COMPACT CONVOLUTIONAL RECURRENT NEURAL NETWORKS VIA BINARIZATION FOR SPEECH EMOTION RECOGNITION

HUAN ZHAO[1*] YUFENG XIAO[1] JING HAN[2] ZIXING ZHANG[1,3]
[1]COLLEGE OF COMPUTER SCIENCE AND ELECTRONIC ENGINEERING, HUNAN UNIVERSITY, CHINA
[2] EMBEDDED INTELLIGENCE FOR HEALTH CARE AND WELLBEING, UNIVERSITY OF AUGSBURG, GERMANY
[3]GROUP ON LANGUAGE, AUDIO & MUSIC, IMPERIAL COLLEGE LONDON, UK

INTRODUCTION

Despite the great advances, most of the recently developed automatic speech recognition systems focus on working in a *server-client* manner. The following issues struggle to satisfy the increasing demand for a succinct model that run smoothly in embedded devices like smartphones:

- High computational cost
- Privacy protection
- Limited network bandwidth

In this paper, we proposed a *binarization* approach to cope with the raised problem. In doing this, the model can be stored with less disk storage, and can be processed in less computational complexity.

RESULTS

Approach	IEMOCAP	Emo-DB
DNN-ELM [2]	51.2	71.6
3-D ACRNN [3]	64.2	81.5
Full-precision CRNN	62.4	80.1
BCRNN	61.9	79.7

Table 1: Performance comparison in term of Unweighted Average Recall (UAR [%]) between the proposed BCRNN with the baseline system and other state-of-the-art systems on the IEMOCAP and Emo-DB.

Approaches	Model size (MB)
DNN-ELM [2]	12.33
3-D ACRNN [3]	323.46
Full-precision CRNN	105.48
BCRNN	4.34

Table 2: Model size comparison between the proposed Binary Convolutional Recurrent Neural Network (BCRNN) with its original full-precised system and other state-of-the-art systems.

CONCLUSION

- Comparable results but with a high model size compression rate
- Complex convolution operations are largely accelerated by simple binary operations.

REFERENCES

[1] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

[2] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Proc. INTERSPEECH*, Singapore, 2014.

[3] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444, Oct. 2018.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation of China under Grant 61772188 and National Key R&D Program of China Grant 2018YFC0831800.

THE PROPOSED MODEL

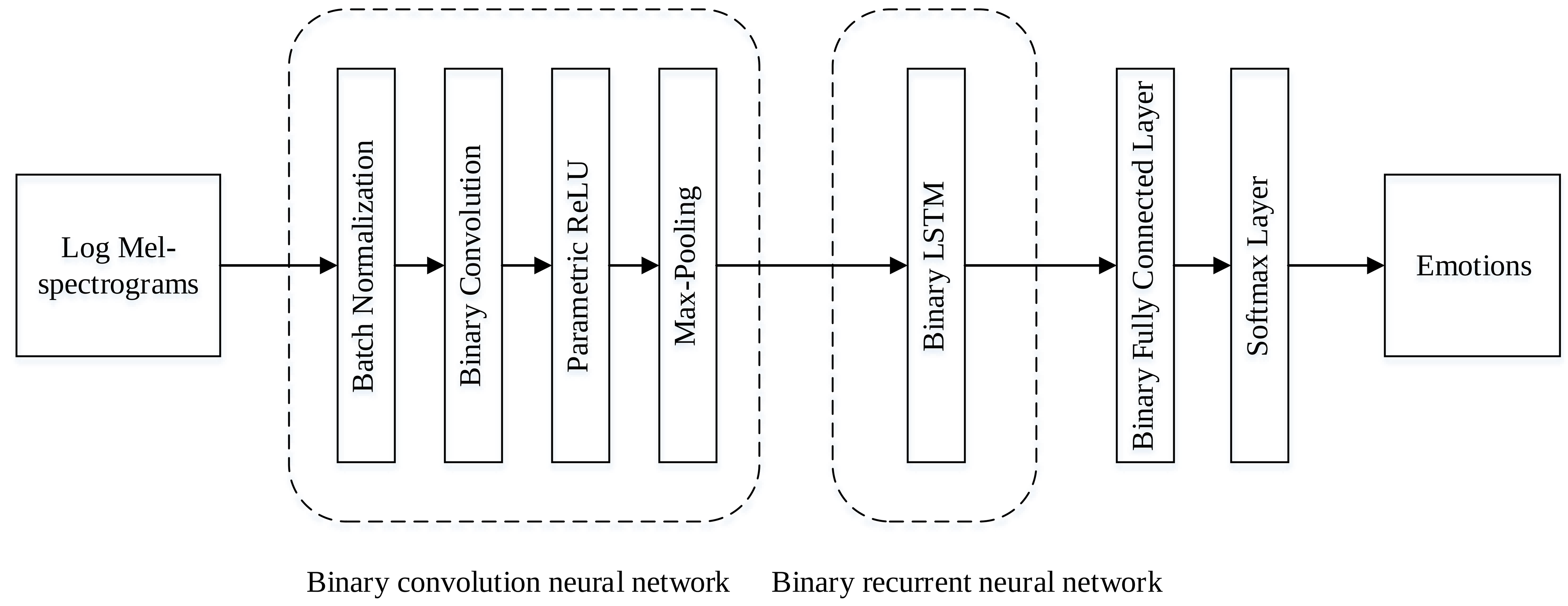

Figure 1: The framework of the proposed compact convolutional recurrent neural network via binarization for speech emotion recognition, which consists of a binary CNN, a binary LSTM-RNN, and a binary fully-connected network.

For **binarization**, we employ the deterministic binarization function as the previous work in [1] .

$$b = \mathrm{sign}(x) = \begin{cases} +1 & \text{if} \quad x \geq 0, \\ -1 & \text{otherwise}, \end{cases} \tag{1}$$

Then, a scaling factor α is introduced to approximate $\mathbf{X}$ by $\alpha\mathbf{B}$. Mathematically, L2 loss function is minimized to obtain an optimal $\alpha^\star$.

$$\alpha^\star = \frac{\mathbf{X}^{\mathrm{T}}\,\mathrm{sign}(\mathbf{X})}{n} = \frac{\sum |X_i|}{n}. \tag{2}$$

BCNN is different from CNN which conducts binary convolution in the convolutional layer. The convolution between $\mathbf{W}$ and $\mathbf{I}$ can be approximated by the binary convolution operation:

$$\mathbf{I} * \mathbf{W} = (\mathrm{sign}(\mathbf{I}) * \mathrm{sign}(\mathbf{W})) * \beta\mathbf{K}. \tag{3}$$

where $\mathbf{K}$ is a scaling factor matrix of input $\mathbf{I}$ and β is a scaling factor of weight $\mathbf{W}$.
BRNN is derived from traditional LSTM. The mathematical expression of LSTM structure can be expressed as:

$$\begin{aligned} \mathbf{d}_t &= [\mathbf{x}_t, \mathbf{h}_{t-1}] \\ \mathbf{I}_t, \mathbf{F}_t, \mathbf{O}_t, \mathbf{G}_t &= \mathbf{W}\mathbf{d}_t \\ \{\mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t\} &= \sigma(\{\mathbf{I}_t, \mathbf{F}_t, \mathbf{O}_t\}) \\ \mathbf{g}_t &= \tanh(\mathbf{G}_t) \\ \mathbf{c}_t &= \mathbf{f}_t \cdot \mathbf{c}_{t-1} + \mathbf{i}_t \cdot \mathbf{g}_t \\ \mathbf{h}_t &= \mathbf{o}_t \cdot \tanh(\mathbf{c}_t), \end{aligned} \tag{4}$$

Then, similarly as in the BCNN model, scaling factors α and β are introduced to approximate the term $\mathbf{W}\mathbf{d}_t$ in Eq. (4) by $\alpha\mathbf{W}^b\beta\mathbf{d}_t{}^b$.
In **backward propagation**, since the gradient for *sign* function is problematic as the derivative of it is zero almost everywhere, we follow previous work in [1] and compute it using the straight-through estimator approach. The gradient $\frac{\partial C}{\partial q}$ can be obtained by:

$$g_r = g_q 1_{|r| \leq 1}, \tag{5}$$

where C is the loss function, and the gradient is canceled when r is too large.