ATTENTIONAL GATED RES2NET FOR MULTIVARIATE TIME SERIES CLASSIFICATION

Chao Yang* *Xianzhi Wang** *Lina Yao*† *Guodong Long*‡ *Jing Jiang*‡ *Guandong Xu*$*$

* School of Computer Science, University of Technology Sydney, Sydney, Australia
† School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
‡ Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, Australia
$*$ Data Science Institute, University of Technology Sydney, Sydney, Australia

ABSTRACT

Multivariate time series classification is a critical problem in data mining with broad applications. We design a novel convolutional neural network architecture, Attentional Gated Res2Net, for robust multivariate time series classification. AGRes2Net uses hierarchical residual-like connections to achieve multi-scale receptive fields and to capture multi-granular temporal patterns. It further employs the gated mechanism to harness inter-relationship between feature maps. We propose two types of attention modules, namely channel-wise attention and block-wise attention, to leverage the multi-granular temporal patterns. Our experiments on six benchmark datasets demonstrate that AGRes2Net not only outperforms several baselines and state-of-the-art methods but also improves the classification accuracy of existing models when used as a plug-in.

Index Terms— multivariate time series classification, convolutional neural networks, attention mechanism, deep learning

1. INTRODUCTION

Multivariate time series classification is the problem of predicting the label for a sequence of signals given a series of sequences with known labels. It is one of the most challenging tasks that has demonstrated significance in various domains, such as activity recognition [1], disease diagnosis [2], and automatic device classification [3]. Deep learning techniques, such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), gain popularity for their abilities to handle massive data and complicated, nonlinear relations via automatic feature extraction and representation learning [4].

Among existing deep learning models for multivariate time series classification, Long Short-Term Memory (LSTM)-based models achieve state-of-the-art performance on many datasets [5], but consume extra training time. In comparison, CNN can better harness graphics processing units (GPUs) for parallel training but still faces challenges in capturing long-term temporal patterns.

While existing studies heavily focus on classification accuracy, less attention is paid to the robustness of multivariate time series classification. Given time-series sequences of various lengths, existing models generally require careful structural adjustment to gain satisfactory performance. Based on the above, we aim to develop a model that can generalize on different sequence lengths and variable numbers while attaining classification accuracy and training efficiency. In this regard, we propose a novel CNN-based architecture, Attentional Gated Res2Net (AGRes2Net), for multivariate time series classification. Since our model is based on CNN, it preserve the capability to scale to parallel computing settings for accelerated training.

We make the following contributions in this paper:

- We propose AGRes2Net for robust multivariate time series classification. AGRes2Net can capture not only temporal patterns of varied ranges but also exploit interrelationship between variables. It is robust for time-series sequences of varied lengths with varied numbers of variables.

- We propose two attention mechanisms, channel-wise attention and block-wise attention, to fully leverage the multi-granular temporal information. Channel-wise attention performs better on datasets with more variables, while block-wise attention is more suitable for avoiding overfitting on datasets with fewer variables.

- Our experiments on six benchmark datasets show the superior performance of AGRes2Net to several baselines and state-of-the-art methods. We also use LSTM-FCNs[5] as an example to incorporate AGRes2Net as a pluggable component to demonstrate the effectiveness of AGRes2Net in improving existing models' performance.

2. RELATED WORK

Traditional time series classification methods are largely based on machine learning techniques, such as K-Nearest

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

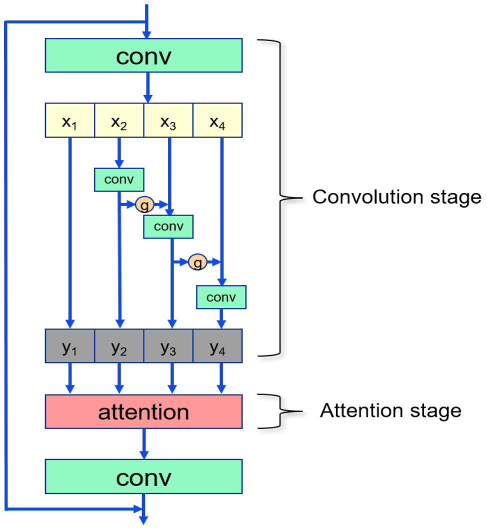

Fig. 1. The structure of AGRes2Net. We suppose there are 4 groups of input feature maps to ease illustration.

Neighbors, Support Vector Machines [6] and Random Forest [7]. Recent studies have tried CNNs [8, 9] for time series classification. As CNN faces difficulty in learning long-range dependencies, it is combined with RNN [10, 3] to exploit the advantages of both CNN and RNN. In particular, LSTM-FCNs [5] construct CNN and RNN in parallel and achieves state-of-the-art performance on several benchmark datasets. A limitation with RNN-based architectures is that they cannot fully leverage the power of GPUs, leading to extended training time.

Our work is fundamentally based on Res2Net [11], a CNN backbone specially designed for feature extraction with multi-scale receptive fields. Res2Net differs from CNN by replacing the standard convolutional filters with s groups of filters. Each group of filters has *w* channels. The groups are connected in a hierarchical residual-like style to achieve multi-scale receptive fields, making it suitable for multivariate time series analysis. Gated Res2Net [9] (or GRes2Net for short) enhances Res2Net by controlling the information flow with the gated mechanism. However, it cannot leverage multi-glandular temporal patterns and tends to forget the information from the low-level temporal sequences during computation. In this paper, we aim to address the limitations of GRes2Net by enforcing robust multivariate time series classification.

3. ATTENTIONAL GATED RES2NET

We propose Attentional Gated Res2Net (AGRes2Net) for Multivariate time series classification. AGRes2Net includes two stages, convolution and attention, as shown in Fig. 1.

3.1. Convolution stage

The computational process of AGRes2Net starts by feeding the input to a convolutional layer for channel expansion. Then we have the input feature map, denoted by $\mathbf{X}$. Following that, $\mathbf{X}$ is divided into groups along the channel, resulting in $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_s\}$, where s is the number of groups. Then, the groups are sent to s groups of hierarchical residual-like convolutional filters for feature extraction, where we use the gated mechanism to control the information flow between groups. The Gated mechanism is proven successful for capturing long-range dependency in time series [15, 16]. Recurrent neural networks are prone to use gates [17, 18] to determine whether the information should be forgotten.

We calculate the value of each gate $\mathbf{g}_i$ based on the prior output feature maps $\mathbf{y}_{i-1}$ and the current input feature maps $\mathbf{x}_i$:

$$\mathbf{g}_i = \tanh\left(a\left(\text{concat}\left(a(\mathbf{y}_{i-1}), a\left(\mathbf{x}_i\right)\right)\right)\right). \tag{1}$$

where a represents either fully-connected layers or convolutional layers, and *concat* is the concatenation operation. Specifically, we set a convolutional layers in our model.

During the convolution stage, we obtain the $\mathbf{y}_i$ as follows:

$$\mathbf{y}_i = \begin{cases} \mathbf{x}_i & i = 1 \\ \mathbf{K}_i\left(\mathbf{x}_i\right) & i = 2 \\ \mathbf{K}_i\left(\mathbf{x}_i + \mathbf{g}_i \cdot \mathbf{y}_{i-1}\right) & 2 < i < s \end{cases} \tag{2}$$

where $\mathbf{K}_i$ denotes convolution and $\mathbf{y}_i$ denotes the output of $\mathbf{K}_i(\cdot)$.

3.2. Attention stage

We propose two types of attention modules, namely channel-wise attention and block-wise attention, to capture multi-granular temporal patterns within multivariate time series.

3.2.1. Channel-wise attention

The channel-wise attention module captures the relations among the results of the convolution stage, i.e., the feature maps of each $\mathbf{y}_i$, while $\mathbf{i} \in \{\mathbf{1,2,3\ldots s}\}$ and **s** is the maximum number of groups mentioned in the previous section. Suppose $\mathbf{y}_i$ contains ***j*** channels, and $\mathbf{j} \in \{\mathbf{1,2,3\ldots J}\}$, where **J** is the maximum number of channels in $\mathbf{y}_i$. Let $\mathbf{h}_{i,j}$ indicate the feature map of the jth channel of $\mathbf{y}_i$. We use three fully-connected layers to learn the query, key and value of the $\mathbf{h}_{i,j}$, donated by $\mathbf{q}_{i,j}$, $\mathbf{k}_{i,j}$ and $\mathbf{v}_{i,j}$.

For another feature map $\mathbf{h}_{m,n}$, we use the same way to obtain $\mathbf{q}_{m,n}$, $\mathbf{k}_{m,n}$ and $\mathbf{v}_{m,n}$ of the $\mathbf{h}_{m,n}$. Then, we calculate the **attention** as follows:

$$\textbf{attention}\left(\mathbf{q_{i,j}}, \mathbf{k_{m,n}}\right) = \frac{\mathbf{q_{i,j} k_{m,n}^T}}{\sqrt{\mathbf{J}}} \tag{3}$$

Once computed, we update the feature map of every channel according to its relations with all the other feature maps.

Table 1. Experimental datasets

Dataset	# Classes	# Variables	Length	Train-Test ratio	SOTA method
MSR Action [12]	20	570	100	48:52	MALSTM-FCN[5]
Ozone [13]	2	72	291	50:50	MLSTM-FCN[5]
AREM [13]	7	7	480	50:50	MALSTM-FCN[5]
LP5 [13]	5	6	15	39:61	Weasel+ muse[14]
EEG [13]	2	13	117	50:50	MLSTM-FCN[5]
Gesture Phase [13]	5	18	214	50:50	MLSTM-FCN[5]

As those feature maps contain temporal information within various ranges, the channel-wise attention module can capture various levels of temporal features and dependencies. Specifically, the updated $\mathbf{h}_{i,j}^{\sim}$ is calculated as follows:

$$\mathbf{h_{i,j}^{\sim}} = \sum_{\mathbf{s}} \sum_{\mathbf{J}} \mathbf{Softmax} \left(\frac{\mathbf{attention} \left(\mathbf{q_{i,j}}, \mathbf{k_{m,n}} \right)}{\sum_{\mathbf{s}} \sum_{\mathbf{J}} \mathbf{attention} \left(\mathbf{q_{i,j}}, \mathbf{k_{m,n}} \right)} \right) \mathbf{v_{m,n}} \tag{4}$$

3.2.2. Block-wise attention

The block-wise attention module regards each $\mathbf{y}_i$ as an individual block that contains the temporal information at a certain granularity level. Instead of separating the channel information, the block-wise attention module feeds $\mathbf{y}_i$ to fully-connected layers to calculate the query, key, and value of each block and then conducts self-attention based on the results.

4. EXPERIMENTS

4.1. Experimental Setup

We conducted experiments on six public multivariate time series datasets that consist of sequences of varied lengths (shown in Table 1). We strictly follow the same data preprocessing steps as taken by state-of-the-art (SOTA) methods for a fair comparison. In particular, we normalize each dataset to zero mean and unit standard deviation, and we apply zero padding to cope with sequences with differed lengths.

We select several competitive baselines and SOTA methods, as illustrated below, to compare with our model:

- **Res2Net** [11] and **GRes2Net** [9]: they are the basic Res2Net and Gated Res2net, which are introduced in Section 2.
- **Res2Net+SE** and **GRes2Net+SE**: these methods additionally incorporate Squeeze-and-Excitation Block (SE) [19] to take full advantage of attention modules.
- **SOTA**: they represent state-of-the-art methods on the experimental datasets as summarized in Table 1.

We use *accuracy* as the evaluation metric, given that all SOTA methods use this single metric for the experimental datasets. We train our model for 500 training epochs using the Adam [5] optimizer. We initialize the learning rate to 0.001 and adjust it to 1/10 of itself after every 100 epochs. We also use dropout to avoid overfitting. We average five runs of models to evaluate models' performance reliably.

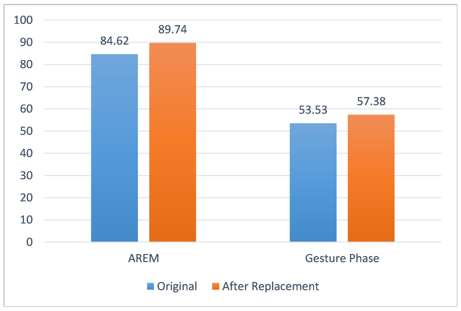

Fig. 2. Performance of LSTM-FCNs before and after incorporating AGRes2Net. The blue bar represents the accuracy (%) of the vanilla LSTM-FCNs; the right bar shows the accuracy of LSTM-FCNs with FCN replaced by AGRes2Net.

4.2. Comparison with Other Methods

Table 2 shows the performance comparison of all the methods on six datasets. Our proposed model outperformed all the baseline methods on all datasets, demonstrating the model's superior robustness on various lengths of time-series sequences. Block-wise attention performed better than channel-wise attention on two datasets, AREM and LP5, which contain the least number of variables among the datasets. Thus, block-wise attention is more suitable for datasets that contain a limited number of variables.

Our model outperforms the two baselines incorporated with the SE module (Res2Net+SE and GRes2Net+SE), revealing the advantages of our attention modules over the SE module. Global average pooling may cause information loss in the SE module, while our attention modules can better utilize the feature maps information.

4.3. Effectiveness as a Pluggable Component

We use LSTM-FCNs [5], the SOTA architecture on most of our experimental datasets, as the backbone to demonstrate

Table 2. Accuracy (%) of compared methods on six benchmark datasets. AGRes2Net-CAtt and AGRes2Net-BAtt represent our proposed models with channel-wise attention and block-wise attention, respectively. The SOTA methods are WEASEL+MUSE [14] for LP5 dataset and LSTM-FCNs [5] for all the other datasets.

Method	MSR Action	Ozone	AREM	LP5	EEG	Gesture Phase
Res2Net	74.57	79.89	76.92	56.42	57.81	58.59
Res2Net+SE	71.82	82.64	84.62	57.99	54.69	58.98
GRes2Net	73.01	80.34	82.05	68.40	60.94	66.01
GResNet+SE	80.37	83.90	87.17	63.28	64.06	66.41
SOTA	74.74	81.50	84.62	71.00	65.63	53.53
AGRes2Net-CAtt	**82.70**	**86.20**	89.74	73.26	**67.19**	**69.53**
AGRes2Net-BAtt	73.32	84.94	**92.31**	**73.96**	**67.19**	64.45

Table 3. Performance of AGRes2Net-CAtt under varied numbers of groups of input feature maps (denoted by s) on AREM dataset.

	Train				Test			
s	Accuracy (%)	Recall (%)	Precision (%)	F-score	Accuracy (%)	Recall (%)	Precision (%)	F-score
4	79.07	77.98	75.34	0.77	79.49	70.07	71.71	0.71
8	81.39	84.52	82.5	0.83	74.35	68.71	82.68	0.75
16	86.05	88.7	83.97	0.86	82.05	77.56	86.48	0.82
32	88.37	90.48	89.39	0.89	89.74	83.63	90.56	0.87
64	97.67	98.21	98.41	0.98	76.92	96.19	84.52	0.90

Table 4. Ablation test on EEG dataset.

Model	Accuracy (%)	Recall (%)	Precision (%)	F-Score
Res2Net	57.81	57.13	58.82	0.58
Res2Net+Gates	59.38	59.43	59.43	0.59
Res2Net+Attention	60.94	61.05	64.17	0.63
AGRes2Net-CAtt	**67.19**	**67.50**	**68.33**	**0.68**

Table 5. Ablation test on AREM dataset

Model	Accuracy (%)	Recall (%)	Precision (%)	F-Score
Res2Net	76.92	74.69	76.39	0.76
Res2Net+Gates	82.05	75.51	87.62	0.76
Res2Net+Attention	87.18	81.63	89.29	0.83
AGRes2Net-BAtt	**92.31**	**87.62**	**95.71**	**0.89**

the effectiveness of our model in improving the classification accuracy of existing deep learning architectures. We replace the original convolutional modules of LSTM-FCNs with AGRes2Net. Fig. 2 shows the replacement improved the accuracy of LSTM-FCNs on two benchmark datasets. We omit to show the results on the other datasets as they lead to the same conclusion.

4.4. Parameter Sensitivity Study

We explore the impact of the number of groups of the input feature maps s on the performance of AGRes2Net, we give the results in Table 3. Generally, constructing more groups of filters can force the model to concentrate on a broader range of temporal information. With the s increasing, the convolutional layers become wider, leading to more receptive fields of different sizes. Our experimental results on AREM dataset show our model that uses channel-wise attention performed best at $s = 64$. We omit to show results on other datasets due to the limited space.

4.5. Ablation Study

We conduct ablation studies to explore the effectiveness of the gate mechanism and the attention module, respectively. To this end, we compare the model without gates and attention module (i.e., the vanilla Res2Net), the model with only gates (i.e., GRes2Net), the model with only attention, and the model with both gates and attention (i.e., AGRes2Net). We use channel-wise attention and EEG dataset as an example to show the ablation study results (Table 4). According to the results, both components help improve the model's accuracy while the attention mechanism plays a greater part than gates.

5. CONCLUSION

We propose a novel model, AGRes2Net, for multivariate time series classification under various sequence lengths. AGRes2Net comprehensively leverages gate and attention mechanisms to capture temporal dependencies at various ranges. Our experiments on six datasets demonstrate it outperforms several baselines and it state-of-the-art methods, and improves existing models when used as a plugin.

6. REFERENCES

[1] Zhibin Yu and Minho Lee, "Real-time human action classification using a dynamic neural model," *Neural Networks*, vol. 69, pp. 29–43, 2015.

[2] R Chitra and V Seenivasagam, "Heart disease prediction system using supervised learning classifier," *Bonfring International Journal of Software Engineering and Soft Computing*, vol. 3, no. 1, pp. 01–07, 2013.

[3] Lei Bai, Lina Yao, Salil S Kanhere, Xianzhi Wang, and Zheng Yang, "Automatic device classification from network traffic streams of internet of things," in *2018 IEEE 43rd Conference on Local Computer Networks (LCN)*. IEEE, 2018, pp. 1–9.

[4] Kurt Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural networks*, vol. 4, no. 2, pp. 251–257, 1991.

[5] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford, "Multivariate lstm-fcns for time series classification," *Neural Networks*, vol. 116, pp. 237–245, 2019.

[6] Argyro Kampouraki, George Manis, and Christophoros Nikou, "Heartbeat time series classification with support vector machines," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 512–518, 2008.

[7] Mustafa Gokce Baydogan, George Runger, and Eugene Tuv, "A bag-of-features framework to classify time series," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2796–2802, 2013.

[8] Bendong Zhao, Huanzhang Lu, Shangfeng Chen, Junliang Liu, and Dongya Wu, "Convolutional neural networks for time series classification," *Journal of Systems Engineering and Electronics*, vol. 28, no. 1, pp. 162–169, 2017.

[9] Chao Yang, Mingxing Jiang, Zhongwen Guo, and Yuan Liu, "Gated res2net for multivariate time series analysis," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[10] Chao Yang, Wenxiang Jiang, and Zhongwen Guo, "Time series data classification based on dual path cnn-rnn cascade network," *IEEE Access*, vol. 7, pp. 155304–155312, 2019.

[11] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[12] Wanqing Li, Zhengyou Zhang, and Zicheng Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 9–14.

[13] Dheeru Dua and Casey Graff, "UCI machine learning repository," 2017.

[14] Patrick Schäfer and Ulf Leser, "Multivariate time series classification with weasel+ muse," *arXiv preprint arXiv:1711.11343*, 2017.

[15] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever, "An empirical exploration of recurrent network architectures," in *International Conference on Machine Learning*, 2015, pp. 2342–2350.

[16] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[17] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[18] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.