Sign Language Transformer using Spatial Representations

M. MAAHIR, DR. K ANITHA SHEELA, N. SATHWIK, J SANDHYA
Department of Electronics and Communication Engineering
Jawaharlal Nehru Technological University Hyderabad
Kukatpally, Hyderabad, Telangana 500085
INDIA

Abstract: Sign Language recognition has been studied and designed by many currently existing models but only a few models exist that focus on translation. We believe translation is very much needed for efficient communication between the disabled and abled. The proposed model not only does the recognition but also translates the sign language to understandable spoken language.

The proposed project implements an end-to-end sign translation system that is capable of simultaneously learning both signs to gloss and gloss to text during the training process. This is done using transformers which consist of encoders and decoders, the encoder is used for recognition of sign language called gloss. Using an encoder transformer, the system understands the language by using the spatial-temporal nature of the language. Spatial understandings from the encoder are then sent through the decoder for translation tasks. The resultant of the decoder is the grammatically accurate sign language conversions We use PHOENIX-2014T dataset which consist of continuous sign videos of weather news reported in Germany with gloss and text. The proposed model is capable of sign to gloss, sign to text and sign to gloss to text.

Keywords:—Transformer, continuous sign language recognition and translation, encoder transformer, decode transformer, word embeddings, spatial embeddings, Attention mechanism.

Received: April 2, 2024. Revised: September 13, 2024. Accepted: October 15, 2024. Published: December 4, 2024.

1 Introduction

In India according to the National association of the deaf (NAD-India) there were about 18 million deaf and partially deaf people in 2016. These people use a type of language that solely depends on hand gestures and movements and expression from face and body. These languages are called sign languages. They bridge the gap between disabled and abled. Apart from regional differences sign language has its own grammatical rules which do not necessarily translate to spoken language; they are non-monotonic meaning there is no one to map. For effective communication exacting grammatical translation is very much required. We designed a sign language translation model using an encoder-decoder transformer based on an attention mechanism which is capable of sign to spoken language.

The major work on bridging the gap between sign language and spoken language was focused on recognition alone. The drawback of this is in terms Continuous sign language recognition [1,2] does not provide exact meaning of what the singer is saying meaning to say the sign language sentences do not align accurately with spoken language sentences. Though gloss conveys the overall meaning, they lack spoken grammar which might alter the meaning in spoken language. To overcome these, we came up with a sign language transformer which not only recognizes the gloss but also translates it to spoken language.

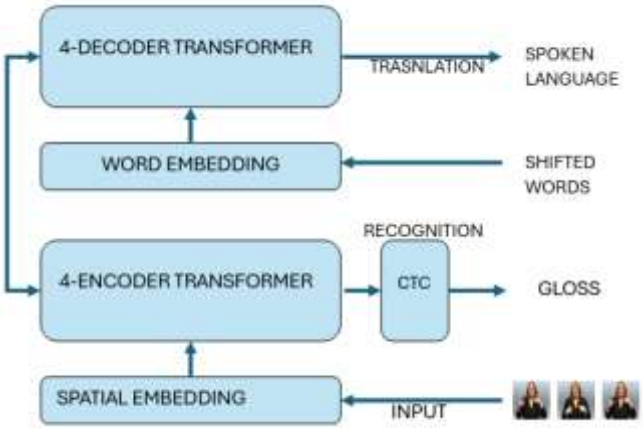

Figure 1: A brief overview of sign language recognition and translation transformer.

The main objective of the sign language transformer is to convert the sign to spoken language which comes under spatio-temporal machine translation task [3] such system should be able address the issues like sign segments but though there have been automatic sign segmentation [4,5,6,7,8] they are not implemented in translation task so far and still remain a challenging task, it is a tedious task in terms of computer vision as the model should be able to know the singer in 3D space and to tackle motion blur and rate of speed of different signs and singers.

Apart from spatial-temporal issues the other major drawback is lack of dataset. To our best knowledge the only available dataset is PHOENTX14T with good vocabulary and sign sentences. According to [3] sign language translation is a Neural Machine Translation [19] task. From the finding of [3] it was proved that using gloss based mid-level representation improves the sign language translation performance drastically. In our problem we use transformer network rather than Recurrent Neural Network (RNN) sign to gloss to text architecture whose accuracy is affected by ability to recognize gloss

For hardware, flexibility and understanding purposes we have manipulated the dataset. We have replaced the German gloss and text to English and used dataset correction techniques for efficiency and better performance. A brief explanation is shown in Figure 1.

We designed a encoder-decoder transformers network for both recognition and translation. We first implemented an encoder transformer called Sign language Recognition Transformer (SLRT). SLRT is input through a spatial embedding layer which lets the model learn the spatiotemporal nature of videos. SLRT uses attention mechanism [11] for contextual understanding and CTC loss [9] for alignment of features to glosses. The representation from SLRT is sent through the sign language translation Transformers (SLTT). Like spatial embeddings word embedding are given to SLTT which consist of masked attention mechanisms followed by encoder-decoder attention mechanisms. The translation task works autoregressive it predicts one word at a time based on current and past outputs. The final output from the SLRT is gloss representation and SLTT is spoken language.

The contribution in this paper can be summarized as:

- Preparing and manipulating PHOENIX-2014T dataset for efficiency and fast processing.
- Designing sign language recognition and translation transformers.
- Use of different metrics to evaluate the accuracy of the model.

The rest of the paper is organized as follows: In section 2 we mentioned the already existing work in sign recognition and translation. In section 3 we talk about the data processing, working of transformers and steps needed for sign to gloss and gloss to text task. In section 4 we talk about the dataset and its implementation. We share our quantitative and qualitative results in 5 and 6. In section 7 we conclude and mention possible future improvement.

2. Related Work

There has been constant progress in the field of sign language recognition for almost two decades. The first ever recognition used glove-based motion tracking but soon this turned out to be costly and then as computer vision [17] started to get advanced recognition from visual perspective became an ideal choice. One such model is [10] hidden Markov model (HMMs) on American Sign Language (ASL). Majority of the work from the past two decades focused on sign language recognition but not on translation tasks because of their lack of availability of proper dataset on (Continuous Sign language) CSL, inability to recognize signs accurately as they do not have the pauses and end as we do in speech. and sign and spoken language are non-monotonic which adds up to more complexity to decode. The next subsection talks about the related work done in the field of recognition of gloss and spoken language.

2.1 Sign Language Recognition

Sign language is the mixture of manual features and non-manual features. Manual features are the moment of hands different from the past, shapes and poses, on the other hand facial expressions such as smiling, moment of eyebrows are non-manual features. As deep learning and machine learning started, models like CNNs, RNNs, and other counterparts got better at recognition of these features in 2D, and 3D space. The development of Recurrent Neural Network has brought in a new approach which is capable of outputting a sequence based on the input sequence which is called sequence to sequence approach. RNN added temporal features which were a breakthrough in sequence-to-sequence task, but its limited window lacks to relate as sequences get bigger. The number of glosses is not equal to the number of frames which cannot be annotated, this makes it a weakly annotated dataset. To address this Connectionist Temporal Classification (CTC) [16] [9] can be used which tries to align the input and output sequence; it plays an important role in terms of sign language recognition (SLR). The recent advances have brought in 3D convolutional networks that are good at capturing spatiotemporal aspects of the sign language, Gesture Recognition with Cameras that combining RGB and depth information from the camera model, and Skeleton-Based Approaches which use skeletal representations of the hand and body which can be less sensitive to background. The implementation of the attention mechanism allowed the model to focus on relevant parts of the sign sentences which resolved the dependencies in sign language issues.

2.2 Sign language Translation

Sign language Translation was only conceptual proven due to lack of a huge dataset, Sign language sentence and spoken language do not have one to one mapping and unlike audio and text there has been no sign level segmentation. Recent development in encoder decoder networks, Attention mechanism [11] and availability of annotated dataset is feasible to some extent. As of now there are no available models that can be converted to spoken language translations just from the sign videos. Camgoz et al. [3] With the help of Neural Machine Translation (NMT) [17, 19] approached in an end-to-end sign language video to spoken language sentence translation called the Sign2Text model. Later it was proved in [3] addition of gloss intermediaries drastically increased the performance which were based on CSLR [12] and attention based NMT [13] methods. The current state of the art for sign translation is done using the Transformers as transformers are capable of parallelism.

processing does not have RNN window size limitations. Transformers are good speech recognition, sentence summarization which come in sequence-to-sequence tasks. Due to its potential in sequence-to-sequence modelling and parallel processing they are ideal choices for sign language recognition and translation tasks.

2.3 Dataset:

There have been quite a few datasets developed for sign recognition but the majority of them only contain word level sign videos which is not useful for Continuous sign language recognition (CSLR). There have been few datasets designed for CSLR but they are limited in vocabulary and sentences. We have mentioned a few of the public dataset we could find on sign language in table 1. As per now the dataset that well suits our problem is PHOENIX-2014T [14]. PHOENIX-2014T consists of consecutive gloss and text of sign language interpretation on German weather forecast from 2009-2011 of 9 different singers who made 1066 different signs which corresponds to 2887 different words. The resolution of the videos is 210 by 260 pixels recorded at 25 frames per second.

3. Transformers and Prerequisites

The overall idea is to convert signed video to spoken languages but the direct two-step process is not practically possible one way to design a translation system is to use CSLR [12] techniques and then use NMT [13] for translation. But from the finding in CSLR [12] the overall accuracy depends on the

Table 1: Information regarding different datasets on sign language

Dataset	LANGUAGE	RESOLUTION	SINGERS	VIDEOS	SIGNS
PHOENIX-2014T(ours) [21]	DGS	210x260	9	8,257	1,066
PHOENIX-2014T [22]	DGS	210x260	9	6,841	1081
SIGNUM [23]	DGS	776x578	25	15,075	455
BSL-1K [24]	BSL	1280x720	7	2,73,000	1064
MS-ASL [25]	ASL	1920x1080	222	25,513	1000
WL-ASL [26]	ASL	1280x720	119	21,083	2000

extent to the model's ability to learn the relation between the gloss and sign. To overcome this [15] proposed a novel approach to learn from spatial and temporal data of the sign in videos. From the inspiration on [15] implemented a transformer by manipulating dataset and parameters to meet our hardware 3 for detailed representations of the transformer network. In simple terms transformers are required to find out conditional probabilities of p(G|V) and p(S|V) If G=$(g_1 \dots g_N)$ is the gloss sequence on N words, S = $(w_1 \dots w_U)$ is spoken language sentence with U words and V = $(I_1 \dots I_T)$ is sign video of T frames requirements and understanding refer Figure 3

3.1 Embedding

We use embeddings as a lookup table for the encoder and decoder as transformers are not capable of maintaining recurrence. Embedded transformers can keep track of frames and videos.

3.1.1 Spatial Embedding

As our proposed model uses the spatial and temporal nature of the videos for translation, it requires these spatial and temporal features to be extracted before passing to SLRT (sign language Recognition Transformer). Spatial and Temporal are extracted using pretrained CNNs and Pytorch Embedding class for extracting and holding the features in sequential order.

S_t = Spatial Embedding I_t

$$\widehat{S_t} = S_t + Positional\ Embeddings(t) \quad (1)$$

3.1.2 Word Embeddings

Word embeddings are used in the decoder transformer where the objective is to text-to-text translation. Word embeddings are similar to spatial embeddings, the only difference being the use of a linear layer rather than CNNs which acts as a lookup table.

W_u = Word Embedding W_u

$$\widehat{W_u} = W_u + Positional\ Embeddings(u) \quad (2)$$

3.2 Sign Language Recognition Transformer

Give a sign video the SLRT should predict the respective glosses. To assist this, we make use of spatial embedding to provide the spatial representation of frames before passing through SLRT. The SLRT begins with attention mechanism [11] using which the model learns the contextual understanding between the frames. The way this works is by using query, key and value variables. Query is the current frame we are trying to focus; key is information about all other frames and value is details these frames consist of. A simple dot product gives us the dependencies of particular features we are interested in getting the relation between the signs. Mathematically speaking SLRT tries to predict the probability of $\hat{S}_i$ giving all S in a video.

$$RT_t = SLRT\left(\hat{S}_t \middle| \hat{S}_{1:T}\right) \quad (3)$$

So far, the encoder is capable of only predicting the spatial representations. But the end goal of the encoder is to predict gloss to achieve this; gloss intermediaries using CTC are added after SLRT processing. [9,16]. As, sign and gloss both rely on spatial and temporal features they both have one to one mapping so it makes sense to map every frame to every sign moment it represents to do this every frame must be annotated one way to do this is by cross entropy loss but this method tedious and rarely used. So, the best approach is

to use sequence to sequence loss functions, one of such is CTC [16] . CTC is used to align gloss to frame. and CTC calculates the gloss probabilities for a given spatial representation of frames(video) from SLRT.

$$p(G|V) = \sum_{p \in P} p(p|v) \quad (4)$$

p is the current path and P is all possible paths

The work of CTC is not just aligning the text and gloss but to enhance the transformer network to accurately extract the features during the training process. For the performance increment we use an add and normalization layer next to every step. The spatial representations are then sent to the translation transformer for decoding.

3.3 Sign Language Translation Transformer

Unlike other encoder decoder networks on translation which train on output of CSLR this causes a bottleneck on ability for transformers to learn. To alleviate this spatial representation of the encoder is used. This starts with getting the word into in word embeddings, as explained in the encoder part transformers lack the recurrence so have the positional and ordered words linear layer and positional encoding layers are used. Linea layer and positional encoding as a lookup table which helps in training processes where the model tries to learn. The positional encoded data is then sent to the Masked Multi-Headed attention mechanism. The words are converted into a sequence of vectors (like a lookup table) into a word embedding layer using a linear layer. These are then sent through masked self-attention to limit the ability to learn the contextual understanding of the future tokens. Masked self-attention focuses on current and past tokens. In the next step the encoder-decoder multi-headed attention mechanism takes in spatial representation of the encoder and masks the attention self-attention layer which helps the decoder to train on the relationship between the two components.

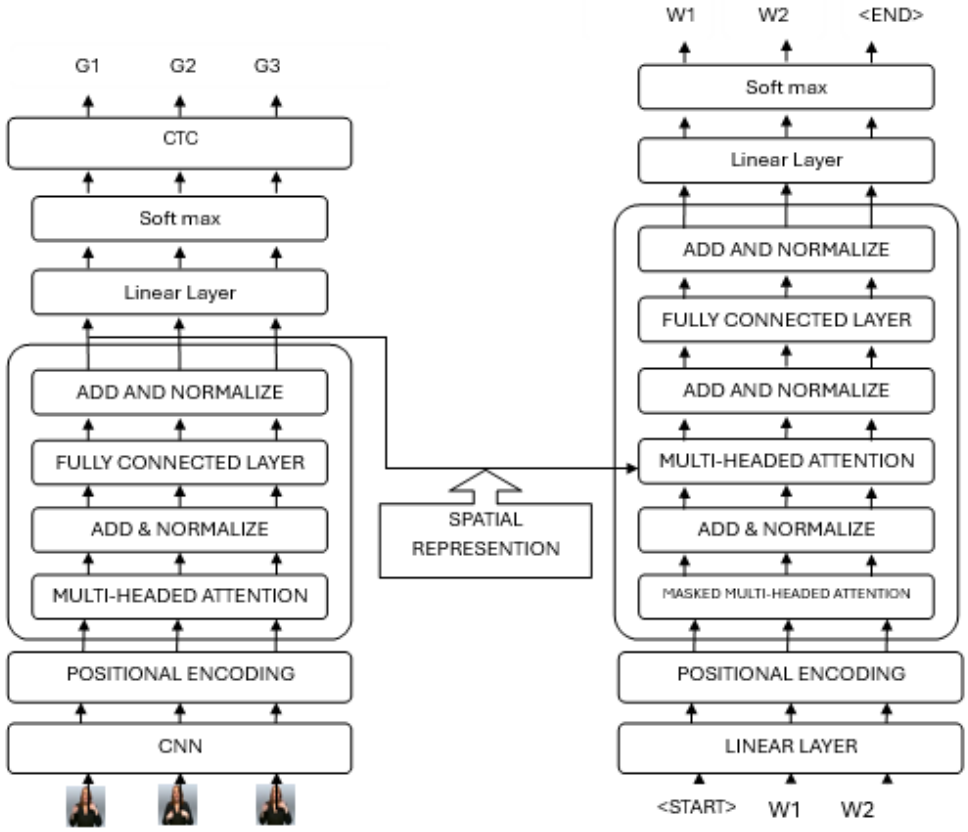

Figure 2: Detailed architecture of sign language recognition and translation transformer.

Like SLRT every layer is followed by an add and normalization layer to increase the performance. The decoder gives current token probability given all previous tokens and spatial representations of SLRT. Mathematically SLTT:

$$TT = SLTT\left(W_u \middle| W_{1:T,} RT_{1:T}\right) \quad (5)$$

And the ability of the decoder to produce one word recursively can be formulated as probability of spoken language given sign videos.

$$p(G|V) = \prod_{u=1}^{U} p(W_i|TT_u) \quad (6)$$

4. Implementation and Dataset

This explanation explains the dataset used, settings and parameters used for both recognition and translation.

4.1 Dataset

We used PHOENIX2014T dataset [17] that contain 2000 sign videos for training, 400 for validation and 400 for testing. In this paper the dataset has been altered for flexibility and understanding. As machine learning is independent of language, we translated the gloss and text into the German sentence this way we could on the go notice the changes between hypothesis and reference sentence (This step is optional) For the loading and training efficiency we used pytorch to convert the video frames to tensor (figure 3). As transformers require and are capable of parallel processing and require reference and translation data parallelly we have stored them in two files, and this affects the processing speed. Finally, we have filtered grammatical errors due to translation from German to English, by ratio, tokenized and converted the whole glossary to lowercase letters. Finally, JoeyNMT [15,18] allows character level, sub-word level and word level translation for the sake of this paper we used word level.

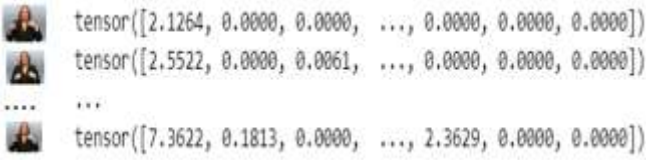

Figure 3: Tensor representation of sign frames.

4.2 Implementation

Architecture details: we used JoeyNMT v1.0[15, 18] to implement the Transformers. For spatial embedding layers to extract the spatial and temporal nature pretrained network ImageNet and data augmentation to discard similar frames are used.

Transformer: Similar setting is used in both SLRT and SLTT. The hidden units are 512 and 8 heads in each layer. methos and 0.1 dropout rate to avoid over-fitting. The total architecture consists of 4 such layers.

Optimization: The learning rate of 0.0002, batch size of 16, eval metric for early stopping, Adam optimizer and plateau scheduler that reduces learning rate based on early stopping metric are set. The check points are saved at every 500 steps. We used a NVIDIA RTX 4090 for training and testing with the above settings. Decoding: For decoding greedy search is used to convert gloss to text. We used 0 to 5 beam search decoding widths and alpha called length penalty [19] for normalization value of 2. The combination of beam width and penalty are saved and used on test data.

5.Qualitative Results

This section shows the results of sign to gloss, gloss to text and sign to text outputs of the model. We show the three-reference text and the respective model outputs (Table 2). The model performs well given the limitation of reduced dataset and misspelled words in German to English conversions. The model struggles to predict nouns due to the obvious fact that as few nouns only appear once or twice in the data. There is trouble due to frames per second which cause motion blur the network can deal with these motion effects.

For a clear frame the model performs very well. The improvement of FPS and definition video recording significantly enhances the feature detection process. Overall, the translation task is adequate for the given dataset.

gloss hypothesis: central region cloud tomorrow, rain, thunderstorms
gloss reference : friday southeast sun friday then more cloud, could of rain gloss hypothesis: friday southeast sun friday then more could of rain
gloss reference: fifteen degrees sauer land region only six degrees gloss hypothesis: to for region only
gloss reference : thursday german country variable mostly rain possible strong wind gloss hypothesis: thursday loc of changeable rain off
text reference : the wind from the east blows moderately in the northeast and is partly fresh. text hypothesis : the wind blows weakly to possible on the north sea.
text reference : tomorrow in the north it will initially be very cloudy with a few drops. text hypothesis : tomorrow it can be strongly approved with heavy storm
text reference : during the day it rains in the north and west and sometimes thick fog in the south as well as clouds. text hypothesis : tomorrow of the rain in the north and east of the sowing the wind can be individual little bit.
text reference : and we have a storm with us every day. text hypothesis : and we have a storm with us every day.
text reference : on friday in the southeast half it was still partly friendly while the clouds text hypothesis : on friday in the southeast half it was still partly friendly while the clouds.
text reference : tomorrow in the north it will initially be very cloudy with a few drops. text hypothesis : tomorrow it can be strongly approved with heavy storm

Table 2: Generated gloss and text cases by the model

6. Quantitative Results

In the section we share our results and setup of encoder and decoder. As the model is based on sequence-to-sequence tasks it requires a particular metric to know about how well the model is predicting from the desired result. This section explains evaluation metrics in detail, result, valuation and train and validation loss on recognition and translations followed by the respective graphs of training and validation. The summary of sign to text is mentioned in (Table 3).

6.1 Evaluation Metrics:

WER (Word Error rate) For Recognition Word Error rate (WER) is one of the widely used metrics. Word error rate is calculated based on tokens in a sentence. It compares the predicted output sentence tokens and the reference sentence token. WER calculated by comparing the two sentences (R and H) and aligning by where the model inserted new words (I), Deleted words (D) and substitution (S) of different tokens in place of respective reference words to all the words(N) present in the sentence. Figure 5 mentions the graphs related to I, D, S and WER.

$$WER = \frac{I + D + S}{N} \quad (7)$$

For Translation three metrics namely BLEU (Bilingual Evaluation Understudy) scores of 0-4 n-grams, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Rouge score and CHRF (Character n-gram F-score) score are measured. These metrics require precision and recall. If R is a reference sentence, H is a hypothesis sentence and n-grams are the number of overlapping words in R and H. Precision measures how many words in the H are relevant to the R, while recall measures how many of the relevant words in the R are present in the H.

$$Precision = \frac{\text{overlapping n} - \text{grams}}{\text{otal n} - \text{grams in the hypothesis}} \quad (8)$$

$$Recall = \frac{\text{overlapping n} - \text{grams}}{\text{Total n} - \text{grams in the reference}} \quad (9)$$

BLEU (Bilingual Evaluation Understudy) is used in evaluating precision of machine generated translation. It says how many words in the prediction are present in the reference sentence. BLEU works by comparing n-grams (n is the number of words consecutive words taken into consideration) in the candidate translation to those in the reference translations. For our model we achieved a bleu score of 14.70 (Figure 8) and bleu 1, bleu 2 bleu 3 and bleu 4 (Figure 4) of 37.86, 28.63, 18.93 and 14.70 respectively for 9000 steps.

$$BLEU\_N = \frac{overlapping\ n\_grams}{\text{Count of all n\_grams in hypothesis}} \quad (10)$$

$$BLEU = BP \times e^{\left(\frac{1}{n}\sum_{i=0}^{n} logP_i\right)} \quad (11)$$

- BP is the brevity penalty, which penalizes hypotheses that are shorter than the reference.
- P_i is the precision for n-grams of length i.
- n is the maximum length of the n-grams

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is used for evaluating recall of the quality of reference. Rouge measures how many of the words in reference appear in prediction. The model has resulted a rouge score of 63.57 (Figure 7)

$$ROUGE\_N = 2 \times \left(\frac{precision_n \times recall_n}{precision_n + recall_n}\right) \quad (12)$$

Figure 6: Validation loss of Recognition (left) & Translation (right)

$$n = word\ level\ n-grams$$

CHRF (Character n-gram F-score) is used for character level evaluation of translation models unlike words used in BLEU and Rouge. CHRF calculates precision, recall, and F-score based on the counts of overlapping character n-grams between the candidate translation and the reference translations. The F-score combines precision and recall providing a single measure of translation quality. The model has achieved 66.63 (Figure 7) CHRF score for 9000 steps.

$$CHRF = 2 \times \left(\frac{precision_c \times recall_c}{precision_c + recall_c}\right) \quad (13)$$

$$c = character\ level\ n-grams$$

Training and Validation plots:

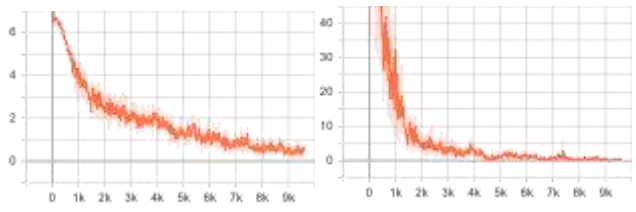

Figure 4: Training loss of Recognition (left) & Translation (right)

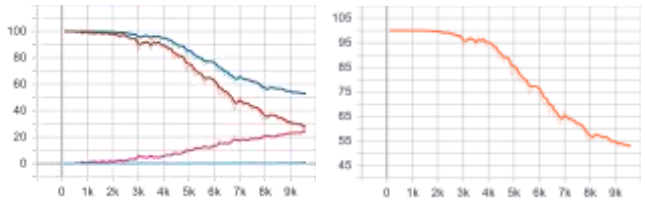

Figure 5: Word Error Rate and I, D, S rate (left) & WER (right) valid.

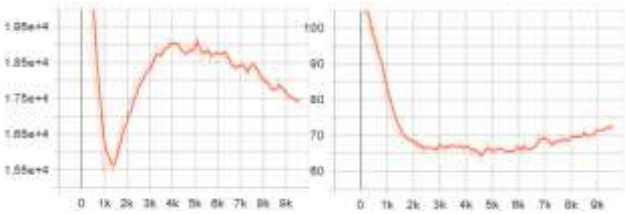

Figure 6: Validation loss of Recognition & Translation

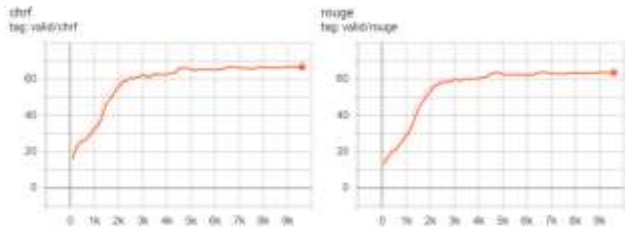

Figure 7: CHRF (left) & Rouge (right) Validation Scores.

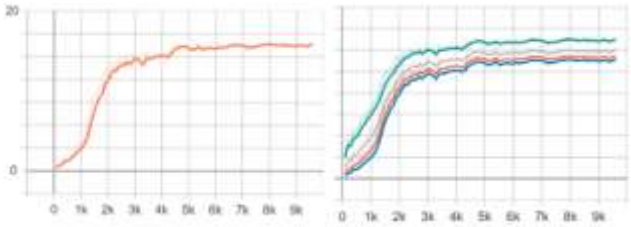

Figure 8: Bleu 4 score (left) & Bleu 1-4 scores (right) of validation.

The training loss is a weighted sum of the recognition loss and the translation (figure 4). The validation losses (Figure 6) for recognition and translation are calculated in the same manner as the training loss. In summary we have used 4 different metrics to measure our model's performance WER for recognition and BLEU, ROUGE AND CHRF for translation. The model result based on the precision metrics (table 3) shows the recognition and translation in terms of WER and BLEU Scores

SIGN to Text	WER	BLEU1	BLEU2	BLEU3	BLEU4
DEV	56.36	38.52	29.13	21.89	16.46
TEST	58.47	37.86	28.63	18.93	14.70

Table 3: WER and BLEU scores of Signs to Text Translation

7. Conclusions

In this paper we mentioned prior techniques that only focus on recognition but not translation. It also addresses the currently existing issues like limited dataset and use of CSLR and NMT techniques whose accuracy totally depends on the ability to recognize gloss at encoder that causes bottlenecks to develop a translation system. In this paper as language is independent of machine learning tasks for the flexibility and understanding the German gloss and text are converted to English, with data augmentation, use of separate files for reference and hypothesis we noticed considerable increase in performance. We approached recognition and translation as spatial-temporal tasks and addition of gloss intermediaries in the form of CTC loss boosted the performance compared to sign to gloss and gloss to text model. With the extracted spatial features translation encoder has been implemented that can predict one word at a time for respective sign videos.

The model's overall accuracy is impacted by dataset, therefore, a better datasets and higher resolution, being able to understand non-manual features, use of updated sequence to sequence loss functions and other variants of transformers may increase the performance.

Acknowledgement:

This paper owes its existence to the invaluable support provided by the Department of Electronics and Communications Engineering and the JNTUH Technology Business Incubator at Jawaharlal Nehru Technological University Hyderabad. Their generous assistance, along with the provision of essential hardware and software resources, has played a pivotal role in making this research endeavor possible.

References

[1] O. Koller, H. Ney, and R. Bowden. Deep Hand: How to Train a CNN on 1 million Hand Images When Your Data Is Continuous and Weakly Labelled. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] O. Koller, S. Zargaran, and H. Ney. Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNNHMMs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural Sign Language Translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[4] Mark Borg and Kenneth P Camilleri. Sign Language Detection in the Wild with Recurrent Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019.

[5] Shujjat Khan, Donald G Bailey, and Gourab Sen Gupta. Pause detection in continuous sign language. International Journal of Computer Applications in Technology, 50, 2014.

[6] Pinar Santemiz, Oya Aran, Murat Saraclar, and Lale Akarun. Automatic Sign Segmentation from Continuous Signing via Multiple Sequence Alignment. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), 2009.

[7] Frank M Shipman, Satyakiran Duggina, Caio DD Monteiro, and Ricardo Gutierrez-Osuna. Speed-Accuracy Tradeoffs for Detecting Sign Language Content in Video Sharing Sites. In Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS), 2017.

[8] Neva Cherniavsky, Richard E Ladner, and Eve A Riskin. Activity Detection in Conversational Sign Language Video for Mobile Telecommunication. In Proceedings of the International Conference on Automatic Face and Gesture Recognition (FG), 2008.

[9] Alex Graves, Santiago Fern´andez, Faustino Gomez, and J¨urgen Schmidhuber. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In Proceedings of the ACM International

[10] C. Vogler and D. Metaxas. Parallel Hidden Markov Models for American Sign Language Recognition. In IEEE Interna- tional Conference on Computer Vision (ICCV), 1999.

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems.

[12] M. Li, T. Zhang, Y. Chen, and A. J. Smola. Efficient Mini-batch Training for Stochastic Optimization. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014.

[13] M.-T. Luong, H. Pham, and C. D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015.

[14] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In CVPR, 2018.

[15]. Camgoz, Necati Cihan, et al. "Sign language transformers: Joint end-to-end sign language recognition and translation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[16] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep Audio-visual Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2018.

[17] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural Sign Language Translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[18] Julia Kreutzer, Joost Bastings, and Stefan Riezler. Joey NMT: A minimalist NMT toolkit for novices. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations, 2019.

[19] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation. arXiv:1609.08144, 2016.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Dr. K Anitha Sheela and J Sandhya have done data selection, methodology preparation, dataset tuning and testing.

M. Maahir and N. Sathwik have done model preparation, Implementation, training, optimization, preparation of summary on model.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.