SignSense A mode of communication for deaf and mute people

Vedant Kambli , Sahil Paleja , Advika Lad , Tasmay Sawant , Aruna Gawade, Nilesh Rathod, Komal Patil

Artificial Intelligence and Machine Learning Department , SVKM Dwarkadas J. Sanghvi College of Engineering Affiliated to Mumbai University, Mumbai India.

Abstract

In the technological realm, there has been a lot of advancement in the field of Artificial Intelligence (AI), machine learning and deep learning which help in communicating with the masses easily using numerous smart devices and tools. Even in these times Deaf and mute people face challenges while conveying their thoughts, words or views, be it in the office meetings or while talking with friends and family. On education front, deaf students are unable to understand subjects and gain knowledge, which turns as a major hurdle in their learning, overall personal growth and development. The smart AI tools are often limited by things like internet connectivity, specialized bulky or uncomfortable gadgets which might not be very practical for day-to-day communications. To resolve this problem, the paper aims on developing an application that is powered by Artificial Intelligence that can run on their mobile phones without internet connectivity to help in converting speech to sign language for deaf people to understand, and Sign language to text or Audio for mute people to speak and convey their thoughts.

This system integrates a fine-tuning algorithm, which catches on instances where the model might've made a wrong prediction and uses a human feedback method to check whether the said words were correct or incorrect. This helps the model to understand and learn the user's behavior with respect to communication using hand signs and improve over time. Overall, this aids deaf people of all ages to overcome communication barrier and promote their career and ***academic development. (technology word,GRU, result)***

Abstract result.

Keywords: Artificial intelligence, Sign language conversion, Deep Learning, Transformer, Deaf Education, day-to-day communication

Introduction

In a world that thrives on communication, the challenges faced by individuals with hearing and speech impairments are often overlooked. Deaf and mute individuals encounter barriers when attempting to convey their thoughts, words, or sentiments to the broader society. Despite the advancements in artificial intelligence (AI), machine learning, and deep learning, existing technologies for communication are hindered by limitations such as dependence on internet connectivity and the impracticality of specialized, bulky gadgets in day-to-day interactions.

This paper addresses the pressing need for an innovative solution that empowers the deaf and mute community to communicate seamlessly without relying on continuous internet connectivity or cumbersome devices. We propose the development of a mobile application driven by AI, designed to run offline on standard mobile phones. This application aims to bridge the communication gap by converting speech to text for the deaf and sign language to text or audio for the mute.

Additionally, users can leverage smart notes for quick and concise communication in dynamic environments, such as railway stations, bus stops, and crossroads.

The significance of this system becomes evident in environments where internet connectivity is slow or unavailable, rendering existing communication systems ineffective. Unlike wearable technologies with bulky designs and separate charging requirements, our system ensures accessibility without compromising on comfort or style. Furthermore, we introduce a fine-tuning algorithm that learns from user feedback, continuously improving the accuracy of the system in interpreting sign language.

This paper not only fills a crucial void in the current technological landscape but also presents a solution that evolves over time, adapting to the unique communication patterns and habits of individuals using sign language. By exploring the shortcomings of existing systems and proposing an innovative offline solution, our work strives to make meaningful contributions towards empowering the deaf and mute community in their communication endeavors.

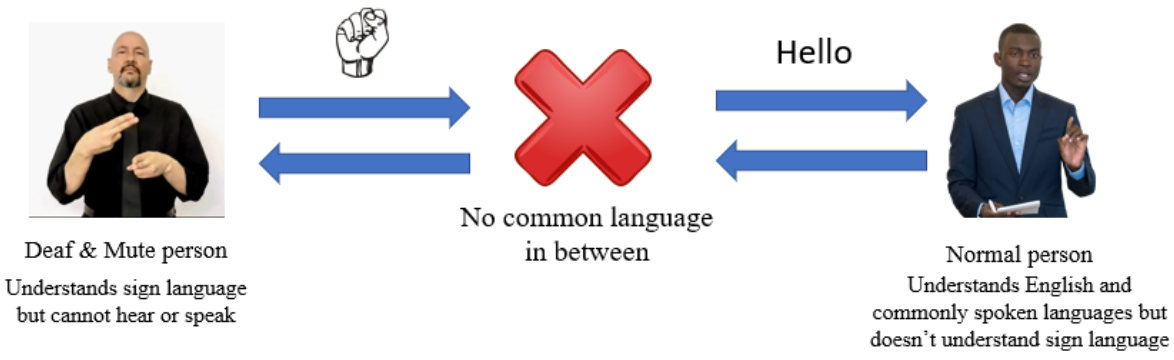

Fig.1

Problem in the everyday life of a deaf and mute person where he is unable to communicate his thoughts to normal people.

The potential for transforming the educational system and promoting inclusivity and accessibility through the integration of sign language and speech is substantial. Educational institutions can create an inclusive environment by providing equal access to learning materials for students with hearing impairments through the provision of conversion capabilities. This method not only meets the special needs of these students, but it also fosters peer collaboration and communication through a variety of expression channels. Offering a variety of learning modalities enhances the cognitive development of students with hearing impairments. Teachers can communicate more effectively, fostering clearer understanding and engagement. Moreover, the conversion of sign language to speech is in compliance with legal requirements to give students with disabilities equal access to education. It is becoming more and more possible to integrate such solutions into educational settings thanks to technological advancements in areas like computer vision and natural language processing. To ensure that education becomes more inclusive and adaptive to diverse learning styles and needs, successful implementation requires staff training, continual support, and collaboration with the deaf and hard-of-hearing community.

Examples of cases in education system:

1. The school used cutting edge technology to implement a system that converts sign language to speech. Through the use of sign language, which was translated into spoken language for the remainder of the class, the student was able to express themselves through this system. Additionally, students were given sign language translations of teachers' spoken instructions. This method improved the class's overall learning experience while also facilitating smooth communication.

2. The school district established a collaborative partnership with neighborhood associations that advocate for the community's hard-of-hearing members. Collectively, they offered insightful perspectives on communication styles, cultural quirks, and preferences. In order to ensure that the technology matched the unique needs of the students and educators, community members were also involved in the testing and improvement of the technology. By working together, a more community-supported and culturally aware sign language to speech conversion system was produced. This case study illustrated how crucial it is to include members of the deaf and hard-of-hearing community in the design and execution of inclusive educational technologies in order to ensure their long-term viability and efficacy.

Case 1: Deaf person communicating using traditional ways

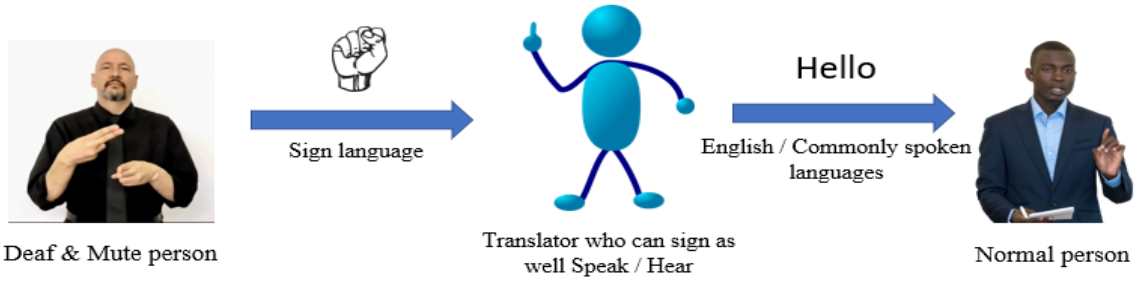

Fig.2

Case 2: Normal person communicating using traditional ways

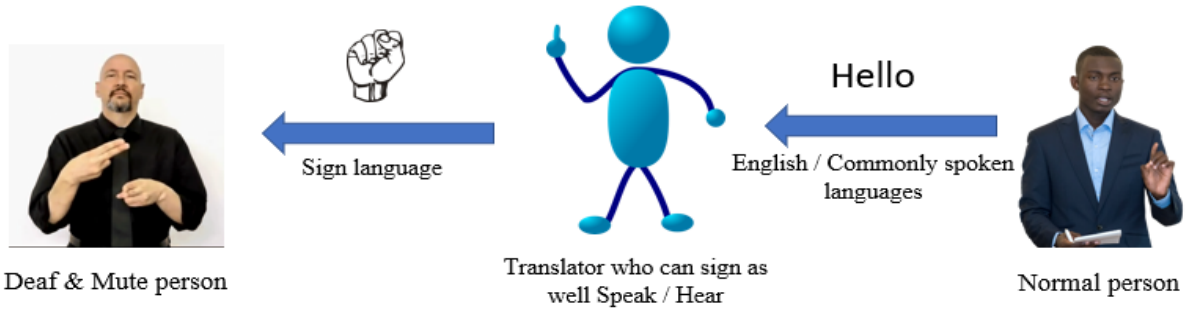

Fig.3

Case 3: Normal person communicating using Sign Sense

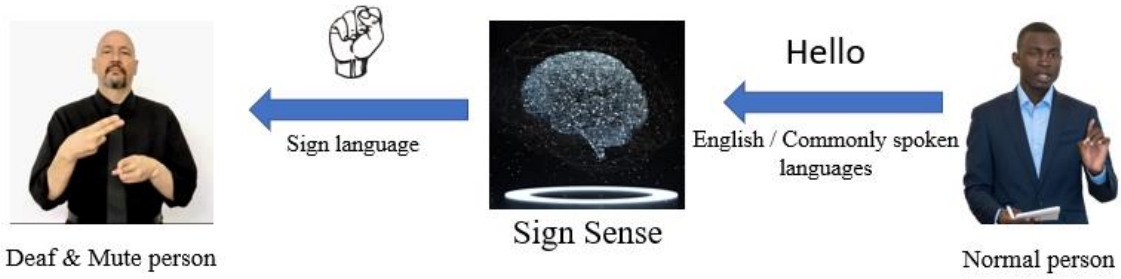

Fig.4

Case 4: Deaf Person communicating using Sign Sense

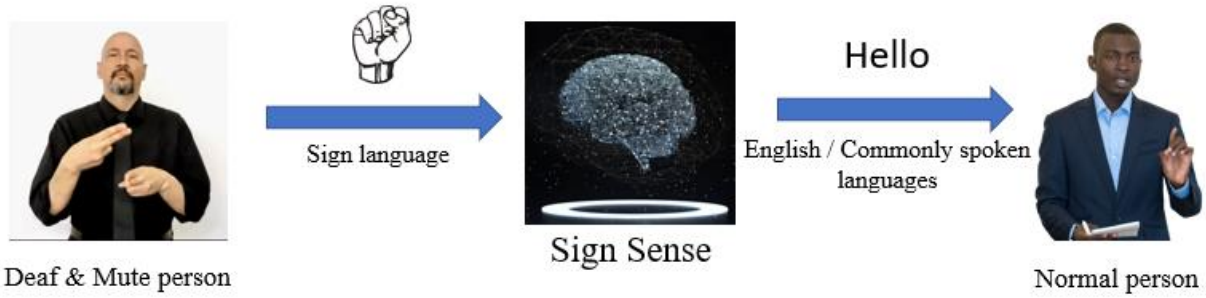

Fig.5

Methodology

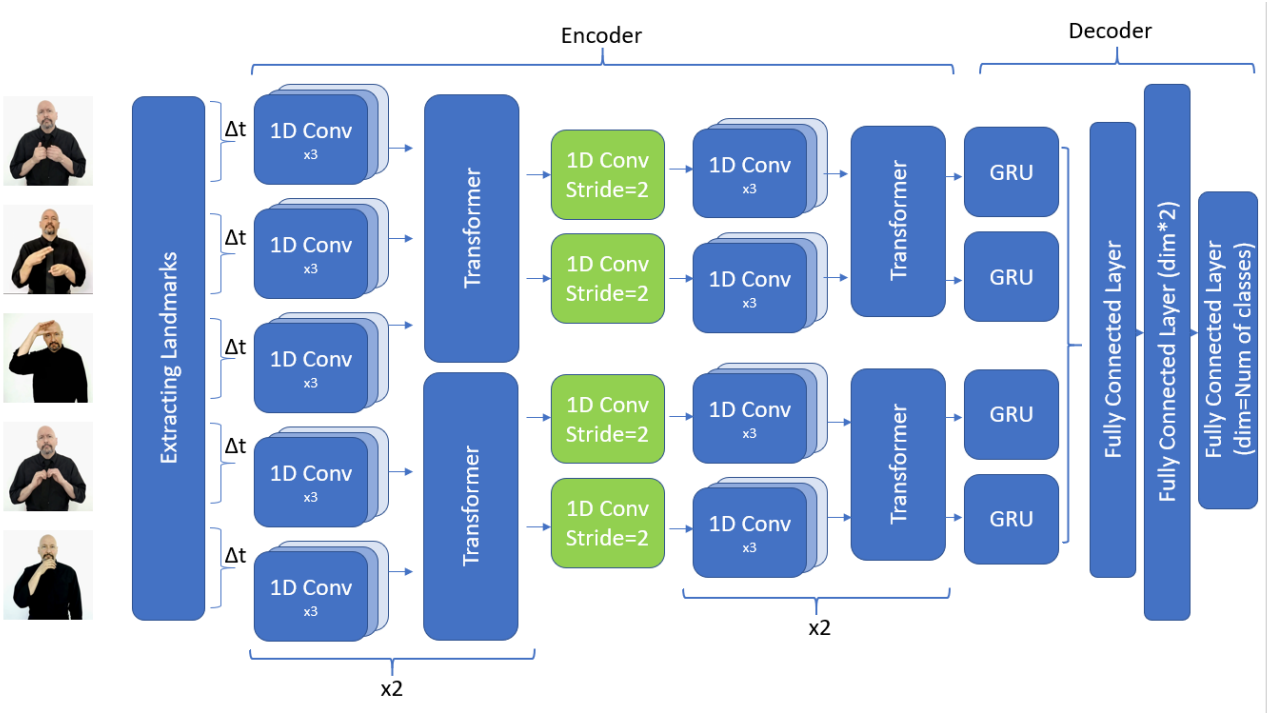

Fig.6

Data Preprocessing

The preprocessing pipeline is a meticulously crafted sequence of steps aimed at optimizing the model's ability to glean meaningful insights from sequential video data. At its core, the utilization of landmarks, spanning facial features such as left-right hand, eye, nose, and lips landmarks, forms the foundation for detailed facial motion analysis. Each landmark serves as a unique identifier, and the normalization process, anchored around the 17th landmark located in the nose, ensures a consistent reference point across different facial structures. This normalization not only facilitates

the model's robustness to translation and scale variations but also establishes a standardized coordinate system, laying the groundwork for accurate motion feature extraction.

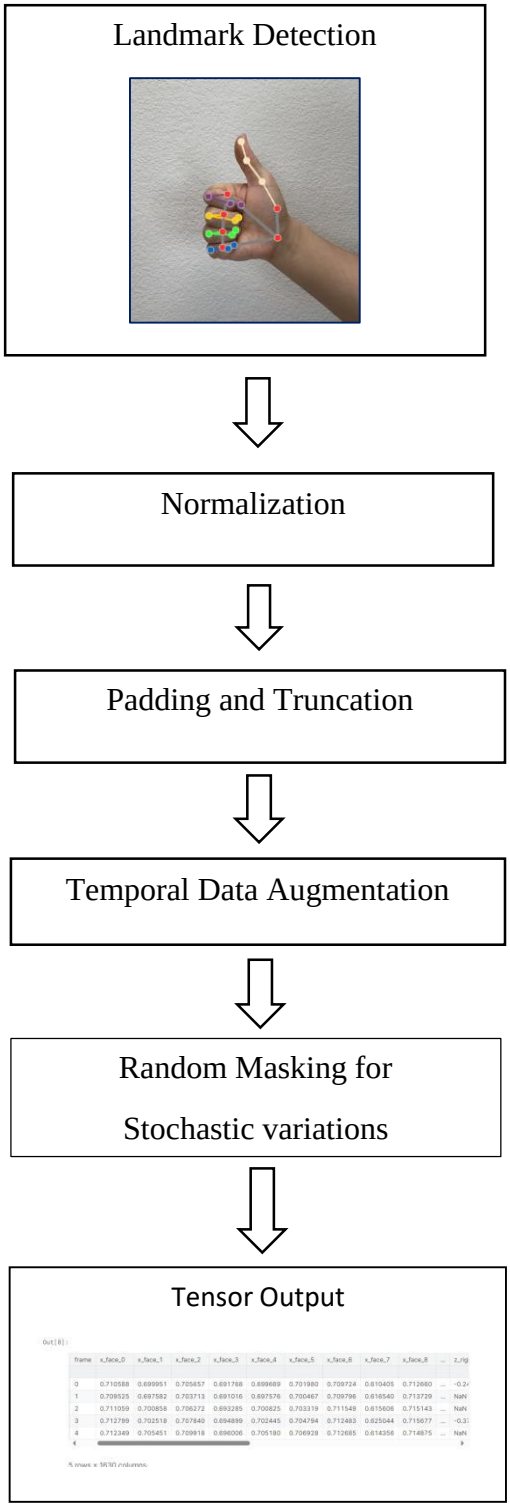

Fig.7

To address the challenge of variable-length inputs, a sophisticated masking strategy is implemented during training This strategy involves both padding and truncation, with a specified maximum sequence length of 786. A critical aspect of the model involves handling variable-length input efficiently. while for inference, only truncation is applied. Causal padding is used to maintain accurate masking indices.

Temporal augmentation techniques further enrich the dataset by introducing variability in the temporal domain. Random resampling, with variations ranging from 0.5x to 1.5x of the original length, simulates different temporal scales and durations. This augmentation strategy exposes the model to a diverse set of temporal patterns, enhancing its ability to generalize across various video lengths and speeds. Random masking introduces stochastic variations, promoting robustness by training the model to cope with missing or occluded information in the video sequences. This meticulous preprocessing ensures that the model is well-equipped to discern intricate temporal and spatial dynamics, ultimately enhancing its performance on the targeted video- based task.

Model Creation

The Deep learning model chosen as the solution for this problem is a custom-made Encoder-Decoder architecture. [38]

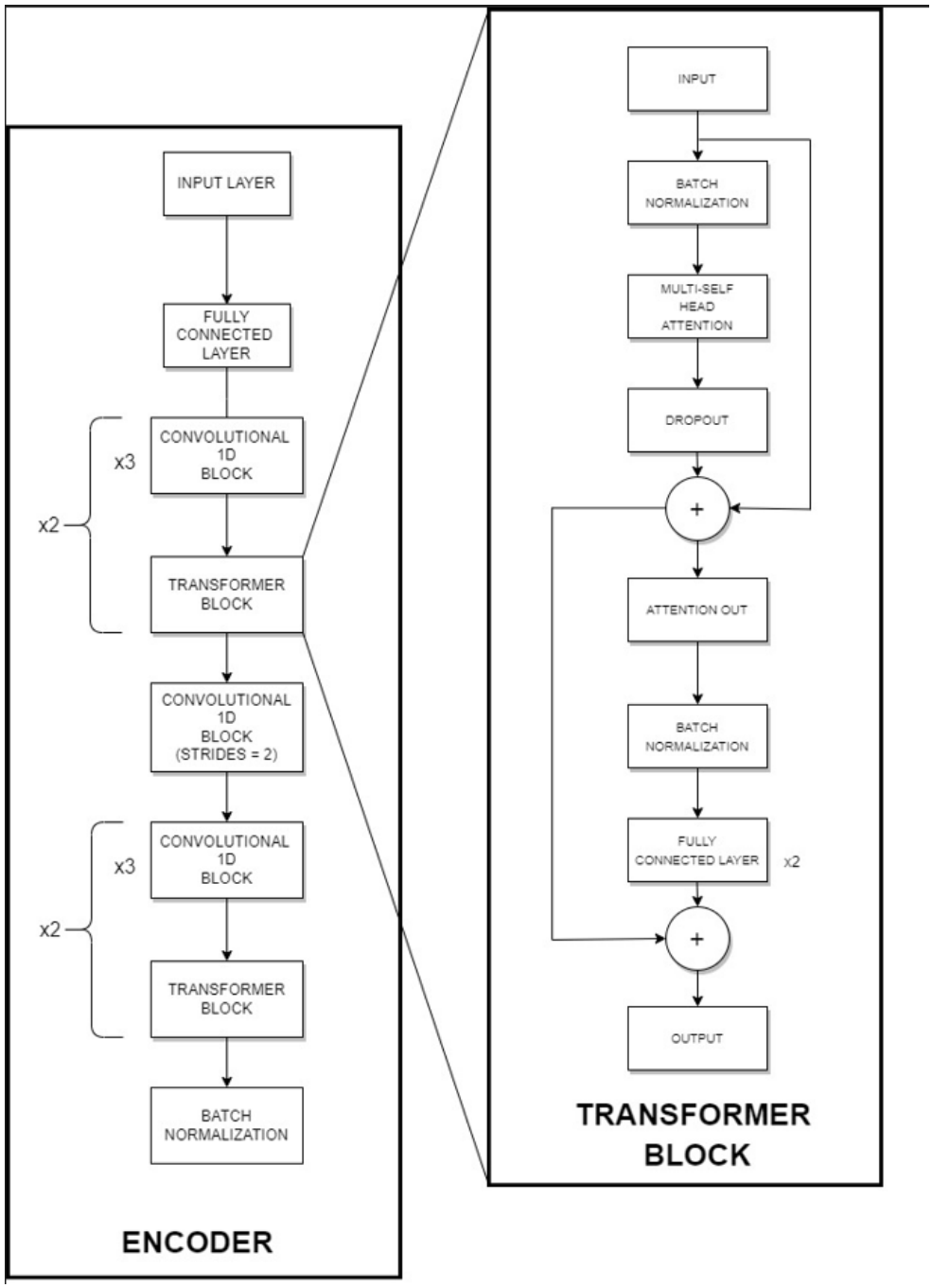

Fig.8

Encoder

The proposed Encoder model is designed for sequence-based classification tasks, accommodating variable-length input sequences. The architecture combines convolutional and transformer blocks to capture both local and global dependencies within the input sequences. A flexible design allows

for the adjustment of the model's capacity based on the specified dimension. Key components include masking for variable-length sequences, batch normalization for stabilization, and late dropout for regularization.

The reason for using one-dimensional CNN with transformer block is that in case of high inter frame correlation, 1 dimensional convolution networks perform better. Since they use convolutional operations along the temporal dimension to capture features and dependencies in the sequence.

In case of inter frame correlation and sequential data (Landmarks), one dimensional convolution might be more efficient than transformers.

Inter-frame correlation refers to the relationship or similarity between adjacent frames or elements in a sequence. For example, in video data, adjacent frames often exhibit strong correlation as they represent consecutive moments in time.

Convolutional Neural Networks (CNNs) are well-suited for capturing local patterns and dependencies in sequential data. They operate by sliding convolutional filters over the input sequence, allowing them to capture local correlations effectively. Transformers, on the other hand, excel at capturing long-range dependencies and global context in sequential data. They achieve this through self-attention mechanisms that enable each element in the sequence to attend to all other elements, allowing for effective modeling of inter-frame correlations across long distances.

Therefore, this paper thought of exploring a hybrid structure consisting of one-dimensional convolutions along with transformers to capture both, short-term as well as long term dependencies and patterns.

Conv1DBlock.The following are the reason for using Conv1DBlock

1. Size Amplification: /heading size

The Conv1DBlock, a foundational component, experiences a substantial augmentation in its expand ratio, transitioning from 2 to 4. This deliberate enlargement aims to broaden the receptive field of the convolutional operation. The heightened expand ratio empowers the block to discern intricate spatial features within input sequences, contributing to a more nuanced understanding of complex patterns. Size and Depth of the Encoder: The Conv1DBlock, serving as a fundamental building block in the Encoder, undergoes significant enhancements. Its expand ratio increases from 2 to 4, enlarging the receptive field for better spatial feature capture. The depth of the Encoder expands from 8 to 17 layers, fostering a more profound hierarchical representation of sequential data. Padding and Stride, The shift from 'causal' to 'same' padding simplifies operations, and the introduction of an output stride of 2 enhances the model's ability to capture high-level features, albeit with additional masking logic. Batch Normalization: A Batch Normalization (BN) layer is strategically introduced at the input of the Conv1DBlock to stabilize training, crucial for scenarios involving adversarial weight perturbation and an extended number of training epochs.

Depth Augmentation:

2. Layer Proliferation: The Encoder's depth undergoes a significant expansion, evolving from a prior configuration of 8 layers to a more intricate 17-layer architecture. This profound increase in depth facilitates the creation of a more detailed hierarchical representation of sequential data. The additional layers progressively distill abstract and nuanced features, empowering the Encoder with a richer understanding of input sequences. Padding Strategy: 3. Shift from 'Causal' to 'Same': The conventional 'causal' padding strategy is replaced with 'same' padding. This modification

simplifies the architecture by aligning it with standard convolutional operations. 'Same' padding enables the convolutional layers to operate on input sequences without considering future data points, potentially enhancing computational efficiency and overall model behavior.

Output Stride Introduction:

4 .Down-Sampling Logic: The introduction of an output stride of 2 incorporates a down-sampling effect within the Encoder. While necessitating additional logic for masking, this stride configuration holds promise for improving the model's ability to capture high-level features. By down-sampling input sequences, the Encoder gains a broader perspective, facilitating the extraction of more abstract and global patterns.

Batch Normalization Integration:

5. Stability Enhancement: A Batch Normalization (BN) layer is strategically introduced at the input of the Conv1DBlock. This addition aims to stabilize the training process, particularly in scenarios where adversarial weight perturbation (AWP) and an increased number of training epochs may introduce instabilities. BN enhances the model's capacity to learn robust and consistent features from input sequences, contributing to overall training stability. These nuanced modifications collectively represent a meticulous design strategy, aiming to refine the Encoder's feature extraction capabilities. The Conv1DBlock's size amplification and depth augmentation, coupled with the shift in padding strategy, introduction of output stride, and incorporation of BN, contribute to a more efficient and stable encoding process.

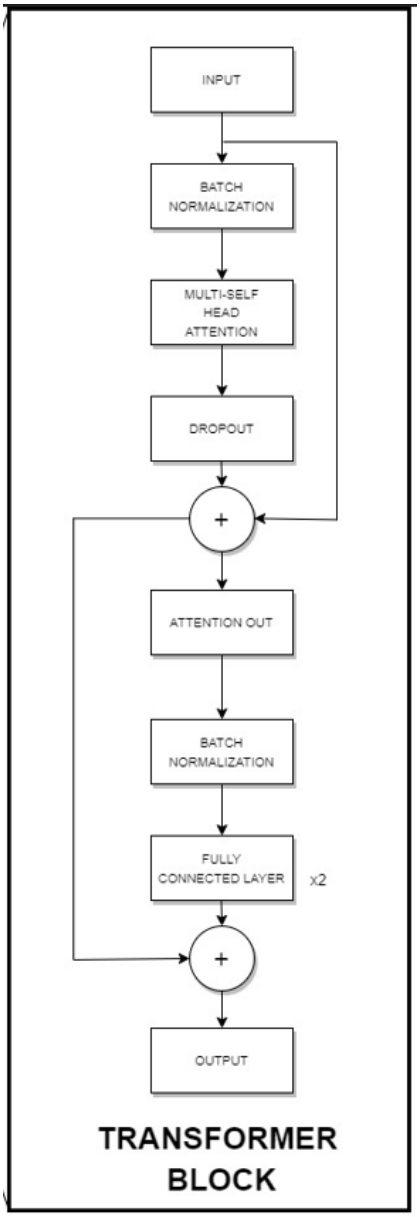

Fig.9

Decoder

The system used a single GRU layer followed by one FC layer. It performs the task of aligning variable-length input sequences with variable-length output sequences.

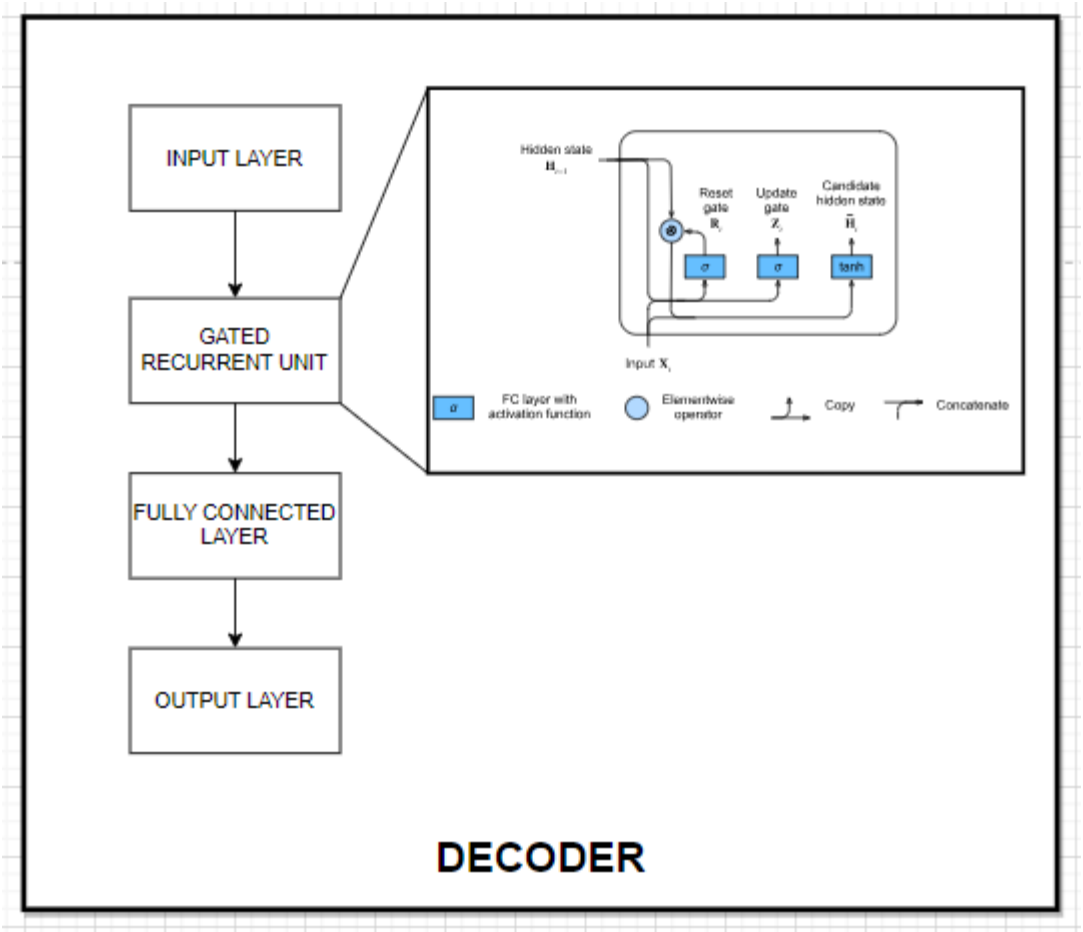

Fig.10

Gated Recurrent Unit (GRU)

The Gated Recurrent Unit (GRU) is a variant of recurrent neural networks (RNNs) designed to overcome certain challenges associated with modeling long-range dependencies in sequential data. The GRU consists of units with hidden states that evolve over time, and it incorporates two crucial gating components: the update gate and the reset gate. Sign language is inherently dynamic, with signs unfolding over time, and capturing the temporal nuances is essential for accurate interpretation. GRUs excel at this task by utilizing these gating mechanisms.

The GRU reduce the gating signals to two from the LSTM RNN model. The two gates are called an update gate Zt and a reset gate rt .the GRU model is presented in the form :

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$
$$\tilde{h}_t = g(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$$

with the two gates presented as [20]

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

It observes that the GRU RNN with less external gating signal in the interpolation . This saves one gating signal and the associated parameters. In essence, the GRU RNN has 3-folds increase in parameters in comparison to the simple RNN. Specifically, the total number of parameters in the GRU RNN equals 3×(n2 + nm +n). In various studies, it has been noted that GRU RNN is comparable to, or even outperforms, the LSTM, Moreover, there are other reduced gated RNNs, e.g. the Minimal Gated Unit (MGU) RNN.

CTC Classifier

The CTC classifier is particularly well-suited for sequence-to-sequence tasks, such as converting sign language gestures to text. CTC is designed to handle variable-length alignments between input and output sequences, making it effective for tasks where the temporal alignment may not be one-to-one.

CTC Loss:

$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} p_t(y_t|\mathbf{x})$$

where pt is the softmax output distribution, and pt(yt|x) is a conditional probability of output yt at t-th time step given input sequence x.

Let B is a mapping function from a path y to a label sequence l by removing the repeated and blank labels. The conditional probability of l can be expressed as the sum of probabilities of all the corresponding paths:

$$p(\mathbf{l}|\mathbf{x}) = \sum_{\mathbf{y}:\mathcal{B}(\mathbf{y})=\mathbf{l}} p(\mathbf{y}|\mathbf{x})$$

The CTC loss is defined as the negative log probability of correctly labelling the sequence:

$$CTC(\mathbf{l}|\mathbf{x}) = -log(p(\mathbf{l}|\mathbf{x}))$$

CTC allows the model to learn to align the input and output sequences without requiring explicit alignment information during training.

The model predicts a sequence of characters at each timestep, and CTC handles the mapping between the predicted sequence and the target sequence.

Fully Connected Layer

After the GRU processes the sequential input data, the output from the last time step is passed through a Fully Connected layer. The purpose of this FC layer is to transform the high-dimensional output from the GRU into a format suitable for the subsequent CTC classifier. The FC layer

introduces weighted connections between the GRU output and the CTC, allowing the model to learn a mapping between the temporal features and the desired output sequence.

Result

```
Target    : tracee roberson
Prediction: tracee roberson
-------------------------------------------------------
Target    : https://www.bridgat.com/
Prediction: https://www.bridgat.com/
-------------------------------------------------------
Target    : /destructionguepesfrelons
Prediction: /destructionguepesfrelona
-------------------------------------------------------
Target    : turgut-gozutok/305459/bowman
Prediction: turgut-gozutk/305459/bowman
-------------------------------------------------------
Target    : 165-685-2563
Prediction: 1765-685-2563
-------------------------------------------------------
```

Fig.11

Evaluation prediction with testing data

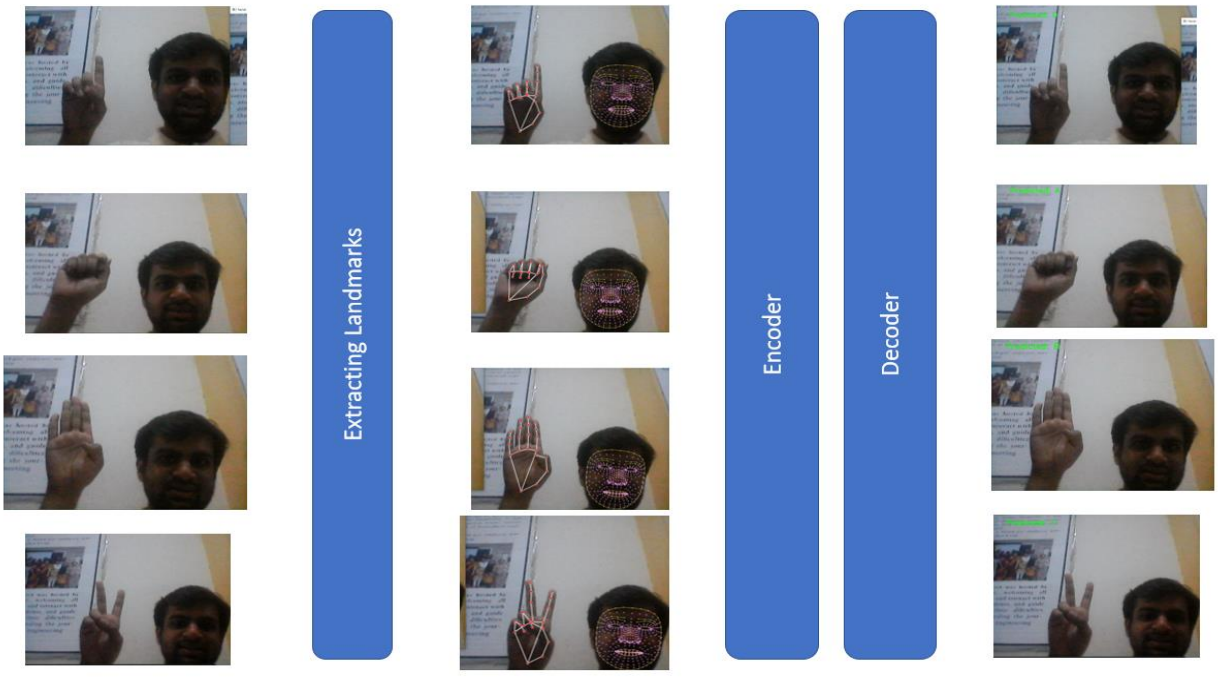

Fig.12

Conclusion

In conclusion, the proposed system, “Sign Sense”, presents a promising solution for bridging the gap between the hearing-impaired and those unfamiliar with sign language.

The Encoder-Decoder architecture, featuring 1D convolutional blocks, transformer blocks in the encoder, and a GRU layer connected to a fully connected layer in the decoder with a CTC (Connectionist Temporal Classification) classifier, represents a sophisticated and effective model for interpreting sign language and converting it to text. This architectural choice allows for the capture of intricate temporal dependencies, crucial in understanding the nuances of communication.

One of the standout features of our system is its offline functionality, ensuring accessibility in environments where internet connectivity is unreliable or absent. This addresses a critical gap in existing technologies that heavily rely on continuous internet access.

In addition to technical innovation, our work recognizes the importance of user comfort and style. By steering clear of bulky wearable technologies with separate charging requirements, our mobile application is designed for practical, day-to-day use without compromising on aesthetics.

This system can be used in educational institutes by the teachers to help the students understand the subject as the students will be able to communicate their doubts clearly which will improve their understanding in the subject. They will feel more confident in themselves while expressing their thoughts and views, also will not feel deprived of opportunities as compared to the abled people.

Disclosure Statement

Disclosure of interest

The authors report no conflict of interest

References:

1) P. Pryandi, M. Bayu Dewantara, H. L. Hendric Spits Warnars, A. Ramadhan, N. Noordin and F. Hanis Abdul Razak, "Smartphone Application for the Deaf and the Deaf Caring Community," 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2023, pp. 767-773, doi: 10.1109/ICICT57646.2023.10134458.
2) P. Golîmba and L. Stanciu, "Android Application to Support Communication Between Romanian Hearing and Deaf Peoples," 2022 IEEE 16th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, 2022, pp. 000169-000172, doi: 10.1109/SACI55618.2022.9919511.
3) R. Rastogi, S. Mittal and S. Agarwal, "A novel approach for communication among Blind, Deaf and Dumb people," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2015, pp. 605-610.
4) R. Ganguly, R. Das, R. Bose, S. Sindal and T. K. Rana, "Intelligent Gloves for Deaf and Dumb People," 2021 5th International Conference on Electronics, Materials Engineering &

Nano-Technology (IEMENTech), Kolkata, India, 2021, pp. 1-4, doi: 10.1109/IEMENTech53263.2021.9614717.

5) Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, “Con- ´ nectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” ACM/IEEE Supercomputing Conference (SC), vol. 148, pp. 369–376, 2014.
6) V. Setiawan et al., "An Interactive Sign Language Based Mobile Application for Deaf People," 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2023, pp. 1635-1641, doi: 10.1109/ICOEI56765.2023.10125821.
7) A. Wisesa, W. Andriyani, T. Suprawoto and Hamdani, "Development of Learning Media for The Deaf Using a Webcam," 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2022, pp. 160-165, doi: 10.1109/ISRITI56927.2022.10052934.
8) J. Dhruv and S. Kumar Bharti, "Real-Time Sign Language Converter for Mute and Deaf People," 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), Gandhinagar, India, 2021, pp. 1-6, doi: 10.1109/AIMV53313.2021.9670928.
9) S. Tovide, W. D. Tucker and O. O. Ajayi, "SignSupport: An Emergency Mobile Application for the Deaf," 2022 IST-Africa Conference (IST-Africa), Ireland, 2022, pp. 1-13, doi: 10.23919/IST-Africa56635.2022.9845605.
10) M. R. Chilukala and V. Vadalia, "A Report on Translating Sign Language to English Language," 2022 International Conference on Electronics and Renewable Systems (ICEARS),Tuticorin,India,2022,pp.1849-1854,doi:10.1109/ICEARS53579.2022.9751846.
11) S. Tornay, M. Razavi and M. Magimai.-Doss, "Towards Multilingual Sign Language Recognition," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6309-6313, doi: 10.1109/ICASSP40776.2020.9054631.
12) Kim, Suyoun, Takaaki Hori, and Shinji Watanabe. "Joint CTC-attention based end-to-end speech recognition using multi-task learning." 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017.
13) Alex Graves and Navdeep Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014, pp. 1764–1772.
14) Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al., “Deep speech: Scaling up endto-end speech recognition,” arXiv preprint arXiv:1412.5567, 2014.
15) Yajie Miao, Mohammad Gowayyed, and Florian Metze, “EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding,” in 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015, pp. 167–174.
16) Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “End-to-end continuous speech recognition using attention-based recurrent NN: First results,” arXiv preprint arXiv:1412.1602, 2014.

17) Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, “End-to-end attention-based large vocabulary speech recognition,” arXiv preprint arXiv:1508.04395, 2015.
18) Liang Lu, Xingxing Zhang, and Steve Renals, “On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5060–5064.
19) William Chan and Ian Lane, “On online attention-based speech recognition and joint mandarin character-pinyin training,” Interspeech 2016, pp. 3404–3408, 2016.
20) Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
21) Hori, Takaaki, Shinji Watanabe, and John R. Hershey. "Joint CTC/attention decoding for end-to-end speech recognition." Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017.
22) William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. arXiv preprint arXiv:1508.01211 .
23) Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton, “Chainer: a next-generation open source framework for deep learning,” in Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS), 2015.
24) Matthew D Zeiler, “Adadelta: an adaptive learning rate method,” arXiv preprint arXiv:1212.5701, 2012.
25) Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio, “On the difficulty of training recurrent neural networks,” arXiv preprint arXiv:1211.5063, 2012.
26) Sepp Hochreiter and Jurgen Schmidhuber, “Long short-term ¨ memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
27) John Garofalo, David Graff, Doug Paul, and David Pallett, “CSR-I (wsj0) complete,” Linguistic Data Consortium, Philadelphia, vol. LDC93S6A, 2007.
28) Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014.
29) Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” Signal Processing Magazine, IEEE, vol. 29, no. 6, pp. 82–97, 2012.
30) Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton, “Acoustic modeling using deep belief networks,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, pp. 14–22, 2012.
31) Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In Conference on Empirical Methods on Natural Language Processing (EMNLP). volume 4, pages 230–237.
32) Alex Graves, Santiago Fernandez, Faustino Gomez, ´ and Jurgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In International Conference on Machine learning (ICML). pages 369–376.

33) Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In Advances in Neural Information Processing Systems (NIPS). pages 577–585.
34) Herve Bourlard and Nelson Morgan. 1994. ´ Connectionist speech recognition: A hybrid approach. Kluwer Academic Publishers.
35) steven Bird. 2006. NLTK: the natural language toolkit. In Joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL) on Interactive presentation sessions. pages 69–72.
36) Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 .
37) Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, “End-to-end attentionbased large vocabulary speech recognition,” arXiv preprint arXiv:1508.04395, 2015.
38) https://www.kaggle.com/competitions/asl-fingerspelling/discussion/434588

Ctc

Deaf

Convo

Trans

Dl rel paper