Emotion Classification Using Deep Neural Networks

Joseph Boulis, Mohamed Hemdan, Ahmed Osama

Computer Science and Engineering Department, The American University in Cairo

Abstract

One of the major challenges that is found in the computer vision field is to detect the facial emotions of humans. Recently, in computer vision and machine learning, it's possible to detect emotion from video or image accurately. Hence, building on top of the numerous techniques from recent research, we demonstrate a state-of-the-art 75.4% accuracy on the FER2013 test set. In our project, we experimented with several implementations of deep learning models based on transfer learning for the problem facial expression recognition (FER). Thus, the proposed model is proven to be effective for facial emotion detection.

Dataset

The dataset used in this project is the Fer2013. Fer2013 is a well-studied dataset that contains 32,298 facial Greyscale images of different expressions with size restricted to 48×48, and the main labels of it can be divided into 7 types: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral. 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 training set consists of 28,709 examples and test set consists of 3,589 examples.

Models

Baseline Model:

A vanilla CNN was implememted using four 3x3x32 same-padding, ReLU filters, interleaved with two 2x2 MaxPool layers, and completed with a FC layer and softmax layer in addition to batchnorm and fifty percent dropout layers to properly address high variance and increase the accuracy from 53.0% to 64.0%.

Transfer Learning:

We used ResNet50, SeNet50 and VGG16 as the pre-trained models. For all of them, the original output layers were removed and 50% dropout is applied. For ResNet50 and SeNet50, all the layers, expect for the last 5 layers, are frozen, bsides two fully connected layers of size 4096 and 1024 with 50% dropout and a softmax output layer are added. Regarding the VGG16, the entire model is frozen and a fully connected layer of size 1024 with 50% dropout and a softmax output layer are added.

Ensemble:

After ensembling four models (ResNet50 with/without class weights, SeNet50 with/without class weights, VGG16 with/without class weights, and baseline model), we were able to achieve our highest accuracy of 75.4%.

Methods

Problems found in the dataset:

1) **Class imbalance:** problem is found in this dataset since the target class's frequency is highly imbalanced. The occurrence of images with Disgust emotion, as shown in Table(1), is very low compared to the other classes present. In other words, there is a bias or skewness against this minority class.
2) **Intra-class variation**: image variations occur between different images of one class. For instance, in the angry class, images vary from sketching to painting to cartoonistic to realistic images. Moreover, many changes in facial pose and subtle differences between expressions can be found.
3) **Occlusion**: the existence of multiple images with occlusion, where one or more objects come too close and seemingly merge or combine with the targeted image to be classified which lead to wrongly classifying the occluded face.
4) **Outliers**: some of the images in the dataset do not represent any emotion and act as outliers as result of the technique used to collect the dataset.

Methods to solve solve some of the mentioned problems:

1) **Class Weighting**: To solve this problem, we applied class weighting to reduce class imbalance, resulted in dropping misclassification rate from 61% to 34% for "disgust" class.

2) **Data Augmentation**: ±10 degree rotations, ±10% image zooms, and ±10% horizontal/vertical shifting, horizontal mirroring.

3) **Test-Time Augmentation**: TTA with horizontal flip and seven augmented images improved test accuracy by 1.7% on the five-layer model.

	Training	Test
Angry	3995	491
Disgust	436	55
Fear	4 4097	528
Happy	7215	879
Sad	4830	594
Surprise	3171	416
Neutral	4965	626

Table 1: Frequency of the samples for the Training and Test Datasets

EVALUATION

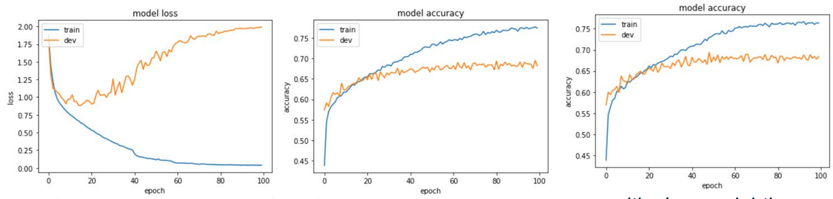
We experimented with a simple CNN model as our baseline model and some pre-trained models like the ResNet50, SeNet50 and VGG16 models. We used Transfer learning to fine tune these models to the Facial Emotion Recognition task. We used Class weighting to counter the problem of class imbalance and ensemble classification to increase the accuracy

The figure below shows the accuracy and cross entropy loss on the test set for all our chosen models. Results showed that there is no much gain obtained from class weighting. On the other hand, the results showed that ensembling the output of all models increases the accuracy.

The ensemble model reaches an accuracy of about 75.4\% which is quite better than the current solutions, 75.2\%.

Model Name	Accuracy		CE Loss	
	Using class weights	Without class weights	With class weights	Without class weights
VGG16	69.2%	70.23%	0.8557	0.8456
SeNet50	71.56%	72.54%	1.8335	1.866
ResNet50 model	72.4%	72.65%	0.7922	0.7965
Baseline Model	66.37%		0.9045	
ensemble	75.5%		0.6977	

Due to the large complexity of the models we are using and the small dataset size we are training the models, our model experiences overfitting, as shown in the figures. We can see that in most of the graphs, the validation accuracy stops improving early at epoch 20-40.



CE loss and accuracy of VGG without class weighting with class weighting

There are still challenges when it comes to emotion classification that even humans are still struggling with. Humans cannot distinguish some pictures to which class they belong. For example, here are some pictures that may have several labels, as shown in figures below. As these pictures have confused the labelers, it will for sure confuse the model.

Sadness and Fear

Surprise and Happiness

Future Work

We want to use more advanced techniques like Facial Landmark detection and alignment to reach a better input representation to our model and hence increase our model's learning capacity. In addition, we want to integrate a psychological emotion representation model like the Circomplex model to analyze emotions in terms of valence and arousal, so the model predicts them and infers the probability of each emotion. Furthermore, we want to counter the problem of small training data size. We will try to use more auxiliary data like AffectNet to combine with our FER13 dataset. This can prevent our model from over fitting.

Currently, we are developing an application that detects users' emotions using text and voice. We want to integrate our work in the facial emotion detection problem inside our application to accurately predict the emotion.

Conclusion

We started the project with the aim that we get an accuracy better than the current solutions. We have experimented with different pre-trained CNN models like ResNet50, SeNet50 and VGG16 and tried to fine tune them to our emotion classification problem. We have also experimented with the idea of class weighting to counter the problem of class imbalance. We have also built a simple CNN model, as our baseline. By combining the output of all these models, we have achieved an accuracy of 75.4\% which is better than the current techniques.

REFERENCES

[1] Faiza Abdat, Choubeila Maaoui, and Alain Pruski. 2011. Human-computer interaction using emotion recognition from facial expression. In 2011 UKSim 5th European Symposium on Computer Modeling and Simulation. IEEE, 196–201.

[2] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al

[3] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. IEEE transactions on affective computing (2020).