Fast-convergence GRU Model for Indian Sign Language Recognition

Barathi Subramanian, Bekhzod Olimov, Jeonghong Kim*
Dept. of Computer Science and Engineering, Kyungpook National University
e-mail: jhk@knu.ac.kr

Summary

Sign language recognition faces several challenges such as accurate tracking of hand gestures, occlusion of hands, and high computational cost. Currently, these challenges are addressed using the advancements in deep learning methods. Nevertheless, these larger complex approaches cannot manage long-term sequential data and are distinguished by poor information processing and learning efficiency in capturing useful information. To solve these problems, we propose an integrated MediaPipe-optimized gated recurrent unit (MOPGRU) model for Indian sign language (ISL) recognition. Specifically, we improved the update gate of the standard GRU cell by multiplying it by the reset gate to discard the redundant information from the past in one screening. Additionally, we replace the hyperbolic tangent activation in standard GRUs with exponential linear unit (ELU) activation and SoftMax with Soft sign activation in the output layer of the GRU cell. Thus, our proposed model achieved better prediction accuracy with higher learning efficiency than other sequential models.

1. Introduction

Sign language is a visually interactive language used by people who have difficulties in hearing. Any form of effective communication should be understandable for both the informant and the recipient. In the real-life situations, there is always a communication barrier between the normal hearing community and the people with hearing impairments. Similar to spoken languages, sign language also differs from one geographic region to others; therefore, it is hard for many individuals to understand gesture-based communication that leads to misinterpretation between people [1].These issues can also cause more serious problems like learning difficulties, social isolation, and depression for the individuals with hearing disabilities.

From the World Health Organization report, approximately 500 million people worldwide suffer from hearing loss[2], [3]. Because of the high prevalence of the hard-of hearing community population, there has been an increased interest in eliminating communication obstacles faced within the hard-of-hearing community and other people with normal hearing [4]. Hence, Sign language recognition (SLR) develops an assistive system that automatically converts an input sign into its corresponding speech or text [5]. SLR system is not only useful for overcoming the communication gap between hearing and non-hearing communities but also creates a new path for human-computer interaction-based applications. [6].However, the major challenge in developing a continuous SLR system is finding a modeling prototype that acquires the sign gesture and its corresponding text. Therefore, in this study we propose a SLR system for one of most complex language: Indian Sign Language (ISL) because of the following reasons[7]:

- It uses both single and double hands for portraying gestures.
- It includes both static and dynamic gestures.
- It uses similar gestures to represent different things.

Considering these drawbacks of ISL, we propose MediaPipe-optimized gated recurrent unit (MOPGRU) SLR system that diminishes the problem of hand occlusion and lower learning efficiency by adjusting the output of the update gate using the reset gate integrated with an open-source framework called the MediaPipe Holistic pipeline[8] . The proposed model helps to understand the gestures (made by the hearing difficulty persons) for the normal people without having prior knowledge about sign language. Furthermore, we replace the hyperbolic tangent (Tanh) activation function in the candidate memory state with an ELU activation function to overcome the vanishing and exploding gradient problem. We also change the activation function in the output layer of each GRU cell as Softsign

instead of SoftMax to achieve a faster, simpler, and cost effective SLR system.

As a result of this modified setup, we observed that our proposed model extracted the desired features with higher learning efficiency than other sequential models along with lower training time. We created our own dataset by capturing images in real-time for our experimental study.

The rest of the paper is organized as follows. Section 2 contains detailed explanation of the proposed methodology. Section 3 represents experimental results. Finally, Section 4 concludes this study with the conclusion and future work.

2. Proposed Methodology

To accurately recognize the sign gestures and translate them into text, our proposed method consists of three stages: The first stage involves the data preprocessing and feature extraction, followed by the next stage with data cleaning and labelling and concluding stage covers the gesture recognition. Figure 1. shows an overview of the proposed method.

2.1. Stage 1: Data Preprocessing and Feature extraction

Data preprocessing and feature extraction is carried over by the MediaPipe framework. Here, features from the face, hands, and body are extracted as keypoints and landmarks using built-in data augmentation techniques from sequence of input frames taken from a web camera. We apply a multistage pipeline from MediaPipe, called MediaPipe Holistic. The workflow of stage 1 is briefly discussed below:

- The human pose and subsequent landmark model were estimated using BlazePose's poses detector. Then, three ROI crops for the face and hands (2×) were derived from the inferred pose landmarks, and a recrop was employed to improve the ROI.
- Next, the corresponding landmarks were estimated. To achieve this, the full-resolution input coordinates were cropped to the ROIs for task specific hand and face models.
- Finally, all landmarks were combined to yield the full 540+ landmarks.

2.2. Stage 2: Data Cleaning and Labelling

In this stage, the extracted keypoints from stage 1 are saved in a file to identify and the null entries are removed from the data, after which data labelling follows. After stage 1, the extracted features, that is, the landmark points (21 * 3+21 * 3+33 * 4+468 * 3 =1662) per frame are flattened, concatenated and stored in a file to check and remove any null entries from the data. Data cleaning is important since it prevents failed detection of features, which occurs when a blurred image is sent to the detector and leads to a null entry into the dataset [9], [10]. Thus, when training occurs with this noisy data, the prediction accuracy is reduced, and bias may occur. To fit the obtained data for the next stage of training, testing and validation, labels are created for each class and their corresponding frame sequences are stored.

2.3. Stage 3: Gesture Recognition

In stage 3, the cleaned and labelled gestures are trained and classified by our MOPGRU model for ISL recognition with the translated sign gestures in the form of text on the screen.

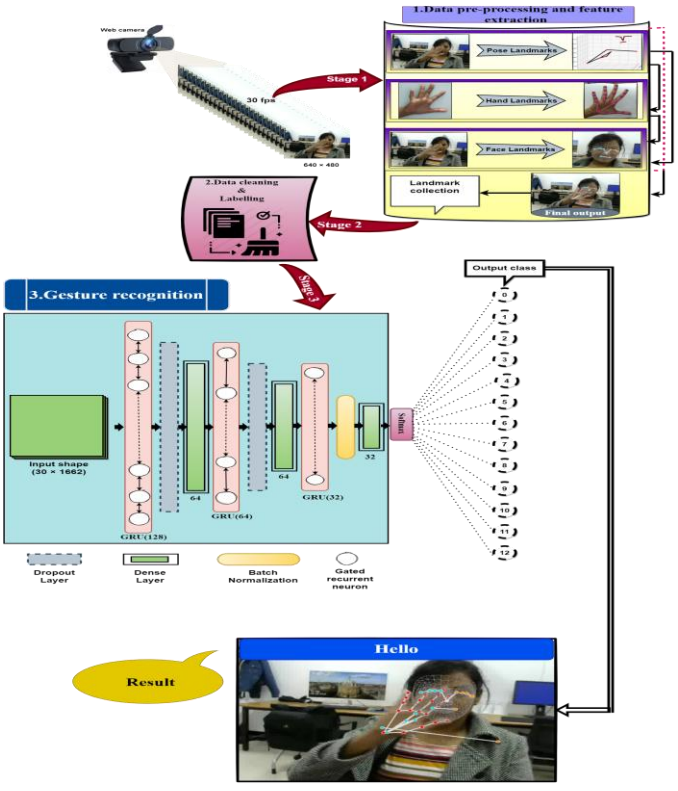


Figure 1. General overview of the proposed method

In the proposed MOPGRU model, the cleaned and labelled data from stage 2 are used. The major alteration performed in the standard GRU cell is that its update gate is improved, the candidate memory activation function (tanh) is substituted by the ELU activation function, and the output layer SoftMax activation function by Softsign respectively. Considering the problems, such as low learning efficiency, high computational cost, slow convergence rate and incapable of dropping out the unwanted information in one screening with the complex state of time series data of the standard GRU model, we propose a MOPGRU neural network model by improving the update gate: that is, modifying the original update gate input to update gate multiplied by reset gate. Thus, the update gate is adjusted by obtaining feedback from the output of the reset gate. The opposite effects caused by unnecessary information is profoundly avoided by refining the present input information by the reset gate attains faster convergence and efficient in learning.

The other change implemented in the standard GRU cell was replacing the candidate memory tanh activation function with the ELU function. The use of the tanh activation function in training feed-forward connections, especially in standard GRU, is ineffective as shown in the performance decline when the network has deeper connection [11]. Compared to other activation functions, ELU achieves higher accuracy and faster convergence in less training time and computational complexity since the negative ELU value permits the mean unit to shift toward 0 [12].

Softmax activation is specifically used to normalize the outputs and convert the weighted sum values to probabilities that sum to one by exponentiating the features and scale with the sum of the exponents. Because of its exponential nature, it is computationally expensive and time-consuming to train the model. Hence, we incorporated the Softsign activation function

3.Experimental Results

3.1. Dataset Description:

A real-time dataset of 30 videos for each sign gesture with 30 frames was created, each with a size of 640 × 480, using a web camera in various directions and different lighting conditions. We separated the collected dataset in a ratio of 70:15:15 to form the corresponding training, testing, and validation datasets.

3.2. Experimental Simulation:

The simulation was conducted using Python 3.7 version on a desktop computer with 32 GB RAM and an Intel Core i7 processor with a frequency of 3.60 GHz, running on Windows 10 Pro with a 64-bit operating system. The input image was captured using a web camera with a resolution of 720pixels/30 fps of RGB images. The first layer of GRU accepts a sequence of landmark keypoints (1662) extracted from the frames of each video with the time step of 30 as each video contains 30 frame sequences. For the hidden units at the first time-step, the input-hidden vectors and all elements of the initial hidden state were set to 0. The training data were set to 100 epochs, with 13 gestures in each batch. To optimize the network, the Adam optimization method was used and was set to 10- 4 with exponential decay rates of 0.9 and 0.999, respectively.

To evaluate the performance of our MOPGRU model, we used the mean squared error (MSE), mean absolute error (MAE), and R-squared or coefficient of determination metrics, as listed in Table 1.By comparing the models with respect to both testing accuracy and loss as shown in Figure 2 and 3 , we conclude that our proposed MOPGRU model achieved the highest accuracy of 95% and has a faster convergence speed with least loss.

Table 1.Performance comparison on MAE, MSE, R-Squared

Network Model	MAE	MSE	R-Squared
Simple RNN	4.10	28.90	-1.38
LSTM	0.75	4.95	0.59
Standard GRU	0.44	1.38	0.83
BiGRU	0.40	2.50	0.79
BiLSTM	0.85	5.35	0.56
MOPGRU	0.22	1.34	0.88

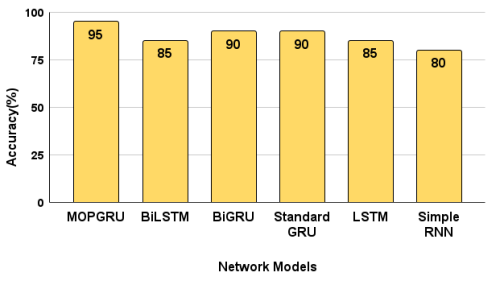


Figure 2. Model comparison in terms of Test accuracy (in %)

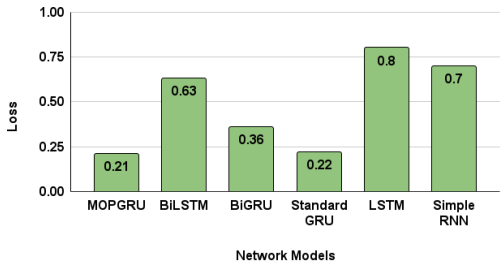


Figure 3. Model Comparison in terms of Test loss

4.Conclusion

In this study, we proposed a MOPGRU model for ISL recognition. Specifically, we modified the update gate of the standard GRU cell by multiplying its output by the reset gate. With our improved update gate mechanism, the output of the reset gate re-screens, the information and removes the unwanted information in the data, thereby giving more attention to the important information. We have cost-effectively implemented the model along with improved learning efficiency, prediction accuracy, and information processing capability of the standard GRU neural network.

참고문헌

[1] R. A. Kadhim and M. Khamees, “A real-time american sign language recognition system using convolutional neural network for real datasets,” *TEM J.*, vol. 9, no. 3, pp. 937- 943, 2020, doi: 10.18421/TEM93-14.

[2] T. R. Gadekallu *et al.*, “Hand gesture classification using a novel CNN-crow search algorithm,” *Complex & Intelligent Systems*, vol. 7, no. 4. pp. 1855- 1868, 2021, doi: 10.1007/s40747-021-00324-x.

[3] N. B. Ibrahim, H. H. Zayed, and M. M. Selim, “Advances, Challenges and Opportunities in Continuous Sign Language Recognition,” *J. Eng. Appl. Sci.*, vol. 15, no. 5, pp. 1205- 1227, 2019, doi: 10.36478/jeasci.2020.1205.1227.

[4] O. Koller, “Quantitative Survey of the State of the Art in Sign Language Recognition,” no. August 2020, 2020, [Online]. Available: http://arxiv.org/abs/2008.09918.

[5] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri, “A Modified LSTM Model for Continuous Sign Language Recognition Using Leap Motion,” *IEEE Sens. J.*, vol. 19, no. 16, pp. 7056- 7063, 2019, doi: 10.1109/JSEN.2019.2909837.

[6] B. Kanisha, V. Mahalakshmi, M. Baskar, K. Vijaya, and P. Kalyanasundaram, “Smart communication using tri-spectral sign recognition for hearing-impaired people,” *J. Supercomput.*, vol. 78, no. 2, pp. 2651- 2664, 2022, doi: 10.1007/s11227-021-03968-1.

[7] N. K. Bhagat, Y. Vishnusai, and G. N. Rathna, “Indian Sign Language Gesture Recognition using Image Processing and Deep Learning,” *2019 Digit. Image Comput. Tech. Appl. DICTA 2019*, 2019, doi: 10.1109/DICTA47822.2019.8945850.

[8] F. Zhang *et al.*, “MediaPipe Hands: On-device Real-time Hand Tracking,” 2020, [Online]. Available: http://arxiv.org/abs/2006.10214.

[9] B. Olimov, S. Karshiev, E. Jang, S. Din, A. Paul, and J. Kim, “Weight initialization based-rectified linear unit activation function to improve the performance of a convolutional neural network model,” *Concurr. Comput. Pract. Exp.*, vol. 33, no. 22, pp. 1- 11, 2021, doi: 10.1002/cpe.6143.

[10] B. Olimov, J. Kim, and A. Paul, “DCBT-Net: Training Deep Convolutional Neural Networks with Extremely Noisy Labels,” *IEEE Access*, vol. 8, pp. 220482- 220495, 2020, doi: 10.1109/ACCESS.2020.3041873.

[11] C. Gulcehre, M. Moczulski, M. Denil, and Y. Bengio, “Noisy activation functions,” *33rd Int. Conf. Mach. Learn. ICML 2016*, vol. 6, pp. 4457- 4466, 2016.

[12] S. H. Djork-Arne Clevert, Thomas unterthiner, “fast and accurate.pdf,” in *Fast and accurate deep network learning by exponential linear unit(ELUs)*, 2016, pp. 1- 14.