LIGHT-SERNET: Speech Emotion Recognition

Karan Mundhra
Indian Institute of Technology Kanpur
karanm21@iitk.ac.in

October 22, 2024

1 Implementation Details

This paper introduces LIGHT-SERNET, a lightweight fully convolutional neural network (CNN) designed for speech emotion recognition (SER) on devices with limited computational resources. The model processes speech using Mel-frequency cepstral coefficients (MFCCs) and employs three parallel convolutional paths to capture both time- and frequency-based features. With just a 0.88 MB model size, it achieves competitive performance on the IEMOCAP and EMO-DB datasets.

Input Pipeline	**Body Part I**	**Body Part II**	**Head: Classifier**
- Normalize audio signals	- Path 1: 9×1 CNN	- LFLBs	- Dropout (0.3)
- Extract MFCCs	- Path 2: 1×11 CNN	- Convolution	- Fully Connected Layer
- Apply FFT, Mel-scale filter	- Path 3: 3×3 CNN	- ReLU	- Softmax Classification
- Select 40 MFCC coefficients	- Concatenate Paths	- GAP	

The model is trained using the Adam optimizer with an initial learning rate of 10^{-4}, decaying every 20 epochs after epoch 50. Regularization techniques include batch normalization after each convolutional layer, a dropout rate of 0.3 before the softmax layer, and L2 weight decay with a rate of 10^{-6} to reduce overfitting. Training runs for 300 epochs with a batch size of 32.

Github Repository: https://github.com/karanm21/EE798R_term_project

2 Dataset Description

The model is evaluated on two datasets:

- **IEMOCAP**: A multimodal dataset with 12 hours of audio-visual recordings, featuring both scripted and improvised emotional speech. It contains 5,531 labeled utterances with emotions such as happiness, sadness, anger, and neutral.
- **EMO-DB**: A German emotional speech dataset containing 535 utterances across 7 emotion classes, recorded by 10 actors.

Downloaded the IEMOCAP dataset from **Kaggle** and the EMO-DB dataset from **Berlin Database of Emotional Speech**. Implemented EMO-DB dataset on Google Colab and the other one on server.

3 Results

The tables below shows the performance of LIGHT-SERNET on the IEMOCAP and EMO-DB datasets, evaluated using cross-entropy (CE) and focal loss (F-Loss). Metrics include unweighted average recall (UAR), weighted accuracy (WA), and F1-score (F1).

Table 1: Performance of LIGHT-SERNET on IEMOCAP dataset

Dataset	Input Length	Loss Type	UAR (%)	WA (%)	F1 (%)
IEMOCAP (improvised)	3 seconds	F-Loss	67.79	67.14	66.98
IEMOCAP (improvised)	3 seconds	CE-Loss	64.00	67.45	65.27
IEMOCAP (improvised)	7 seconds	F-Loss	66.51	70.34	68.04
IEMOCAP (improvised)	7 seconds	CE-Loss	67.95	71.02	69.48
IEMOCAP (scripted+improvised)	3 seconds	F-Loss	64.23	67.14	65.55
IEMOCAP (scripted+improvised)	3 seconds	CE-Loss	67.56	66.84	67.45
IEMOCAP (scripted+improvised)	7 seconds	F-Loss	70.80	70.08	70.76
IEMOCAP (scripted+improvised)	7 seconds	CE-Loss	71.51	70.57	71.22

Table 2: Performance of LIGHT-SERNET on EMO-DB dataset

Dataset	Input Length	Loss Type	UAR (%)	WA (%)	F1 (%)
EMO-DB	3 seconds	F-Loss	93.71	94.02	93.84
EMO-DB	3 seconds	CE-Loss	94.04	94.02	93.98

3.1 Example Calculation of Metrics

Let's calculate the performance metrics for the IEMOCAP (scripted+improvised) dataset (input length of 7 seconds, focal loss results):

	precision	recall	f1-score	support
angry	0.7775	0.7316	0.7539	1103
sadness	0.7023	0.7574	0.7288	1084
neutral	0.6528	0.7090	0.6798	1703
happiness_excited	0.7059	0.6339	0.6680	1636
accuracy			0.7008	5531
macro avg	0.7096	0.7080	0.7076	5531
weighted avg	0.7031	0.7008	0.7007	5531

Figure 1: Classification Report for IEMOCAP (scripted+improvised, 7 seconds, Focal Loss)

Unweighted Average Recall (UAR):
The unweighted average recall is the mean of the recall values for each class:

$$UAR = \frac{0.7316 + 0.7574 + 0.7090 + 0.6339}{4} = \frac{2.8319}{4} = 0.7080\,(70.80\%)$$

Weighted Accuracy (WA):
The overall accuracy provided in the report is 70.08%.

F1-score (F1):
The macro F1-score is the average of the F1-scores across all classes:

$$F1 = \frac{0.7539 + 0.7288 + 0.6798 + 0.6680}{4} = \frac{2.8305}{4} = 0.7076\,(70.76\%)$$

Table 3: Summary of Metrics for IEMOCAP (7 seconds, Focal Loss)

Metric	Value
Unweighted Average Recall (UAR)	70.80%
Weighted Accuracy (WA)	70.08%
F1-score	70.76%

3.2 Confusion Matrix Plot

Below is the confusion matrix plot for the IEMOCAP (scripted+improvised) dataset using 7-second inputs and focal loss.

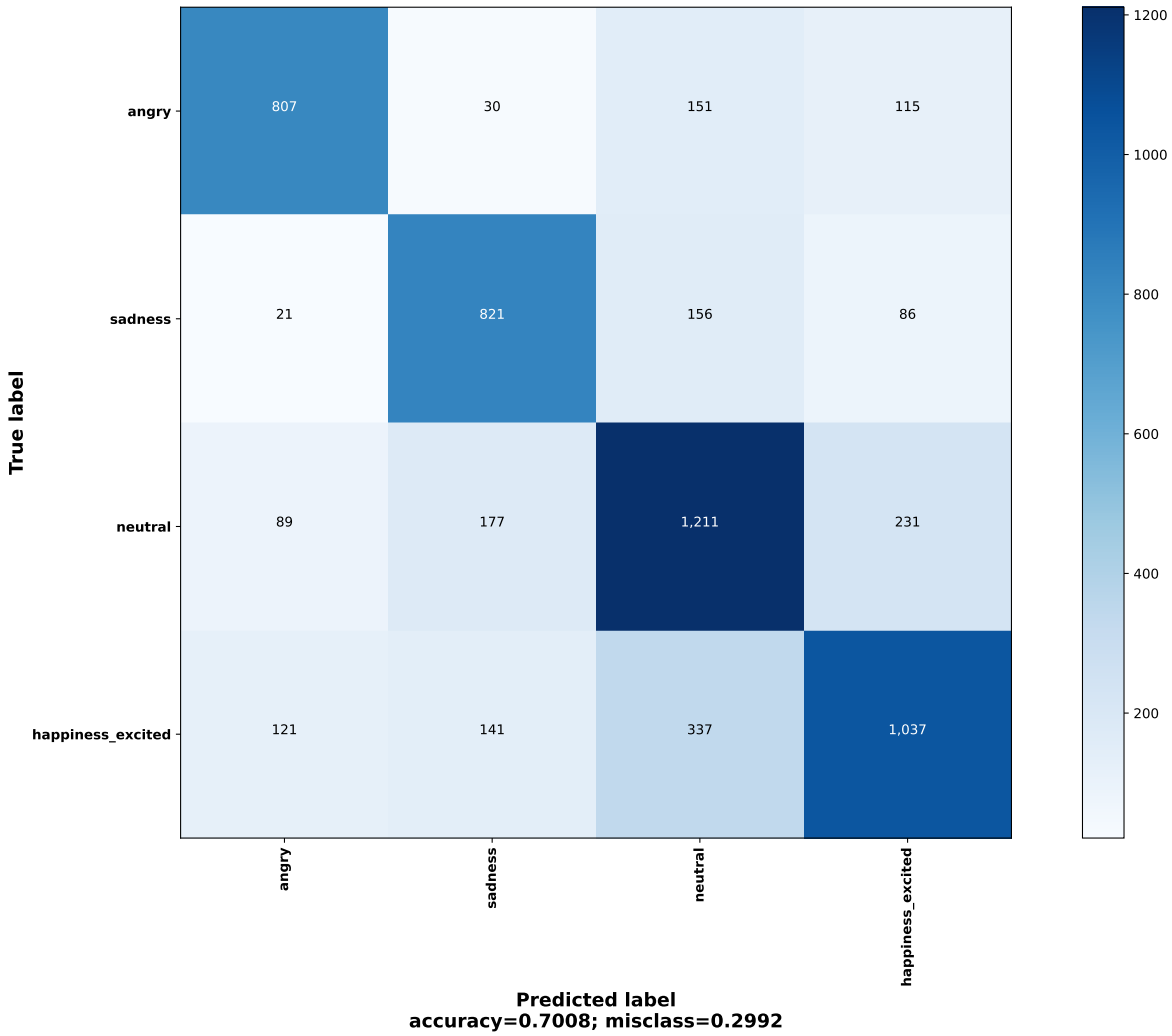

Figure 2: Confusion Matrix for IEMOCAP (scripted+improvised) with 7-second input (Focal Loss)