CS 351 Project: Real-time emotion detection

Muhammad Riaz Ul-Haq
CS
2017306
u2017306@giki.edu.pk

Abdul Raheem Zaidi
CS
2017009
u2017009@giki.edu.pk

***Abstract*—We present our method of identifying emotion using convolutional neural networks. We focus on classifying 7 emotions using facial imagery in real time. The system contains a HAAR cascade face detection module and a 5-layer CNN trained on the FER2013 dataset. Our proposed model achieves a 61.49% and 62.25% accuracy on the validation and test set respectively.**

I. Introduction

In his 1971 paper, Paul Ekman identified 6 universal facial expressions - anger, disgust, fear, happiness, sadness and surprise [1]. Even today, researchers aim to achieve reliable accuracy for this set of emotions.

Humans express what's in their mind in form of speech, gestures and emotions. Building a system that understands human facial emotions in real time can unlock a wide range of applications.

It can be integrated into other systems, for instance, ATMs could be set up such that they won't dispense money when the user is scared or appears threatened.

In the gaming industry, emotion-aware games can be developed which could vary the difficulty of a level depending on the player's emotions.

Emotions can also be assessed while a viewer watches ads to see how they react to them. This is especially helpful since ads do not usually have feedback mechanisms apart from tracking whether the ad was watched and whether there was any user interaction.

Software for cameras can use emotion recognition to take photos whenever a user smiles.

This project proposes a model that is aimed at real-time facial emotion recognition.

II. Literature Review

Yu and Zhang [2] used a five layer CNN to achieve a 0.612 accuracy. They pre-trained their models on the FER-2013 dataset and then fine-tuned the model on the Static Facial Expressions in the Wild 2.0 (SFEW) dataset. They also chose to use stochastic pooling layers over max pooling layers citing its better performance on their limited data.

Kahou et al. [3] used a CNN-RNN architecture to train a model on individual frames of videos as well as static images. They made use of the Acted Facial Expressions in the Wild (AFEW) 5.0 dataset for the video clips and a combination of the FER-2013 and Toronto Face Database for the images. Instead of using long short term memory (LSTM) units, they used IRNNs which are composed of rectified linear units (ReLUs). These IRNNs provided a simple mechanism for dealing with the vanishing and exploding gradient problem. They achieved an overall accuracy of 0.528.

Tan et al. [4] proposed a neural network model to classify a group image into a particular emotion - positive, neutral or negative. The model consists of two convolutional neural networks - the first is based on group images and the second is based on individual facial emotions. The facial emotion CNN comprises of two CNNs - one for aligned faces which is trained using the ResNet64 model using the Webface dataset and the other for non-aligned faces which is trained using the ResNet34 model on the FER+ dataset. The group images are trained using VGG19 model on the Places and ImageNet datasets. The validation set consists of 2068 images - combined from all of the datasets used for training and the model achieved an accuracy measure of 80.9.

III. Methodology

Classification of human emotion requires several input characteristics, visual signals, audio signals, pose/body language and speech context. We have chosen to simplify our approach to classification by focusing on visual input as the primary source of facial emotion classification.

The task of classifying emotions in real time can be divided into two components:

- detection of facial features from the frame;
- inference of emotion from the features of a single frame.

A. Face Detection

We use a HAAR Cascade Classifier for face detection from a webcam video source. The classifier uses a pre-trained model provided by OpenCV. The region of the detected faces is cropped and processed to be fed to the emotion detection module.

B. Emotion Detection

We use a CNN implemented in Keras with a Tensorflow backend. The dataset selected for training is FER2013, which is a recognized public dataset that has been used for this problem by previous researchers and students. It contains 35887 grayscale, labeled images; all images are 48×48. FER2013 categorizes images into 7 emotions: happy, sad, fear, disgust, neutral, angry and surprise.

10% of the total images (3589 images) in the dataset are used for the validation set and another 10% is used for the test set.

C. Data Pre-Processing

Frames obtained from the video source are sent to the face detection module which checks if a face is present in the frame. If it detects a face, the region of the detected area is cropped and fed to the data pipeline for pre-processing.

The image is converted to grayscale since all images used during training are grayscale. The pixel data is then divided by 255 to scale it to a value $v \in [0, 1]$. The resulting image is then resized to a 48×48 image which is sent to the CNN for inference.

IV. Experimental Results

A. Proposed Model

We propose a 5 layer CNN to classify emotions using facial images. We evaluate the model on the basis of categorical accuracy:

$$\text{Categorical Accuracy} = \frac{\text{No. of correctly identified images}}{\text{Total number of images}}$$

The model was able to achieve an accuracy of 62.25% on the test set. A visualization of the performance of the model during training is given in **Figure 1**. The model fits well, but converges near 0.6.

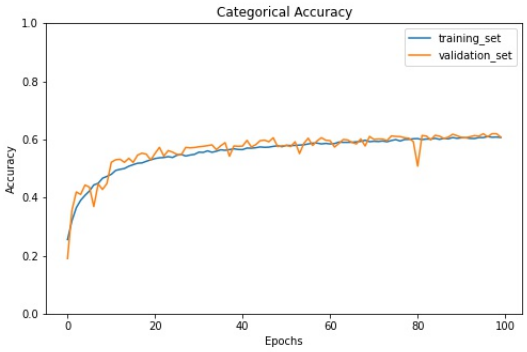

Fig. 1. Accuracy of the model during training for 100 epochs

Our model is similar to the one proposed by [2]. Differences include the choice of using average pooling instead of stochastic pooling, the presence of a large 8×8 filter and the absence of a fully connected layer.

B. Data Augmentation

Due to the small size of the FER2013 dataset, it is hard to achieve good accuracy without increasing the size and complexity of the model. We combat this by generating augmented images from the training set. For each batch in an epoch, images are selected and modified by random amounts of predefined rules. Images may be rotated up to 40° and horizontally or vertically flipped.

Augmenting the data increased our accuracy on the test set from 57.81% to 62.25%.

C. Comparison with Deeper Neural Networks

Deeper neural networks are able to localize and extract more features at the cost of being computationally expensive. **Figure 2** shows a comparison of test set accuracy of our model against VGG, Res-Net, Inception and DeepEmotion.

For use in real-time emotion detection on non-specialized devices, the model has to be small. A 30fps camera will have to send 30 frames to face detection model followed by the emotion detection model within a second, and the output has to be sent back to the video stream. The emotion detection model becomes the performance bottleneck in the pipeline. Our model provides better real-time performance due to its computational simplicity at the cost of accuracy.

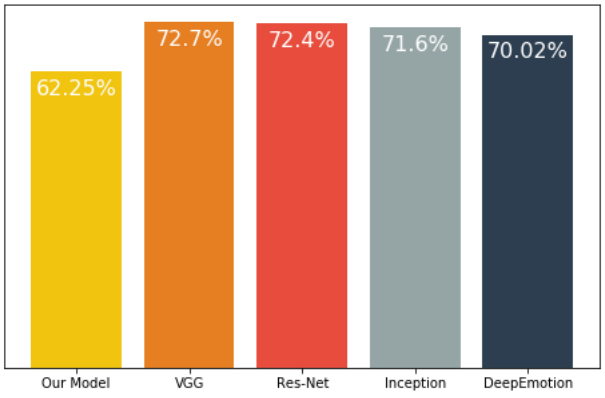

Fig. 2. Test set accuracy comparison

D. In Action

The source files for the complete project can be found at the GitHub page. The confusion matrix of the test set accuracy gives us an idea of which emotions are likely to be identified correctly:

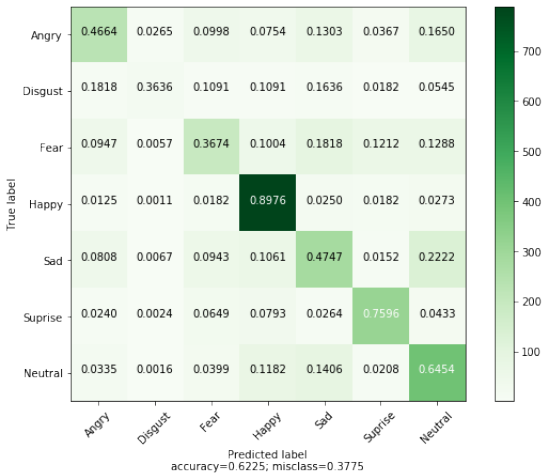

Fig. 3. Confusion matrix of the test set accuracy

V. Conclusion

We have presented a method of real-time facial emotion recognition. The proposed model is able to achieve good accuracy on the FER2013 dataset. It is not computationally intensive therefore the model is suitable for usage in a real-time application.

References

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of Personality and Social Psychology*, vol. 17, no. 2, p. 124–129, 1971.

[2] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI 15*, 2015.

[3] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI 15*, 2015.

[4] L. Tan, K. Zhang, K. Wang, X. Zeng, X. Peng, and Y. Qiao, "Group emotion recognition with individual facial emotion cnns and global image based cnns," *Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017*, 2017.