The Effect of Different Stimulus on Emotion Estimation with Deep Learning

Süleyman Serhan Narlı[1] , Yaşar Daşdemir[1], Serdar Yıldırım [2]

[1]İskenderun Technical University, Engineering and Natural Sciences Faculty, Computer Engineering, Hatay-TURKEY

suleyman.narli.mfbe18@iste.edu.tr

yasar.dasdemir@iste.edu.tr

[2]Adana Science and Technology University, Engineering Faculty, Computer Engineering, Adana-TURKEY

syildirim@adanabtu.edu.tr

Abstract

Emotion recognition system from brain signals is important for human-computer and human-machine interaction. There are many different methods to extract important attributes from these signals. Some of these methods are classification by machine learning, classification by artificial neural networks and classification by deep neural networks. In this study, the effects of these 3 different stimuli (Audio, Video, Audio and Video) on the effect on emotion estimation were investigated.

For this study, 1125 samples (375 Audio, 375 Video, 375 Audio-Video) EEG signals were recorded from 14 channels with a sampling frequency of 128 Hz. In this experiment, signals indicating brain activation for 10 different emotions were recorded by considering the emotion evaluations of the participants. The participants were primarily played back emotionally with audio recordings, then the video was played without audio and finally, video and audio were watched together.

In the preprocessing step MARA method and Independent component analysis (ICA) were used to eliminate the artifact in the signals. A specific model was created by using deep learning for feature extraction.

The effect of different stimuli on performance was investigated for emotion recognition.

Keywords: EEG, Deep Learning, Audio stimuli, Video stimuli, Audio-Video stimuli, Convulutional Neural Network.

Introduction:

Emotions are a physiological mechanism that affects our decisions in our daily lives and directs us. For this, it is important to investigate this area, there are many emotion estimation algorithms.

There are many studies in this area by using machine learning (artificial intelligence) by processing the signals obtained from the brain.

There are limited number of researches and articles that make estimation of emotions with convolution process, artificial neural networks.

In addition, it is the data obtained from the outside world which affects emotions.

What kind of stimulation is our feelings more affected?

Which stimulant affect our emotions is more effective in predicting emotion when our brain perceives events?

In this study, the effects of different stimulus types on emotion recognition will be investigated.

Related Work:

The success of convolutional neural networks on EEG signals depends on the data set(Lin, Li, and Sun 2017). The success of convective neural networks on EEG signals depends on the data set.

Convolutional neural networks are frequently used in image processing (Bhattarai et al. 2017), there is not much study on the processing of EEG signals by this method. To process the EEG signals using the CNN (Convolutional Neural Network) method, it is correct to change the dimensional setting of the data set (Zeng et al. 2018), for this we need to convert the data set to image format and apply a convolution process (Zeng et al. 2018), there is a 3-layer RGB plane in image format and the structure of EEG signals it is similar to this structure on the basis of frequency because the number of time, frequency, and number of channels during the recording of EEG signals can be considered as a 3D plane (Acharya et al. 2018).

The way the data set is created in the emotion recognition system is very important,

The fact that the emotion stimulus type is an audio, video (Mohammadi, Frounchi, and Amiri 2017) or Audio-Video produces different results. There are many different used data sets(Liu and Sourina n.d.) and the most commonly used is the Deap dataset. In the content of the data set used in this study, the emotions affected by the different types of stimuli were recorded (Dasdemir, Yildirim, and Yildirim 2017) that is, only data, only video and audio-video are recorded together.

1- Dataset description:

The data set used in the study was recorded with Emotiv EPOC (Emotiv Systems Inc., San Francisco, USA). [1]

EEG signals from 25 volunteers were recorded over 14 channels for 60 seconds, with a total of 15 different clips for different emotion excitations, which were played in 3 different formats; Audio, Video and Audio-Video. As a result, 45 different stimuli were used for each subject. The data were recorded at 128 frequencies.(Table 1)

Table 1. List of Stimulus

Stimuli Type	Number of Volunteers	Number of Clips	Number of Records	Total Records: 1125
Audio	25	15	375	
Video	25	15	375	
Audio - Video	25	15	375	

Values for the Valence - Arousal - Dominance axes were calculated for the recorded data.

In this study, Valence> 0.5 for 1, Valence <0.5 for 0, was used.

2- Data Preprocessing:

During recording with the EPOC Emotiv device the raw data may be recorded loudly (blinking, muscle movements, eye movements, etc.)

With independent component analysis, some of these noises can be destroyed.

In this study, MARA (Multiple Artifact Rejection Algorithm), an open source EEGLAB attachment (Winkler et al. 2011), was used for automatic artifact rejection using ICA.

MARA is a supervised machine learning algorithm that solves a binary classification problem: It works as "accept or reject" the stand-alone component.

3- Methodology:

3.1. Editing the Data Set:

The data set used represents 1125 samples in total (Table 2).

Table 2. Sample list

Total number of clips	Number of Volunteers	Total Number of Channels	Sampling Sate	Total recorded time	Raw data length
45	25	14	128	60	1125 x 14 x 7680

For the 1125 samples in the data set; 375 x 14 x 7680 Audio, 375 x 14 x 7680 Video and 375 x 14 x 7680 Audio-Video.

The data set was re-arranged for the learning model to be used and the new data set was defined as 3D; 1 x 14 x 120 x 64, thus, the new data set was made ready for processing.

3.2. Convolutional Neural Network (ConvNet):

A simple ConvNet is a layer array, and each layer of a ConvNet converts one volume activation to another with a differentiable function. We use three main layer types to create ConvNet architectures: Convex Layer, Pool Layer, and Fully Linked Layer (exactly as seen in Normal Neural Networks).

The parameters of the CONV layer consist of a series of learnable filters. Each filter is small (along width and height), but extends across the full depth of the inlet volume.

3.2.1. Convolutional Layer:

The Conv layer is the basic building block of a Convolution Network that performs most of the calculation heavy lifting operations.

Convolution is the first layer that extracts properties from an input image. Convolution maintains the relationship between pixels by learning image properties using small data input frames.

Image matrix is a mathematical process that takes two inputs as filter or kernal.

The convolution of an image with different filters can be applied by applying filters such as edge detection, blur and sharpening.

3.2.2. Non-Linear (ReLU) :

ReLU stands for Straightened Linear Unit for a nonlinear operation. Output) (x) = max (0, x).

ReLU's goal is to introduce nonlinearity in our ConvNet. Because real-world data does not want ConvNet to learn, there will be non-negative linear values.

Other non-linear functions such as tanh, sigmoid or eLU can also be used instead of RELU. The majority of the data scientists use ReLU for their performance better than the other two groups.

In this study "elu" was used as activation.

3.2.3. Pool Layer:

The pool layers section will reduce the number of parameters when the images are too large. Spatial pooling also called sub-sampling or downsampling, which reduces the dimensionality of each map, but preserves important information. Different types of spatial pooling can be:

Max Pooling, Average Pool, Sum Pooling

Maximum pooling takes the largest element in the rectified property map. Taking the largest element can also take the average pool. The sum of all items in the property mapping is called ball pooling.

3.2.4. Fully Connected Layer:

The layer, which we call the FC layer, flattened our matrix into the vector and fed it to a completely bound layer, such as neural networks.

3.2.5. Normalization process:

In order to increase the stability of a neural network, batch normalization normalizes the output of the previous activation layer by subtracting the batch mean and dividing by the standard deviation of the batch. However, after these change / scale activation outputs by some randomly initiated parameters, the weights in the next layer are no longer optimal. SGD (stochastic gradient lowering) eliminates this normalization if there was a way to minimize the loss function.

3.3. Deep Learning Network:

A model was developed for the creation of a deep learning network, and this model was run with Google Colaboratory (an open source service provided by Google) using the python (keras) software language.(Figure 1-2)

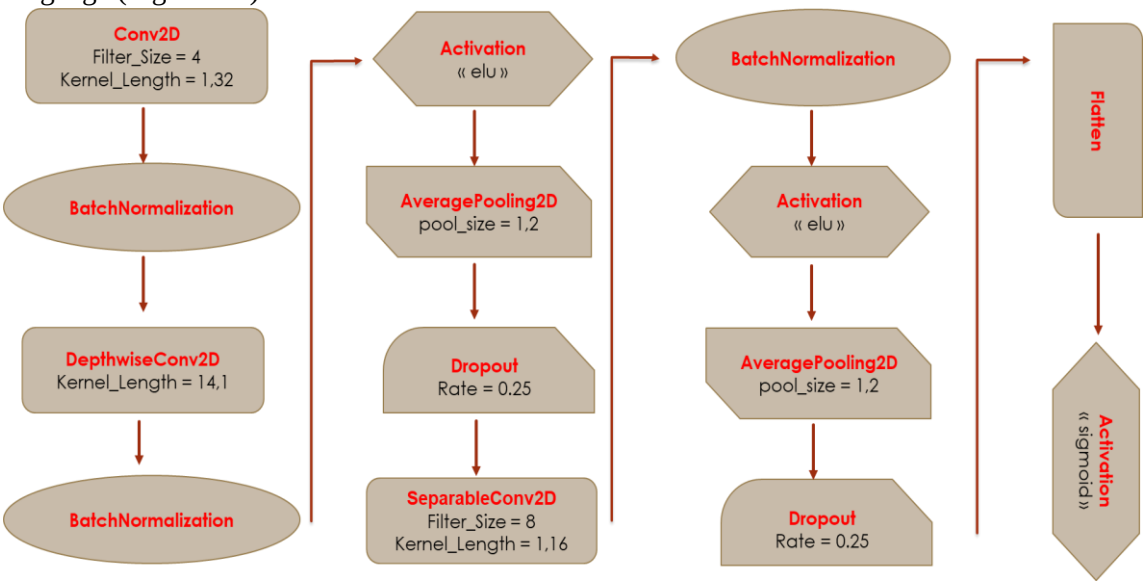

Figure 1- Schema of the CNN model

Layer (type)	Output Shape	Param #
InputLayer	(None, 14, 128, 60)	0
Conv2D	(None, 14, 128, 4)	7680
batch_normalization	(None, 14, 128, 4)	56
depthwise_conv2d	(None, 1, 128, 16)	224
batch_normalization	(None, 1, 128, 16)	4
Activation	(None, 1, 128, 16)	0
average_pooling2d	(None, 1, 64, 16)	0
Dropout	(None, 1, 64, 16)	0
separable_conv2d	(None, 1, 64, 8)	384
batch_normalization	(None, 1, 64, 8)	4
Activation	(None, 1, 64, 8)	0
average_pooling2d	(None, 1, 8, 8)	0
Dropout	(None, 1, 8, 8)	0
Flatten	(None, 64)	0
Dense	(None, 1)	65
sigmoid (Activation)	(None, 1)	0

Total params: **8,417**
Trainable params: **8,385**
Non-trainable params: **32**

Figure 2- Model evalute result

4. The Conclusion:

According to the data used in the study, the performance rate of the Video-type stimulus seems to be better than the others. The results can be improved by changing the parameters of the model and at this point, it is important to show the difference between them by training different stimulus types on the same model, Validation performance is shown in Table 3.

According to the results obtained, the audio stimuli performance curve and other types are shown below: (Figure 3-4-5)

Audio Stimuli Rate :

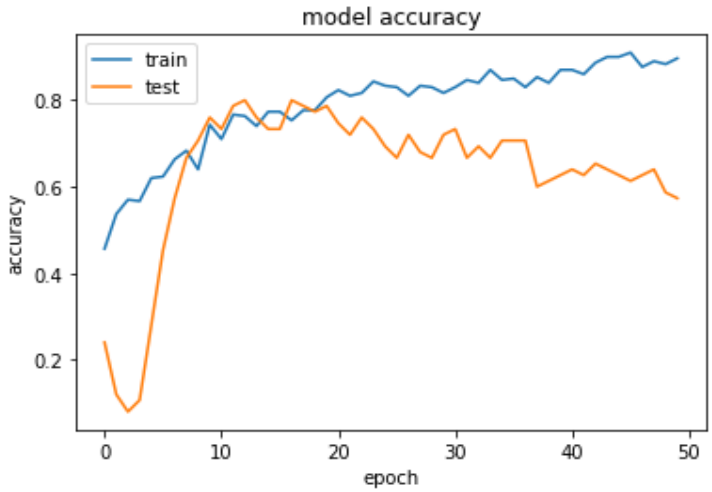

Figure 3- Audio stimuli Accuracy

Video Stimuli Rate :

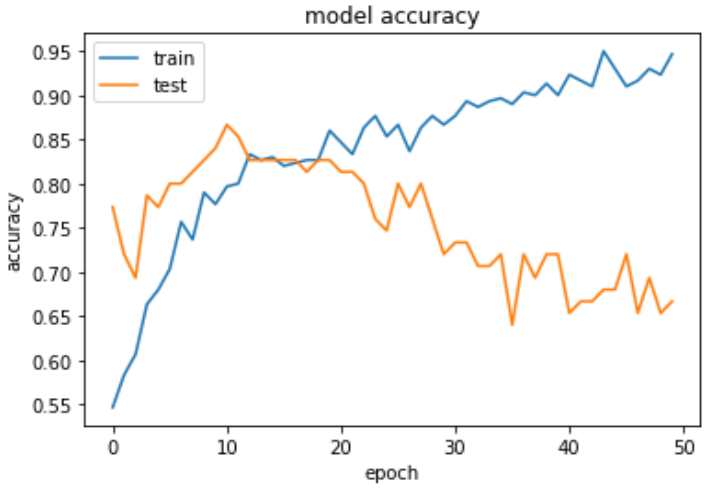

Figure 4- Video stimuli accuracy

Audio – Video Stimuli Rate :

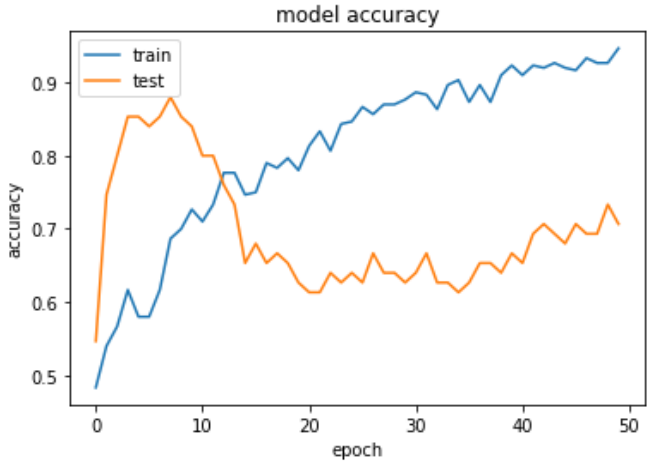

Figure 5- Audio-Video stimuli accuracy

Table 3 – Comparing accuracies between stimulus

Stimuli	**Train Accuracy**	**Validation Accuracy**
Audio	**%82,83**	**%65.30**
Video	**%83,87**	**%58,28**
Audio - Video	**%78**	**%59,44**

References

Acharya, U Rajendra et al. 2018. “Deep Convolutional Neural Network for the Automated Detection and Diagnosis of Seizure Using EEG Signals.” *Computers in Biology and Medicine* 100(September 2017): 270–78. https://doi.org/10.1016/j.compbiomed.2017.09.017.

Bhattarai, Smrity et al. 2017. “Digital Architecture for Real-Time CNN-Based Face Detection for Video Processing.” : 1–26.

Dasdemir, Yasar, Esen Yildirim, and Serdar Yildirim. 2017. “Analysis of Functional Brain Connections for Positive–negative Emotions Using Phase Locking Value.” *Cognitive Neurodynamics* 11(6): 487–500.

Lin, Wenqian, Chao Li, and Shouqian Sun. 2017. “Image and Graphics.” 10667(January 2018). http://link.springer.com/10.1007/978-3-319-71589-6.

Liu, Yisi, and Olga Sourina. “EEG Databases for Emotion Recognition.”

Mohammadi, Zeynab, Javad Frounchi, and Mahmood Amiri. 2017. “Wavelet-Based Emotion Recognition System Using EEG Signal.” *Neural Computing and Applications* 28(8): 1985–90.

Zeng, Hong et al. 2018. “RESEARCH ARTICLE EEG Classification of Driver Mental States by Deep Learning.” *Cognitive Neurodynamics* 12(6): 597–606. https://doi.org/10.1007/s11571-018-9496-y.