EEG-based Emotion Recognition – Evaluation Methodology Revisited

Sławomir Opałka[0000−0001−9329−544X], Bartłomiej Stasiak[0000−0002−8937−1988], Agnieszka Wosiak[0000−0001−6124−1236], Aleksandra Dura[0000−0003−0428−017X], and Adam Wojciechowski[0000−0003−3786−7225]

Institute of Information Technology, Łódź University of Technology
ul. Wólczańska 215, 93-005 Łódź, Poland
corresponding author: bartlomiej.stasiak@p.lodz.pl

Abstract. The challenge of EEG-based emotion recognition had inspired researchers for years. However, lack of efficient technologies and methods of EEG signal analysis hindered the development of successful solutions in this domain. Recent advancements in deep convolutional neural networks (CNN), facilitating automatic signal feature extraction and classification, brought a hope for more efficient problem solving. Unfortunately, vague and subjective interpretation of emotional states limits effective training of deep models, especially when binary classification is performed basing on datasets with non-bimodal distribution of emotional state ratings. In this work we revisited the methodology of emotion recognition, proposing to use regression instead of classification, along with appropriate result evaluation measures based on mean absolute error (MAE) and mean squared error (MSE). The advantages of the proposed approach are clearly demonstrated on the example of the well-established and explored DEAP dataset.

Keywords: EEG · Emotion recognition · Regression · Classification · CNN

1 Introduction

In recent years, electroencephalography (EEG) became an emerging source of bioelectrical signals providing invaluable information concerning various aspects of human brain activity. Among them, emotional states and their manifestations in the EEG signal still hide many secrets and raise expectations in several research domains: medicine, psychology and human computer interaction.

EEG signal reflects a complex brain activity spectrum, requiring advanced signal processing methods and feature extraction methodologies to be meaningfully interpreted. Although recent advancements in deep learning techniques have revealed a potential for EEG signal classification in several domains [1], emotion analysis still seems to be one of the most challenging ones, especially due to the subjective evaluation process, intrinsic noise and acquisition channels crosstalk [2]. Data acquisition for emotion recognition tasks typically involves

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

eliciting specific emotions in subjects, e.g. by watching video clips, appropriately selected by experts. EEG is recorded during these sessions and emotion self-assessment is usually conducted after each video clip.

The current research on deep learning has shown outstanding results in computer vision and image processing [4]. Due to the nature of the EEG signal on one hand and the fundamental principles and characteristics of deep learning tools and methods on the other hand, it can be expected that these methods will become the mainstream research technique for EEG signal processing in the near future [5]. Especially the deep learning approach for the EEG-based emotion recognition problem seems to leave considerable room for improvement.

The main motivation for our contribution regards the problem of ambiguities inherent to the self-assessment (ground-truth ratings) of the subject's emotional state and their impact on the machine learning strategies applied. We propose a novel approach for emotional state recognition with a convolutional neural network (CNN), which is based on regression rather than on binary classification. Our evaluation methodology puts more emphasis on the self-assessment nuances and it also lets us interpret the results in a more suitable and understandable manner.

2 Previous work

One of the first studies, originating the discussion on emotional state classification and becoming an inspiration for the research community was published by Koelstra et al. [6]. The authors have introduced a Database for Emotion Analysis using Physiological signals (DEAP) which encompasses a set of psychophysiological parameters acquired from users who were watching specially selected music video clips.

DEAP dataset has originated numerous experiments on quantitative classification of emotional dimensions, namely *valence* (quality of emotion from unpleasant to pleasant) and *arousal* (emotion activation level from inactive to active) in accordance with Russell's valence-arousal scale [3], as well as *dominance* (from a helpless and weak feeling to an empowered feeling) and personal impressions encoded in *liking* parameter.

Extensive review of DEAP-based experiments was published by Roy et al. [8]. The authors concluded that the EEG signal suffers from considerable limitations that hinder its effective processing and analysis. Due to low signal-to-noise ratio (SNR), non-stationary characteristics and high inter-subject variability, signal classification becomes a big challenge for real-life applications.

An important source of inspiration for our research was the paper by Craik et al. [9]. The authors have reported a systematic review on EEG-related tasks classification. One of the reviewed aspects concerned input formulation for a CNN-based deep learning solution to the emotional state classification problem, which became a supportive context for the present study. Despite of the fact that most of the authors addressed the problem of signal artifact removal, inherent signal characteristics are hardly addressed. It is difficult to identify persistent

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

noisy channels or noise that is sparsely presented in multi-channel recordings. Manually processed data was highly subjective and rendering it was difficult for other researches to reproduce the procedures. Surprisingly, one of the main findings reported by Craik et al. [9] was that – according to the authors' best knowledge – there were no studies demonstrating that deep learning can achieve results comparable with classification methods based on handcrafted features. Especially convolutional neural networks, although popular in image processing domain, can hardly be found in the domain of EEG-based emotional state recognition. In fact, just a few authors examined CNN architecture with frequency domain EEG spectrograms prepared as an input [10–13]. They have concentrated on motor impairment [10], mental workload [11, 13] and sleep stage scoring [12], rather than on emotional state analysis/classification.

There were also several neural architectures containing convolutional layers, which were employed for examining the DEAP dataset and which might therefore be treated as a meaningful reference. In their both works Li et al. [14, 15] proposed hybrid neural architectures interpreting wavelet-derived features, with the CNN output connected to LSTM RNN modules, but they achieved low classification accuracy. On the other hand, noticeable difference in classification accuracy could be attributed to alternative input formulation. Yanagimoto et al. [16] directly used signal values as inputs into a neural network, while Qiao et al. [17] and Salama et al. [18] converted the input data into Fourier maps and 3D grids respectively, considerably improving the classification accuracy. Nevertheless, CNN application for processing spectrogram-based EEG data in the frequency domain seems to be a quite novel strategy of outstanding and unexplored potential.

In the context of valence-arousal dimensions and deep learning-based classification, Lin et al. [19] proposed an interesting multi-modal approach for the DEAP dataset, achieving for bi-partitioned classification of valence and arousal 85.5% and 87.3% respectively. It must be noted, however, that the obtained accuracy is greater than many other DEAP-based experiments, mainly due to considering all the available physiological signals rather than just EEG.

Frequency-domain representation of the EEG signal, collected with numerous electrodes (32 channels in the case of DEAP) and split into frequency subbands, puts high demands on the neural network model, which must deal with this highly-dimensional input data in conditions of the limited number of subjects, stimuli, recording time, etc. Feature extraction is therefore an important processing step applied by most authors to reduce the complexity of the neural network and to let it learn effectively with low generalization error. It should be noted that data dimensionality reduction may also be obtained by appropriate selection of the EEG electrodes, e.g. on the basis of channel cross-correlation.

An extensive and detailed analysis of the most suitable approach for EEG signal feature extraction was provided by Nakisa et al. [20]. The authors considered different time-domain, frequency-domain and time-frequency domain features in the context of the selected datasets [6, 7]. Four-class (High/Low Valence/Arousal) emotion classification process, based on a probability neural net-

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

work (PNN) and 30 features selected with 5 alternative evolutionary computation algorithms, provided average accuracy reaching 65% within 100 iterations. Supportive conclusion regarded DEAP-oriented EEG channels selection. The authors noticed that FP1, F7, FC5, AF4, CP6, PO4, O2, T7 and T8 were the most relevant electrodes in the context of emotion recognition. The above findings and the current progress in deep learning-based feature extraction and classification [21] led to an observation that automatically selected features from a limited number of channels (9 for DEAP) may result in more spectacular results, particularly given that traditional classifiers like SVM-related ones reach 78% for bi-partitioned valence and arousal classes [22] on 16 frequency and time domain features and respectively 63% and 58% for 3-partitioned valence and arousal classes [23] on 11 frequency and time domain features. High accuracy of non-deep learning approaches was also reported by other authors [24].

Inherent emotional dimensions in the DEAP dataset were subjectively quantified by subjects, who assigned numerical ratings between 1 and 9. However, most of the authors aggregated emotional dimensions into two groups: high (greater or equal 5) and low (lower than 5) within each dimension respectively. In consequence, the mean values of valence and arousal within four quadrants on the Low-High/Arousal-Valence plane were as follows: Low-Arousal-Low-Valence (LALV: 4.3 $\pm$1.1; 4.2 $\pm$0.9), High-Arousal-Low-Valence (HALV: 5.7 $\pm$1.5; 3.7 $\pm$1.0), Low-Arousal-High-Valence (LAHV: 4.7 $\pm$1.0; 6.6 $\pm$0.8), High-Arousal-High-Valence (HAHV: 5.9 $\pm$0.9; 6.6 $\pm$0.6). It must be noted that these means were relatively close to each other (although statistically different), as compared to the whole range of the possible values. Additionally, considering the fact that the authors of DEAP selected only 40 video clips – out of initial 120 – namely the most extreme ones in the context of valence/arousal, unambiguous quantitative differentiating of emotional dimensions becomes a real challenge requiring in-depth analysis.

The above findings, addressing the ambiguity in emotional state classification, suggest revisiting deep learning-based approaches. Specifically, the *quantization* of the emotional dimensions and reformulating the problem from binary (low-high) classification into the *regression* task will be covered in the following part of this paper.

3 Materials and Methods

The DEAP database [6], introduced in the previous section, was published by Queen Mary University of London for the purpose of emotion analysis on the basis of physiological signals, including EEG. The data consists of recordings of 32 participants who were watching music videos. Forty 1-minute long videos were carefully selected to induce emotions falling into 4 general categories: LAHV, LALV, HAHV and HALV, as described above. The recorded data comprises 32 EEG channels conforming to the international 10-20 system and 8 additional channels representing various physiological signals, including EOG (horizontal and vertical eye movements), EMG (activity of zygomaticus major and trapez-

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

ius muscles), GSR, respiration, blood volume pressure and temperature. In our research only the EEG channels were used.

Every participant rated every video he/she had watched, in terms of four distinct emotional qualities: valence, arousal, dominance and liking. Each rating was expressed as a real number from the range $[1, 9]$. The two first qualities (valence, arousal) defined our research goal: to predict the participant's rating on the basis of the EEG signal recording.

3.1 EEG data preprocessing

The EEG signal recorded from a participant watching one film comprises 8064 time-domain samples per electrode, which corresponds to 63 seconds at sampling frequency of 128 Hz. Power spectral density is computed for each frame of size 128 samples (one second, *von Hann* window applied) with 50% overlap (hop-size: 64 samples). This yields a spectrogram with 125 time points (frames) and frequency resolution of 1 Hz. The frequencies below 4 Hz and above 45 Hz are rejected (in fact, they are already filtered out in the originally preprocessed DEAP dataset). The spectrogram values are then scaled logarithmically so that they fit within $[0, 1]$ range. The first 5 frames are rejected, and the spectrogram is rescaled along the frequency axis (antialiasing filter applied) to the final shape of 20×120 (frequency $\times$ time). An example is presented in Fig. 1.

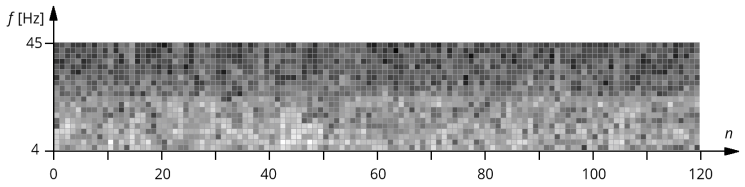

Fig. 1. Spectrogram of the first EEG channel (Fp1 electrode) recorded for the first participant watching the first video (n denotes the frame index)

Each spectrogram is cut into chunks of 10 frames each (which corresponds to ca 5 seconds) with overlap of 50%. This yields 23 chunks for a single spectrogram. The corresponding chunks coming from all the 32 electrodes (i.e. the chunks representing the same time range within all 32 spectrograms) are grouped together, forming the single *object* (input tensor) to be recognized by the network. Every recording of a single film watched by a single user is therefore represented by separate 23 fragments (input tensors) of size $32 \times 20 \times 10$ (electrodes $\times$ frequency bands $\times$ time frames). All these 23 fragments have the same target value (defining our training goal) which is simply equal to the participant's rating of the film under consideration. Each tensor is an individual input object for a convolutional neural network described in the next section.

We decided to apply a 4-fold crossvalidation scheme with fixed division into the training, testing and validation subsets. Every experiment was based on the

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

EEG data from 40 films of *a single participant only*. It was therefore repeated for all 32 participants individually (and for each of the 4 folds) and averaged results are reported (Sect. 4).

Considering the data from a single participant, 10 films were included in the testing set, another 10 films – in the validation set and the remaining 20 films were used for training. Complete films were always used, i.e. we did not mix fragments (input tensors) from different films. In Fig. 2 the assignments of individual films to particular subsets (in each of the 4 folds) are shown in the bottom 4 rows. `Ts` denotes the testing set, `Vd` – the validation set and the blank fields indicate the films used for training. The reason for these particular assignments

Film idx	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Valence	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
Arousal	H	H	H	H	H	H	H	H	H	H	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	H	H	H	H	H	H	H	H	H	H
Fold 1	Ts	Ts	Ts	Ts	Ts						Vd	Vd	Vd	Vd	Vd						Ts	Ts	Ts	Ts	Ts						Vd	Vd	Vd	Vd	Vd					
Fold 2						Ts	Ts	Ts	Ts	Ts						Vd	Vd	Vd	Vd	Vd						Ts	Ts	Ts	Ts	Ts						Vd	Vd	Vd	Vd	Vd
Fold 3	Vd	Vd	Vd	Vd	Vd						Ts	Ts	Ts	Ts	Ts						Vd	Vd	Vd	Vd	Vd						Ts	Ts	Ts	Ts	Ts					
Fold 4						Vd	Vd	Vd	Vd	Vd						Ts	Ts	Ts	Ts	Ts						Vd	Vd	Vd	Vd	Vd						Ts	Ts	Ts	Ts	Ts

Fig. 2. Film-to-subset assignments in the individual folds

is explained in the upper part of the table in Fig. 2. Most of the first 20 films have on average higher valence ratings (`H`) than the last 20 films (`L`). As for the arousal, the first and the last 10 films tend to be rated higher than the middle 20. Therefore, the chosen assignment yields more balanced testing set (and also the validation and the training ones), containing 5 `H`'s and 5 `L`'s both for valence and arousal, irrespective of the fold number (although the ratings of individual participants may occasionally deviate from this simple `H`/`L` distinction).

3.2 Convolutional neural network model

Having analyzed the extensive review presented by Roy et al. [8], we decided to use a simple, yet effective CNN architecture shown in Fig. 3.

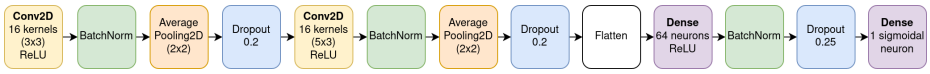

Fig. 3. Applied model architecture

The architecture is implemented as `Sequential` model with *Keras* interface for *TensorFlow* library. Apart from the dropout layers, there are some additional elements (not shown) aimed at generalization properties enhancement: `L2` kernel regularizer (regularization coefficient: 0.01) in both convolutional layers and a GaussianNoise layer ($\sigma = 1.5$) applied before the first convolutional layer.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

4 Experiment objectives and design

Polarity inference is the most fundamental task in many emotion recognition or sentiment analysis problems. *Does she like me or not? Are they interested or bored?* We often tend to ignore the possible shades of gray between the extremes. The DEAP dataset construction principles seem to support this view, provided that the 40 films had been deliberately selected (out of the initial collection of 120 videos) to maximize the strength of the elicited emotions. More precisely, the database contains these films, which happened to lie closest to the 4 extreme corners in the valence-arousal 2D space, as rated by at least 14 volunteers per film in a preliminary step of video material selection [6].

Using the collected videos in the actual experiments, targeted at emotion recognition from the physiological signals, follows naturally the same principle. A typical approach found in most research works, including also the original paper by Koelstra et al. [6], aims at *classification* of the videos in two classes: low and high, with respect to any of the four aforementioned emotional qualities. For example, if the valence rating of a film exceeds 5, the film is automatically included in the "high valence" class ("low valence" in the opposite case). This approach seems natural, straightforward and valid. However, taking into account the actual data collected from the subjects participating in DEAP database construction, it is probably overly simplistic, as we demonstrate below.

The first thing to consider is the histogram of the participants' ratings (Fig. 4). Its unusual shape results from the fact, that the participants usually tried to give

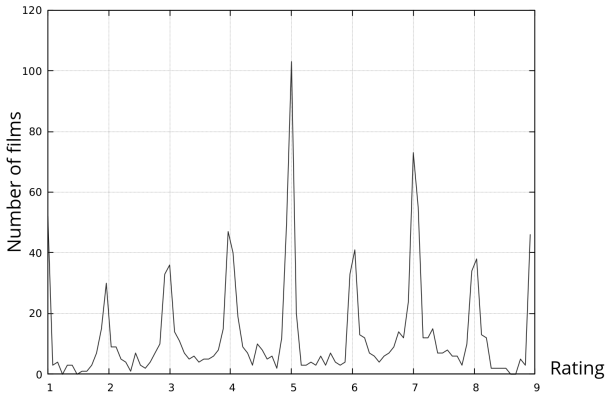

Fig. 4. Histogram of *valence* ratings

integer scores, although the available input method was based on a continuous scale (ranging from 1.0 to 9.0, as mentioned before). Now we can clearly see the problem of "neutral responses": over 100 ratings were very close to 5 or – if we take a slightly broader tolerance – 221 ratings fell within the range $[4.5, 5.5]$. Considering the total number of ratings ($1280 = 40$ films evaluated by 32 users),

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

this accounts for over 17% ratings which were probably intended to mean "no opinion". Splitting the data with a hard threshold of 5 into positive/negative valence ratings will inevitably lead to significant confusion, irrespective of the particular machine learning or classification approach.

Moreover, assigning the same class ("low") for films labelled "1" and "4", while separating the "4"s ("low") from the "6"s ("high") seems also arguable. Although the emotional valence rating is highly subjective and it probably does not follow any simple linear scale or distance measure suggested by Fig. 4, predicting the actual *rating* instead of the "low-high" quantization seems much more appropriate.

Finally, we have to take into account that the participant's rating is the result of an intrinsic decision-making process based not only on purely emotional reactions but on many other premises as well. They may include prior knowledge and personal attitude towards the video content, the general worldview, the social, political and cultural background and – last but not least – the comparison with the previously watched videos and the ratings given. These factors may easily change the final rating of the current video within certain limits, independently on the actual emotions deducible from the recording of the physiological signals. This change may be relatively small in terms of sheer numbers, but in some cases it may easily shift the film from the "low" to the "high" class, or *vice versa*.

This, again, supports the claim that *prediction of the participant's rating* (i.e. *regression*) should be the preferred approach to the analysis of the DEAP dataset (and other collections based on similar data acquisition principles). It should be noted that increasing the number of classes, as an alternative to regression, would probably not be as effective, because we could not directly represent the relations between consecutive classes on the ordinal scale in our machine learning approach (and in the evaluation procedures).

4.1 Experimental validation

In a single experiment (for a single participant), the training set for the convolutional neural network described in Sect. 3.2 included 460 input objects (20 films $\times$ 23 input tensors), according to Sect. 3.1. For each dimension of the input tensor, the mean μ_{train} and standard deviation σ_{train} within the training set were computed and used for data normalization of all the three sets: training (Tr), testing (Ts) and validation (Vd):

$$Tr_{norm} = \frac{Tr - \mu_{train}}{\sigma_{train}} \tag{1}$$

$$Ts_{norm} = \frac{Ts - \mu_{train}}{\sigma_{train}} \tag{2}$$

$$Vd_{norm} = \frac{Vd - \mu_{train}}{\sigma_{train}} \tag{3}$$

The goal of the supervised training was to obtain proper regression, i.e. to minimize the mean squared error (MSE) between the output of the last layer (a single neuron with a unipolar sigmoidal activation function, Sect. 3.2) and the target participant's rating value (ground-truth). The target value was divided by

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

10.0 to make it fit within the range [0.1, 0.9]. The network was trained with Adam optimizer [8]. After numerous preliminary experiments, the maximum number of epochs and the batch size were set to 600 and 100, respectively. The validation dataset (Sect. 3.1) was used to select the best model in terms of validation MSE minimization.

Apart from this, two additional experiments were done. In one of them ("regression_opt"), no validation set was used (or, more precisely, it was merged with the training set for the total number of 30 training films) and the optimal model was selected on the basis of the MSE value for the testing set. In this case, the obtained results may be interpreted as the theoretical maximum that might be reached, provided that the optimal stopping criterion is known beforehand. The other experiment is based on high/low classification instead of regression, so it is basically a reference for comparison of the results. In this case, all the participant's ratings and the network outputs are thresholded (below $0.5 \rightarrow 0$; greater or equal $0.5 \rightarrow 1$). It is worth to note, however, that this binary representation of the targets is used for the training process only. The CNN, once trained, is tested and evaluated in the same way as in the case of the regression-based experiments, as described in the following section.

4.2 Evaluation metrics

Following the non-binary formulation of the training target, also the evaluation methods, used for the analysis of the testing set results, should be defined in a more "fine-grained" way. Mean square error (MSE) and mean absolute error (MAE) between the CNN outputs and the participant's ratings seem to be reasonable measures, telling us how big the discrepancy is on average. We also used a standard binary classification metric (CLS) based on thresholding both the outputs and targets with the fixed threshold of 0.5 and simply counting the objects for which the match occurred.

These three measures (MSE, MAE and CLS) were computed in two ways: for individual input tensors representing the film fragments (23 independent results for a single film) and for the whole films. In the latter case, the arithmetic mean of the outputs obtained for all the 23 input objects representing the same film was treated as the final prediction and compared with the target.

As an additional form of result presentation, we computed the percentage of films within a given range of MAE values (with respect to the true participant's rating). This computation was done on the whole-film basis, as defined above.

Apart from the objective evaluation measures, individual results were also carefully inspected and verified manually. This allowed us to detect a significant problem, responsible for degradation of the results in many cases. It occurred especially for these users and folds for which most of the ratings in the training set were close to the middle 5. In such cases the network tended to get stuck in the local minimum, producing non-diversified results, also close to 5 (Fig. 5, left). This problem may be viewed as a direct consequence of resigning from the "hard" binary classification approach which had forced the network to decide on either the high or low value at the output.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

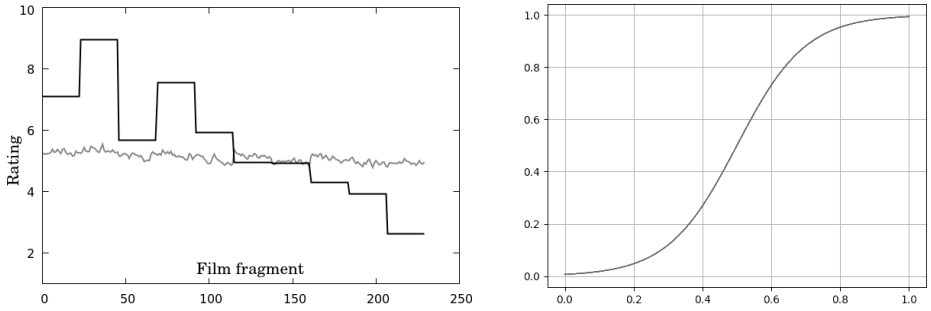

Fig. 5. Left: participant's ratings (black) and network output (grey) for all 230 fragments from the test set (user 10, fold 2; each sequence of 23 consecutive fragments represents one film); **Right**: the sigmoidal function used to transform the target values in order to force the network to produce more diversified output values

As a remedy, a special scaling of the targets in the training set, with a sigmoidal function, was applied (Fig. 5, right). This "soft alternative" to the binary classification drew the target values more to the extremes, encouraging the network to leave the "mid-level comfort zone" during training, while preserving the relative order of the rating values. The result for the same user/fold pair is presented in Fig. 6. Naturally, it should be remembered that training the network

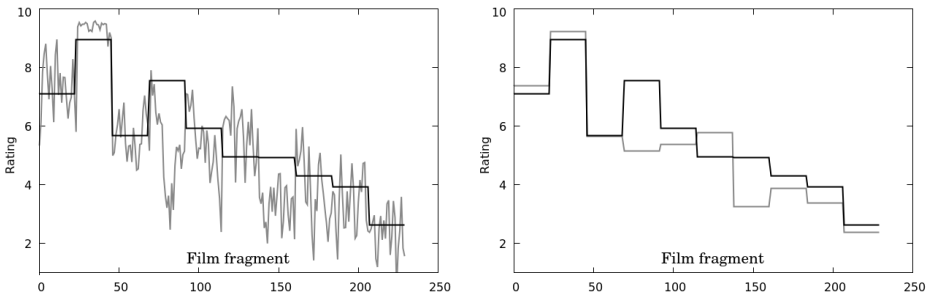

Fig. 6. Left: training result for the same dataset as in Fig. 5, but with the targets transformed with the sigmoidal function; **Right**: the same result but with the mean output ratings per film

with target values transformed with a sigmoidal function requires that in the testing phase its outputs are transformed with the inverse function, before any comparison or evaluation is performed.

4.3 Results

All the training sessions were independently done for the valence and for the arousal ratings. In every experiment, the training was repeated three times and

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

the mean values of the evaluation metrics are reported in Table 1 (valence) and Table 2 (arousal).

Table 1. Results (valence)

	Film fragments			Whole films		
	MSE	MAE	CLS	MSE	MAE	CLS
Classification	0.128	0.298	58.9%	0.096	0.252	60.0%
Regression	0.052	0.181	59.2%	0.046	0.172	60.2%
Regression_opt	0.038	0.153	67.5%	0.032	0.143	71.4%

Table 2. Results (arousal)

	Film fragments			Whole films		
	MSE	MAE	CLS	MSE	MAE	CLS
Classification	0.134	0.307	59.6%	0.106	0.268	60.9%
Regression	0.050	0.173	60.4%	0.046	0.165	62.2%
Regression_opt	0.035	0.145	68.2%	0.031	0.137	70.3%

Figures 7 and 8 reveal how many films were rated sufficiently close to the ground-truth in terms of MAE. For example, the last column in the last group of Figure 7 tells us that the absolute difference between the CNN output and the participant's rating was less than 4.5 in 98.8% of all the films. It is worth noting that this result (and all other presented results) is averaged over individual results from 384 training sessions (3 repetitions × 32 participants × 4 folds) and that the CNN output is in fact the mean of 23 outputs for 23 different fragments of the same film.

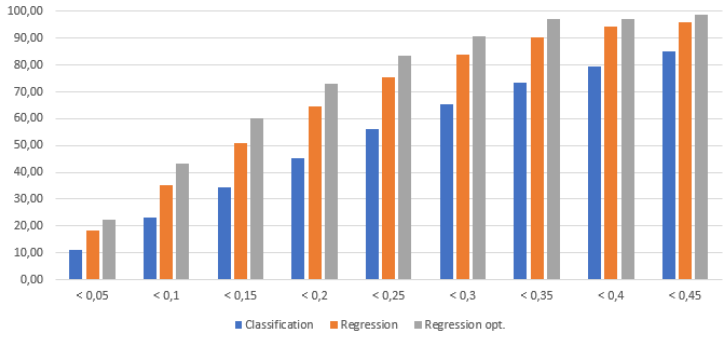

Fig. 7. Percentage of films within a given MAE range (valence)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

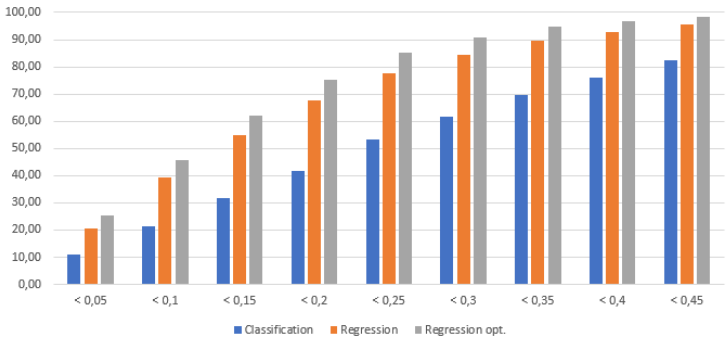

Fig. 8. Percentage of films within a given MAE range (arousal)

4.4 Discussion

The obtained results clearly show the advantages of the proposed approach. Although the binary classification accuracy (CLS) is only slightly better for the regression-based model than for the classification-based one (60.2% vs 60% and 62.2% vs 60.9% for the valence and arousal, respectively), the MSE and MAE values are definitely smaller (roughly two times smaller). In fact, this outcome is understandable when we take into consideration that the thresholding, used for producing the binary class labels, inevitably leads to discarding all the nuances present in the participants' ratings. Nevertheless, the practical usefulness of the obtained small MSE and MAE values seems quite clear, when we note, for example, that over two-thirds (three-fourths) of the films will get the arousal rating prediction within ± 2 (± 2.5) from the ground-truth, if we apply our regression-based CNN training approach.

Considering the binary classification accuracy itself, it has to be agreed upon that the obtained result, slightly exceeding 60% (or 70% for the optimally determined early stopping criterion), is not very impressive. One potential reason for that is the relatively small training set, especially when we consider the huge dimensionality of the input space.

Increasing the number of training examples may be obtained in several ways, e.g. by including the data from other participants or by increasing the number of spectrogram chunks, either by shortening them or by increasing the overlap. The first option (training on the data coming from many participants) would also be the most general and useful one. However, the heterogeneity of the EEG characteristics among the participants, poses significant problems in obtaining good generalization properties of the CNN models. The second option (generating more training objects from the EEG signal) is related to the question of the optimal range and resolution of the input data both in terms of the frequency content (e.g. how many bands and what frequency range should be analyzed) and the temporal characteristics (e.g. duration of the analyzed film fragments).

Instead of increasing the training set size, we may also search for dimensionality reduction. The EEG signals coming from adjacent electrodes are usually

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

quite significantly correlated, and some EEG channels may be more useful in emotion analysis that the others. Similarly, some frequency ranges might probably be excluded from the input data or, at least, represented with decreased resolution. These are just a few examples of the research directions that will be considered in our future work.

5 Conclusion

In the presented work, we compared two Russell's emotional state evaluation methodologies in the task of valence/arousal prediction on the basis of the EEG signal. We confronted state-of-the-art binary classification with our regression-based approach. Our motivation was supported by the detailed analysis of the representative DEAP dataset, highlighting the pitfalls and difficulties resulting from simple high/low label assignment. We also proposed new evaluation metrics (MSE/MAE) conforming to the reformulated emotion recognition task. Subject-oriented experimental evaluation of the proposed methodology, based on a convolutional neural network trained on EEG signal spectrograms, revealed the improvement in the obtained results, both in terms of the new metrics and binary classification accuracy. The CNN trained to perform the regression task yielded much higher target rating prediction rates (with respect to the binary classification), with the difference reaching 26.2 percentage points (arousal) and 19.5 percentage points (valence), for MAE tolerance range of ± 0.2.

Future works will concentrate on analysis of automatic vs handcrafted EEG signal feature selection, including both EEG channels and frequency range selection. We will also investigate potential solutions for improvement of the generalization properties of the proposed CNN model.

As a final conclusion, we encourage the research community to revise the evaluation methodology used in emotional state recognition tasks and to consider regression as more appropriately reflecting the subjective nature of emotional state ratings reported by the users.

References

1. Opałka, S.; Stasiak, B.; Szajerman, D.; Wojciechowski, A. Multi-Channel Convolutional Neural Networks Architecture Feeding for Effective EEG Mental Tasks Classification. Sensors 2018, 18, 3451.
2. Teplan M., Fundamentals of EEG measurement, Meas. Sci. 2, 1–11, 2002.
3. Russell J. A., A circumplex model of affect, Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178, 1980.
4. Lazarek J., Pryczek M., A Review of Point Cloud Semantic Segmentation Methods, Journal of Applied Computer Science 26(2), 99-105, 2018.
5. Li, G., Lee, C. H., Jung, J. J., Youn, Y. C., Camacho, D., Deep learning for EEG data analytics: A survey. Concurrency and Computation, 32(18), 2020:e5199.
6. Koelstra S., Muehl C., Soleymani M., Lee J.-S., Yazdani A., Ebrahimi T., Pun T., Nijholt A., Patras I. DEAP: A Database for Emotion Analysis using Physiological Signals. IEEE Trans. Affect. Comput. 2012;3:18–31. doi: 10.1109/T-AFFC.2011.15.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40

7. M. Soleymani, J. Lichtenauer, T. Pun and M. Pantic, A Multimodal Database for Affect Recognition and Implicit Tagging, in IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 42-55, Jan.-March 2012, doi: 10.1109/T-AFFC.2011.25.
8. Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., Faubert, J., Deep learning-based electroencephalography analysis: a systematic review. Journal of neural engineering, 16(5), 2019.
9. Craik, A., He, Y., Contreras-Vidal, J. L., Deep learning for electroencephalogram (EEG) classification tasks: a review. J. of neural engineering, 16(3), 031001, 2019.
10. Vrbancic G., Podgorelec V., Automatic classification of motor impairment neural disorders from EEG signals sing deep convolutional neural networks, Elektron. Elektrotech. 24, 3–7, 2018.
11. Kuanar S., Athitsos V., Pradhan N., Mishra A. and Rao K. R., Cognitive analysis of working memory load from EEG, by a deep recurrent neural network, 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp 352–5, 2018.
12. Vilamala A., Madsen K. H. and Hansen L. K., Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring, 2017 IEEE 27th Int. Workshop on Machine Learning for Signal Processing (MLSP), 2017.
13. Jiao Z., Gao X., Wang Y., Li J. and Xu H., Deep Convolutional Neural Networks for mental load classification based on EEG data, Pattern Recog. 76, 582–95, 2018.
14. Li X., Song D., Zhang P., Yu G., Hou Y. and Hu B., Emotion recognition from multi-channel EEG data through convolutional recurrent neural network, 2016 IEEE Int. Conf. Bioinformatics Biomedicine, pp 352–9, 2016.
15. Li Y., Huang J., Zhou H. and Zhong N., Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks, Appl. Sci. 7, 1060, 2017.
16. Yanagimoto M. and Sugimoto C., Recognition of persisting emotional valence from EEG using convolutional neural networks, 2016 IEEE 9th Int. Workshop Computational Intelligence Applications, pp 27–32, 2016.
17. Qiao R., Qing C., Zhang T., Xing X. and Xu X., A novel deep-learning based framework for multi-subject emotion recognition, Proc. of ICCSS, 181–5, 2017.
18. Salama E. S., El-khoribi R. A., Shoman M. E. and Shalaby M. A., EEG-based emotion recognition using 3D convolutional neural networks, Int. J. Adv. Comput. Sci. Appl. 9, pp 329–37, 2018.
19. Lin, W., Li, C., Sun, S., Deep Convolutional Neural Network for Emotion Recognition Using EEG and Peripheral Physiological Signal. In International Conference on Image and Graphics, pp. 385–394, 2017.
20. Nakisa, B., Rastgoo, M. N., Tjondronegoro, D., Chandran, V., Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors. Expert Systems with Applications, 93, 143-155, 2018.
21. Asghar, M.A., Khan, M.J., Fawad, Amin, Y., Rizwan, M., Rahman, M., Badnava, S., Mirjavadi, S.S., EEG-Based Multi-Modal Emotion Recognition using Bag of Deep Features: An Optimal Feature Selection Approach. Sensors 19(23), 5218, 2019.
22. Yin, Z., Wang, Y., Liu, L., Zhang, W., Zhang, J. Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination. Frontiers in Neurorobotics, 11, 2017.
23. Menezes, M.L.R., Samara, A., Galway, L., Sant'Anna, A., Verikas, A., Alonso-Fernandez, F., Wang, H., Bond, R. Towards emotion recognition for virtual environments: An evaluation of EEG features on benchmark dataset. Pers.&Ubiq. Comp., 1–11, 2017.
24. Wang X-W, Nie D, Lu B-L. Emotional state classification from EEG data using machine learning approach. Neurocomputing 129:94-106, 2014.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_40