Analysis of Live Emotionally Intelligent Bot

Mohit Dayal[1] , Jitender kumar[2]
[1]Chandigarh University, Mohali ,Punjab, India.
[2]Rawal Institute of Engineering and Technology, Faridabad Haryana, India.

E-Mail: mohitdayal.md@gmail.com

Abstract

Emotion is a human way of expressing feelings. Emotions tell how well the social interactions of the user went in the recent past. This can prove to be valuable information to improve and enhance user experience in chatbots. This information will make bots not just intelligent but emotionally intelligent by understanding the user's social life. The information can help in providing more relevant answers to the user. The user messages in chatbots do not provide any information about the user's state and emotions. There is a need for some other data source that could provide valuable information about the user's emotions.

In this project, we have developed an emotionally-intelligent bot application based on a convolution model. The goal is to improve the performance of bots by using the facial emotions of the user. The convolution model is trained using TensorFlow on 48x48 pixel gray-scale images of faces. The assessment of the performance of bot is also discussed. The outcomes conclude a significant increase in user satisfaction.

Keywords: Facial Emotion, CNN, Convolution, TensorFlow, HMM Model

Introduction

The various sources of emotional information include facial expressions, tone of voice, gestures, etc. On a platform like a mobile application, the facial expressions can be easily extracted from the user's device camera. Moreover, there have been major developments in facial expression recognition after the introduction of Capsule Networks. Also, there are a lot of extensively researched facial recognition models: HMM Model, Convolution Networks. Emotions often mediate and facilitate interactions among human beings. Thus, understanding emotion often brings context to seemingly bizarre and/or complex social communication Emotion can be recognized through a variety of means such as voice intonation, body language, etc. However, the easier, more practical method is to examine facial expressions. There are seven types of human emotions shown to be universally recognized across different cultures: anger, disgust, fear, happiness, sadness, surprise, contempt. Interestingly, even for complex expressions where a mixture of emotions could be used as descriptors, the cross-cultural agreement is still observed. Therefore a utility that detects emotion from facial expressions would be widely applicable. Such an advancement could bring applications in medicine, marketing, and entertainment. The task of emotion recognition is particularly difficult for two reasons: There does not exist a large database of training images and classifying emotion can be difficult depending on whether the input image is static or a transition frame into a facial expression. The latter issue is particularly difficult for real-time detection where facial expressions vary dynamically. Most applications of emotion recognition examine static images of facial expressions. We investigate the application of convolutional neural networks (CNNs) to emotion recognition in real-time with a video input stream. Given the computational requirements and complexity of a CNN, optimizing a network for efficient computation for frame-by-frame classification is necessary. Besides, accounting for variations in lighting and subject position in a non-laboratory environment is challenging. We have developed a system for detecting human emotions in different scenes, angles, and lighting conditions in real-time. The result is a novel application where an emotion-indicating emoji is superimposed over the subjects' faces.

Literature Review

Previous works are focused on eliciting results from unimodal systems. Machines used to predict emotion by only facial expressions [1] or only vocal sounds [2]. After a while, multimodal systems that use more than one feature to predict emotion has more effective and give more accurate results. So, the combination of features such as audio-visual expressions, EEG, body gestures has been used since. More than one intelligent machine and neural networks are used to implement the emotion recognition system. The multimodal recognition method has proven more effective than unimodal systems by Shiqing et al. [3]. Research has demonstrated that deep neural networks can effectively generate discriminative features that approximate the complex non-linear dependencies between features in the original set. These deep generative models have been applied to speech and language processing, as well as emotion recognition tasks [4-6]. Martin et al. [7] showed that the bidirectional Long Short Term Memory(BLSTM) network is more effective than the conventional SVM approach.; In speech processing, Ngiam et al. [8] proposed and evaluated deep networks to learn audio-visual features from spoken letters. In emotion recognition, Brueckner et al. [9] found that the use of a Restricted Boltzmann Machine (RBM) before a two-layer neural network with fine-tuning could significantly improve classification accuracy in the Inter speech automatic likability classification challenge [10]. The work by Stuhlsatz et al. [11] took a different approach for learning acoustic features in speech emotion recognition using Generalized Discriminant Analysis (GerDA) based on Deep Neural Networks (DNNs). Yelin et al. [12] showed three-layered Deep Belief Networks (DBNs') give better performance than two-layered DBNs' by using audiovisual emotion recognition process. Samira et al [13] used a Recurrent neural network combined with Convoluted Neural Network (CNN) in an underlying CNN-RNN architecture to predict emotion in the video. Some noble methods and techniques also enriched this particular research. They are more accurate, stable and realistic. In terms of performance, accuracy, reasonability, and precision these methods are the dominating solutions. Some of them are more accurate but some are more realistic. Some take much time and require greater computation power to produce a more accurate result but some compromise accuracy over performance. The idea of being successful might differ but these solutions are the best possible till now. 4 Y. Fan, X. Lu, D. Li, and Y. Liu. proposed a method for video-based emotion recognition in the wild. They used CNN-LSTM and C3D networks to simultaneously model video appearances and motions [16]. They found that the combination of the two kinds of networks can give impressive results, which demonstrated the effectiveness of the method. In their proposed method they used LSTM (Long Short Term Memory) - a special kind of RNN, C3D – A Direct Spatio-Temporal Model and Hybrid CNN-RNN and C3D Networks. This method gives great accuracy and performance is remarkable. But this method is much convoluted, time-consuming and less realistic. For this reason, efficiency is not that impressive [16]. Wei-Long Zheng and Bao-Liang Lu proposed EEG-based effective models without labeled target data using transfer learning techniques (TCA-based Subject Transfer) [18] which is very accurate in terms of positive emotion recognition than other techniques used before. Their method achieved 85.01% accuracy. They used to transfer learning and their method includes three pillars, TCA-based Subject Transfer, KPCA-based Subject Transfer, and Transductive Parameter Transfer. For data preprocessing they used raw EEG signals processed with a bandpass filter between 1 Hz and 75 Hz and for feature extraction, they employed differential entropy (DE) features. For evaluation, they adopted a leave-one subject-out cross-validation method. Their experimental results demonstrated that the transductive parameter transfer approach significantly outperforms the other approaches in terms of the accuracies, and a 19.58% increase in recognition accuracy has been achieved. Though this achievement is limited to positive emotion recognition only. This method is limited in terms of negative and neutral emotions recognition. Yet a lot of improvement needed to recognize negative and neutral emotion more accurately [18].

Problem Statement

This project will investigate facial expression which contains the information of images in the form of pixels. Ultimately we will be predicting facial expression in live time using computer vision library-opencv2 and Affectiva SDK in Android Studio. This report will describe the work carried out during the iterative process of data preparation, modelling and evaluation including data formatting, consistency or other quality issues, integration with open cv files, libraries used, and other data manipulation techniques that were used during the work. The project is under research by several companies such as Samsung. This problem statement is also under improvement due to model improvement. This main aim is to integrate live emotion detection in an android chatbot that will smartly reply according to the detected emotions and show the user the best result according to the emotion.

There are two major issues related to non-emotional bots:

1. Noise

The content provided by these bots may or may not be relevant to user in their current emotional state. Many of the answers are noise for the user.

2. Non-Engagement

If the bots can’t relate to the user’s feelings, they can’t connect and conversation soon dies off.

Need of the Project

The social network of humans has grown significantly with developments in social networking. These developments have brought many changes in human life, one being emotional effects. There has been an increase in emotional instability in recent years. The cases of depression, rage, suicides are direct outcomes of these changes.

The solution like emotionally intelligent chatbots can provide a personal bot not just customized to a user’s personality but also responsive to the user’s current emotional state. This solution can provide responses that can selectively counter emotional instability. Below response to users message shows the improvement that Emotional intelligence can bring into conversations:

User: I got just the passing marks in today’s test.

Sad: Keep working hard! You will get results.

Neutral: You should give more time to studies then.

Fear: Don’t worry! At least you passed the test

Dataset Description

The data consists of 48x48 pixel greyscale images of faces. The faces have been automatically registered so that the face is more or less centred and occupies about the same amount of space in each image. The task is to categorize each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). Fer.csv contains two columns, "emotion" and "pixels". The "emotion" column contains a numeric code ranging from 0 to 6, inclusive, for the emotion that is present in the image. The "pixels" column contains a string surrounded in quotes for each image. The contents of this string a space-separated pixel values in row-major order. Fer.csv contains only the "pixels" column. The training set consists of 28,709 examples. This dataset was prepared by Pierre-Luc Carrier and Aaron Courville, as part of an ongoing research project. This dataset was prepared by Pierre-Luc Carrier and Aaron Courville, as part of their research project. The dataset is made available on kaggle website which contains 48x48 pixel grayscale images of faces. The data is divided into training, validation and test sets. The training set contains 28709 example images while the test set contains 3589 example images.

The images are provided such that the face takes approximately the same space in all images and is self-centered.

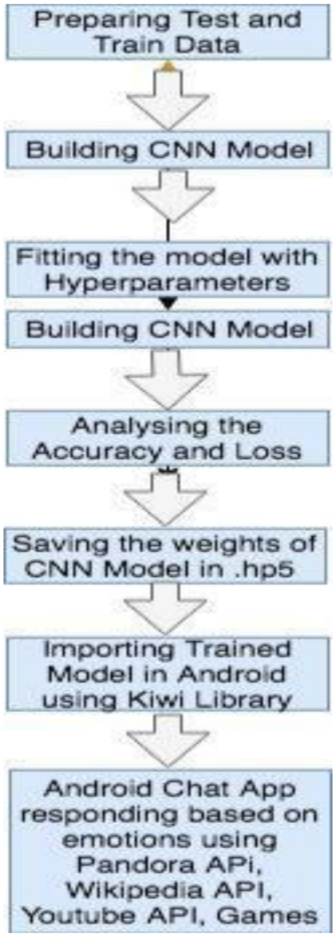
The training set contains two columns: emotion and pixels. The "emotion" column has labels from 0 to 6, representing the emotion that is present in the image. The "pixels" column has a string for each image that contains space-separated pixel values in row-major order.

The seven emotion labels under consideration are as follows:

Angry, Neutral, Sad, Happy, Surprise, Fear, Disgust

Research Methodology

1. The data was prepared by dividing it into a training set and testing set.
2. The CNN model was build using TensorFlow backend on training data.
3. The model was fitted on hyperparameters to extract the face from the images.
4. The trained model was tested using a testing data set to find accuracy and loss.
5. The weights of the model were saved in .hp5 file.
6. The model was imported as a python model in the Android project using kiwi library.
7. The real-time camera captures the frame after 20 sec and sends an image to a trained model to detect the emotion.
8. The chatbot on android uses a combination of Pandora API, YouTube API, Wikipedia API, and use emotion to respond to the user.


Preparing Test and Train Data
Building CNN Model
Fitting the model with Hyperparameters
Building CNN Model
Analysing the Accuracy and Loss
Saving the weights of CNN Model in .hp5
Importing Trained Model in Android using Kiwi Library
Android Chat App responding based on emotions using Pandora APi, Wikipedia API, Youtube API, Games

Convolutional Neural Networks (ConvNets or CNNs) are a category of neural networks that have proven very effective in areas such as image recognition and classification. ConvNets have been successful in identifying faces, objects, and traffic signs apart from powering vision in robots and self-driving cars. CNNs uses a variation of multilayer perceptrons designed to require minimal preprocessing. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics. Convolutional networks were inspired by biological processes in which the connectivity pattern between neurons is inspired by the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual filled known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field. CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered. This independence from prior knowledge and human effort in feature design is a major advantage.

CNN makes use of kernels to extract features from images. Kernels can be of any function used to capture the local dependencies in an image. They use convolution, pooling, classification, and non-linearity.

Performance Analysis

A set of 20 questions were asked by 20 users each from the non-emotional bot and emotionally intelligent bot. The user's satisfaction in terms of relevance was measured to find the average and percentage relevance of the responses made by these bots. The comparison shows a significant increase in the relevance of responses by the bot when emotionally intelligent:

Relevant Answers (Average)	Relevant Answers (Percentage)
Non-Emotional Bot 13.2	66
Emotional Bot 18.8	94

Conclusion

The results of the research show that the relevance of the content and response can be significantly increased by exploiting the information obtained from user facial expressions. The emotionally intelligent chatbot also makes the user more engaged.

Future Scope

This project can be extended by using the capsule network for facial emotion recognition. This model can provide better accuracy in emotion recognition. Also, the bot application can be extended to take more emotion data sources like user voice tone, gestures. This could further improve emotion detection by the bot.

References

1. Gil Levi, Tal Hassner; Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns, SC / Information Sciences Institute, the Open University of Israel, 2014.
2. Kun Han, Dong Yu, Ivan Tashev; Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine; Department of Computer Science and Engineering, The Ohio State University, Columbus,43210, OH, USA; Microsoft Research, One Microsoft Way, Red mond,98052, WA, USA,2014.
3. Shiqing Zhang, Xiaohu Wang, Gang Zhang, Xiao Ming Zhao; Multimodal Emotion Recognition Integrating Affective Speech with Facial Expression; Institute of Image Processing and Pattern Recognition Taizhou University Taizhou 318000 CHINA, Hunan Institute of Technology Hengyang 421002 CHINA, Bay Area Compliance Labs. Corp. Shenzhen 518000 CHINA, 2014.

4. S L Happy; Anjith George; Aurobinda Routray, “A Real-Time Facial Expression Classification System Using Local Binary Patterns.,” 2012 IEEE.
5. A. Mohamed, G.E. Dahl, and G. Hinton, ―Acoustic modeling using deep belief networks,‖ Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, pp. 14– 22, 2012.
6. G. Sivaram and H. Hermansky, ―Sparse multilayer perceptron for phoneme recognition,‖ Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, pp. 23– 29, 2012.
7. Martin W¨ollmer, Angeliki Metallinou, Florian Eyben, Bj¨ornSchuller, Shrikanth Narayanan; Context Sensitive Multimodal Emotion Recognition fro m Speech and Facial Expression using Bidirectional LSTM Modeling; Institute for Human-Machine Communication, TechnischeUniversit¨atM¨unchen, Germany Signal Analysis and Interpretation Lab (SAIL), University of Southern California, Los Angeles, CA, 2010.
8. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A.Y.Ng, ―Multimodal deep learning,‖ in Proceeding soft he 28th International Conference on Machine Learning (ICM L), 2011, pp. 689–696.
9. R. Brueckner and B Schuller, ― Likability classification - a not so deep neural network approach,‖ in Proceedings of Inter speech, 2012.
10. B. Schuller, S. Steid l, A. Batliner, E. N¨oth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F.Eyben, T. Bocklet, et al., ―The inter speech 2012 speaker trait challenge,‖ Inter speech, Portland, Oregon, 2012.
11. A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, ― Deep neural networks for acoustic emotion recognition: raising the benchmarks,‖ in Acoustics, Speech, and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011, pp. 5688– 5691.
12. Samira Ebrahimi, Vincent Michalski, Kishore Konda,Goethe Roland Memisevic, Christopher Pal―Recurrent Neural Networks for Emotion Recognition in Video‖,KahouÉcolePolytechnique de Montréal, Canada ; Universität Frankfurt, Germany; Université de Montréal, Montréal, Canada; 2015.
13. A. Yao, D. Cai, P. Hu, S. Wang, L. Shan, and Y. Chen; HoloNet: towards robust emotion recognition in the wild, 2016.
14. Yelin Kim and Emily Mower Provos, Data-driven framework to explore patterns (timings and durations) of emotion evidence, specific to individual emotion classes; University of Michigan Electrical Engineering and Computer Science, Ann Arbor, Michigan, USA; 2016.
15. Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks. Proceeding ICMI 2016 Proceedings of the 18th ACM International Conference on Multimodal Interaction, Pages 445-450, Tokyo, Japan — November 12 - 16, 2016.
16. Zixing Zhang, Fabien Ringeval, Fabien Ringeval, Eduardo Coutinho, Erik Marchi, and Björn Schüller, Semi-Supervised Learn ing (SSL) technique
17. Wei-Long Zheng1 and Bao-Liang Lu, Personalizing EEG-Based Affective Models with Transfer Learning, Center for Brain-like Computing and Machine Intelligence, Department of Computer Science and Engineering, Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Brain Science and Technology Research Center, Shanghai Jiao Tong University, Shanghai, etc.

Source of Support: Nil
Conflict of Interest: None