Enhancing Facial Emotion Recognition with Deep Neural Networks: A Transfer Learning and Data Augmentation Approach

Joseph Boulis
The American University in Cairo
Cairo, Egypt
Joseph_boulis@aucegypt.edu
+201288101673

Mohamed Hemdan
The American University in Cairo
Cairo, Egypt
mashrafhemdan@aucegypt.edu
+201097660870

Ahmed Osama
The American University in Cairo
Cairo, Egypt
ahmedosama001@aucegypt.edu
+201119856617

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.

KEYWORDS

FER13, emotion detection, CNN, deep learning

Figure 1: samples of the FER dataset

1 INTRODUCTION

Facial emotion identification is the process of identifying expressions that convey basic emotions such as fear, happiness, and sadness, etc. It can be used in many applications, including digital advertising, online gaming, customer feedback evaluation, and healthcare [2] [1]. Thanks to developments in computer vision, high emotion detection accuracy has been achieved in images taken under controlled conditions and in consistent environments (e.g. frontal faces and posed expressions), making this a solved problem with 98.9% accuracy. However, distinguishing basic expressions in natural conditions is still challenging due to variations in head pose, illumination, and occlusions [7].

Recently, with the advent of deep learning, computer vision research aimed to improve classification accuracy remarkably in such scenarios, exceeding human level performance. As a result, numerous groundbreaking applications have been developed in the fields of sociable robotics, medical treatment, driver fatigue surveillance, and many other human-computer interaction systems [6].

The purpose of this paper is to understand the FER problem better and improve the performance of emotion recognition models. To achieve such a goal and improve the overall accuracy, a couple of techniques found in the literature were utilized, including data augmentation, class weighting, ensembling, and transfer learning (fine-tuning and trained hyperparameters of ResNet50, SeNet50, and VGG16).

In this paper, we refer to the GitHub repository for the DeepEmoNet Facial Emotion Recognition project implementation [1].

[1]https://github.com/josephhany/DeepEmoNet-Facial-Emotion-Recognition

2 RELATED WORK

The competition of FER13 was started by Goodfellow et al which is considered one of the most famous competitions in the field of emotion detection especially for images. Fellow had built the FER2013 dataset as a Kaggle competition. This inspired scientists and researchers in the field of machine learning and computer vision to design and build better and efficient emotion detection systems. All three winners in this competition used discriminatively trained CNNs with image translations [3]. The first place winner, Yichuan Tang, used the L2-SVM loss function and the primal goal of SVM as the loss function for training to achieve a 71.2 percent accuracy [9]. This was a brand-new development at the time, and it performed admirably on the competition dataset.

In a new survey research on FER, S. Li and W. Deng shed light on the present state of deep-learning-based FER techniques [6]. Furthermore, Pramerdorfer and Kampel [7] analysed six current state-of-the-art studies and ensembled their networks to achieve 75.2 percent test accuracy on FER2013, which is the most accurate classification obtained in a published publication to our knowledge.

Zhang et al. used auxiliary data and additional features to achieve an accuracy of 75.1 percent among the six papers: a vector of HoG features was generated from face patches and processed by CNN's first fully connected layer. In addition, they used facial landmark registration which has advantages even in difficult situations [10]. Finally, Kim et al. used face registration, data augmentation, extra features, and ensembling in their study with the second best accuracy [5].

3 DATASET

The dataset used in this project is the Fer2013, shown in Fig(1). 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 the public test set consists of 3,589 examples [3].

4 METHODS

After doing Exploratory Data Analysis, we figured out that four major problems exist.

First, class imbalance problem is found in this dataset since the target class's frequency is highly imbalanced. To be specific, 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.

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

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

Second problem is the Intra-class variation, where 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.

Third, 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.

Fourth, as a result of the technique used to collect the FER2013 dataset, namely through gathering the results of a Google image search of each emotion and synonyms of the emotions, some of the images in the dataset do not represent any emotion and act as outliers.

Below we some of the proposed solutions will be listed.

4.1 Data Augmentation

Since the dataset has low number of training examples, a good solution is to augment the dataset with appropriate data augmentation (DA) techniques, which are typically based on either geometric transformation or generative adversarial networks. Given our limited computational power and time, we decided to augment the data using commonly used geometric transformation techniques in existing FER papers. As a result, we achieved the best results with the following parameters: horizontal mirroring, ±10 degree rotations, ±10% image zooms, and ±10% horizontal/vertical shifting.

4.2 Class Weighting

In order to solve the problem of class imbalance, we had to modify the current training algorithm to take into account the skewed distribution of the classes. This was achieved by using the class weighting technique in which we give different weights to both the majority and minority classes. The difference in weights will influence the classification of the classes during the training phase. The aim is to penalize the misclassification made by the minority class by setting a higher class weight and at the same time reducing weight for the majority class. Thus, we assigned weights inversely proportional to the number of samples found in each class. As a result of this technique, for the disgust class, we were able to drop the misclassification rate from 61% to 34%.

4.3 Ensembling Test-Time Augmentation (TTA)

A soft voting ensemble involves summing the predicted probabilities for class labels and predicting the class label with the largest sum probability. As a result of performing an ensemble with soft voting of seven models, our highest test accuracy significantly improved from 73.2% to 75.4%. Likewise, Test-Time Augmentation was applied with horizontal flip and seven augmented images which resulted in improving the test accuracy of our five-layer model by 1.7%.

5 MODELS

5.1 Baseline Model

Our first approach to understand the problem is to develop a deep learning model from scratch. Hence, we built a vanilla CNN using four 3x3x32 same-padding, ReLU filters, interleaved with two 2x2 MaxPool layers, and completed with a FC layer and softmax layer. In addition, we added batchnorm and fifty percent dropout layers to properly address high variance. Accordingly, our accuracy was raised by 11% from 53.0% to 64.0%.

5.2 Transfer Learning

After noticing that the FER2013 dataset is a small and imbalanced dataset, we thought of employing transfer learning to boost the accuracy of our model. Testing this idea, we used the Keras VGG-Face library and the pre-trained models ResNet50, SeNet50, and VGG16 to explore the effect of using transfer learning models on the results. Throughout the training period, we recolored and resized the 48x48 grayscale images in FER2013 to meet the input criteria of these new networks, which required RGB images no smaller than 197x197 pixels.

5.2.1 . ResNet50 Fine Tuning

ResNet50 is a variant of the ResNet model which has 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer. It is a widely used ResNet model of which we can load a pre-trained version for the network trained on more than a million images from the ImageNet database.

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	

Figure 2: The Accuracy and Cross-Entropy Loss of the models we have used with and without class weighting. Clearly, the ensemble model gives the highest accuracy

With that being known, ResNet50 was the first pre-trained model we decided to explore. Our first experiment was to reproduce the work done by Brechet et al [4]. Hence, we added two fully connected layers (of sizes 4,096 and 1,024 pixels) to the original output layer, as well as a softmax output layer with seven emotion classes.

Regarding the input, the first 170 layers were frozen, but the rest of the network could still be trained. We utilised stochastic gradient descent as our optimizer with a learning rate of 0.01 and a batch size of 32. After 122 epochs of training the model with SGD at a 0.01 learning rate and a batch size of 128, we reached an accuracy of 73.2% on the test dataset.

5.2.2 . SeNet50 Fine Tuning

Another pre-trained model that we explored was the SeNet50. Since SeNet50 is a deep residual network with 50 layers and has an architecture which is very similar to that of ResNet50, not much time was spent on tuning it. Hence, the same parameters used for the ResNet50 were used with the SenNet50 by which we reached 72.5% accuracy on the test dataset.

5.2.3 . VGG16 Fine Tuning

The last transfer learning model we experimented on was the VGG16. Compared to ResNet50 and SeNet50, VGG16 is a shallower model, yet contains more parameters and is more complex. In this experiment, we kept all the pre-trained layers frozen. However, we added two fully connected layers, with sizes 4096 and 1024, each with fifty percent dropout. As a result, on the test dataset, we got a 70.2% accuracy after 100 epochs of training with the Adam optimizer.

6 RESULTS

We first experimented with a simple CNN model and hyper-parameter tuned it to act as our baseline model when compared to the target model. We have also used some pre-trained models like the ResNet50, SeNet50 and VGG16 models. We have used Transfer learning to fine tune these models to the Facial Emotion Recognition task. We have also used Class weighting to counter the problem of class imbalance. Each model is trained twice: with and without class weighting. In order to get better accuracy, we have used the ensemble classification by combining the output of all of these models and use majority voting to decide on the class.

Figure (2) shows the accuracy and cross entropy loss on the test set for all our chosen models. The figure also showed the accuracy

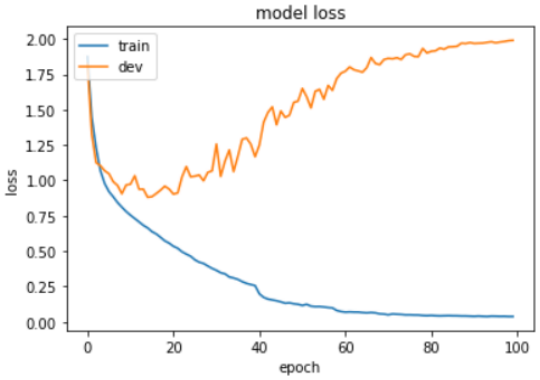

Figure 3: Validation and Test Cross Entropy Loss using the SeNet50 Model and using Class Weigting

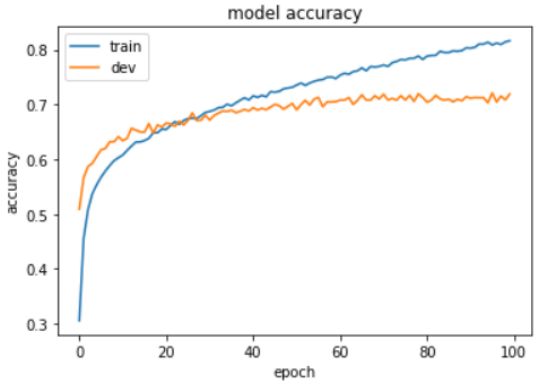

Figure 4: Validation and Test Accuracy using the ResNet50 Model without using class weighting

and loss with and without class weighting is used. Results showed that there is no much gain obtained from class weighting which indicates that this class imbalance may not be the largest contributor to this accuracy level. On the other hand, the results showed that combining 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. According to a paper by Pramerdorfer and Kampel [8], they had achieved an accuracy of about 75.2% by assembling six different classifiers/ networks which is one of the well-known models tackling the problem of Facial Emotion Detection.

6.0.1 . Challenges

Looking at the learning curves of our model, we have found that our model sometimes experiences overfitting as shown in figure (3).

Due to the large complexity of the models we are using and the small dataset size we are training the models, over-fitting is most probable. This causes the validation accuracy to flatten out at early epochs, as shown in the accuracy graphs in these figures (4, 5, 6, 7, 8, 9). We can see that in most of the graphs, the validation accuracy stops improving early at epoch 20-40.

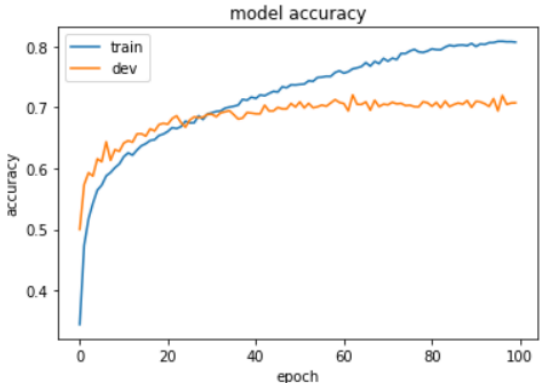

Figure 5: Validation and Test Accuracy using the ResNet50 Model using class weighting

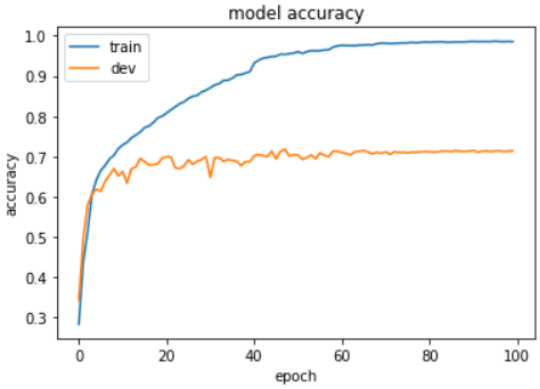

Figure 6: Validation and Test Accuracy using the SeNet50 Model without using Class Weigting

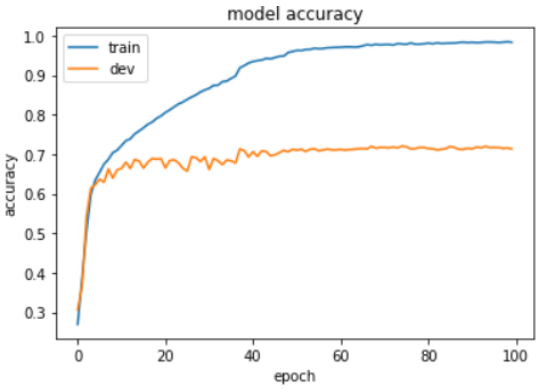

Figure 7: Validation and Test Accuracy using the SeNet50 Model using Class Weighting

There are still challenges when it comes to emotion classification which 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 (10, 11). As these pictures have confused the labelers, it will for sure confuse the model.

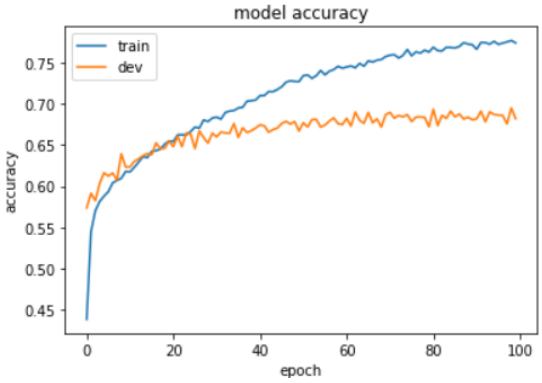

Figure 8: Validation and Test Accuracy using the VGG16 Model and without using Class Weigting

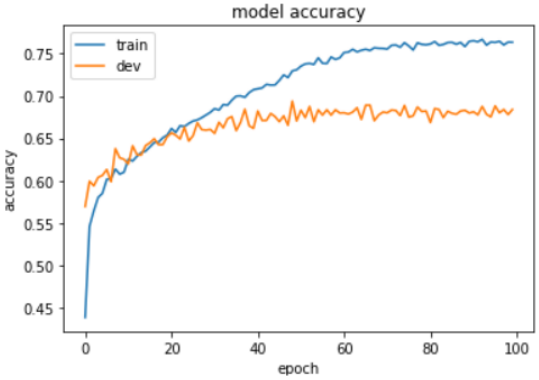

Figure 9: Validation and Test Accuracy using the VGG16 Model and using Class Weigting

Figure 10: This woman can be labeled as experiencing sad or fear

7 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

Figure 11: This picture shows a man that is either happy or surprised

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.

8 FUTURE RECOMMENDATIONS

In this section, we explain the future work we aim to do to improve our accuracy. First, we aim 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 future work, we would like to explore more the psychological background of emotions and emotion representation. In specific, We want to integrate a psychological emotion representation model like the Circomplex model to our model to analyze emotions in terms of valence and arousal. Instead of training the model to predict the emotions in terms of classes, which limits the number of emotions, the model can be trained to predict the valence and arousal depicted on the input face and then we can infer the probability of each emotion using the Circomplex model.

From our research recently, we have found that the problem of small training data size can be mitigated by using auxiliary data. We can merge the FER13 dataset with other data sets like AffectNet to increase the training size of our data and prevent our model from over fitting.

As a community deployment project, we are currently 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 so that the application can use the text, image, and voice to accurately predict the emotion.

ACKNOWLEDGMENTS

This project is part of course work for CSCE 4604 (Practical Machine Deep Learning) course. Thanks for Professor Mohamed Moustafa for providing supervision and feedback throughout the course.

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] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. 2003. Real time face detection and facial expression recognition: development and applications to human computer interaction.. In *2003 Conference on computer vision and pattern recognition workshop*, Vol. 5. IEEE, 53–53.

[3] 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. 2013. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*. Springer, 117–124.

[4] Amil Khanzada, Charles Bai, and Ferhat Turker Celepcikay. 2020. Facial expression recognition with deep learning. *arXiv preprint arXiv:2004.11823* (2020).

[5] Bo-Kyeong Kim, Suh-Yeon Dong, Jihyeon Roh, Geonmin Kim, and Soo-Young Lee. 2016. Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 48–57.

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

[7] Christopher Pramerdorfer and Martin Kampel. 2016. Facial expression recognition using convolutional neural networks: state of the art. *arXiv preprint arXiv:1612.02903* (2016).

[8] Christopher Pramerdorfer and Martin Kampel. 2016. Facial Expression Recognition using Convolutional Neural Networks: State of the Art. arXiv:1612.02903 [cs.CV]

[9] Yichuan Tang. 2013. Deep learning using support vector machines. *CoRR, abs/1306.0239* 2 (2013).

[10] Zhanpeng Zhang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. 2015. Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision*. 3631–3639.