Depression Detection Using Visual Cues

Aditya Shah[1], Saurabh Mota[1], Ashit Panchal[1]
[1]*Department of Computer Engineering, Shah and Anchor Kutchhi Engineering College, Mumbai, India*

Abstract

Depression is one of the very common mental disorders in the world. Many times, people suffer from depression without knowing. Hence there is a need for automated depression detection. We propose a system which serves as a decision-making backing system for professionals, based solely on features that are extracted from facial expressions and features by interpretation of visual cues of depression. It will predict the scales of the Beck Depression Inventory-II (BDI-II) from visual expressions. Deep learning architectures have a very high accuracy in image recognition and classification. Hence, we propose using deep learning techniques for obtaining better accuracy. Convolutional Neural Network (CNN) is a technique of deep learning that is used for image classification. We propose to use CNN for automated depression detection. The proposed method will be tested on the "Audio/Visual Emotion Challenges 2014" (AVEC2014) dataset.

Keywords: *Machine Learning, Deep learning, Image Preprocessing, Depression Detection, Convolutional Neural Network, Beck Depression Inventory – II.*

1. Introduction

Depression is the sentimental expression of a state where one stays preoccupied with self with less hope of achieving something (personal needs/preferences) due to self-helplessness and self-powerlessness. If the personal preferences or ambition is fulfilled actually or in imagination, the result is exhilaration. This helplessness state in which development is halted predisposes to depression; is how Bibring E. defines depression [1]. It means that depression is an emotional state wherein a person feels a perpetual feeling of helplessness and powerlessness in achieving a particular goal that he/she has fixated on for an extended period of time. There has been a lot of attempts at decoding the mystery of depression detection involving solutions like detecting depression using sentiment analysis, speech evaluation, visual signs, questionnaire response evaluation and many more. One of these methods that has a promising scope in the field of depression detection is using visual cues. Visual cues are the visible signs that a person gives when presented with a stimulus. Such cues are studied and then evaluated in determining whether a person is depressed or not.

Due to the advancement in the sector of machine learning and Artificial Intelligence, many systems take different parameters as input and can give quite desirable results on the level of depression. Using the technology that deals with that specific input type, we can obtain desirable output, thus fulfilling the purpose of depression detection. It can be done by using machine learning methods like SVM and using deep learning methods like neural networks like CNNs and LSTMs. Since deep learning is known to give good results with higher accuracy than the machine learning algorithms, using this technology, we can get one step closer to solving the depression detection problem.

Deep learning is a sub-category of Artificial Intelligence. The applications of machine-learning have grown manifolds in recent years. This has created a need for more and improved machine learning strategies that aid the accuracy and practical viability of machine learning applications. To achieve this, the algorithms and techniques used for machine learning have undergone a massive revamp in terms of

time and space complexities, approaches in multiple dimensions and usage of numerous smaller strategies to generate a single more robust and efficient methodology.

Machine learning algorithms needed a good representation of data, and this was not an easy task to represent raw data. Deep learning is the tool that processes on this raw data by carrying out feature engineering techniques that require minimum domain knowledge and minimal human effort, and this results in an enhancement in the processes of machine learning. It consists of layered structure of data representation where higher-level features are obtained at higher layers, and lower level features are obtained at lower layers.

As deep learning has made tremendous advancements and remarkable performance in various applications, the widely used dominions of deep learning are business, science and government. This further includes adaptive testing, biological image classification, computer vision, natural language processing, cancer detection, face recognition, speech recognition, handwriting recognition, object detection, stock market analysis, smart city and many more [2].

The deep learning process has two stages: Training and Testing. The training part involves running huge dataset with distinct data through a deep learning network so that a network can adjust its weights according to the input received throughout the training stage. The next stage: testing stage is the stage where the "trained" model of a deep learning network is tested for accuracy and various other metrics like false-positives etc. by verifying the output of the network and the actual expected output.

Deep learning involves using large datasets, which depending on the quality of the dataset can be used for optimizing the overall performance of the network.

2. Related Work

- **A. Pampouchidou et al, "Facial geometry and speech analysis for depression detection."**

The existing system is based on unique features extracted from facial expression geometry and speech by inferring from non-verbal indicators of depression [3]. The system has been tested both in gender independent and gender-based modes and with dissimilar fusion methods. The algorithms were evaluated for several amalgamations of parameters and classification schemes, on the dataset provided by the AVEC2013 and AVEC2014. The framework achieved a precision of 94.8% on the self-evaluated dataset of AVEC2013 [11] and AVEC2014. The optimal system performance was obtained using the nearest neighbour classifier on the decision fusion of geometrical features in the gender independent mode, and audio-based features in the gender-based mode single visual and audio decisions were combined with the OR binary operation. The system reported a high success rate in detecting persons displaying high levels of depressive symptomatology (94.8%), combined with a substantial false positive detection rate (60.6%).

However, we did not implement classification on the basis of gender. The classification was based solely on their facial features, irrespective of their gender and age [3].

- **W. C. D. Melo et al, "Combining Global and Local Convolutional 3D Networks for Detecting Depression from Facial Expressions"**

Deep learning architectures, different fusion of Convolution-3D network are projected to maximize the accuracy, where considering the subject's whole facial region i.e global region and the area of the eye i.e. local region, Spatio-temporal features were extracted. [6]. This allows the system to highly focus on a local facial region that is highly important for depression analysis. The authors also integrate 3D Global

Average Pooling to well-organizedly encapsulate Spatio-temporal features and reducing number of model parameters and potential-overfitting by removing fully-connected layers (FC layers) [6]. Through distribution learning, the author introduced a deep learning architecture for predicting accurate depression levels [6]. It depends on an expectation loss function that allows evaluating the underlying data distribution over depression levels, where expected values of the distribution are enhance to obtain the ground-truth levels. Even under label uncertainty, accurate predictions of depression levels are produced [6].

- **M. Valstar et al, "Avec 2014"**

Different visual features are extracted from facial expression images. The deep learning method is implemented to obtain important visual features from the frames of facial expression [7]. We incorporated the use of deep learning concepts in our implementation that helped us gain a substantial amount of accuracy as depicted further in the document. The authors started a competition based on a locally available baseline system to process audio and visual data. Regarding the visual cues, for every video frame, alginment and detection of faces are performed.

In our case, we use a fully connected layer with four tensors that correspond to 4 classes of BDI-II scale, for the purpose of classification. The dataset has a sub challenge for depression detection on the BDI-II scale [7].

- **S. Dargan et al, "A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning."**

Deep Learning : One of the sub category of machine learning is deep learning which is a technique relative to brain structure and function, which is termed as artificial neural networks. Deep learning is most efficient, supervised, time and cost-efficient hence one of the most popular machine learning technique [4]. Images, text or sound are directly used to perform classification in deep learning techniques. Training of deep learning models is done by using large labelled data and the neural network architecture which is used learns the features directly from the dataset without manual feature extraction. The various Neural Networks (NNs) like Recursive NN, Recurrent NN, Convolutional NN, Deep Belief Network, Deep Boltzmann Machine, Generative Adversarial Network, Variational Autoencoder, etc are different types of deep learning architectures [2].

Convolutional Neural Networks : CNNs are categories of NN which have been used in applications like the classification of images, image recognition, identifying objects, etc. CNN detects the necessary features from the image data using corresponding filters and extracting the required features for prediction. Convolution is termed as a mathematical process on two objects to obtain an outcome that expresses how the form of one object is modified by the opposite.

Through this computation, we detect a specific feature from the input image and obtain the result having information about such features collectively called a feature map.

- **J. Jeong, "The Most Intuitive and Easiest Guide for CNN"**

The images are going to be processed for feature extraction by adding multiple convolutional layers and pooling layers. And there will be FC layers heading to the layer for SoftMax (for a multiclass case) or sigmoid (for a binary instance) function. As the layers go deeper and deeper, the features that the model deals with, become more complex [5].

Various types of CNN are:

- LeNet [12]
- AlexNet [13]
- ZfNet [14]
- Vgg16 [15]
- GoogleNet [16]

- **Simonyan, K, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition."**

VGG16 : VGG16 (as shown in fig. 1) is a type of CNN model which was proposed by K. Simonyan and A. Zisserman from the University of Oxford. The model is trained on ImageNet. It's architecture has 19 layers. It is a large network with 138M parameters. This particular network architecture was the runners up of the ILSVRC-2014 competition with an error rate of 5.1% considered as a top-5 error rate. It refines AlexNet by replacing large kernel size of 11 and 5 to small kernel size, i.e. 3x3.

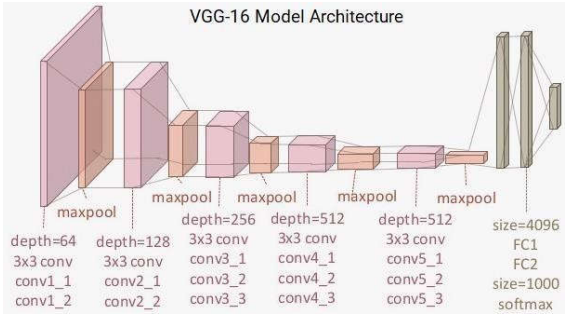

Fig. 1 VGG16 Model Architecture

3. Experimental Methodology

Dataset : The model proposed in the paper is used to predict the level of depression in the volunteer using visual cues. To evaluate the model's performance, AVEC2014's depression sub-challenge dataset is used. The aim of the sub-challenge is to categorise the volunteer according to the Beck Depression Inventory (BDI) score. The BDI scores are classified into 4 categories as: 0-13 (no depression), 14-19 (mild depression), 20-28 (moderate depression), and 29-63 (severe depression).

The AVEC2014 dataset has 300 videos recorded by the volunteers using the webcams and microphones [7]. Each volunteer performs two tasks:

Northwind – In this task the volunteers record the video in German language and read aloud an excerpt of the fable "Die Sonne und der Wind" (The North Wind and the Sun).

Freeform - In this task the volunteers answer the question from memory from their childhood in German language.

In both the tasks, the recordings are split into three parts: training, development and test set with each part having 50 videos. To train the model, training and development sets are used from both tasks as training data, and the test sets are used to evaluate the performance of the model [7].

Pre-Processing : The preprocessing in the proposed method is to extract faces in image form. The goal is to provide facial images as input to the proposed deep learning architecture. Frames are extracted from the videos. Frames are extracted at a constant rate of 4 frames per second (f.p.s). In [8], it is concluded that a microexpression while telling a lie or truth is about ≤0.40s and ≤0.50s on the face. Hence by taking four fps, we do not miss any vital microexpression which can be used. The images are distributed in the respective folder according to the labels given in the dataset. The network is provided with a frontal face input image.

The Multi-Task Cascade Convolutional Network was used in this method to detect and align faces jointly [9]. As a result, after resizing and rescaling, each image has a size of 224×224. This is the input to the proposed model. VGG16 takes an input image of size 224x224. Hence all the extracted images have a size of 224x224, which is given as the input to the model.

The first layer of the model is the conv1 layer that takes 224 x 224 RGB image as input. After this the image is passed through a stack of convolutional layers which have a very small receptive field of 3x3 filters. The spatial padding of convolutional layer input is such that the spatial resolution is preserved after convolutional, i.e. the padding is 1-pixel for 3×3 convolutional layers, and the convolutional stride is fixed to 1 pixel. Max-pooling (Spatial Pooling) is carried out by five max-pooling layers. As shown in Fig.2 some convolutional layers are followed by max-pooling layers but not all the conv. layers are followed by max-pooling. Max-pooling of stride 2 is performed over a 2×2-pixel window.

The stack of convolutional layers are followed by Three Fully-Connected (FC) layers with different depths in this architecture: the first two FC layers have 4096 channels; and the third FC layer performs a 4 - way BDI-II depression scale-based classification and thus contains four channels. The soft-max layer is the last layer. The configuration of the fully connected layers is the same in all networks. All hidden layers are equipped with the Rectified Linear Unit (ReLU) non-linearity.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
dense_1 (Dense)	(None, 4)	16388

Total params: 134,276,932
Trainable params: 134,276,932
Non-trainable params: 0

Fig. 2 Implementation VGG16 Model

Algorithm :

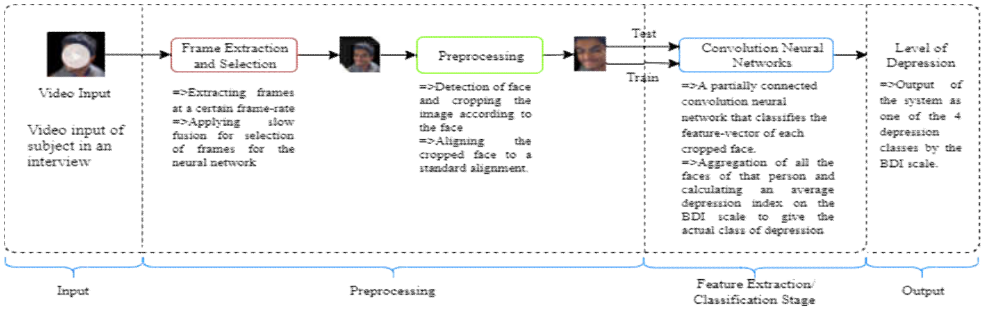

Fig. 3 Flow Diagram

Fig 3 shows the complete algorithm and flow of the process. It can be delineated as follows:

1. Input: Video of a subject in an interview
2. Extraction of images from the video at a fixed frame rate.
3. Detection of face and adjustments by cropping and aligning of the face.
4. These photos are passed into the CNN, and the CNN model is trained.
5. The testing of the model reveals the level of depression based on the BDI-II scale for the test subject.

4. Performance Measures

The above-described model is trained on the AVEC 2014 database used in the AVEC depression detection challenge of 2014. The training of this model was carried out by setting all the parameters of the model to be trainable and using optimizer 'Adam'. The final classification task was done by a fully connected layer with four vectors, each corresponding to a class of the BDI-II scale of depression.

One of the most commonly used regression loss function is Mean Square Error (MSE) (refer the below equation) which is the sum of squared distances between our target variable and predicted values.

$$MSE = \frac{\Sigma_{i=1}^{n}(y_i - y^p{}_i)^2}{n}$$

One of the other loss functions used for regression models is Mean Absolute Error (MAE) (refer the below equation) which is the sum of absolute differences between the target and predicted variables. As a result, it measures an average magnitude of errors in a set of predictions, without considering their directions.

$$MAE = \frac{\Sigma_{i=1}^{n}|y_i - y^p{}_i|^2}{n}$$

5. Results and Conclusions

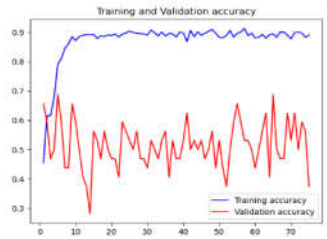

Fig. 4 Training and Validation Accuracies

Fig. 5 Training and Validation Losses

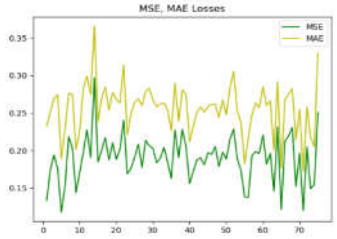

Fig. 6 Validation MSE and Validation MAE

We obtained the maximum Validation accuracy of **68.75%** (fig. 4) and Validation loss of **2.88** (fig. 5) with the above configuration of network and conditions for classification. **Validation MSE** measured at around **11.78** and **Validation MAE** measured at about **18.94** (fig. 6)was obtained during the training.

We proposed an architecture of a CNN model based on the VGG16 CNN to detect the levels of depression as depicted in the Beck Depression Inventory-II (BDI-II). The level of depression is determined based on the features extracted from the facial images of the subjects when they are asked to do specific tasks. The VGG16 based model was trained on a large dataset of images extracted from the videos of the abovementioned subjects' task performances. The last layers of the VGG16 model are attached with a 4-way FC (Fully Connected) layer to handle the task of classification of the feature vectors passed to it.

References

1. Bibring, E. (1953). The mechanism of depression. In P. Greenacre (Ed.), Affective disorders; psychoanalytic contributions to their study (p. 13– 48). International Universities Press.
2. S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, "A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning," Archives of Computational Methods in Engineering, Jan. 2019.
3. Pampouchidou, O. Simantiraki, C.- M. Vazakopoulou, C. Chatzaki, M. Pediaditis, A. Maridaki, K. Marias, P. Simos, F. Yang, F. Meriaudeau, and M. Tsiknakis, "Facial geometry and speech analysis for depression detection," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017.
4. "What Is Deep Learning? How It Works, Techniques & Applications," How It Works, Techniques & Applications - MATLAB & Simulink. [Online]. https://www.mathworks.com/discovery/deeplea rning.html.
5. J. Jeong, "The Most Intuitive and Easiest Guide for CNN," Medium, 17-Jul-2019. [Online]. https://towardsdatascience.com/themostintuitive-and-easiest-guide-forconvolutionalneural-network-3607be47480.
6. W. C. D. Melo, E. Granger, and A. Hadid, "Combining Global and Local Convolutional 3D Networks for Detecting Depression from Facial Expressions," 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019.
7. M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014," Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC 14, 2014.
8. D. Matsumoto and H. C. Hwang, "Microexpressions Differentiate Truths From Lies About Future Malicious Intent," Frontiers in Psychology, vol. 9, 2018.

9. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks,” IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.
10. Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M. C., & Sahli, H. (2017, October). “Multimodal measurement of depression using deep learning models”. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (pp. 53-59).
11. M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. “AVEC 2013: the continuous audio/visual emotion and depression recognition challenge.” In Proc. of the 3rd ACM international workshop on Audio/visual emotion challenge, ACM, 2013, pp. 3-10.
12. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
13. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks” Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.
14. Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." In European conference on computer vision, pp. 818-833. Springer, Cham, 2014.
15. Simonyan, K, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
16. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.