Offline Automatic Speech Recognition System Based on Bidirectional Gated Recurrent Unit (Bi-GRU) with Convolution Neural Network

S. Girirajan and A. Pandian*

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur 603203, India
E-mail: giririajans.cse@gmail.com; pandiana@srmist.edu.in
**Corresponding Author*

Received 10 February 2022; Accepted 31 March 2022;
Publication 05 July 2022

Abstract

In recent years, the usage of smart phones increased rapidly. Such smartphones can be controlled by natural human speech signals with the help of automatic speech recognition (ASR). Since a smartphone is a small gadget, it has various limitations like computational power, battery, and storage. But the performance of the ASR system can be increased only when it is in online mode since it needs to work from the remote server. The ASR system can also work in offline mode, but the performance and accuracy are less when compared with online ASR. To overcome the issues that occur in the offline ASR system, we proposed a model that combines the bidirectional gated recurrent unit (Bi-GRU) with convolution neural network (CNN). This model contains one layer of CNN and two layers of gated Bi-GRU. CNN has the potential to learn local features. Similarly, Bi-GRU has expertise in handling long-term dependency. The capacity of the proposed model is higher when compared with traditional CNN. The proposed model achieved nearly

Journal of Mobile Multimedia, Vol. 18_6, 1659–1676.
doi: 10.13052/jmm1550-4646.1869

5.8% higher accuracy when compared with the previous state-of-the-art methods.

Keywords: Bi-GRU, CNN, recurrent neural network, automatic speech recognition, MFSC, 2-dimensional convolution network.

1 Introduction

In the modern digital world, humans are more closely connected with technologies like electronic gadgets, smart phones, robots, etc. Such a device can be controlled with natural human speech with the help of an automatic speech recognition (ASR) system. Smart devices like Google Voice Assistant, Amazon Alexa, Apple Siri, etc., are commonly used by people in their day-to-day environment. These devices follow client-server architecture. Usually, an ASR system requires high computing power and a complex model to achieve high accuracy. So these requirements are placed in the cloud environment. The input audio speech signal is passed through smart devices like smartphones. Then the collected audio signals are forwarded to the cloud server. Basic preprocessing, as well as recognition of text transcript for the corresponding audio signal, is carried out in the cloud server. Then the recognized text transcript is returned to the smartphones. By using such client-server architecture, we are able to reach high performance, but along with that, we face some limitations. The limitations are listed below.

- It requires high bandwidth for Internet connectivity.
- It requires more secured connection since we use client-server architecture. Data security plays a vital role in that type of architecture [1]. It reduces electricity costs and OPEX.

To overcome these issues, we need an offline ASR system. The already available offline ASR system is able to recognize some simple commands due to hardware limitations of smartphones. The accuracy of those offline ASR systems is very less compared with online ASR.

By considering the above issues and drawbacks, we design an ASR system that runs entirely in local smartphones with hardware compatibility. The proposed offline ASR system combines convolution neural network (CNN) and recurrent neural network (RNN) models. The local features are learned by using CNN and long-term dependency is handled with the help of bidirectional gated recurrent unit (Bi-GRU). It is an RNN model. The proposed model attained 91.1% accuracy in Linguistic Data Consortium for

Indian Languages (LDC-IL) speech corpora which are relatively 5.8% high when compared with the previous state-of-the-art methodology.

Our proposed work was performed in two aspects.

- Initially, by using an uncomplicated neural network with fewer layers, we designed a CNN + RNN model. Then it was compared with the previous such hybrid models [2, 3]. The accuracy we obtained with the proposed model was much higher when compared with the existing hybrid model.
- To choose a feature that contains high importance, we acquired the gated CNN model in the second phase. Improve corporate image.

2 Related Work

Image processing is a major domain where we use CNN widely. After the high usage of the deep neural network (DNN) model, most of the image-based tasks such as classifying and identifying an object in an image are carried out with the help of CNN with high accuracy [4–7]. CNN is very powerful in accepting the interesting field, sub-sampling, splitting the weight with modification in data. The potential of CNN is high in finding the similarities between the local data. This capability makes the CNN outperform speech-related tasks and also in natural language processing.

Initially proposed RNN does not provide a well-efficient result in time series data. But it suffers a lot due to gradient descent and vanishing problem. Based on the drawback which we faced in RNN, researchers developed a model called long short-term memory (LSTM) [8] that provides a better solution. By using LSTM, we are able to overcome the long-term dependencies in learning. With the further enhancement in LSTM, a new model called gated recurrent unit is developed with reduced gates when compared with LSTM. Most of the speech-related as well as natural language processing based research problems are solved by using LSTM and GRU with high accuracy.

Neural network concepts are used widely in almost all major domains including speech recognition. At an earlier stage, the speech recognition system uses hidden Markov model (HMM) for the system that identifies some specific keywords. Compared with the performance of HMM, neural network concepts like fully connected DNN achieved 40% higher accuracy. But the fully connected DNN has major issues in identifying unstable data. Speech data have a lot of variance especially in frequency as well as time. With a

huge collection of training data, we are able to handle this invariance in data by using fully connected DNN [9].

To overcome the issues we faced in fully connected DNN, an alternate model is proposed, which is CNN. This model has been widely used in image-processing-related tasks and also it shows a consistent improvement in speech recognition as well. In the works [10–14], researchers proposed CNN to recognize the keyword spotting that gave better performance compared with fully connected DNN by minimizing the scale of the model. In the works [1] and [9], researchers proposed dilated convolution model along with transfer learning to recognize the speech command for multi-scale input. In the works [15] and [16], researchers achieved an accuracy of around 95% in LDC-IL Speech Corpora data set by using CNN.

The performance of CNN is outstanding when compared with fully connected DNN. But CNN also contains some drawbacks. By using CNN, we cannot focus the complete speech on both frequency and time. Since we need to have more number of layers in CNN and because the filter's shape is limited and also CNN is able to learn the features that are local in nature, deeper model is required to focus on a large scope. RNN performs well when compared to CNN while dealing with long-term dependencies of sequence data.

Hybrid convolution RNN proposed in [2] shows better performance. It consists of 32 layers for CNN and 1 layer for RNN. By the hybrid concept, we can achieve better learning capability in terms of local features as well as long-term dependencies. GRU is a type of RNN model combined with RNN which consists of 15 layers giving high accuracy. Based on the model used in [3], we propose the new model that combines RNN with gated CNN. Since the previous works are done with extensive and problematical network structures, the LDC-IL Speech Corpora data set based speech command data set is used for testing the model. The accuracy we obtained with the proposed model is 91.2%, which is relatively 5.8% high when compared with previously proposed methodologies.

3 Our Work

The proposed hybrid model that combines the gated CNN with Bi-GRU is shown in Figure 1. As described above, CNN is widely used in image-processing-related tasks. In such a case, we cannot directly input the speech signal to CNN. Since speech consists of 1D feature but image consists of features like static, delta, and delta-delta, these features are grouped together

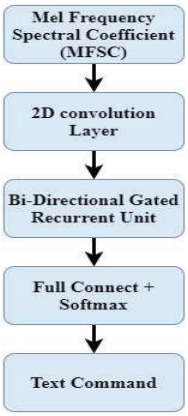

Figure 1 Gated CNN with Bi-GRU.

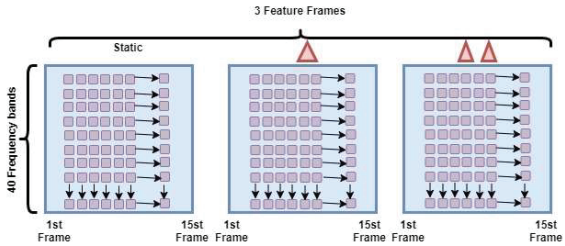

Figure 2 Input speech features for 2D CNN.

to form feature maps. To overcome the above issue, we extracted the features from the speech signal by using locality preserving projection by considering both frequency and time. We do not face any issue with time while generating feature maps. On the other hand, frequency creates a problem due to the usage of the mel-frequency cepstral coefficient (MFCC). We cannot maintain locality by using discrete cosine transform. So we decided to use computed energy directly without using DCT. That can be described as mel-frequency spectral coefficient (MFSC) feature. These frames that are generated by MFSC take the place of each speech frame. For each frequency band, acoustic energy distribution will be described.

Features that are generated by using MFSC will be converted into feature maps and then they can be used for CNN. MFSC features are distributed for both frequency and time. It was arranged in three 2D feature maps as shown in Figure 2. Frequency and temporal variations are normalized by 2D convolution.

Increased feature maps from 2D convolution are then passed to the multilayer gated CNN for feature selection and are integrated at the concatenation layer. Since we use the gating concept in CNN, it will make attention only to the features with high importance by avoiding the features that are unimportant.

After extracting features that have high importance by using gated CNN, the model will train with Bi-GRU to generate the feature vector for the given speech signal. In the end, based on the feature vector generated by Bi-GRU, the model will predict the speech command.

3.1 Preprocessing

In the proposed model, we have used MFSC for the given speech fragment. We gave speech as an input to CNN. So based on this consideration, we fixed 102 ms as a size for the frame window, and then we used 20 triangular bandwidth filters as well as 30.5 ms as the offset. Based on this, we got the dimension for f as 20 for each speech fragment. The speech duration is considered majorly for number of frames t.

3.2 Convolutional Neural Network

To carry out image-processing-based tasks, the widely used methodology is CNN. It can learn the local features of the given image with high accuracy. So we decided to use CNN in speech-related tasks especially to learn local features of the given speech signal.

In CNN, if we increase the number of layers, then it is able to recognize the more number of features in the local scope. Due to this advantage, we have used multilayer CNN to examine the feature of speech in various frames.

Frames that are generated by MFSC are used as the input for CNN. To perform the MFSC feature extraction, three steps need to be followed as shown in Figure 3. The sample rate for the input speech signal is fixed as 16 kHz and we fixed 102 ms for each block to extract the features. Hann window is used to compute the short time discrete Fourier transform (STFT) at 30.5 ms duration with 50% overlap through which we got 40 frames. For zero padding, each frame is considered with a length of 1024.

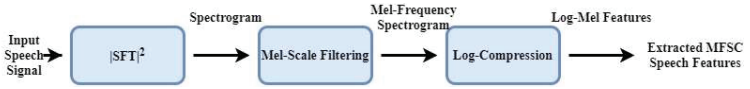

Figure 3 Feature extraction stages in MFSC for giving input to a 2D convolution layer.

Filterbanks of 20 triangular bandpass filters were applied. Finally, we applied the logarithmic to gain the 2D feature matrix of size 20 × 49 for each fragment.

3.3 Two-Dimensional Convolution

In the proposed model, 2D convolution is the first part of CNN. To fix up this model, we required three parameters. Each parameter has some specific advantage that is described below.

i. Parameter 1: $m = 1$, used to learn features in local time sequence.
ii. Parameter 2: $h = f$, used to describe different feature maps by considering the different frequency.
iii. Parameter 3: $r > f$, used to generate many feature maps by recombining frequency.

3.4 Gated Convolutional Neural Network

In the proposed model, gated CNN is the second and third layers, used to learn more speech-based local features. The authors in [17] proposed the gated CNN architecture that is shown in Figure 4.

The mathematical foundation of gated CNN is similar to that of the LSTM multiplication gate that has been described in the following equation:

$$y = \tan h(K_f * x) \odot \sigma(K_g * x). \tag{1}$$

In the above equation, K_f and K_g denote the kernels for each convolution layer. σ denotes the sigmoid function. * denotes the operation carried out by convolution and $\odot$ denotes the multiplication operation.

Learning the local features and signifying the capacity are carried out with high performance in a gated CNN when compared with CNN. We can have

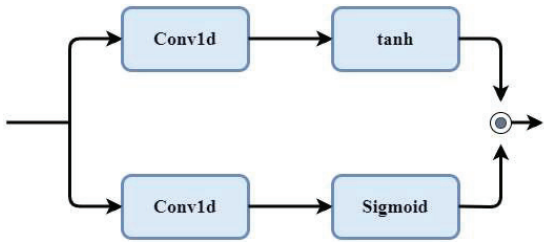

Figure 4 Gated convolutional neural network.

more non-linear operations as well as multiplication in gated CNN. Also, we can multiply respective elements with tan*h* and σ to gain self-attention [18]. Overfitting risk is reduced by regularization using batch normalization. It is difficult to normalize the speech signal by mean and variance. So mini-batch is generated using the mean and variance of each audio file followed by depth-wise separable convolution and point-wise convolution.

3.5 Recurrent Neural Network

In speech signals, attributes and contents are highly dependent on time order. A similar feature that appears in different time orders will have different meanings. So such local features are not recognized with high accuracy by using CNN. To overcome this issue, we included RNN in the proposed model.

Most of the natural language processing based task is carried out by using RNN since it performs well to learn local features and handle long-term dependencies. In recent years, RNN is used to learn features from continuous speech signals for large vocabulary [10, 11]. To learn the local features of the speech signals in an efficient way, we fixed the Bi-GRU layer after the CNN network. The structure of RNN and bi-directional connectivity of GRU is shown in Figures 5 and 6.

To set up the connection between the previous state and the current state is uncertain in the RNN model. A step involved in implementing GRU follows. The activation of the memory cell A_t at the time *t* could be a linear interpolation of the previous initiation A_{t-1} and the activation candidate A'_t at the time *t*, r_t is the reset gate and z_t is the update gate. The *W* terms indicate matrices of weight [9].

$$z_t = \sigma(W^{(\mathrm{z})} B_t + U^{(z)} A_{t-1}) \tag{2}$$

$$r_t = \sigma(W^{(\mathrm{r})} B_t + U^{(r)} A_{t-1}) \tag{3}$$

$$A'_t = \tanh(W B_t + r_t \bigodot U A_{t-1}). \tag{4}$$

4 Experimental Analysis

4.1 Data Set

This Tamil Speech Recognition database was collected in Tamil Nadu and contains the voices of 450 different native speakers who were selected according to age distribution (16–20, 21–50, and 51+), gender, dialectical

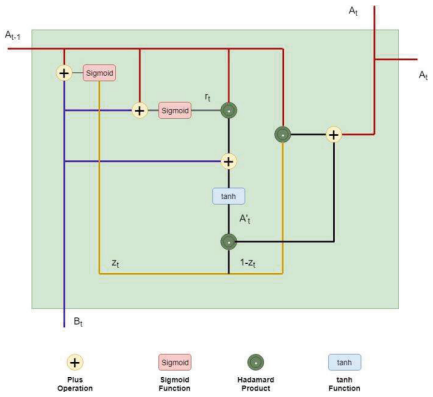

Figure 5 Gated recurrent unit.

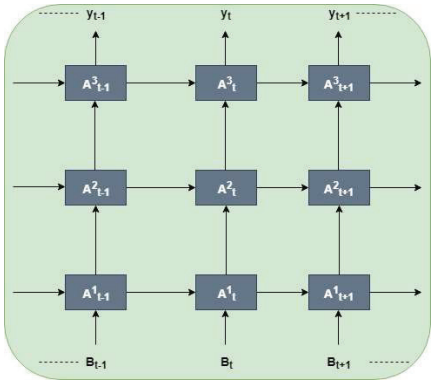

Figure 6 Bi-directional RNN (Bi-RNN).

regions, and environment (home, office, and public place). Each speaker recorded a news text in a noisy environment through a recorder having an inbuilt microphone. The recordings are in stereo recording and the extracted channel is also included in the specific files. It includes an audio file, text files, and NIST files which were saved as .ZIP files. All the speech data are transcribed and labeled at the sentence level. These speech audios are monolingual and are made by form and function words, command and control words, phonetically balanced vocabulary, proper names, and most frequent 1000 words. The total duration of Tamil speech is 87 hours 3 minutes 24 seconds. The sampling rate of these speech signals is 16 kHz and has a sampling resolution of 16 bit [28].

4.2 Experimental Settings

The proposed model is analyzed from various aspects like CNN network structure and hybrid CNN with Bi-GRU and then evaluated against the already available model. Various experiments were conducted against the model that contains different structures. These models are described below.

i. *C-n-G-m*-BGRU: To recognize the speech commands by adjusting the layers in the model. In this model, *n* denotes the number of layers in the 2D convolution model and *m* denotes the number of layers in the gated GRU model.
ii. Transfer Learning Network [15]: We have used an already trained model with 121-layer to recognize the LDC-IL Speech Corpora data set.

Limited epochs are used to train the models. Almost all the models gave better performance within 100 epochs. Based on the performance of the model during training, we have selected the model. Later, the model that has been selected has undergone 10 experiments repeatedly. For the final evaluation, the average of those 10 experimental results is considered. In the proposed model, to map the input with corresponding output, max pooling, ReLU non-linearity, and dropout regularization techniques are used.

The entire baseline and proposed neural network models were trained using the PyTorch-Kaldi toolkit. This toolkit uses the PyTorch deep learning framework for implementing the neural network models and Kaldi for feature extraction and decoding. Therefore, our proposed models were implemented using the PyTorch deep learning framework and run on the PyTorch-Kaldi toolkit. The training process was accelerated using Nvidia Tesla M40 GPU on a single machine.

5 Result Analysis

5.1 Effect of Layers in Gated Bi-GRU

Experiments were made with the different numbers of gated layers in CNN and the training and testing accuracies were observed. We have built the different models shown in Table 1.

The accuracy of each model with different layers is measured for both testing and training data set. An experiment is conducted using an LDC-IL data set. The decision cost function (DCF) is used to measure the performance of the proposed work. Equation (5) describes the calculation of the

Table 1 Effect of layers in gated CNN

Model	Training Accuracy %	Testing Accuracy %
1-CNN-2-Gated Bi-GRU	91.4	91.1
1-CNN-5-Gated Bi-GRU	90.9	90.5
1-CNN-7-Gated Bi-GRU	90.2	90.0
1-CNN-9-Gated Bi-GRU	89.2	88.7
1-CNN-10-Gated Bi-GRU	88.7	88.4

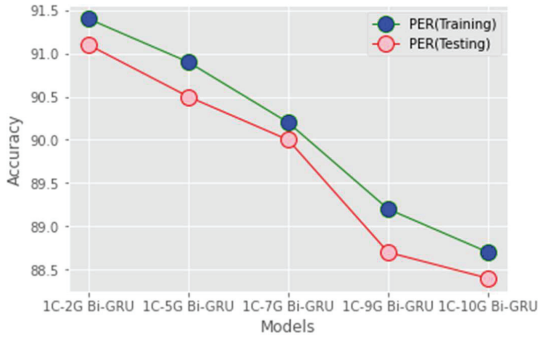

Figure 7 Effect of layers in gated CNN.

DCF score.

$$\begin{aligned} \mathrm{DCF}(\theta) = {} & 0.75 * \text{Probability of False Negative} \\ & + 0.25 * \text{Probability of False Positive.} \end{aligned} \tag{5}$$

From the above table, we concluded that the model with more gated layers gave less accuracy when compared with the model with minimum gated layers. In the proposed work, Adam optimization is used with a learning rate of 0.001. We also tried with 20 and 50 gated layers, but the performance of those models is diverging.

Since we used a limited amount of data for training, if we increase the gated layers, it struggles a lot in learning the local features with such limited data. For the proposed model by using a 16-ms frameshift, the average time frame processing is calculated as 10.2 ms. From the above table, we came to the conclusion that the model with 1-CNN-2-Gated Bi-GRU gave better performance and accuracy percentage when compared with other models. So the remaining part of the implementation is carried out with the 1-CNN-2-Gated Bi-GRU model.

Table 2 Effect of gated convolution

Model	Training Accuracy %	Testing Accuracy %
1-CNN-2-Gated Bi-GRU	91.4	91.1
3-CNN-0-Gated Bi-GRU	87.7	87.7
1-CNN-5-Gated Bi-GRU	90.9	90.5
6-CNN-0-Gated Bi-GRU	87.4	87.2
1-CNN-7-Gated Bi-GRU	90.2	90.0
8-CNN-0-Gated Bi-GRU	84.0	83.7

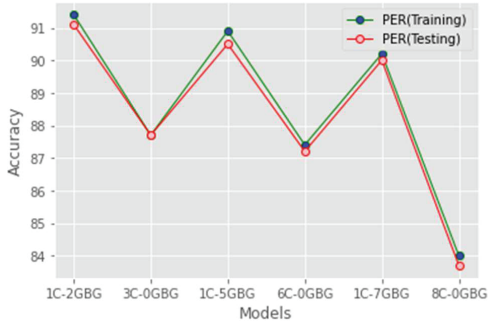

Figure 8 Effect of gated convolution.

5.2 Effect of Gated Convolution

Experiments were conducted to compare the performance of conventional and gated CNN. To do this, we placed the gated CNN in 1-CNN-2-Gated Bi-GRU, 1-CNN-5-Gated Bi-GRU, and 1-CNN-7-Gated Bi-GRU models along with conventional CNN. Comparison between gated and conventional CNN is shown in Table 2. From the observation, we concluded that gated CNN predicts the model with more accuracy when compared with conventional CNN.

5.3 Effect of CNN and RNN

Experiments were conducted to compare the performance of combined CNN and RNN with only CNN and only the RNN model. Table 3 shows the accuracy of 1-CNN-2-Gated Bi-GRU and only CNN and only RNN model for both testing and training phase. From the table, we concluded that the proposed CNN with RNN model gave better performance when compared with only CNN and only RNN model. By using the proposed model, we are able to recognize the speech command with high accuracy.

Table 3 Effect of CNN and RNN

Model	Training Accuracy %	Testing Accuracy %
1-CNN-2-Gated Bi-GRU	91.4	91.1
0-CNN-0-Gated Bi-GRU	66.0	62.1
1-CNN-2-Gated Full Connect	81.8	81.6

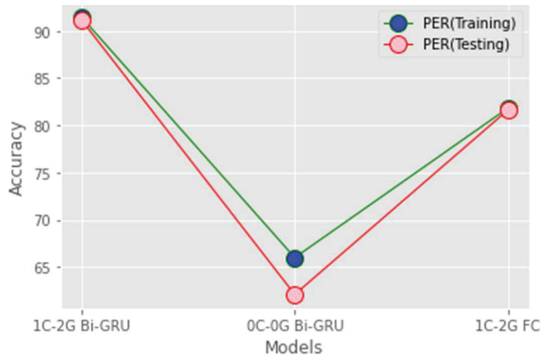

Figure 9 Effect of CNN and RNN.

0-CNN-0-Gated Bi-GRU: In this model, we removed the CNN; so the model contains only the RNN structure.

1-CNN-2-Gated Full Connect: In this model, we replaced the full connect structure insist of RNN; so the model contains only CNN structure.

Already some research work is carried out by combining CNN and RNN, but such models are designed with more number of layers with complex structures. In [15], they have used 32 CNN layers along with 1 RNN layer. Similarly, in [33] and [34], the researchers used eight CNN layers along with seven RNN layers. The proposed model is trained within 100 epochs and each epoch takes only 18 seconds; on the other hand, for the testing purpose, each epoch takes only 1 second.

6 Conclusion

In this work, we have designed the 1-CNN-2-Gated Bi-GRU model to recognize the Tamil speech command on the mobile device. By using this model, we are able to learn local features more accurately with the help of CNN structure and to learn features based on long-term dependencies with the help of RNN. Gated CNN structure helps in the improvement of model capacity.

When compared with only CNN or only RNN model or previous hybrid CNN with RNN models, our proposed model recognizes speech command more accurately. Since we used the least number of the layer, it makes our model simple and easy to access. Our proposed model is able to recognize speech commands with 91.1% accuracy which is nearly 5.8% high compared with an existing model.

References

[1] Chen, G., Parada, C., Heigold, G.: Small-footprint keyword spotting using deep neural networks. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. pp. 4087–4091. IEEE (2014)

[2] Arik, S.O., Kliegl, M., Child, R., Hestness, J., Gibiansky, A., Fougner, C., Prenger, R., Coates, A.: Convolutional recurrent neural networks for small-footprint keyword spotting. arXiv preprint arXiv:1703.05390 (2017)

[3] Zhang, Y., Chan, W., Jaitly, N.: Very deep convolutional networks for end-to-end speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. pp. 4845–4849. IEEE (2017)

[4] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[5] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)

[6] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[7] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)

[8] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)

[9] Sainath, T.N., Parada, C.: Convolutional neural networks for small-footprint keyword spotting. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)

[10] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al.: Deep speech 2: End-to-end speech recognition in English and mandarin. In: International Conference on Machine Learning. pp. 173–182 (2016)
[11] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to end speech recognition. arXiv preprint arXiv:1412.5567 (2014)
[12] Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for largescale audio classification. In: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. pp. 131–135. IEEE (2017)
[13] Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C.L.Y., Courville, A.: Towards end-to-end speech recognition with deep convolutional neural networks. arXiv preprint arXiv:1701.02720 (2017)
[14] Wang, Y., Getreuer, P., Hughes, T., Lyon, R.F., Saurous, R.A.: Trainable frontend for robust and far-field keyword spotting. In: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. pp. 5670–5674. IEEE (2017)
[15] McMahan, B., Rao, D.: Listening to the world improves speech command recognition. arXiv preprint arXiv:1710.08377 (2017)
[16] Warden, P.: Launching the speech commands dataset. Google Research Blog (2017)
[17] van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. In: Advances in Neural Information Processing Systems. pp. 4790–4798 (2016)
[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. ArXiv e-prints (Jun 2017)
[19] Li H-J, Wang Z, Pei J, Cao J, Shi Y (2020) Optimal estimation of low-rank factors via feature level data fusion of multiplex signal systems. IEEE Ann Hist Comput 01:1–1
[20] Li H-J, Wang L, Zhang Y, Perc M (2020) Optimization of identifiability for efficient community detection. New J Phys 22(6):063035
[21] Zhao P, Hou L, Wu O (2020) Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. Knowl-Based Syst 193:105443

[22] Zhang, Yg., Tang, J., He, Zy. et al. A novel displacement prediction method using gated recurrent unit model with time series analysis in the Erdaohe landslide. Nat Hazards 105, 783–813 (2021).
[23] Afif, M., Ayachi, R., Said, Y. et al. An Evaluation of RetinaNet on Indoor Object Detection for Blind and Visually Impaired Persons Assistance Navigation. Neural Process Lett 51, 2265–2279 (2020). https://doi.org/10.1007/s11063-020-10197-9
[24] H. Sadr, M. M. Pedram and M. Teshnehlab, "A robust sentiment analysis method based on sequential combination of convolutional and recursive neural networks", Neural Process. Lett., vol. 50, no. 3, pp. 2745–2761, Dec. 2019.
[25] J. Chen, H. Jing, Y. Chang, Q. Liu "Gated recurrent unit based recurrent neural network for remaining useful life prediction of nonlinear deterioration process" Reliability Engineering & System Safety, 185 (2019), pp. 372–382
[26] P. Huang, X. Xie and S. Sun, "Multi-view opinion mining with deep learning", Neural Process. Lett., vol. 50, no. 2, pp. 1451–1463, Oct. 2019.
[27] Y. Deng, L. Wang, H. Jia, X. Tong, F. Li, "A sequence-to-sequence deep learning architecture based on bidirectional gru for type recognition and time location of combined power quality disturbance" IEEE Transactions on Industrial Informatics (2019)
[28] A. Gharehbaghi, P. Ask, A. Babic "A pattern recognition framework for detecting dynamic changes on cyclic time series" Pattern Recognition, 48(3) (2015), pp. 696–708
[29] Guo, N. Li, F. Jia, Y. Lei, J. Lin "A recurrent neural network based health indicator for remaining useful life prediction of bearings" Neurocomputing, 240 (2017), pp. 98–109
[30] J. Wu, K. Hu, Y. Cheng, H. Zhu, X. Shao, Y. Wang, "Data-driven remaining useful life prediction via multiple sensor signals and deep long short-term memory neural network", ISA Transactions, 97 (2020), pp. 241–250
[31] Choudhary, N. LDC-IL: The Indian repository of resources for language technology. Lang Resources & Evaluation (2021). https://doi.org/10.1007/s10579--020-09523-3
[32] Li S, Chen SF, Liu B (2013) Accelerating a recurrent neural network to finite-time convergence for solving time-varying Sylvester equation by using a sign-bi-power activation function. Neural Process Lett 37:189–205

[33] S. Girirajan, A. Pandian, "Acoustic model with hybrid Deep Bidirectional Single Gated Unit (DBSGU) for low resource speech recognition," *Multimedia Tools Application*, 2022

[34] A. Pandey, D. L. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE International Conference on Acoustics, Speech, & Signal Processing*, 2019, pp. 6875–6879

Biographies

S. Girirajan received B.E degree in Computer Science and Engineering from Asan Memorial Engineering College in 2010. M.Tech degree in Computer Science Engineering from SRM University, India in 2016. He is currently pursuing his Philosophy of Doctorate in Computer Science and Engineering at SRM Institute of science and Technology, Chennai, India and works as Assistant Professor. He has 5+ years of teaching experience. His research interests include Speech Recognition, Machine Learning, Deep Learning and Image Processing.

A. Pandian received the Ph.D degree in Computer Science & Engineering at SRM institute of science and technology, Chennai, India in 2015. He works currently as an Associate Professor for the Department of Computer Science and Engineering at SRM Institute of Science and Technology, Chennai and he has 24 Years of teaching experience. He has participated and presented many Research Papers in International and National Conferences and also published many papers in International and National Journals. His area of interests includes Text Processing, Information retrieval and Machine Learning.