TTIC's WMT-SLT 22 Sign Language Translation System

Bowen Shi
TTI-Chicago
bshi@ttic.edu

Diane Brentari
Univeristy of Chicago
dbrentari@uchicago.edu

Greg Shakhnarovich
TTI-Chicago
greg@ttic.edu

Karen Livescu
TTI-Chicago
klivescu@ttic.edu

Abstract

We describe TTIC's model submission to WMT-SLT 2022 task (Müller et al., 2022) on sign language translation (Swiss-German Sign Language (DSGS) → German). Our model consists of an I3D backbone for image encoding and a Transformer-based encoder-decoder model for sequence modeling. The I3D is pretrained with isolated sign recognition using the WLASL dataset. The model is based on RGB images alone and does not rely on the pre-extracted human pose. We explore a few different strategies for model training in this paper. Our system achieves 0.3 BLEU score and 0.195 Chrf score on the official test set.

1 Introduction

Sign language, a full-fledged natural language that conveys meaning through gestures, is the primary chief of communication among Deaf people. Sign language translation is a task for automatically translating sign languages into written languages. Due to its widespread potential applications, it has recently received growing research interest (Camgoz et al., 2018, 2021).

Existing methods for sign language translation are primarily based on gloss, a transliteration system annotating sign language with symbols from written language. Utilizing gloss usually boosts the performance of current translation systems by a large margin. In the widely used German sign language translation benchmark Phoenix14T (Camgoz et al., 2018), state-of-the-art gloss-based models (Chen et al., 2022) are roughly 15 points better (in Bleu-4) than gloss-free models (Camgoz et al., 2018). However, gloss is more expensive to annotate than written language translation. There have been relatively few amounts of studies for gloss-free sign language translation. Specifically, Orbay and Akarun (2020); Shi et al. (2022) utilize local visual features (e.g., hands) to enhance the translation performance. Those systems require domain-specific training data (e.g., labeled handshape data used in Orbay and Akarun (2020)), which is not always accessible for the target sign language. The fusion of visual features at different scales also increases the complexity of the modeling pipeline.

In this paper, we study a simple model for sign language translation between DSGS and German in a gloss-free setting. Our model uses a 3D convolutional network for visual feature extraction and a Transformer-based encoder-decoder for sequence modeling. It is built on raw RGB images rather than pose keypoints, thus avoiding potential mistakes from pose estimation and remaining fast in inference. We further study the impact of hyperparameters and different pretraining strategies on translation quality. Without ensembling, our model achieves 0.3 Bleu score and 0.195 Chrf score on the official test set.

2 Method

In this section, we describe our method for sign language translation. Our model consists of an Inflated 3D ConvNet (I3D) (Carreira and Zisserman, 2017) for visual encoding and a Transformer-based encoder-decoder model (Vaswani et al., 2017) for sequence modeling, which are described respectively below.

I3D I3D (Carreira and Zisserman, 2017) is a 3D convolutional neural network proposed in action recognition. I3D has previously been explored in sign language processing (Albanie et al., 2020; Li et al., 2020; Vaezi Joze and Koller, 2019) and achieved competitive performance in isolated sign recognition (Li et al., 2020). More formally, given a sequence of image frames $\mathbf{I}_{1:T}$, the I3D model $\mathbf{M}^v$ encodes them into a sequence of visual features $\mathbf{f}_{1:T'}$: $\mathbf{f}_{1:T'} = \mathbf{M}^v(\mathbf{I}_{1:T})$, where T and T' respectively denote the length of video and visual feature sequence. Note due to the temporal stride in convolutional kernels of I3D, T' is not equal to T and is usually several factors smaller.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 989–993
Abu Dhabi, December 7–8, 2022.

To encourage the visual encoder $\mathbf{M}^v$ to capture more signing-related visual cues (e.g., arm movement, handshape, and so on), we pretrain the I3D model with isolated sign recognition on WLASL, a large-scale dataset consisting of isolated American sign language (ASL) signs. Though ASL and DSGS are two different sign languages, visual features regarding body movement are shared. Empirically, we observed considerable gains in isolated sign pretraining. For computational efficiency, the pretrained I3D network $\mathbf{M}^v$ is frozen in translation model training.

Transformer-based encoder-decoder We employ a Transformer-based encoder-decoder (Vaswani et al., 2017) model $\mathbf{M}^{(s)}$ to decode visual feature $\mathbf{f}^{(v)}_{1:T'}$ into text $w_{1:N}$: $w_{1:N} = \mathbf{M}^{(s)}(\mathbf{f}^{v}_{1:T})$. $\mathbf{M}^s$ is a standard sequence-to-sequence model widely used in machine translation (Vaswani et al., 2017; Barrault et al., 2020; Akhbardeh et al., 2021). Thus we only briefly review it here and a more detailed description can be found in Vaswani et al. (2017). Our sequence-to-sequence model $\mathbf{M}^s$ includes a Transformer encoder and Transformer decoder, which are joined via attention. Specifically, the Transformer encoder transforms the visual features $\mathbf{f}^{v}_{1:T}$ into $\mathbf{e}_{1:T}$ by injecting temporal information based on self-attention and positional embedding. The Transformer decoder generates token sequence $w_{1:N}$ in an auto-regressive manner while attending to the encoder output $\mathbf{e}_{1:T}$ through the attention mechanism.

Training loss We use cross-entropy loss for model training. More formally, given the translation pair $(\mathbf{I}_{1:T}, \hat{w}_{1:N})$, suppose the model outputs probability vector $\mathbf{p}(\cdot|\mathbf{I}_{1:T}, \hat{w}_{1:n-1})$ at decoder step n. The loss is then computed as

$$l = -\sum_{n=1}^{N} \log p(\hat{w}_n|\mathbf{I}_{1:T}, \hat{w}_{1:n-1}) \qquad (1)$$

Inference At test time, we use beam search for decoding image sequence $\mathbf{I}_{1:T}$. The beam width and length penalty are hyperparameters tuned using the development set.

3 Experimental Setup

Data We use FocusNews and SRF data to train our translation model. Both FocusNews and SRF consist of DSGS-German pairs, which include 19 and 16 hours (10,136 and 7,071 sequences) of DSGS videos, respectively. The two datasets differ in multiple aspects. For example, FocusNews are live signing from teleprompters by deaf signers based on news from 2008 to 2014, whereas SRF dataset contains news videos from 2020 to 2021 which is interpreted into DSGS by hearing interpreters. Both datasets are incorporated into training. Note that frame rates in videos of FocusNews and SRF differ, we feed the raw videos in FocusNews and SRF without frame rate conversion. To pretrain the visual encoder, we use WLASL (Li et al., 2020), a large-scale isolated sign dataset including $\sim 21k$ pairs of American sign language video clips and English words.

Training We use sentencepiece unigram tokenizer (Kudo, 2018) to tokenize the German translation. The number of subword units is tuned to 18,000 We use a 2-layer Transformer with 512 hidden dimensions and 2048 hidden dimensions for both encoder and decoder. A dropout layer with a zeroing probability of 0.1 is added between the self-attention layer and the feedforward network. The model is trained with Adam (Kingma and Ba, 2015) for 18K steps at a batch size of 32. The learning rate is linearly increased to 0.0008 for 2K steps and decayed to 0 in the remaining steps. The visual backbone I3D is pretrained on WLASL and frozen during translation model training. During isolated sign training, we initialize I3D from a model trained on the action recognition dataset Kinetics (Carreira and Zisserman, 2017). We use SGD with a 0.9 momentum value to train the model for 50 epochs at a batch size of 4. The initial learning rate is 0.001 and is halved if accuracy on the validation set does not increase for 3 epochs. Before feeding into I3D, each isolated sign video is truncated to a 64-frame clip, which is padded with all-zero frames if the length is shorter than 64. Each image frame is resized to 240×240. It is randomly cropped to 224×224 and horizontally flipped at a probability of 0.5 in training. At test time, we only center cropping to every image frame.

Evaluation We evaluate the system using BLEU-{1,2,3,4} (Lin, 2004) and ROUGE (Lin, 2004) scores.

4 Experimental Results

In this section, we report results and conduct some analyses of the translation model on the development set.

Hypothesis	Reference	Bleu-4
das der stand der dinge im moment. gibt es eine grosse aufsp. (that's the state of things at the moment. is there a big sp.)	das der stand der dinge im moment. (that's the state of things at the moment.)	51.56
mit live-untertiteln von swiss txt guten abend, meine damen und herren, willkommen zur "tagesschau" (with live subtitles from swiss txt good evening, ladies and gentlemen, welcome to the "tagesschau")	guten abend, meine damen und herren, willkommen zur "tagesschau". (good evening, ladies and gentlemen, welcome to the "tagesschau".)	51.42
die armee muss ihre arbeit nicht mehr einmal. (the army doesn't even have to do its job anymore.)	doch die bevölkerung macht nicht mit. (but the population does not participate.)	0.00
die französischen roben programm speziell für gehörlose. (the french robes program especially for the deaf.)	bei auf der webseite des sportverbandes können detailliertere informationen nachgelesen werden. (more detailed information can be found on the website of the sports association.)	0.00

Table 1: Qualitative examples produced by our translation system. The sentence within () is the corresponding English translation.

4.1 Main Results

Table 2 shows the performance of our model on the development set compared to the Sockeye baselines reported from the official repo (Müller et al., 2022). Our model outperforms Sockeye baselines, which are models based on the pre-extracted human pose. However, the overall values in different metrics are very low. We further show translation examples produced by our model (see Table 1). We noticed the phrases that are translated correctly by our model are usually duplicate phrases frequently appearing in training (e.g., willkommen zur "tagesschau"). For most of the sentences, the model is unable to capture its meaning generally though many predictions are grammatically correct. Such observation shows that large-vocabulary sign language translation is very challenging.

	Train Data	Rouge	B1	B2	B3	B4
Sockeye (Müller et al., 2022)	FN	-	-	-	-	0.21
Sockeye (Müller et al., 2022)	Srf	-	-	-	-	0.59
Sockeye (Müller et al., 2022)	FN,Srf	-	-	-	-	0.15
Ours	FN,Srf	7.92	8.36	2.92	1.55	**1.02**

Table 2: Performance of our model on development set. The Sockeye baselines are from the official repo (Müller et al., 2022). FN: FocusNews

4.2 Hyperparameter Tuning

Among the set of hyperparameters, we find that the following two hyperparameters have the most significant effect on translation performance: learning rate and the number of layers. We detail their impact on model performance below. Other hyperparameters (e.g., dropout, learning rate schedule) are also tuned in our experiments. However, their impact is relatively negligible and thus not detailed in the paper.

Learning rate We tuned the learning rate among {0.001, 0.002, 0.004, 0.008, 0.016}. As is shown in Table 3, increasing the learning rate consistently improves the model performance across all the metrics. The benefit plateaus around 0.008, which is the optimal value among the set of values we consider.

Number of layers We further tuned the number of Transformer layers (see Table 4). We keep the number of encoder and decoder layers the same and set the hidden/feedforward dimension to 512/2048 in the corresponding experiments of Table 4. Increasing number of transformer layers degrades the performance. This is probably because the 3D convolutional kernels of I3D capture some temporal relations in the video, which reduces the reliance of

LR	Rouge	B1	B2	B3	B4
0.001	7.82	8.38	2.51	1.21	0.76
0.002	6.85	7.25	2.22	1.05	0.69
0.004	6.86	6.08	2.22	1.18	0.82
0.008	**7.92**	**8.36**	**2.92**	**1.55**	**1.02**
0.016	7.54	6.11	2.27	1.35	1.01

Table 3: Impact of learning rate on translation performance

the whole model on Transformer modules to capture sequential information. Furthermore, larger models (i.e., more layers) usually require more training data. The total amount of sign language videos (35 hours) is probably insufficient to train a deep transformer encoder-decoder.

# Layer	Rouge	B1	B2	B3	B4
2	**7.92**	**8.36**	**2.92**	**1.55**	**1.02**
4	7.10	7.01	2.31	1.21	0.82
6	6.17	7.32	1.55	0.52	0.24

Table 4: Impact of Transformer layers on translation performance

4.3 Effect of I3D pretraining

The I3D backbone is pretrained on WLASL. Here we compare three options of I3D pretraining: WLASL, BSL-1K, and Kinetics-400. BSL-1K is a coarticulated sign dataset of 1064 British sign language (BSL) signs (273K video clips in total), collected from BBC videos interpreted into BSL. Kinetics (Carreira and Zisserman, 2017) is the action recognition dataset with 650K videos from 400 human action categories. As is shown in Table 5, pretraining with sign-language specific datasets (WLASL, BSL-1K) consistently outperforms pretraining with general human action videos (Kinetics). This is expected as signing-related visual cues (e.g., handshapes), essential for sign language translation, are better captured in isolated sign datasets. Pretraining with WLASL achieves better results than BSL-1K. Though BSL-1K contains an overall larger number of video clips than WLASL (21K vs. 273K), it has fewer unique signs (1064 vs. 2000). This probably suggests that a sign language corpus with more signing categories will be more beneficial to sign language translation compared to its counterpart with fewer signs.

PT Data	Rouge	B1	B2	B3	B4
WLASL	**7.92**	**8.36**	**2.92**	**1.55**	**1.02**
BSL-1K	6.88	6.19	1.86	0.84	0.69
Kinetics	5.05	4.18	1.02	0.65	0.41

Table 5: Impact of Transformer layers on translation performance

5 Conclusion

This paper describes TTIC's DSGS-German translation system submitted to the WMT-SLT 2022 challenge. Our model consists of an I3D model for visual feature extraction and a Transformer-based encoder-decoder for sequence modeling. The system is based on RGB images alone and remains conceptually simple. Our experiments show that pretraining the visual frontend with isolated sign recognition helps achieve better translation performance. However, the overall translation quality is still in a very low regime. Our future work includes combining pose and RGB-based models for sign language translation.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondrej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*.

Loïc Barrault, Magdalena Biesialska, Ondrej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubesić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos

Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

N. C. Camgoz, B. Saunders, G. Rochette, M. Giovanelli, G. Inches, R. Nachtrab-Ribback, and R. Bowden. 2021. Content4all open research sign language translation datasets. *ArXiv*, abs/2105.02351.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *CVPR*, pages 7784–7793.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. *CVPR*, pages 4724–4733.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *CVPR*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *ACL*.

Dongxu Li, Cristian Rodriguez-Opazo, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *WACV*, pages 1448–1458.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *ACL*.

Mathias Müller, Annette Rios, and Amit Moryossef. 2022. Sockeye baseline models for sign language translation. https://github.com/bricksdont/sign-sockeye-baselines.

Alptekin Orbay and Lale Akarun. 2020. Neural sign language translation by learning tokenization. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 222–228.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. *ArXiv*, abs/2205.12870.

Hamid Vaezi Joze and Oscar Koller. 2019. Ms-asl: A large-scale data set and benchmark for understanding american sign language. In *The British Machine Vision Conference (BMVC)*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.