Light-SERNet: A lightweight fully convolutional neural network for speech emotion recognition

Gianluca Panici 1666962

DEPARTMENT OF COMPUTER, CONTROL, AND MANAGEMENT ENGINEERING ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

Light-SERNet: What is it?

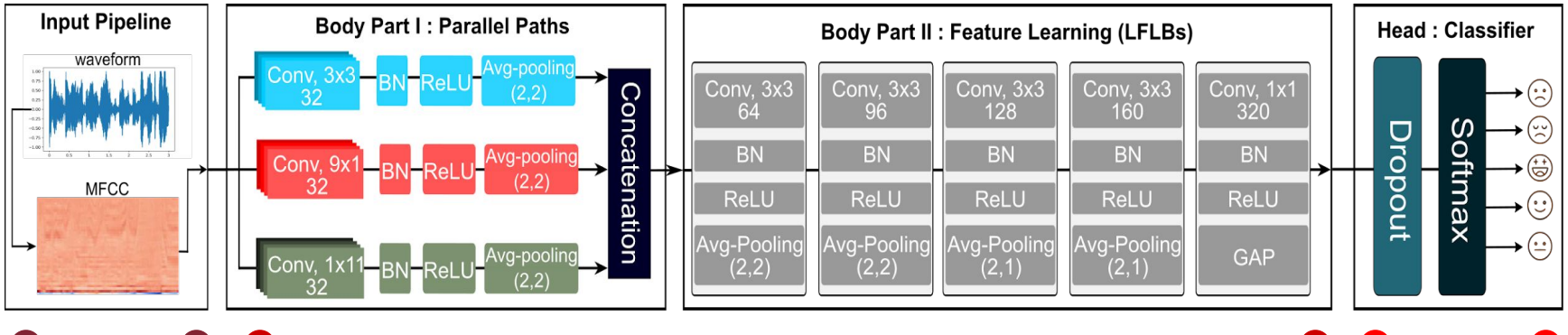
Suitable for low power devices

Light-SERNet is a **lightweight fully convolutional neural network** for **speech emotion recognition**.

CNNs are useful to **divide information** in smaller chunks to analyze and find out some **"hidden" characteristics**.

Speech Emotion Recognition: **SER.**
Feature extractor + classifier

Light-SERNet: Architecture



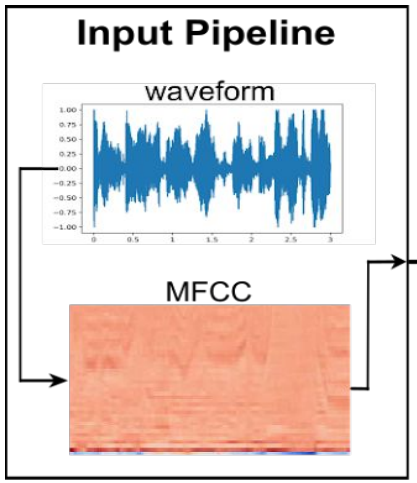
Input Pipeline, for input processing

Body, divided in:

- **Body Part I -** three parallel paths
- **Body Part II -** LFLBs (consecutive layers)

Head, where the classification happens

Architecture: Input Pipeline



INPUT: audio signal

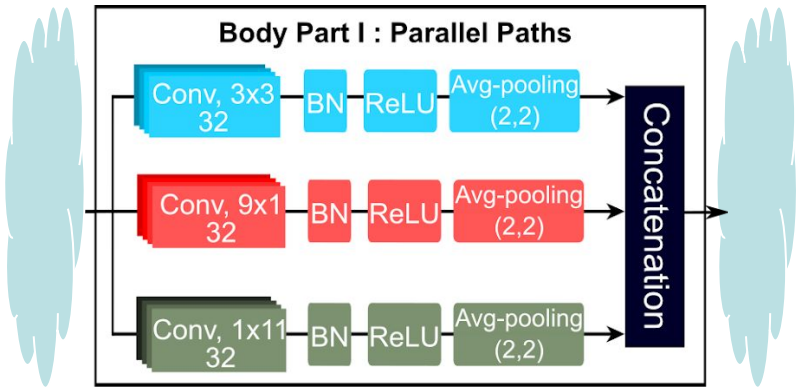
The audio signal is split into 64 ms frames with 16 ms overlaps

A 1024-point Fast Fourier Transform (**FFT**) is applied to each frame

The MFCCs of each frame are calculated using an inverse discrete cosine transform

- Normalization of the audio signal
- Mel Frequency Cepstral Coefficients (**MFCC**) are calculated

Architecture: Body Part I



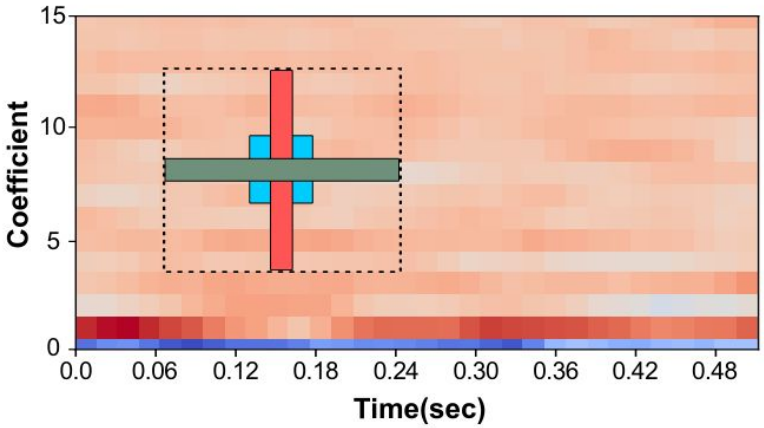
Information is routed in three **parallel paths** in which for each dimension of the multi-dimensional signal a **receptive field** is calculated

Three different **kernel sizes** for features extraction:

- 3 x 3 for spectral-temporal dependencies
- 9 x 1 for spectral dependencies
- 1 x 11 for temporal dependencies

Architecture: Body Part I

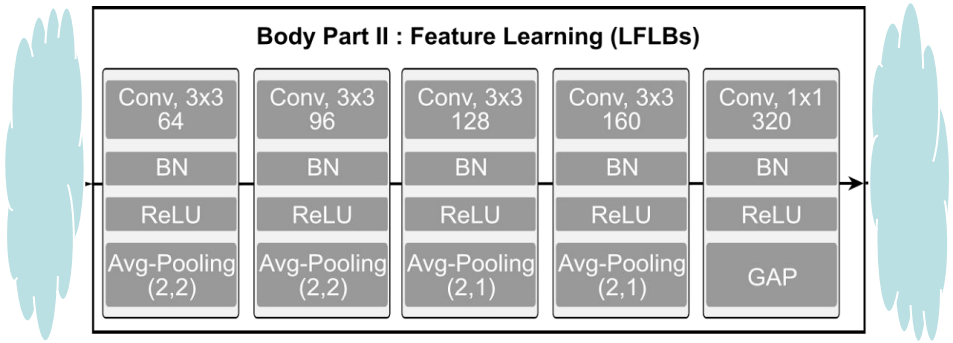
The advantage of using this technique over having only one path with the same receptive field size is to **reduce the number of parameters** and the computational cost of this part of the model



Three different **kernel sizes** for features extraction:

- 3 x 3 for spectral-temporal dependencies
- 9 x 1 for spectral dependencies
- 1 x 11 for temporal dependencies

Architecture: Body Part II

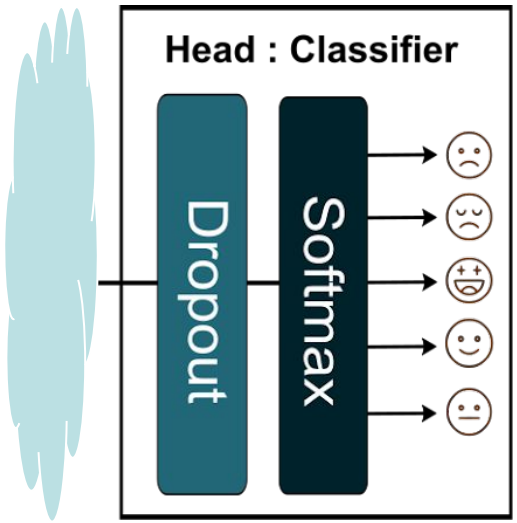


LFLB: collection of **consecutive layers**

LFLB configuration:

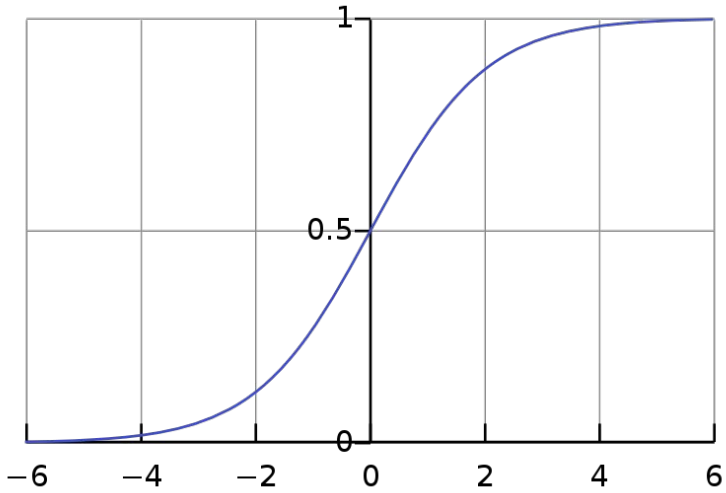
- Convolution Layer
- Batch Normalization Layer (**BN**)
- Exponential Linear Unit (**ELU**) layer
  - Rectified Linear Unit (**ReLU**) layer
- Max Pooling Layer
  - Average Pooling Layer

Architecture: Head



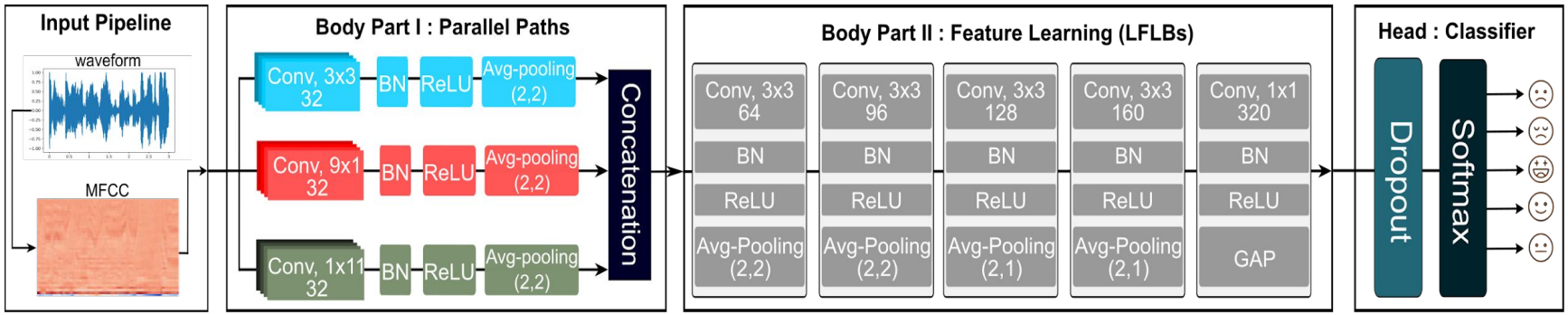
Dropout Layer: to reduce **overfitting**

Softmax Activation Function: reduces computational complexity and the number of parameters



OUTPUT: Class (emotion)

Light-SERNet: Architecture Recap



Light-SERNet: Dataset

EMO-DB is a German-language dataset, recorded by ten professional actors and actresses (five men and five women). The dataset includes 535 emotional utterances in **7 classes**: **anger** (23.7%), **natural** (14.7%), **sadness** (11.5%), **fear** (12.9%), **disgust** (8.6%), **happiness** (13.2%) and **boredom** (15.1%).

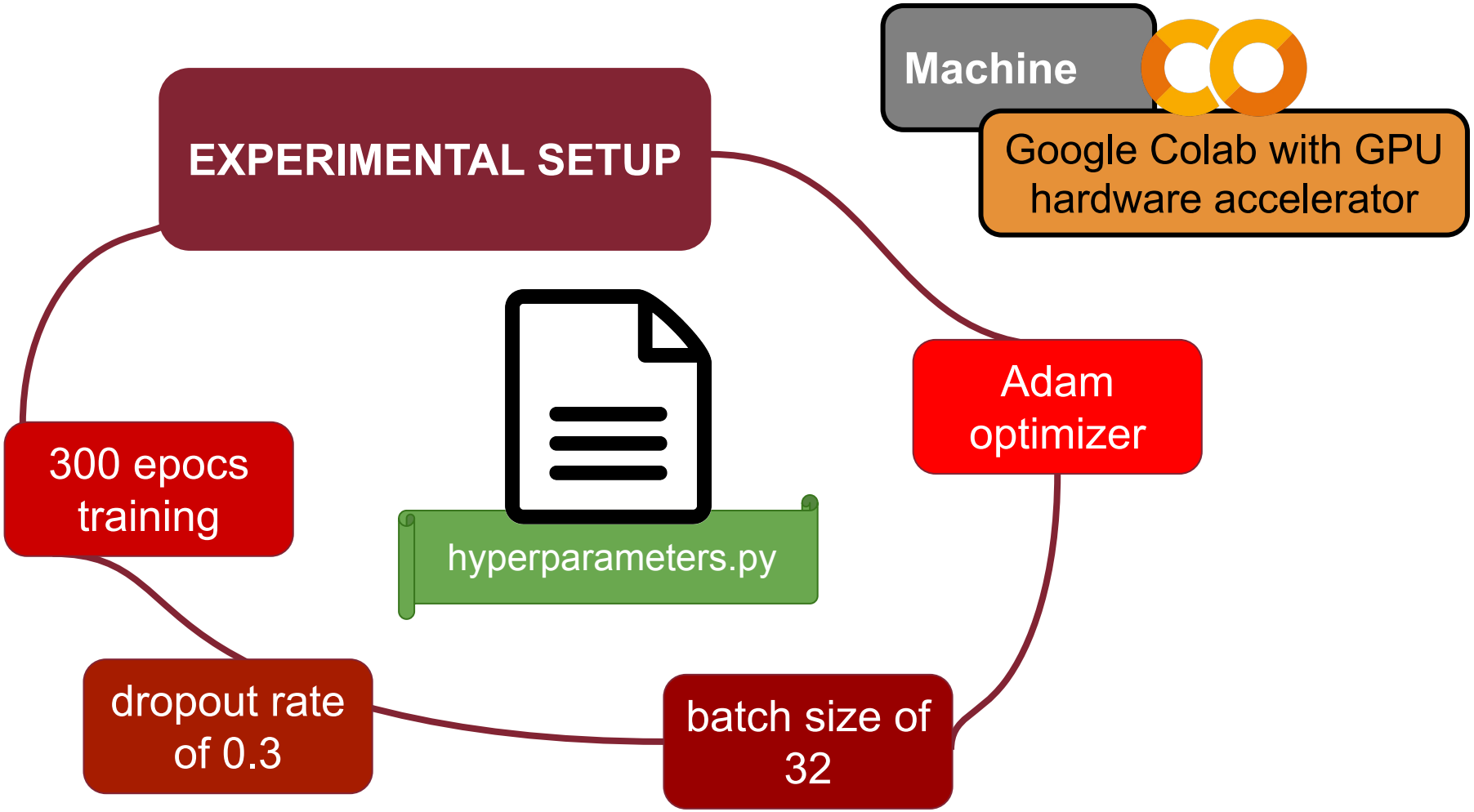
ANGER | **NATURAL** | **SADNESS** | **FEAR**

DISGUST | **HAPPINESS** | **BOREDOM**

Light-SERNet: Dataset

code	text (german)	try of an english translation
a01	Der Lappen liegt auf dem Eisschrank.	The tablecloth is lying on the frigde.
a02	Das will sie am Mittwoch abgeben.	She will hand it in on Wednesday.
a04	Heute abend könnte ich es ihm sagen.	Tonight I could tell him.
a05	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.	The black sheet of paper is located up there besides the piece of timber.
a07	In sieben Stunden wird es soweit sein.	In seven hours it will be.
b01	Was sind denn das für Tüten, die da unter dem Tisch stehen?	What about the bags standing there under the table?
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	They just carried it upstairs and now they are going down again.
b03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	Currently at the weekends I always went home and saw Agnes.
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	I will just discard this and then go for a drink with Karl.
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	It will be in the place where we always store it.

Training and Results Comparison



Training

1

Give the proper INPUT to start the training

-dn
dataset name

-ln
cost function

-id
input durations

-v
verbose for training bar

-at
audio type

-it
type of input

Training

2

Dataset Segmentation

Modeling the input to obtain **same sized** and **normalized** audio files, i.e. files with a **fixed length** and with the **same level** for the entire duration

3

Start the effective training

K-FOLD cross validation with K = 10

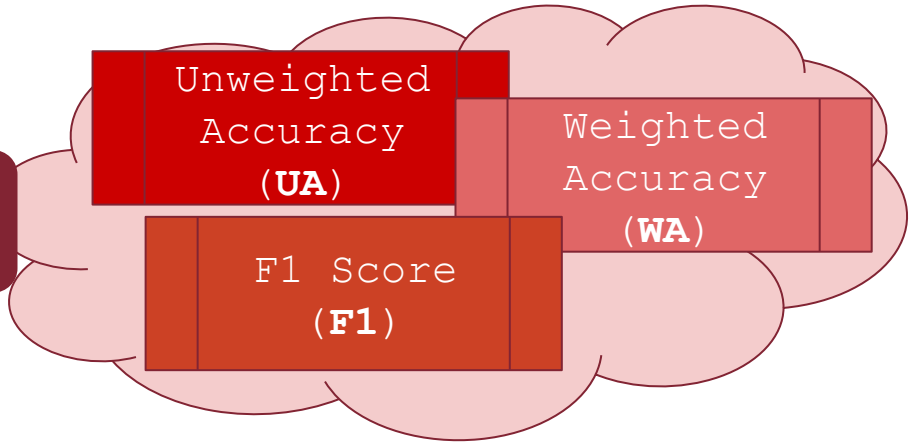
During the train... the best weights are saved

Models with **different degrees of precision** are generated

- Float32
- Float16
- int8

Comparison of Obtained Results

METRICS USED

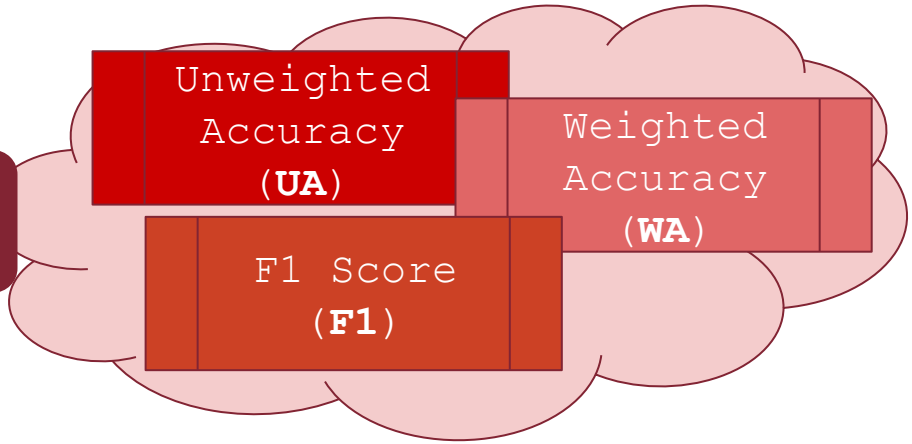


	EMO-DB					
	F-Loss			CE Loss		
	UA	WA	F1	UA	WA	F1
Study results	92.88	93.08	93.05	94.15	94.21	94.16
My results	94.94	95.14	95.10	94.89	95.33	95.27

Consistent results with the study came out from my experiments.

Comparison of Obtained Results

METRICS USED



	EMO-DB					
	F-Loss			CE Loss		
	UA	WA	F1	UA	WA	F1
Study results	92.88	93.08	93.05	94.15	94.21	94.16
My results	94.94	95.14	95.10	94.89	95.33	95.27

In my work the **UA** is higher with the F-Loss.
However the gap is so minimal that can be negligible.

Inference Tests

Model trained with Cross Entropy and saved with Float32 as precision

Used audio files

Audio from EMO-DB dataset
<dataset>

Phrase from dataset, audio recorded by me
<rec>

NOT from dataset, audio recorded by me
<external>

Inference Tests

All the files belonging to the dataset were correctly classified

Inconsistencies in some files recorded by me

WHY?

- Not professional actresses
- Not professional microphone
- background noise

	File Name	Origin	Correct Class	Classified as
1	sadness_external.wav	external	SADNESS	BOREDOM
2	sadness_external_2.wav	external	SADNESS	BOREDOM
3	happiness_external.wav	external	HAPPINESS	HAPPINESS
4	anger_external.wav	external	ANGER	ANGER
5	disgust_rec.wav	recorded	DISGUST	DISGUST
6	boredom_rec.wav	recorded	BOREDOM	BOREDOM
7	anger_rec.wav	recorded	ANGER	HAPPINESS

Conclusions

More and more applications nowadays take advantage of vocal input commands given by the user and to distinguish the emotion of the speaker can completely change the response of the vocal assistant.

Psychological and behavioural studies could be integrated into this processes at marketing level: when an human feels more comfortable when interacting with a machine, the user's satisfaction with using the product increases and so all the involved parts in the interaction get advantage. Users are happy and the company sells more products.

What can be improved is the use of datasets that involve speakers speaking naturally and not actors recorded with professional instruments.

GitHub

References

- Arya Aftab et al. “Light-SERNet: A lightweight fully convolutional neural network for speech emotion recognition”. In: arXiv preprint arXiv:2110.03435 (2021)

- GitHub: PanK0/LIGHT-SERNET. URL: `https://github.com/PanK0/LIGHT-SERNET`