Electronics Research Institute
Sharif University of Technology

icassp 2022
Singapore

Light-SERNet: A Lightweight Fully Convolutional Neural Network for Speech Emotion Recognition

Authors: Arya Aftab, Alireza Morsali, Shahrokh Ghaemmaghami, Benoit Champagne

Presented by: Arya Aftab

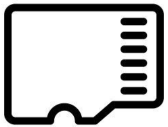
IEEE Signal Processing Society

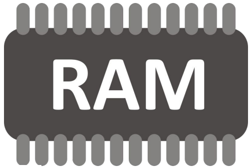
Previous Works

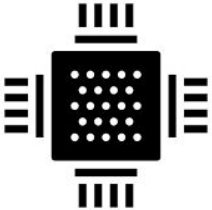
- Employed several parallel paths with large convolutional filters [Yenigalla'18]
- Proposed a 3-D attention-based convolutional recurrent neural network [Chen'18]
- Proposed a combination of dilated residual network and multi-head self-attention [Li'19]
- Quantized the weights of the neural networks [Zhao'19]
- Combined the attention mechanism and the focal loss [Zhong'20]

IoT Devices

- Model Size
- Peak Memory Usage(PMU)
- Computational Cost

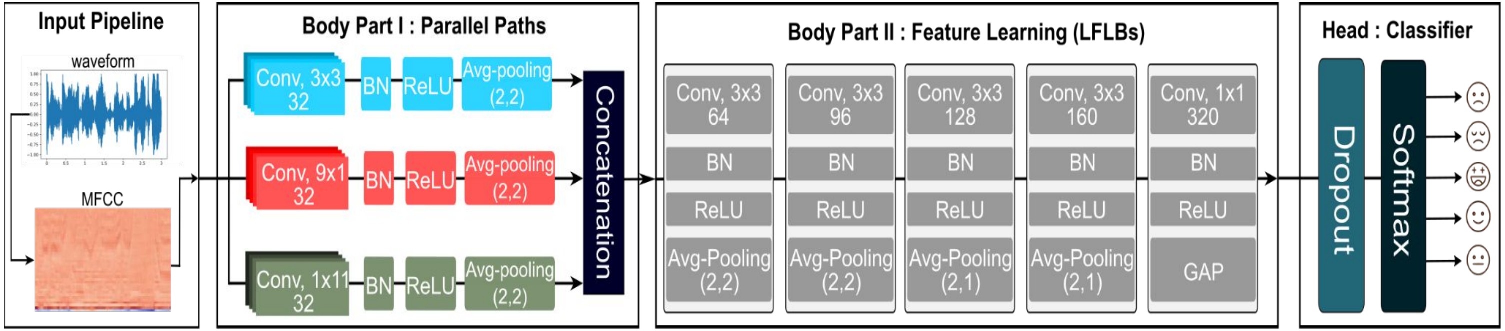






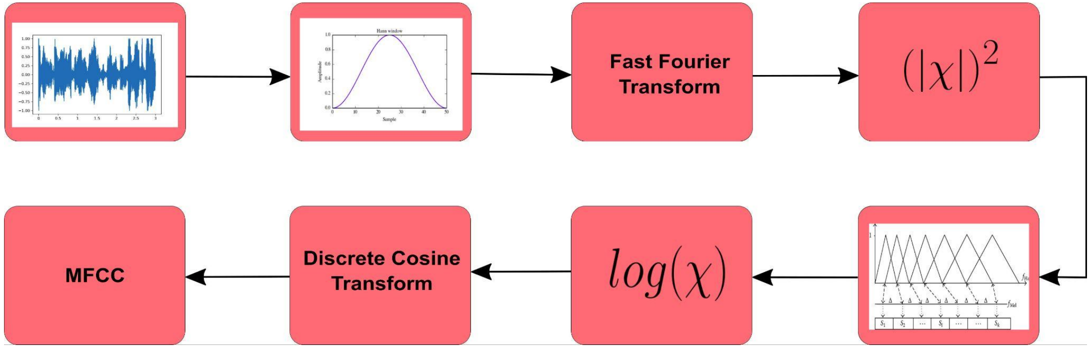
Optimization of Model

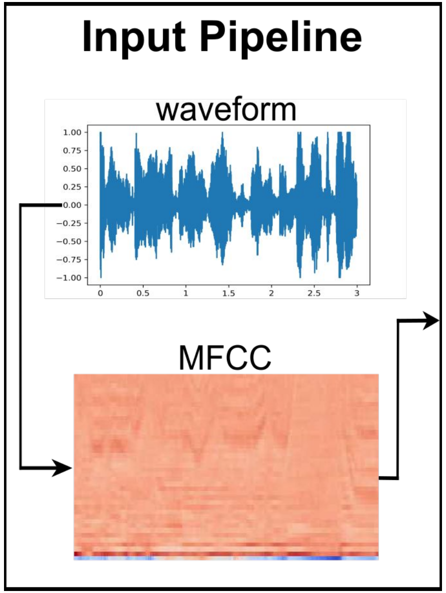
- Input Pipeline
- Feature Extractor
- Classifier



Input Pipeline

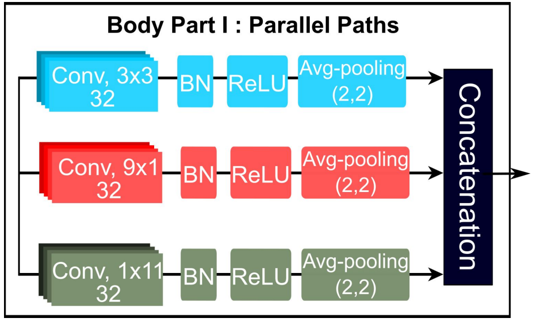
- Input Size
- Window Size
- Input Feature Type
- Number of Features





Feature Extractor

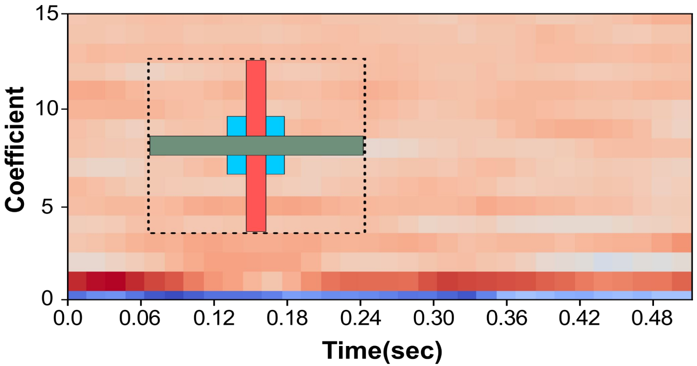
Body Part 1:



Receptive field size:

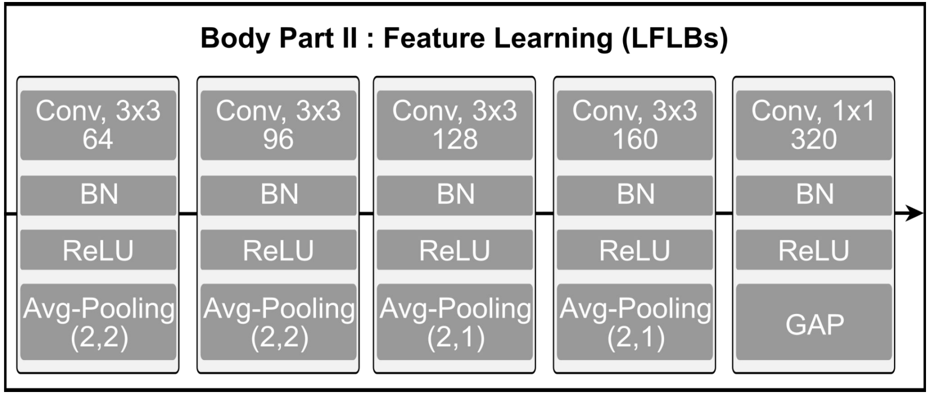
$$r_{l-1} = s_l \cdot r_l + (k_l - s_l) \qquad \text{(Eq. 1}$$

$$r_0 = \sum_{l=1}^{L} \left((k_l - 1) \prod_{i=1}^{l-1} s_i \right) + 1 \qquad \text{(Eq. 2}$$



Feature Extractor

Body Part 2:



Training:

$$
\begin{aligned}
\mu_{\mathcal{B}} &= \frac{1}{m}\sum_{i=1}^{m} x_i \\
\sigma_{\mathcal{B}}^2 &= \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2 \\
\widehat{x}_i &= \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \\
y_i &= \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i)
\end{aligned}
$$

Evaluation:

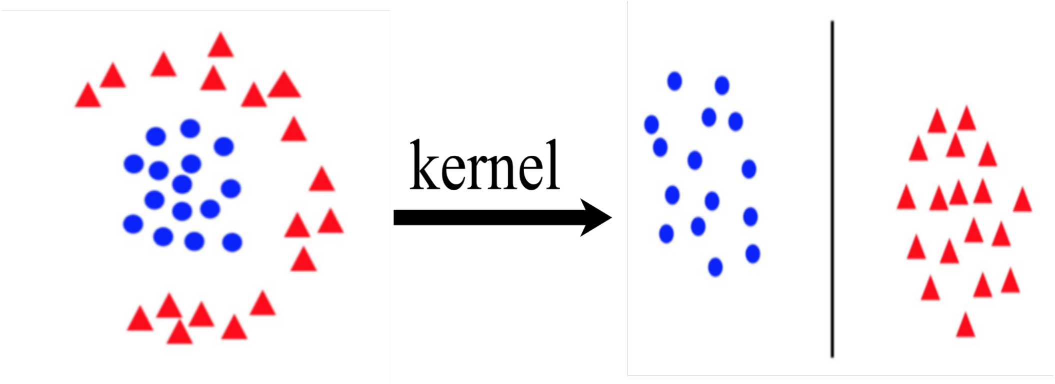
$$
\begin{aligned}
z &= W * x + b \\
\text{out} &= \gamma \cdot \frac{z - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta
\end{aligned}
$$

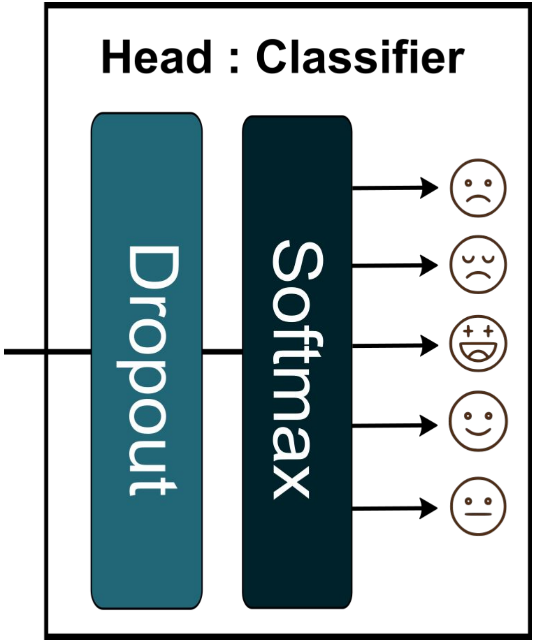
Batch Normalization Folding:

$$
\begin{aligned}
w_{\text{fold}} &= \gamma \cdot \frac{W}{\sqrt{\sigma^2 + \epsilon}} \\
b_{\text{fold}} &= \gamma \cdot \frac{b - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta
\end{aligned}
$$

Classifier

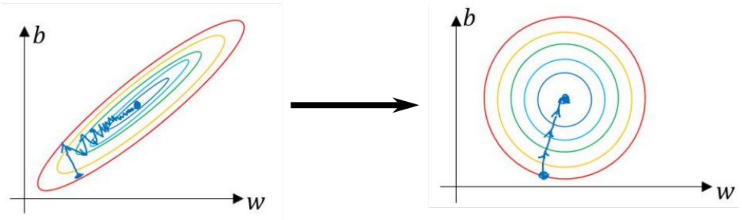
$$Output = W^T f(Input) + b$$



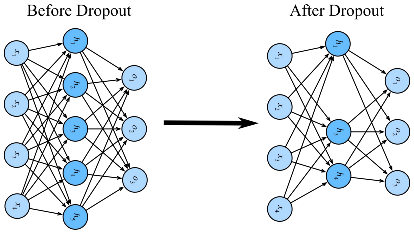


Regularizers

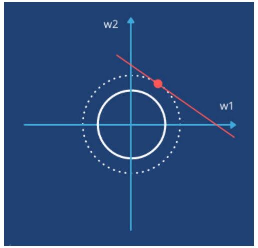
- Batch Normalization



- Dropout



- L2 Regularizer



Comparison

Model on different input lengths and loss function

Table 1: The proposed model performance of different input lengths between CE-Loss and F-Loss on the IEMOCAP (improvised), IEMOCAP (scripted+improvised), and EMO-DB datasets in terms of UA(%), WA(%), and F1(%).

Input Length	IEMOCAP(improvised)						IEMOCAP(scripted+improvised)						EMO-DB					
	F-Loss			CE Loss			F-Loss			CE Loss			F-Loss			CE Loss		
	UA	WA	F1	UA	WA	F1	UA	WA	F1	UA	WA	F1	UA	WA	F1	UA	WA	F1
3 seconds	68.37	77.41	76.01	68.42	76.60	75.44	66.10	65.47	65.42	65.81	65.37	65.40	92.88	93.08	93.05	94.15	94.21	94.16
7 seconds	70.78	79.87	78.84	71.51	78.73	77.86	70.76	70.23	70.20	70.12	69.15	69.09	-	-	-	-	-	-

Comparison

Model on IEMOCAP dataset

Table 2: Comparison of the model size (MB) and performance with those of other methods, on the IEMOCAP (scripted + improvised), in terms of UA, WA, and F1.

Methods	**Size**	**UA(%)**	**WA(%)**	**F1(%)**
Han (2014) [2]	12.3	48.20	54.30	-
Li (2019) [3]	9.90	67.40	-	67.10
Zhong (2020) [4]	0.90	71.72	70.39	70.85
Ours (F-Loss, 7sec)	0.88	70.76	70.23	70.20

Table 3: Comparison of the model size (MB) and performance with those of other methods, on the IEMOCAP (improvised), in terms of UA, WA, and F1.

Methods	**Size**	**UA(%)**	**WA(%)**	**F1(%)**
Chen (2018) [5]	323	64.74	-	-
Yenigalla(2018) [6]	7.20	61.60	71.30	-
Satt (2017) [7]	5.50	62.00	67.30	-
Zhao (2019) [8]	4.34	61.90	-	-
Ours (F-Loss, 7sec)	0.88	70.78	79.87	78.84

Comparison

Model on EMO-DB dataset

Table 4: Comparison of model size (MB) and performance in terms of UA, WA, and F1 with those of other methods on the EMO-DB.

Methods	**Size**	**UA(%)**	**WA(%)**	**F1(%)**
Chen (2018) [5]	323	82.82	-	-
Zhao (2019) [8]	4.34	79.70	-	-
Zhong (2020) [4]	0.90	90.10	91.81	90.67
Ours (CE-Loss, 3sec)	0.88	94.15	94.21	94.16

Conclusions

- Experimental results show that the performance of our model is comparable to that of state-of-the-art models.
- We have proposed a lightweight model that can be used for IoT devices.
- In addition to being lightweight, the other features of our model, such as PMU and computational cost are suitable for IoT devices.
- Due to the use of common layers such as convolution, it can be easily implemented by Tensorflow Lite on devices such as microcontrollers.

References

1- R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, “Dilated residual network with multi-head self-attention for speech emotion recognition,” in IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 6675–6679.

2-Y. Zhong, Y. Hu, H. Huang, and W. Silamu, “A lightweight model based on separable convolution for speech emotion recognition,” in Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, Nov. 2020.

3-M. Chen, X. He, J. Yang, and H. Zhang, “3-d convolutional recurrent neural networks with attention model for speech emotion recognition,” IEEE Signal Process. Lett., vol. 25, no. 10, pp. 1440–1444, Oct. 2018.

4-P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, “Speech emotion recognition using spectrogram & phoneme embedding,” in Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, Sept. 2018, pp. 3688–3692.

5-A. Satt, S. Rozenberg, and R. Hoory, “Efficient emotion recognition from speech using deep learning on spectrograms,” in Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, Aug. 2017, pp. 1089–1093.

6-H. Zhao, Y. Xiao, J. Han, and Z. Zhang, “Compact convolutional recurrent neural networks via binarization for speech emotion recognition,” in IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 6690–6694.