x260	25	CSLR	RGB
PHOENIX14 SI [9]	German	9	1,117	4,667	7.28	210x260	25	CSLR	RGB
PHOENIX14-T [2]	German	9	1,231	8,257	10.53	210x260	25	**CSLR+SLT**	RGB
Signum SI [20]	German	25	780	19,500	55.3	776x578	**30**	CSLR	RGB
Isolated Signum [20]	German	25	455	11,375	8.43	776x578	**30**	Isolated SLR	RGB

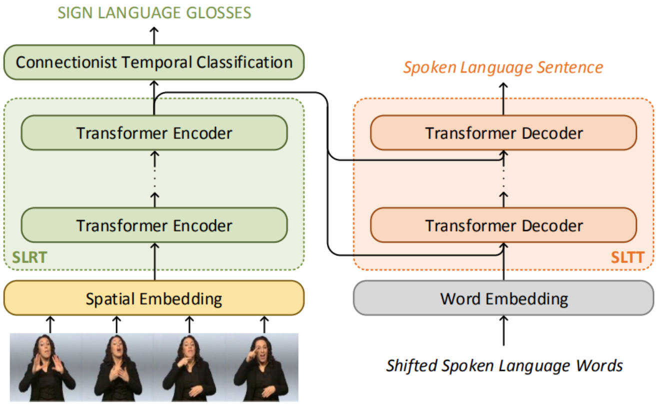

Fig. 2. : An overview of the end-to-end Sign Language Recognition and Translation transformers [3], Sign Language Recognition Transformer (SLRT), a vanilla transformer encoder model trained using a CTC loss, to predict sign gloss sequences. These obtained spatio-temporal representations from SLRT are then fed to the Sign Language Translation Transformer (SLTT), an autoregressive transformer decoder model trained by translation loss to predict one word at a time to generate the corresponding spoken language sentence.

B. CTC Criterion with Label Smoothing

To mitigate the overfitting problem mentioned above, we consider adopting a regularization technique called label smoothing. It introduces noise for the labels and changes the construction of the true probability. In [14], a CTC with label smoothing criterion is used for improving end-to-end speech recognition, and we bring this idea to CSLR.

To do so, we add a regularization term to the CTC objective function which consists of the Kullback-Leibler (KL) divergence between the network's predicted distribution P_n and a uniform distribution $\mathcal{F}$ over labels.

$$\mathcal{L}_{ctc_{new}} = (1-\alpha)\mathcal{L}_{ctc} + \alpha \sum_{t=1}^{T} D_{KL}(P_n||\mathcal{F}), \quad (3)$$

where α is tunable parameter for balancing the weight regularization term and CTC loss.

For the following evaluation, we apply the CTC loss with label smoothing to [3], and modify their joint loss, the overview of their model is shown in Figure 2.

In SLT tasks, the SLT model starts to predict one word at a time until it generates the special end-of-sentence token $< eos >$. By breaking down the sequence-level condtional probability into $p(\mathcal{S}|\mathcal{V})$ the ordered conditional probabilities, and the formula is as follow:

$$p(\mathcal{S}|\mathcal{V}) = \prod_{i=1}^{I} p(w_x|h_x), \quad (4)$$

where $p(w_x|h_x)$ denotes the ordered conditional probability at step x.

As for the translation training loss, we keep cross-entropy loss for each word as:

$$\mathcal{L}_{\mathcal{T}} = 1 - \prod_{x=1}^{X} \sum_{d=1}^{D} p(\hat{w}_x^d) p(w_x^d|h_x), \quad (5)$$

where $p(\hat{w}_x^d)$ is the ground-truth probability of the word w at step n and D is the target language vocabulary size.

The networks are trained by minimizing the joint loss term $\mathcal{L}$, which is the weighted sum of the translation loss $\mathcal{L}_{\mathcal{T}}$ and the recognition loss $\mathcal{L}_{\mathcal{R}}$ as follows:

$$\mathcal{L}_{original} = \lambda_R \mathcal{L}_{ctc} + \lambda_T \mathcal{L}_{\mathcal{T}}. \quad (6)$$

After introducing the CTC loss with label smoothing, the modified joint loss is defined as follows:

$$\mathcal{L}_{new} = \lambda_R((1-\alpha)\mathcal{L}_{ctc} + \alpha \sum_{t=1}^{T} D_{KL}(P_t||\mathcal{F})) + \lambda_T \mathcal{L}_{\mathcal{T}}, \quad (7)$$

where λ_R and λ_T are hyperparameters that determine recognition and translation loss functions relative weight during training, and α is the hyperparameters to decide the weight of label smoothing, we do the ablation experiments to evaluate the effect of λ_R, λ_T, and α.

VI. Evaluation

In this section, we conducted the evaluation on the state-of-the-art sign language model with our proposed criterion. We first go through the framework of the evaluation model, and then introduce the evaluation metrics used to measure the proposed CTC Loss for SL understanding. After that, we discuss the experimental results.

TABLE II
IMPACT OF THE LOSS WEIGHT VARIANTS.

Loss Weights		**w/o Label Smoothing**		**w/ Label Smoothing**	
λ_R	λ_T	WER	BLEU-4	WER	BLEU-4
1.0	1.0	42.69	19.94	63.47	20.33
5.0	1.0	**28.46**	**20.55**	29.61	21.76
10.0	1.0	28.90	21.04	**28.53**	**21.77**
20.0	1.0	32.80	19.67	32.75	19.93

TABLE III
IMPACT OF α VARIANTS

Label Smoothing Weight	**WER**	**BLEU-4**
α = 0.0005	30.03	20.58
α = 0.005	31.44	21.00
α = 0.01	**29.61**	**21.76**
α = 0.1	31.86	21.56

A. Evaluation Model and Set up

We choose the state-of-the-art sign language transformer **(sign2(gloss+text))** [3], a multi-task sign language transformer, that jointly train SLR and SLT models at the same time as the baseline.

In the experiments, we keep the same implementation as the baseline model with three transformer encoder and decoder layers, and each layer has 512 hidden units and 8 attention heads. The batch size, initial learning rate, and dropout rate are set to 32, 1e-3, and 0.1, respectively. To optimize the model, we employ Adam optimizer [15] with β_1 = 0.9, β_2 = 0.998, a learning rate schedule, and early stopping. We use Xavier initialization [10] and train the baseline model with the original loss and our proposed loss from scratch.

B. Evaluation Metrics

As for the evaluation metrics, we follow similar evaluations in speech recognition and MT. The most common measure of CSLR performance is Word Error Rate (WER), thus we use to evaluate the performance of our recognition models. The WER can be computed as:

$$\text{WER} = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}, \tag{8}$$

where S, D, I, C, and N indicate the number of **S**ubstitutions, **D**eletions, **I**nsertions, **C**orrections, and words in the reference, respectively.

We also used BLEU [17] score (n-grams ranging from 1 to 4), the most used MT metric, to evaluate the final performance of total training loss $\mathcal{L}_{new}$.

C. Results

We conducted the experiments to evaluate our proposed CTC loss with label smoothing in SL understanding on the PHOENIX14-T dataset [2], and also we performed the ablation experiments to verify the effects of hyperparameters λ_R, λ_T, and α.

We first set the label smoothing ratio α to 0.01, and evaluated the model by changing the λ_R and λ_T weights with and without label smoothing. Table II shows the results. The BLEU-4 scores get improved after we introduced the proposed SL understanding loss. For model with label smoothing, when λ_R = 10.0 and λ_T= 1.0, the results were the best in both CSLR and SLT.

We also verified the impact of label smoothing ratio α, in the following experiment, we set λ_R and λ_T to 5.0 and 1.0, respectively. Table III illustrates the results of chaning label smoothing ratio α.

D. Discussion

In this section, we evaluated our proposed SL understanding loss, CTC loss with label smoothing, which improved the baseline model's performance in SLT on the mainstreaming PHOENIX14-T dataset.

During experiments, we found that even when the WER is higher or far higher than that without label smoothing ($i.e.$, when WER is 29.61 and 63.47), the SLT model still achieves a better performance, which shows that a perfect CSLR system is not always necessary to lead to better SLT outcomes in this SL Understanding task**(sign2(gloss+text))**. The addition of label smoothing leads SLR model to extract more spatio-temporal representations for the following SLT task.

The results guide us to question the current mainstreaming method (**Sign2Gloss2Text**) in SL Understanding, using glosses as an intermediate representation to get predicted translation text, since glosses themselves are sub-optimal supervisions in SL understanding tasks. The process of converting SL frames into glosses inevitably introduces spatio-temporal information loss to SL Understanding.

Besides, for the first stage, we only tried limited combinations of hyperparameters λ_R, λ_T, and α. Still, more combinations of those hyperparameters need to be explored in future work.

VII. CONCLUSION

In this paper, we go through the current status of CSLR and SLT, including their methods, datasets available in the public, and limitations. We urge that more attention be given to SLT for better and more comprehensive SL Understanding.

Considering the limitation of CTC-based SL Understanding and the uniqueness of sign language, we modify the current training loss with label smoothing. We perform evaluations on the state-of-the-art SL Understanding Transformer model using the mainstream dataset, and our proposed loss leads to better results in terms of the BLEU-4 score.

During experiments, we find that a perfect CSLR model is not necessary for a well-performed SLT model, since choosing glosses as mid-representation in SL understanding may introduce information loss.

End-to-end training, relying less on gloss supervision, is a promising step towards better results in SL Understanding, we will continue the work on end-to-end joint training of the recognition and translation so that the CSLR model can extract more spatio-temporal representation to optimize the SLT model. Besides, employing a less information-losing sign language annotation approach is also worth considering.

References

[1] Nikolas Adaloglou et al. “A comprehensive study on sign language recognition methods”. In: *arXiv preprint arXiv:2007.12530* 2.2 (2020).

[2] Necati Cihan Camgoz et al. “Neural sign language translation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7784–7793.

[3] Necati Cihan Camgoz et al. “Sign language transformers: Joint end-to-end sign language recognition and translation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10023–10033.

[4] Necati Cihan Camgoz et al. “Using convolutional 3D neural networks for user-independent continuous gesture recognition”. In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. 2016, pp. 49–54.

[5] Runpeng Cui, Hu Liu, and Changshui Zhang. “Recurrent convolutional neural networks for continuous sign language recognition by staged optimization”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7361–7369.

[6] Jacob Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[7] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[8] Amanda Duarte et al. “How2Sign: a large-scale multimodal dataset for continuous American sign language”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 2735–2744.

[9] Jens Forster et al. “Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-Weather”. In: *LREC*. 2014, pp. 1911–1916.

[10] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[11] Alex Graves et al. “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks”. In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 369–376.

[12] Jie Huang et al. “Video-based sign language recognition without temporal segmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[13] Hamid Reza Vaezi Joze and Oscar Koller. “MS-ASL: A large-scale data set and benchmark for understanding american sign language”. In: *arXiv preprint arXiv:1812.01053* (2018).

[14] Suyoun Kim et al. “Improved training for online end-to-end speech recognition systems”. In: *arXiv preprint arXiv:1711.02212* (2017).

[15] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[16] Pavlo Molchanov et al. “Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4207–4215.

[17] Kishore Papineni 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, pp. 311–318.

[18] Junfu Pu, Wengang Zhou, and Houqiang Li. “Sign language recognition with multi-modal features”. In: *Advances in Multimedia Information Processing-PCM 2016: 17th Pacific-Rim Conference on Multimedia, Xi’an, China, September 15-16, 2016, Proceedings, Part II*. 2016, pp. 252–261.

[19] William C Stokoe Jr. “Sign language structure: An outline of the visual communication systems of the American deaf”. In: *Journal of deaf studies and deaf education* 10.1 (2005), pp. 3–37.

[20] Ulrich Von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. “The significance of facial features for automatic sign language recognition”. In: *2008 8th IEEE international conference on automatic face & gesture recognition*. 2008, pp. 1–6.

[21] Andreas Voskou et al. “Stochastic transformer networks with linear competing units: Application to end-to-end SL translation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11946–11955.

[22] Yonghui Wu et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: *CoRR* abs/1609.08144 (2016). URL: `http://arxiv.org/abs/1609.08144`.

[23] Kayo Yin and Jesse Read. “Better sign language translation with STMC-transformer”. In: *arXiv preprint arXiv:2004.00588* (2020).

[24] Hao Zhou, Wengang Zhou, and Houqiang Li. “Dynamic pseudo label decoding for continuous sign language recognition”. In: *2019 IEEE International conference on multimedia and expo (ICME)*. 2019, pp. 1282–1287.