ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414489

A NOVEL ATTENTION-BASED GATED RECURRENT UNIT AND ITS EFFICACY IN SPEECH EMOTION RECOGNITION

Srividya Tirunellai Rajamani[1,2], *Kumar T. Rajamani*[3], *Adria Mallol-Ragolta*[1], *Shuo Liu*[1], *Björn Schuller*[1,4]

[1] EIHW – Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[2] Siemens Healthineers, Siemens Healthcare GmbH, Erlangen, Germany
[3] Institute of Medical Informatics, University of Lübeck, Germany
[4] GLAM – Group on Language, Audio, & Music, Imperial College London, UK

srividya.tirunellai@informatik.uni-augsburg.de

ABSTRACT

Notwithstanding the significant advancements in the field of deep learning, the basic long short-term memory (LSTM) or Gated Recurrent Unit (GRU) units have largely remained unchanged and unexplored. There are several possibilities in advancing the state-of-art by rightly adapting and enhancing the various elements of these units. Activation functions are one such key element. In this work, we explore using diverse activation functions within GRU and bi-directional GRU (BiGRU) cells in the context of speech emotion recognition (SER). We also propose a novel Attention ReLU GRU (AR-GRU) that employs attention-based Rectified Linear Unit (AReLU) activation within GRU and BiGRU cells. We demonstrate the effectiveness of AR-GRU on one exemplary application using the recently proposed network for SER namely Interaction-Aware Attention Network (IAAN). Our proposed method utilising AR-GRU within this network yields significant performance gain and achieves an unweighted accuracy of 68.3% (2% over the baseline) and weighted accuracy of 66.9 % (2.2 % absolute over the baseline) in four class emotion recognition on the IEMOCAP database.

Index Terms— gated recurrent unit, attention mechanism, speech emotion recognition, ReLU, AReLU

1. INTRODUCTION

The paralinguistic information embedded in the human voice reveals the emotional state of a speaker [1]. This information is of vital importance in *Human-Human Interaction* (HHI), as we as humans use it to adjust, for instance, the content of our message or the tone of our voice with the aim to smooth the interaction and empathise with our interactant. Thus, in order to better mimic HHI, there is a need to power machines with *Speech Emotion Recognition* (SER) technologies that can help amongst manifold further use-cases to boost the *Human-Computer Interaction* (HCI) experience.

The problem of SER has been widely investigated in the literature. Traditional approaches focused on the extraction of hand-crafted features from acoustic signals, such as pitch and energy among others [2], to capture the salient information from the human voice. These hand-crafted acoustic features are then fed into conventional machine learning techniques, such as *Hidden Markov Models* (HMMs) or *Support Vector Machines* (SVMs) [3, 4]. More recent approaches used these hand-crafted acoustic features or directly the raw audio as input for deep learning techniques, including *Convolutional Neural Networks* (CNNs) [5], *Recurrent Neural Networks* (RNNs) [6, 7, 8], or the combinations of CNNs and RNNs [9].

RNNs, such as *Long Short-Term Memory* (LSTM) [10], and *Gated Recurrent Units (GRU)* [11], capture the temporal dynamics of sequential data. Therefore, such techniques are suitable for SER tasks, as these are able to capture the temporal dependencies of the acoustic features. Attention mechanisms can be used to assist RNNs to focus on the most emotionally salient information [6, 7, 8]. Furthermore, contextual information can also be used to improve the performance of SER systems, as shown in recent works [12, 13]. Yeh et al. [13] successfully exploited contextual information through an *Interaction-Aware Attention Network* (IAAN), which uses previous speaker turns in a two-speaker dialog scenario to learn attention scores for detecting the emotional state of one speaker's utterance.

Current context-aware models for SER only use default LSTM or GRU cells. These do not consider altering the internal architecture of these units, and, therefore, might not obtain the optimal performance, yet. One such alteration can consist of exploring different activation functions, as those used in GRU cells play a direct role in determining the outcomes of the networks. In this work, we propose using *Attention-based Rectified Linear Units* (AReLU) [14] as activation function

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

within the GRU cell, aiming to maximise the exploitation of information from the internal components of the GRU. We hypothesise that the performance of current SER system can be improved by optimising the activation function of the internal RNN cells, and, therefore, use AReLU for such purpose.

The rest of the paper is organised as follows. Section 2 describes the methodology followed. Section 3 presents the experiments performed and analyses the results obtained. Lastly, Section 4 concludes the paper, including some directions for further works.

2. METHODOLOGY

This section introduces the structure of Gated Recurrent Unit (GRU), describes the *Attention-based Rectified Linear Unit* (AReLU) activation function, and presents our proposed novel integration of the AReLU activation within GRU cells (AR-GRU).

2.1. Gated Recurrent Units

Gated Recurrent Units (GRUs) are a type of *Recurrent Neural Networks* (RNN), which use gating mechanisms to control and manage the flow of information between cells in the neural network. The structure of the GRU allows adaptively capturing dependencies from large data sequences, while ensuring that the information from earlier parts of the sequences is not discarded. This is achieved through the gating mechanisms, which regulate the information to be kept or discarded at each time step. GRUs are able to overcome the vanishing gradient problem and are faster to train as compared to LSTMs due to the fewer number of parameters to optimise.

2.2. Attention-based Rectified Linear Unit

The *Attention-based Rectified Linear Unit* (AReLU) is a learnable activation function that exploits an element-wise attention mechanism [14]. AReLU amplifies positive elements and suppresses negative ones with learnt, data-adaptive parameters. Since the attention module within AReLU learns element-wise residues of the activated part of the input, the network training is more resistant to gradient vanishing. AReLU's learnt attentive activation results in well-focused activations of relevant regions of a feature map. With only two extra learnable parameters (alpha and beta) per layer, it facilitates fast network training under small learning rates.

AReLU [14] represented as ($\mathcal{F}(x_i, \alpha, \beta)$) is defined using a combination of an element-wise sign-based attention mechanism $\mathcal{L}(x_i, \alpha, \beta)$ and a standard Rectified Linear Unit $\mathcal{R}(x_i)$, as described in equation 1.

$$\begin{aligned}\mathcal{F}(x_i, \alpha, \beta) &= \mathcal{R}(x_i) + \mathcal{L}(x_i, \alpha, \beta) \\ &= \begin{cases} C(\alpha)x_i & , x_i < 0 \\ (1 + \sigma(\beta))x_i & , x_i >= 0, \end{cases}\end{aligned} \tag{1}$$

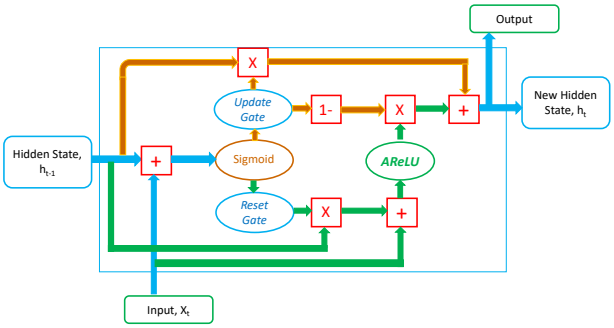

Fig. 1. Our proposed novel AR-GRU Architecture: The classical tanh activation in a GRU is replaced by an Attention-based Rectified Linear Unit.

where $X = \{x_i\}$ is the activation layer's input, $\{\alpha, \beta\} \in R^2$ are learnable parameters, C(·) clamps an input variable into [0.01, 0.99] to prevent α from becoming zero, and σ is the sigmoid activation.

2.3. AR-GRU: Attention-based Rectified Linear Unit within Gated Recurrent Units

The classical activation function in conventional GRUs is the *Hyperbolic Tangent* (tanh). While there are inherent advantages of using the tanh function, it has high computational complexity due to dense activation computations and also is susceptible to the vanishing gradient problem.

The different computational elements of GRU have largely remained static. Adapting the functional units of GRUs could result in significant performance improvements, specifically for tasks as SER. Attention mechanisms have demonstrated significant improvements in the context of deep learning. Attention-based ReLU is one such realisation of a learnable attention mechanism in activation functions. As outlined, in this work, we propose an Attention ReLU activation based GRU unit described in Figure 1.

The integration of the Attention-based ReLU within GRUs helps to capture long range interactions among the features. Capturing long range interactions is of vital importance in speech recognition, and specifically in SER due to the supra-segmental nature of the phenomenon. Hence, the use of AReLU-GRU is expected to help to capture these dependencies, and boost the performance of SER systems in addition to addressing the vanishing gradient problem.

In this work, we empirically determine the optimal initial values of alpha and beta of the AReLU that would best suit SER tasks. The default values of alpha (0.9) and beta (2) in AReLU tend to work best for image recognition and image segmentation tasks, as described in Chen et al. [14]. However, our experiments show that in the context of SER, these default values tend to negatively impact the performance. Our

investigations also demonstrate that using the ReLU activation within GRUs already helps us to significantly boost the performance. Hence, our methodology for optimising the alpha and beta values for AReLU started with making the AReLU resemble a ReLU-like activation function by using an alpha of 0.01 and beta of -4. Since alpha controls the scale factor for the negative values, a value of 0.01 in conjunction with the clamp function would yield the least influence of negative values. Beta controls the scale factor for the positive values and having it as -4 nullifies the scale factor.

Once the integration of the AReLU within the GRU was effective, we then explored amplifying the positive values. This was done by using a beta value of 2 which makes the result of the sigmoid function to be 0.88. Hence, the positive values are scaled by a factor of 1.88. Scaling the positive values this way gave us further boost in performance for the SER task.

AReLU suppresses the negative values. Our experiments demonstrate that suppressing the negative values using the default AReLU weighting parameters (with an alpha of 0.9) negatively impacts the performance. We then explored clamping the negative values and rendering them closer to zero. This ensured that there was no adverse impact in performance due to the negative values.

Although we demonstrate the effectiveness of AR-GRU in the context of SER, it is not limiting in its general applicability in other applications, such as other speech-related or *Natural Language Processing* (NLP) tasks.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Dataset Description

We conducted our experiments to examine the effectiveness of the different activation functions in GRU and BiGRU in the context of SER using the IEMOCAP dataset [15]. This is a benchmark dataset widely used in the field of SER research. This dataset contains 10 speakers and five sessions. Each session comprises of two speakers engaging in different conversational scenarios during their dialogue. In order to compare with previous baseline performances, a four emotion class classification, i. e., anger, happiness, sadness, and neutral, is performed using 5 531 utterances. The distribution of the four emotion classes in the 5 531 utterances are: anger: 19.9 %, happiness: 29.5 %, neutral: 30.8 %, and sadness: 19.5 %.

3.2. Experimental Setup

We use the interaction-aware attention network (IAAN) [13] as the baseline model for our empirical experiments of using different activation functions within GRU and BiGRU. IAAN utilises contextual information and affective influences from previous utterances to model the emotion of the current utterance. It employs a BiGRU for the current utterance of the speaker and two GRUs for the preceding utterances of the speaker and the interlocutor. The acoustic low-level descriptors (LLDs) are extracted using the openSMILE toolkit [16] based on the Emobase 2010 Config, including features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and their statistics in each short frame of an utterance.

We conducted our experiments by employing non-learnable and learnable activation functions within the GRU and BiGRU cells of the IAAN. We evaluate the performance using both unweighted accuracy (UA) and weighted accuracy (WA). We use 5-fold leave-one-session-out (LOSO) cross validation and early stopping by observing the performance on validation set in every 100 training epochs.

3.3. Baseline Methods

We compare our method with the following previous baseline networks:

BiLSTM+ATT[6]: A BiLSTM network which uses an attention-based pooling layer on frame-level features.

MDNN[17]: A multi-path deep neural network which comprises of several local classifiers and a global classifier.

IAAN[13]: A GRU based network that utilises interaction-aware attention to incorporate the influence of contextual information between interlocutors within a transactional frame.

3.4. Results and Analysis

As detailed in section 2, the main novelty of our work is on diverse activation functions' integration within the GRU cells.

We first present the results of utilising ReLU activation units within the GRU cells (R-GRU). Our novel integration of using ReLU within GRU instead of the standard tanh activation function gave us superior performance with UA of 67.7 % and WA of 65.8 %. Our proposed R-GRU shows an improvement of 1.4 % absolute for UA and 1.1 % for WA against the current IAAN baseline. These results indicate that usage of ReLU is highly beneficial and a better suited activation function for GRU especially for SER tasks.

We next progress with further experiments of ReLU-like activation functions being used within GRU cells. ReLU activation is not a learnable activation, and hence has no trainable parameters. This severely limits the usage of the ReLU activation function in broader contexts. Hence, as outlined, we explore using AReLU as a learnable activation within GRU.

We now proceed to detail out five different variants of integrating Attention ReLU within GRU (AR-GRU) and explain in detail each of these variants in this section. The first variant of our proposed solution, AR-GRU (I) is the integration with the default values of alpha of 0.9 and beta of 2 for AReLU but this significantly impaired the performance and we obtained very poor results. We include the results in Table 1 for the sake of completeness. This result demonstrates that default values of alpha and beta parameter do not work for our current intended SER application.

Table 1. The performance of the proposed models in comparison to state-of-the-art (upper part) and different network variants (lower part) on the IEMOCAP corpus for 4-way SER. UAR chance level resembles 25 %.

Model		AReLU parameters alpha	beta	% UA	% WA
BiLSTM +ATT	Mirsamadi et al.(2017)	-	-	58.8	63.5
MDNN	Zhou et al.(2018)	-	-	62.7	61.8
IAAN	Yeh et al.(2019)	-	-	66.3	64.7
R-GRU based network (I)	Experiment 1	-	-	67.7	65.8
AR-GRU based network (I)	Experiment 2	0.9	2.0	35.7	38.7
AR-GRU based network (II)	Experiment 3	0	2.0	66.3	64.7
AR-GRU based network (III)	Experiment 4	0.01	-4.0	66.9	65.4
AR-GRU based network (IV)	Experiment 5	0.01	2.0	67.9	66.6
AR-GRU based network (V)	Experiment 6 : Proposed method	0.01	1.0	**68.3**	**66.9**

The next variant of AR-GRU (II) is the usage of AReLU with alpha of 0.01 and beta of -4. These parameter values make the AReLU very similar to ReLU as explained in the methodology section. With these parameter values, we observe UA of 66.9 % and WA of 65.4 %. This variant of ours gave an improvement of 0.6 % for UA and 0.7 % for WA against the IAAN baseline. The result from this experiment demonstrate that AReLU does help to boost performance, but the ideal parameter values needed to be empirically identified.

To identify the best combinations of alpha and beta, we consider further experiments with varied values of alpha and beta. The next combination explored is an alpha of 0 and beta of 2. This proposed variant AR-GRU (III) nullifies the impact of negative values similar to ReLU. To achieve this effect, the clamp function operating on the alpha parameter of the AReLU is adapted to have zero as the lower threshold. The AR-GRU(III) achieved a UA and WA similar to the baseline IAAN results. This demonstrates that clamping the negative values does not yield any significant contribution in performance. Having a small contribution of negative values does indeed potentially help a GRU cell for SER tasks.

Our next variant of AR-GRU (IV) uses an alpha of 0.01 and beta of 2. Usage of alpha of 0.01 scales the negative values by a factor of 0.01, and hence takes a lower contribution from the negative values, as discovered to be beneficial from the AR-GRU(III) experiments. Having beta of 2 enhances the positive values by a factor of 1.88. AR-GRU (IV) achieves a UA of 67.9 % and WA of 66.6 %. This variant gives an improvement of 1.6 % for UA and 1.9 % for WA against the IAAN baseline.

The previous experiment demonstrates that scaling the positive values positively impacts the GRU performance. The exact magnification factor of the positive values was empirically determined in our final experiment. In this variant of AR-GRU (V), we use an alpha of 0.01 and beta of 1. Using a value of 1 for beta enhances the positive values by a factor of 1.73. AR-GRU (V) achieves the best UA of 68.3 % and best WA of 66.9 %. This variant gives an improvement of 2.0 % for UA and 2.2 % for WA against the IAAN baseline. This combination of alpha (0.01) and beta (1) for AReLU within GRU manifests our novel proposed AR-GRU best suited for SER on the considered task. Table 1 summarises the results from all our diverse experiments in addition to the baseline results for comparison.

4. CONCLUSION AND FUTURE WORK

We demonstrated that our proposed AR-GRU based network with an alpha of 0.01 and beta of 1 boosts performance of GRU for the considered SER task.

The influence of setting the initial values of alpha and beta on the performance of AReLU has been described in the work of [14]. From our experimental results, we conclude that discovering the ideal initial values of alpha and beta is of paramount importance specifically in integrating AReLU within GRU and more so in the context of SER. In our subsequent studies, we intend to carry out an extensive grid-search for optimal values of alpha and beta to reach the best benefits of usage of AReLU within GRU.

Another dimension for further exploration is to experiment with other activation functions within GRU. Recent research in activation function have led to the discovery of several non-learnable activations like SELU [18], EELU [19], Mish [20], and learnable activations such as Comb [21] and PAU [22]. A comparative study of usage of such learnable and non-learnable activations within GRU is an interesting future work. Our experiments demonstrated that a small contribution of the negative values aids in getting improved results. In this context, usage of non-learnable activation functions like Leaky ReLU could be further explored. Other learnable activations that handle negative values similar to AReLU could also be evaluated to analyse the impact on accuracy.

5. REFERENCES

[1] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie, "Emotion Representation, Analysis and Synthesis in Continuous Space: A Survey," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, Santa Barbara, CA, 2011.

[2] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] Björn Schuller, Gerhard Rigoll, and Manfred Lang, "Hidden Markov model-based speech emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[4] Yi-Lin Lin and Gang Wei, "Speech emotion recognition based on HMM and SVM," in *Proceedings of the International Conference on Machine Learning and Cybernetics*, 2005, vol. 8, pp. 4898–4901.

[5] Alif Bin Abdul Qayyum, Asiful Arefeen, and Celia Shahnaz, "Convolutional Neural Network (CNN) Based Speech-Emotion Recognition," in *Proceedings of the IEEE International Conference on Signal Processing, Information, Communication Systems*, 2019, pp. 122–125.

[6] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, 2017.

[7] Ziping Zhao, Zhongtian Bao, Yiqin Zhao, Zixing Zhang, Nicholas Cummins, Zhao Ren, and Björn Schuller, "Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition," *IEEE Access*, 2019.

[8] Yeonguk Yu and Yoon-Joong Kim, "Attention-LSTM-Attention Model for Speech Emotion Recognition and Analysis of IEMOCAP Database," *Electronics*, 2020.

[9] Jianfeng Zhao, Xia Mao, and Lijiang Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[10] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 1997.

[11] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014.

[12] Gaetan Ramet, Philip N. Garner, Michael Baeriswyl, and Alexandros Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *Proceedings of the IEEE Spoken Language Technology Workshop*, Athens, Greece, 2018, pp. 126–131.

[13] Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee, "An Interaction-aware Attention Network for Speech Emotion Recognition in Spoken Dialogs," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, 2019.

[14] Dengsheng Chen, Jun Li, and Kai Xu, "AReLU: Attention-based Rectified Linear Unit," *arXiv preprint*, p. arXiv:2006.13858, 2020.

[15] Carlos Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, 2008.

[16] Florian Eyben, Martin Wöllmer, and Björn Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the ACM Multimedia International Conference*, Florence, Italy, 2010.

[17] Suping Zhou, Jia Jia, Q. Wang, Y. Dong, Yufeng Yin, and Kehua Lei, "Inferring Emotion from Conversational Voice Data: A Semi-Supervised Multi-Path Generative Neural Network Approach," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, Louisiana, USA, 2018.

[18] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter, "Self-normalizing neural networks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, 2017, p. 972–981.

[19] Daeho Kim, Jinah Kim, and Jaeil Kim, "Elastic Exponential Linear Units for Convolutional Neural Networks," *Neurocomputing*, vol. 406, pp. 253–266, 2020.

[20] Diganta Misra, "Mish: A Self Regularized Non-Monotonic Activation Function," in *Proceedings of the 31st British Machine Vision Conference*, 2020.

[21] Franco Manessi and Alessandro Rozza, "Learning Combinations of Activation Functions," in *Proceedings of the 24^{th} International Conference on Pattern Recognition*, Beijing, China, 2018, pp. 61–66.

[22] Alejandro Molina, Patrick Schramowski, and Kristian Kersting, "Padé Activation Units: End-to-end Learning of Flexible Activation Functions in Deep Networks," in *Proceedings of the International Conference on Learning Representations*, 2020.