Facial Expression Recognition Using Prior Attention

Ido Terner
ido.terner@gmail.com

Yair Gat
yairorigat@gmail.com

January 2022

Abstract

Over the past few decades, facial expression recognition has been an active research area, and it's still challenging. Traditional approaches for this problem rely on hand-crafted features, followed by a classifier trained on a database of images or videos. Due to the increased popularity of deep learning in recent years, several works proposed an end-to-end framework for facial expression recognition, using deep learning models. Despite the better performance of these works, there still seems to be great room for improvement. In this work, We propose an approach based on an attention mechanism in the prepossessing phase, even before training our network, which requires low resource consumption and is still able to achieve great performance over previous models on the FERG database.
link to GitHub: https://github.com/iTerner/Facial-Emotion-Detection

1 Introduction

Over the last few years, there is an increase in the desire to use deep learning tools to understand human behavior patterns. One of the most trendy challenges in this field is facial expression recognition. Facial expressions are a form of non-verbal communication. The immediate association when talking about human behavior is facial expression and, it's can tell us a lot about human emotions. Facial expression can be performed using different physical features, e.g. smiling, crying, raising eyebrows, and other nuances and subtlety. Recently, with the "AI winter", neural networks and especially, conventional neural networks (CNNs) are proved as undisputed experts in extracting and learning patterns in images.
Our main goal in this paper is to provide an algorithm that classifies the underlying emotion in the face images in low resource consumption and high accuracy. We use the fact that in the case of facial expression, much of the information comes from two major regions - the mouth region, and the eyes region. We propose an approach based on an attention mechanism in the prepossessing phase, even before training our network. Our attention mechanism creates two blocks,

the first block consists of the eyes and eyebrows, and the second one consists of the mouth. We simply create those blocks using a pre-trained model from the CV2 library which helps us to find those regions.
The challenge we try to solve is popular, and a lot of solutions have already provided sufficient performances. This fact makes this challenge a bit harder. However, our approach performance is remarkable in considering both efficiency and accuracy, based on preprocessing attention mechanism, simple CNNs, and MLPs.

2 Related Works

Earlier works on emotion recognition, rely on the traditional two-step machine learning approach, where in the first step, some features are extracted from the images, and in the second step, a classifier (such as neural network, SVM, or random forest) is used to detect the emotions. Some of the popular hand-crafted features are used for facial expression recognition. Then, a classifier would assign the best emotion to the image. These approaches seemed to work fine on simpler datasets, but with the advent of more challenging datasets (which have more intra-class variation), they started to show their limitation.
With the great success of deep learning, and more specifically convolutional neural networks for image classification and other vision problems, several groups developed deep learning-based models for facial expression recognition. Aneja et al [2] developed a model of facial expressions for stylized animated characters based on deep learning by training a network for modeling the expression of human faces, one for that of animated faces, and one to map human images into animated ones. Zhao [5] proposed an instance-based transfer learning method, which is a weighted ensemble transfer learning framework with multiple feature representations. Mutual information was applied as the smart weighting schema to measure the weight of each feature representation. Feutry in [3] introduced a novel training objective for simultaneously training a predictor over target variables of interest (the regular labels) while preventing an intermediate representation to be predictive of the private labels. The model aims to learn representations that preserve the relevant part of the information while dismissing information about the private labels which correspond to the identity of a person. In addition, Minaee [4] proposed a deep learning approach based on an attentional convolutional network, which is able to focus on important parts of the face.
All of the above works achieve significant improvements over the traditional works on emotion recognition, but they are using the raw data as the input. In this work, we try to modify the data, such as we only use the vital data from the image.

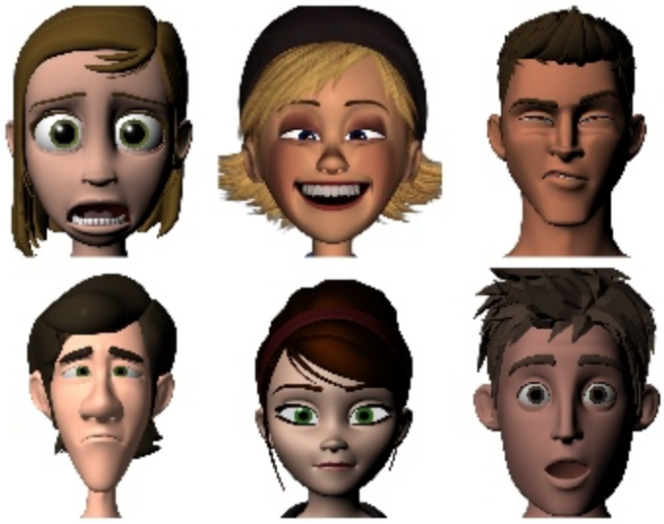

Figure 1: Sample images from the FERG database

3 Methods

We propose an end-to-end deep learning framework, based on attention mechanism in the preprocessing phase and simple CNNs and MLPs that will be described later, to classify the underlying emotion in the face images. Usually improving a deep neural network relies on increasing the network width, manipulating optimization methods, improving regularization, etc. However, we provide another perspective on the problem. Considering the fact that, not all the regions in the face can contribute us to reach classify emotions.

3.1 Attention Mechanism

Surprisingly, our attention mechanism doesn't require us to build any NN. We use the library OpenCV to detect the eyes, eyebrows, and mouth of a given face image. OpenCV provides a real-time optimized Computer Vision library, tools, and hardware, it provides us detected points of each organ in the face image. Then we divide the images into two blocks - the upper block which consists of the eyes and eyebrows, and the lower block which consists of the mouth. We define the blocks as described in the figure below: Figure 2 describes exactly how we create the blocks. Detect the organ points using OpenCV, define the blocks bound by the most extreme points in each block(we also add a margin of safety of 10 units in each axis), and extract those blocks from the original image.
it's worth mentioning that the mechanism described also can help in removing unwanted bias in images, e.g. background, ears, and other features with a low contribution to the task of classifying emotion.

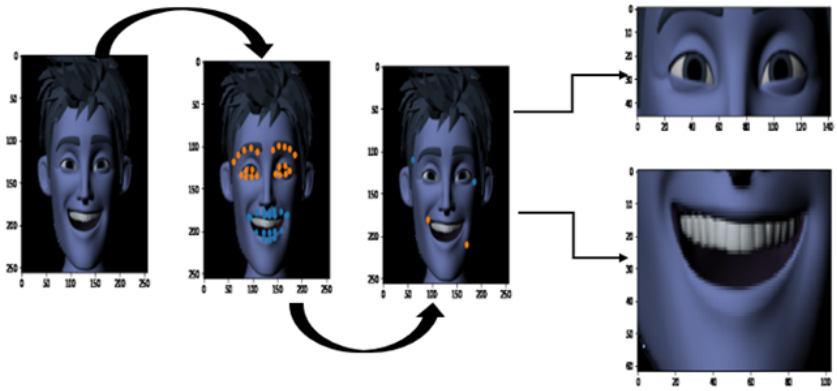

Figure 2: FlowChart of the Attention Mechanism

3.2 Architecture

Figure 3 illustrates the proposed model architecture. The feature extraction part consists of two convolutional layers, each two followed by rectified linear unit (ReLU) activation function, and after the first layer, we use Batch Normalization. They are then followed by a dropout layer and two fully-connected layers, each two followed by a ReLU activation, then, another dropout layer, which ends with another fully connected layer. From that, we extract feature vectors, which we concatenate, and then send to a similar network consisting of a dropout layer and two fully-connected layers, each two followed by a ReLU activation, then, another dropout layer, which ends with another fully connected layer.

This model is then trained by optimizing a loss function using Adam optimizer. The loss function in this work is cross-entropy loss. Adding dropout enables us to train our models from scratch even on very small datasets.

4 Results

In this section, we provide a detailed experimental analysis of our model on the FERG database. We first provide a brief overview of the database used in this work, we then provide the performance of our model on the database and compare the results with some of the promising recent works.

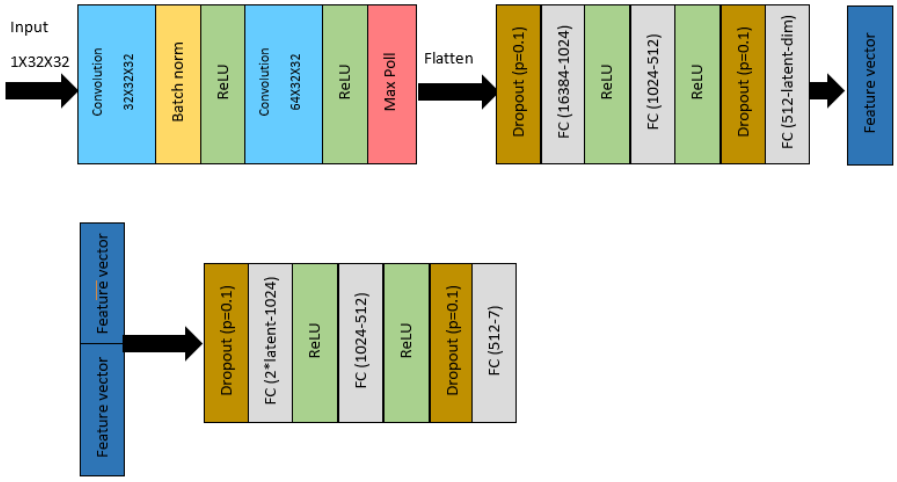

Figure 3: The Architecture of the proposed model. For both the upper image and the lower image we use the same architecture.

4.1 Database

In this work, we provide the experimental analysis of the proposed model on the Facial Expression Research Group Database (FERG). Before diving into the results, we are going to give a brief overview of this database.

4.1.1 FERG

FERG is a database of stylized characters with annotated facial expressions. The database contains 55,767 annotated face images of six stylized characters. The characters were modeled using MAYA. The images for each character are grouped into seven types of expressions [1], (happiness, sadness, anger, fear, disgust, surprise, neutral). Six sample images from this database are shown in Figure 1.

4.2 Experimental Analysis and Comparison

We will now present the performance of the proposed model on the FERG database. We train the model on a subset of that dataset, and validate on the validation set, and report the accuracy over the test set. Before getting into the details of the model's performance on the dataset, we briefly discuss our training procedure. The model was trained for 50 epochs from scratch. For optimization, we used Adam optimizer with a learning rate of 10^{-3}. Each epoch took approximately 55 seconds, which indicate the efficiency of our model. Each image we normalize with random Gaussian variables with zero mean and 0.5

standard deviation. In addition, we resize the image to a size of 32X32.
For the FERG dataset, we use around 44k images for training, 5k for validation, and 5k for testing. Each image we randomly assigned to the train, validation, or test set with a probability of 0.8, 0.1, and 0.1 respectively. We were able to achieve an accuracy rate of around 97.5%. The train and validation loss and accuracy are reported in Figure 5. The confusion matrix on the test set of FERG dataset is shown in Figure 4.
The comparison between the proposed algorithm and some of the previous works on FERG dataset are provided in Table 1.

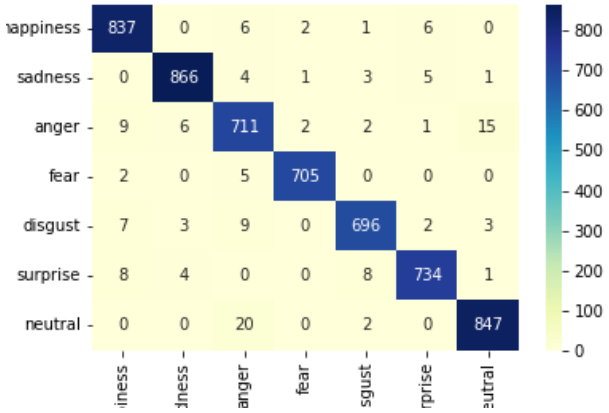

Figure 4: The confusion matrix on FERG dataset.

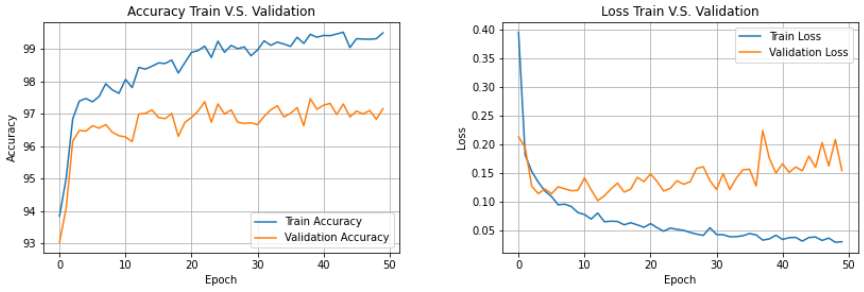

Figure 5: Comparison of the loss and accuracy of the proposed model on the train and validation sets.

Method	Accuracy Rate
DeepExpr [2]	89.02%
Ensemble Multi-feature [5]	97.00%
Adversarial NN [3]	98.02%
Deep Emotion [4]	99.3%
The proposed algorithm	97.50%

Table 1: Classification Accuracy on FERG dataset.

5 Discussion

This work proposes a new framework for facial expression recognition based on attention mechanism during the preprocessing phase. We believe attention is an important piece for detecting facial expressions, which can enable neural networks with less than 10 layers to compete with (and even outperform) much deeper networks for emotion recognition. Also, we provided an extensive experimental analysis of our work on a popular facial expression recognition database and showed promising results.

References

[1] Deepali Aneja, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*, pages 136–153. Springer, 2016.

[2] Deepali Aneja, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. Modeling stylized character expressions via deep learning. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato, editors, *Computer Vision – ACCV 2016*, pages 136–153, Cham, 2017. Springer International Publishing.

[3] Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel. Learning anonymized representations with adversarial neural networks, 2018.

[4] Shervin Minaee and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network, 2019.

[5] Hang Zhao, Qing Liu, and Yun Yang. Transfer learning with ensemble of multiple feature representations. In *2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 54–61, 2018.