Performance and Computational Efficiency Comparison of LSTM and GRU Networks for Sign Language Recognition

[1]Ibrahim CETINKAYA, [2]PROF. DR. TAMER OLMEZ

***Abstract*—**Sign Language Recognition (SLR) is an crucial task to overcome communication issues between sign language users and others. Since sign language consist of complex hand and body gestures, it is challenging problem to build a system that recognize sign language in real time. Recently skeleton based methods are proposed due to avoid complexity between signers and background. In this paper only skeleton keypoints from RGB only video streams has been used. This is the main differences from other approaches. Pretrained Whole-body pose used to extract skeleton keypoints from each video. This paper aims to achieve good performance with only RGB based videos and also compare the two Recurrent Neural Networks (RNN) models, Gated Recurrent Unit and Long Short Term Memory in terms of performance and computational efficiency. Hence proposed systems uses skeleton data extracted from RGB video, Deep Bidirectional Gated Recurrent Unit and Deep Bidirectional Long Short Term Memory model for sign language recognition. Also in this work some algorithmic techniques to train model efficiently e.g., data transformation and augmentation has been used. In terms of GPU Accesing Memory, GPU Utilization , GPU Time Spent Accessing Memory GRU model outperforms LSTM. However LSTM is better in terms of accuracy and training time**.**

***Keywords*—** LSTM , GRU , SLR , Skeleton Keypoints.

I. INTRODUCTION

Sign language is a visual language performed with the dynamic movement of hand gestures, body posture and facial expressions [1]. SLR is a more complex challenge than traditional action recognition. For sign language to be clearly and precisely expressed, both subtle arm/hand motions of whole body motion are necessary. Additionally, emotions can be expressed through facial expression. Similar gestures can even have different meanings based on how many times they are repeated. Sign language recognition can be more difficult since different signers may perform sign language differently. Approaches based on multiple modalities of data [2,3] achieves higher accuracy with complex actions. However some these approaches has high computational cost and for that reason it is harder to train such models.

In this paper only skeleton keypoints from RGB only video streams has been used. This is the main differences from other approaches. We used Whole-body pose to extract skeleton keypoints from each video. Those keypoints are fed into BLSTM and BGRU models to learn spatio temporal relationship. Another challenge is that we used RNN models LSTM and GRU compared in terms of performance and computational efficiency for Sign Language Recognition from skeleton keypoints that are extracted from RGB based videos. Our experiments shows that in terms of training time LSTM outperforms GRU.LSTM also slightly performs better than GRU model. However in terms of computational efficiency GRU is better.

The paper is organized as follows: In Sect. 2, related work briefly discussed. In Sect. 3. methodology of the proposed models are presented. In Sect. 4, dataset and evaluation metrics is shared. In Sect.5 training details shared and In Sect 6, the results of the experiments is discussed..I finish the paper with a short conclusion and future works

II. RELATED WORKS

Sign Language Recognition (SLR) achieves significantprogress and obtained high recognition accuracy in recently years due to the development on practical deep learning architectures and the surge of computational power [4,5,6,7]. [8] incorporates bidirectional recurrence and temporal convolutions together which demonstrates the effectiveness of temporal information in gesture related tasks.

Skeleton based Sign Language Recognition is focuses on spatio temporal relationship between extracted skeleton keypoints coordinates. Skeleton data can be utilized individually to perform efficient action recognition [10,11,12]. On the other hand, it can also be collaborated with other modalities to achieve multi-modal learning aiming for higher recognition performances [13]. RNNs are once popular for modeling skeleton data [12,14].

There exist a few CNN and LSTM based approaches for activity recognition from RGB-only data [15,16]. However most of these works generally focuses on LSTM. In this paper we also investigate computational performance and accuracy of GRU model and compare these two RNN models..

III. MATH

In this section, end to end framework for sign language recognition from RGB videos is presented. Both BGRU and BLSTM models is using same framework. Each step in the framework will be discussed in the following subsections.

[1,2]Istanbul Technical University TURKEY

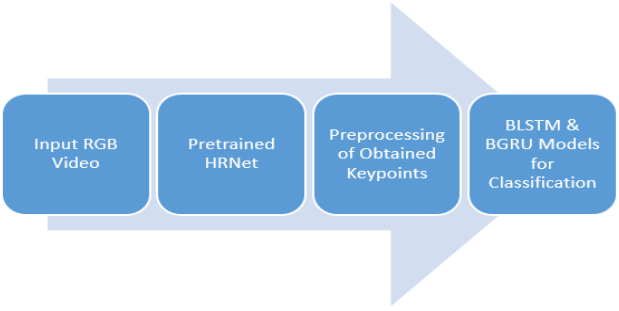

Fig. 1:Overview of proposed method

A. *Overview*

The proposed architecture aims to classify sign language gestures from RGB videos.The general overview of the system can be seen in Figure1.First skeleton keypoints are extracted from RGB raw videos with Whole-body Pose Keypoints. Then skeleton keypoints are preprocessed to improve the quality of the features. After preprocessing is done this data is used to train in our classifiers. BGRU and BLSTM models are used as our classifiers. Overfitting is major problem when deal wih limited data. Therefore some other methods are also applied to prevent overfitting such as dropout, L2 regularization. Each of these steps will be discussed in details in the following subsections.

B. *Whole-body Pose*

Pretrained HRNet [17] is used for whole-body pose estimator provided by MMPose [18] to estimate 44-point whole-body keypoints from the RGB videos. The input of the architecture is raw RGB image and the output is 44 pose keypoints.

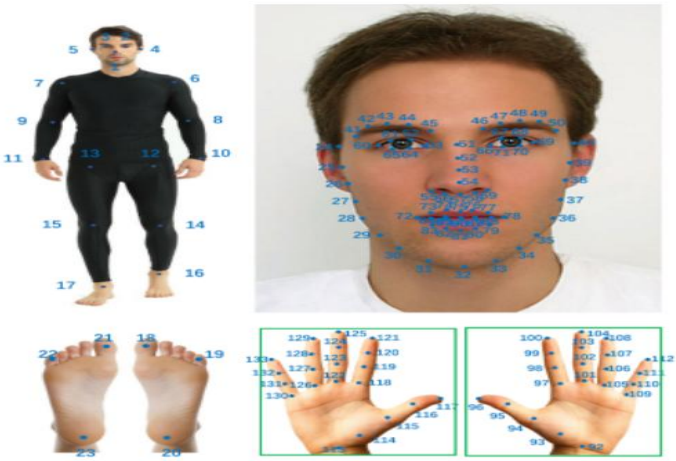

Fig 2:COCO-WholeBody annotation for 133 keypoints

C. *Recurrent Neural Network*

RNNs, recurrent neural networks, are built to process sequential input. Text, audio, video and time series data can be considered as sequential data. RNN generates the current output by utilizing previous information in the sequence. However RNNs have a short term memory issue. It is caused by the vanishing gradient problem. The network does not learn the effect of previous inputs as it processes more steps. Thus the short term memory occurs. To solve this problem two specialised version of RNN were developed; Gated Recurrent Unit (GRU), Long Short Term Memory (LSTM).

C1. Long Short Term Memory (LSTM) – Gated Recurrent Unit (GRU)

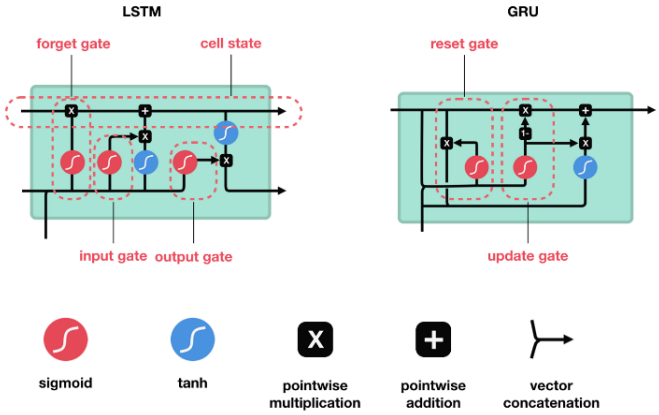

Fig 3:LSTM & GRU Architectures

LSTM and GRU [19] are a Recurrent Neural Networks (RNN) descendant that was specifically designed to adapt long term dependencies when modeling sequential data. RNNs significatly suffers from vanishing gradient problem and this makes RNNs hard to train espicially when modeling long sequences[20]. LSTM and GRU adds non linear gates to solve gradient vanishing problem by maneging information flow, so that long term sequences can be learned. Vanilla LSTM and GRU models only learns from future context in a one way manner.To improve the performance of the models Bidirectional LSTM and GRU models is used[21].This makes models learn both from past and future sequences by using both forward and backward layers. Our experiments shows that using bidirectional models for sign language recognition is better to learn long term dependencies and spatio temporal relationships in sequenced data.

D. *Preprocessing*

The first step of the proposed method is the preprocessing where the video frames are fed into wholebody pose. The output for each video is a matrix of shape $(n_{frames}, (n_{keypoints}, (x,y,c))$. Here n_{frames} is the number of frames in video, $n_{keypoints}$ is the number of keypoints, (x,y) is the coordinates of the keypoints. c is the confidence score of respective keypoints. To simplify the problem, we put a constraint that each video has the same number of frames to obtain same sequence length hence n_{frames} is 16. The confidence scores are also excluded. Afterwards this matrix is flattened and converted to vector of size $n_{keypoints}$*2. So the output vector after preprocessing is $(n_{frames}, (n_{keypoints}, (x,y))$.

E. *Proposed Network Architecture*

The proposed architecture consist of consecutive LSTM and GRU layers with dropout to prevent overfitting.Layer Normalization is also used after layers to keep data normalized. After that the output of these layers are fed into 2 fully connected layers. Rectified Linear Unit (ReLU) is used as activation function.Cross Entropy Loss function utilized to

reduce loss adam optimizer is used. Both of the network is built on same structure. The only difference is LSTM and GRU layers. The proposed architecture especially build as smaller network to reduce computational cost. Only 3 hidden layers are used in both model.

IV. Dataset And Evaluation Metrics

A. *Dataset*

AUTSL dataset to evaluate the performance of the models. AUTSL is an extensive dataset of isolated Turkish sign videos. It includes 226 signs performed by 43 different signers. There are a total of 36,302 video samples. It has 20 different backgrounds with a variety of challenges. Number of training samples are provided in table below.

TABLE I: NUMBER OF SAMPLES IN AUTSL DATASET

	Train	Valid	Test	Total
Num of samples	28,142	4,418	3,742	36,302

B. *Evaluation Metrics*

The proposed architectures evaluated by the performance complexity of the sequences (we used sequence length as complexity measure), accuracy on test set and computational efficiency. In terms of computational efficiency GPU Utilization, GPU Memory Allocated, GPU Time Spent Accessing Memory, Disk Utilization, System Memory Utilization etc. are compared.

V. Training

The proposed architectures are trained Jupyter Notebook with Nvidia 1660TI. Pytorch library is used as deep learning framework. The methods and training parameters are explained in details in section A and section B. Adam optimizer is used to train both model with a fixed learning rate of 1e-5. cross entropy loss is utilized as loss function. Sequence length is fixed as 16 frame.

VI. Discussion

In terms of GPU Utilization , GPU Accesing Memory , GPU Utilization , GPU Time Spent Accesing Memory Gated Recurrent model outperforms Long Short Term Memory model. However LSTM outperforms GRU model in terms of accuracy and training time. GRU model needs much more time for training. LSTM model achieves %79,589 for 6.5 hours of training. GRU model achieves %78,66 for 12.5 hours of training.

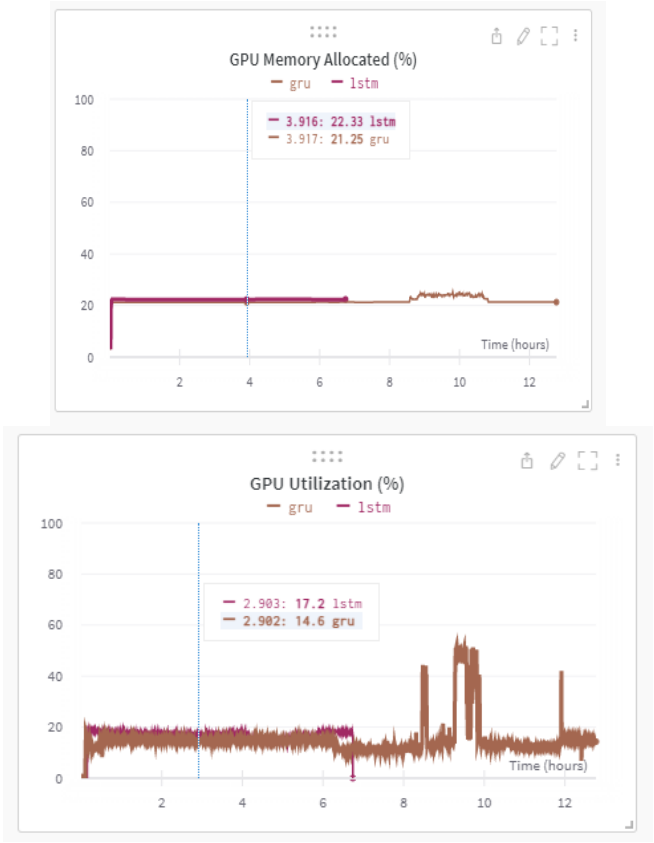

Fig 4: Comparison of GPU Memory Allocated and GPU Utilization

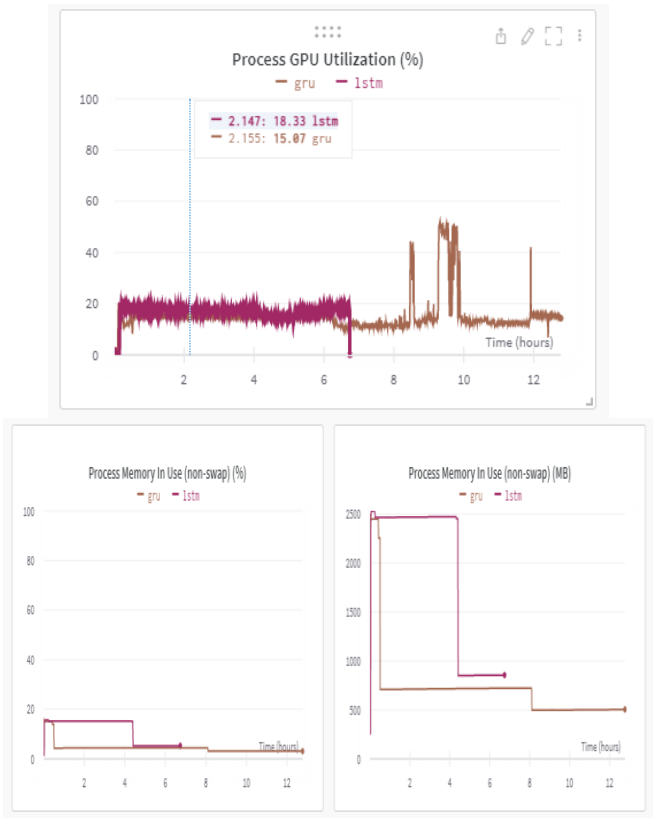

Fig 5:Comparison of Process GPU Utilization and Process Memory In Use

https://doi.org/10.17758/URUAE19.DIR0223105

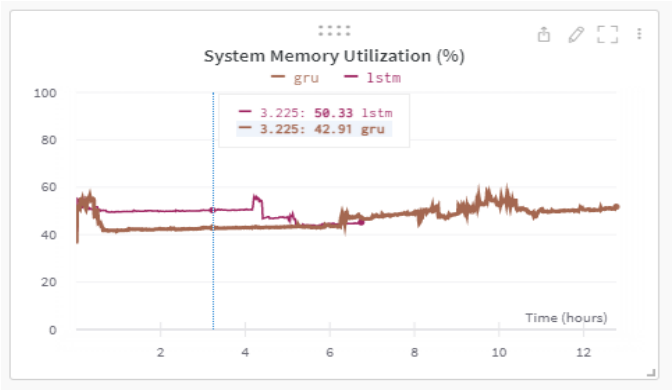

Fig. 6: Comparison of System Memory Utilization

VII. Conclusion And Future Works

This paper aims to achieve good performance with only RGB based videos and also compare the two Recurrent Neural Networks (RNN) models, Gated Recurrent Unit and Long Short Term Memory in terms of performance and computational efficiency. Hence proposed systems uses skeleton data extracted from RGB video, Deep Bidirectional Gated Recurrent Unit and Deep Bidirectional Long Short Term Memory model for sign language recognition.

In this paper RNN models LSTM and GRU compared in terms of performance and computational efficiency for Sign Language Recognition from skeleton keypoints that are extracted from RGB based videos. Our experiments shows that in terms of training time LSTM outperforms GRU.LSTM also slightly performs better than GRU model.

However in terms of computational efficiency GRU is better. In this paper the proposed models trained under fixed sequence length. For further research the models can be trained for different sequence length to determine performance of the models in terms of sequence complexity.

References

[1] Skeleton Aware Multi-modal Sign Language Recognition Songyao Jiang§ , Bin Sun§ , Lichen Wang, Yue Bai, Kunpeng Li and Yun Fu Northeastern University, Boston MA, USA

[2] P. Wang, Z. Li, Y. Hou, and W. Li, “Action recognition based on joint trajectory maps using convolutional neural networks,” in Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016, pp. 102–106. [16] C. Li, Y. Hou, P. Wang, and W. Li, “Joint distance maps based action recognition with convolutional neural networks,” IEEE Signal Processing Letters, vol. 24, no. 5, pp. 624–628, 2017. https://doi.org/10.1109/LSP.2017.2678539

[3] H. Rahmani and M. Bennamoun, “Learning action recognition model from depth and skeleton videos,” in The IEEE International Conference on Computer Vision (ICCV), 2017.arXiv:1506.01497, 2015. 2, 3, 4, 5, 6

[4] Ozge Mercanoglu Sincan, Anil Osman Tur, and Hacer Yalim Keles. Isolated sign language recognition with multi-scale features using LSTM. In 2019 27th Signal Processing and Communications Applications Conference (SIU), pages 1–4. IEEE, 2019. 2, 6 https://doi.org/10.1109/SIU.2019.8806467

[5] Anil Osman Tur and Hacer Yalim Keles. Isolated sign recognition with a siamese neural network of RGB and depth streams. In IEEE International Conference on Smart Technologies, pages 1–6, 2019. 2

[6] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In Proceedings of IEEE Winter Conference on Applications of Computer Vision, pages 1459–1469, 2020. 2

[7] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3595–3603, 2019. 3

[8] Lionel Pigou, Aaron Van Den Oord, Sander Dieleman, ¨ Mieke Van Herreweghe, and Joni Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. International Journal of Computer Vision, 126(2):430–439, 2018. 2 https://doi.org/10.1007/s11263-016-0957-7

[9] Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. IEEE Transactions on Multimedia, 21(7):1880–1891, 2019. 2 https://doi.org/10.1109/TMM.2018.2889563

[10] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3595–3603, 2019. 3

[11] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5457–5466, 2018. 3

[12] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In European conference on computer vision, pages 816–833. Springer, 2016. 3 https://doi.org/10.1007/978-3-319-46487-9_50

[13] Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 2904–2913, 2017. 3 https://doi.org/10.1109/ICCV.2017.316

[14] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In Proceedings of the European Conference on Computer Vision (ECCV), pages 103– 118, 2018. 3

[15] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634. https://doi.org/10.1109/CVPR.2015.7298878

[16] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, “Beyond short snippets: Deep networks for video classification,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4694–4702. https://doi.org/10.1109/CVPR.2015.7299101

[17] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In Proceedings of IEEE Computer Vision and Pattern Recognition, pages 5693–5703, 2019. 6 https://doi.org/10.1109/CVPR.2019.00584

[18] MMPose Contributors. OpenMMLab pose estimation toolbox and benchmark. https://github.com/openmmlab/mmpose, 2020. 6

[19] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997. [20] Y. Bengio, P. Simard, and P. Frasconi, “Learning longterm dependencies with gradient descent is difficult,” IEEE transactions on neural networks, vol. 5, no. 2, pp. 157–166, 1994. https://doi.org/10.1109/72.279181

[20] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” Neural Networks, vol. 18, no. 5, pp. 602–610, 2005. https://doi.org/10.1016/j.neunet.2005.06.042