Real Time Emotion Detection & its Applications in Mental Health Monitoring

Team 4 (Multicollinearity)
Abhishek Baral, Ravitashaw Bathla, Yiran Chen, Nathan Warren, Jiayue Xu

Abstract

In this paper, we reproduced a real-time emotion classifier based on facial expressions from the study conducted by Arriaga et al. (2017) on the FER 2013 dataset. We trained a light-weight convolutional neural network consisting of residual and depthwise convolutional layers which resulted in a reduced number of model parameters for real-time predictions, as claimed by the study. We closely reproduced an accuracy of 65% vs. 66% from the study, which is similar to the human accuracy of 65% (+/-5%). In addition, by updating the labels based on the crowd-sourced FERPlus labels, we achieved an improved accuracy of 81% with our light-weight model, compared to state-of-the-art accuracy 88% which uses a more complex ensemble method. This research allows for further application of this model onto less computationally-abled devices for real-time emotion tracking in assistance of mental illness diagnosis by utilizing facial expressions in relation to an individual's mental state.

Introduction

Mental health has become one of the largest concerns amongst all age groups. Approximately one in five adults suffers from mental illness, and recent studies suggest that only 25 percent of people seek help while the majority are left untreated (Georgescu et al., 2012; NAMI, 2019). Physical health applications are able to track users' fitness by collecting data on heart rate, distance ran, etc. while mental health applications are limited to user input emotion logs without any tracking of users' biomarkers related to emotional wellbeing (Krebs & Duncan, 2015). Research has shown that facial expressions can play an important role in discerning whether an individual is suffering from depression. Studies suggest that individuals that suffer from depression appear to have a "neutral" facial expression that is evaluated as sadder by others, when compared to healthy control groups (Bourke et al., 2010). Other studies suggest facial features from photos of individuals with mental disorders are rated as more depressed, anxious, and angry (Daros et al., 2016). This indicates that there may be common facial features of those who are mentally ill that can be identified through artificial intelligence.

With recent developments in deep convolutional neural networks, several variants of such networks have been introduced, which provide high accuracy for emotion and face detection. However, these deep networks rely on devices with high computational power for training and predictions which pose a limitation to introducing such networks in real-time for mobile devices or devices with limited hardware support. In this paper, we present a real-time emotion detection application using a depthwise separable convolution neural network which is lightweight and can be integrated in several devices lacking significant hardware capability (LeCun et al., 2015).

Background

In the last few decades, there has been a surge in the use of deep neural networks (DNN) for image classification, segmentation, and object detection (Ranganathan et al., 2017; Alom et al., 2018). Such networks, namely VGG16 (Simonyan & Zisserman, 2014), ResNet18 (Zhang et al., 2016), and GoogleNet (Krizheysky et al., 2012) have provided state-of-the-art performance, however the computational cost of training, due to a large number of parameters, is high. Depthwise convolution was introduced by Sifre & Mallat, 2014, which reduces the number of parameters and was subsequently used in the first few layers of Inception (CNN) to reduce the computational cost (Szegedy et al., 2014). Xception network scaled up the depthwise separable filters (Chollet, 2017). MobileNet was introduced for mobile vision application and it is based upon the building blocks of depth wise separable convolutions (Howard et al., 2017).

Initial research on the FER 2013 dataset suggests that the accuracy from Deep Learning using SVM was 71.6% (Tang, 2013), and human accuracy was around 65(+/-)5% (Goodfellow et al., 2016). Further research over the years on these deep learning techniques revealed an accuracy of 74% using learned features from CNN combined with handcrafted features extracted by the bag-of-visual-words with localized SVM (Georgescu et al., 2019). FERPlus dataset was introduced by Microsoft which improved the ground truth labels on the same images from the FER 2013 dataset (Barsoum et al., 2016). The current state-of-the art accuracy for FERPlus is 87.6% using the same techniques in FER 2013 dataset (Georgescu et al., 2019). For real-time emotion detection on FER dataset, Siqeira et. al suggests incorporating CNN based ensemble with shared representations (ESRs) which have been shown to improve performance by reducing the redundancy and computational complexity from varying the branching level of ESRs (Siqueir et al., 2020). Goodfellow, 2013 presents a regression-based CNN approach to understand the spatial features of the facial expressions to understand the emotional intensities that are perceived by an observer and felt by the subject.

Data

Facial Expression Recognition 2013 (FER 2013) is an open source dataset consisting of 35,887 images. Each image was a 48x48 grayscale image of participants facing directly into a camera and produced seven expressions: happy, sad, surprise, fear, disgust, angry, and neutral. The data distribution for different expressions are shown in Fig. 1. The dataset was originally created using a google image search API, which matched a set of 184 emotion related keywords to capture the seven basic expressions. The labels were then manually validated and corrected by human labelers.

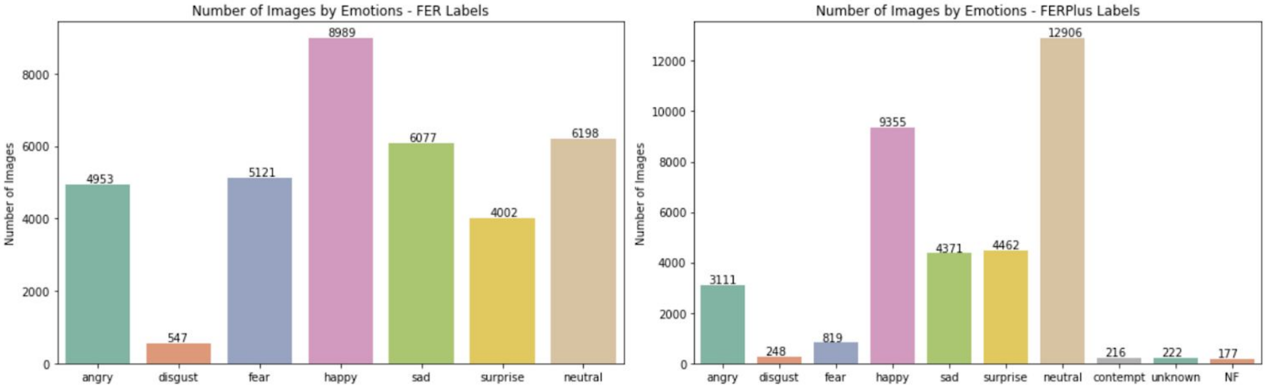


Figure 1: Emotion distribution in FER 2013 (left) and in FERPlus (right)

FERPlus was introduced as an improvement on the labels on the existing FER 2013 dataset. The dataset was relabeled by 10 crowd sourced taggers, where each image is tagged with one single emotion per tagger, and using majority vote, a single label is produced (Barsoum et al., 2016). By crowdsourcing emotion labels for each facial expression, the variability of interpretation is reduced in the FERPlus dataset, resulting in labels that more accurately match the facial expressions in the image (Fig. 2). In case of a tie in votes, the first occurrence of the maximum votes is chosen as the majority vote in the order of how the emotion is arranged in the index. The sequence of the same seven emotions is as follows: neutral, happy, surprise, sad, anger, disgust, and fear.

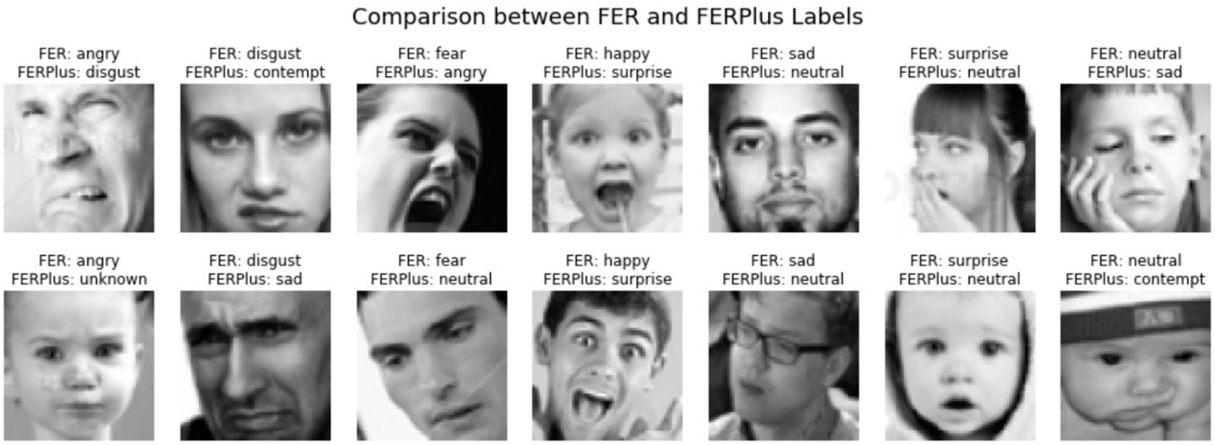


Figure 2: Comparison between FER 2013 and FERPlus labels on the same images, where the FERPlus labels are obtained via majority vote

Challenges with Facial Data

The main challenge faced is that facial recognition raises privacy concerns, and because of this, datasets of facial images are generally difficult to obtain. The second challenge related to training data was the uneven distribution of the emotion labels. In particular, there are few images labelled as disgust which might result in incorrect predictions, due to sparsity. Another challenge relates to the complexity of human facial expressions, as faces may display more than one emotion. It can be seen from fig. 3 that for some expressions, there are ties from the majority vote method suggesting multiple interpretations to the same image. The number of images that consisted of a tie was 1,776 or roughly 5% of the data, which could potentially be misclassified. There exist some images where facial features are covered by hand or watermarks, which makes it hard for the model to recognize emotion features. Facial accessories such as glasses and demographics of the target, could introduce biased judgements to the model which may lead to poor performance in real world applications. There also exists bias towards a prediction of neutral as the majority vote method selects the first label in order in case of a tie. This was also evidenced in figure 1, as the percentage of neutral faces in FERPlus is significantly higher than in the FER dataset.

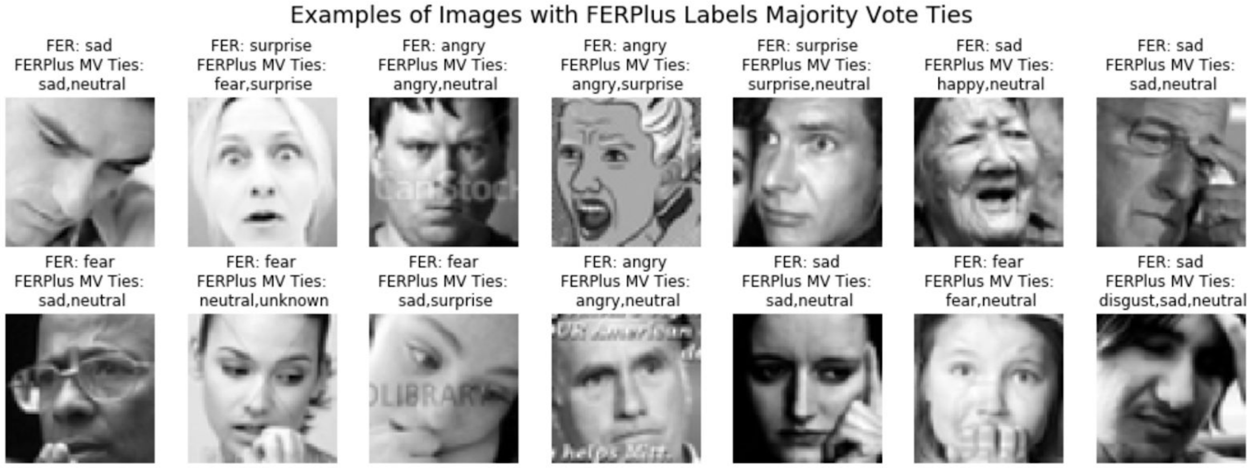

Examples of Images with FERPlus Labels Majority Vote Ties
FER: sad FERPlus MV Ties: sad,neutral
FER: surprise FERPlus MV Ties: fear,surprise
FER: angry FERPlus MV Ties: angry,neutral
FER: angry FERPlus MV Ties: angry,surprise
FER: surprise FERPlus MV Ties: surprise,neutral
FER: sad FERPlus MV Ties: happy,neutral
FER: sad FERPlus MV Ties: sad,neutral
FER: fear FERPlus MV Ties: sad,neutral
FER: fear FERPlus MV Ties: neutral,unknown
FER: fear FERPlus MV Ties: sad,surprise
FER: angry FERPlus MV Ties: angry,neutral
FER: sad FERPlus MV Ties: sad,neutral
FER: fear FERPlus MV Ties: fear,neutral
FER: sad FERPlus MV Ties: disgust,sad,neutral

Figure 3: Example images with FERPlus labels where the majority vote method generates ties

Nearly all the underlying issues that were mentioned above can be remedied by increasing the amount of data needed for training. Specifically, additional data is needed to address concerns of race, age, hand positions, and facial accessories. Based on the limited data, data augmentation could be used to avoid overfitting. Lastly, to address the complexity of human facial expressions, a potential solution could be to return multiple top emotional labels for one facial expression instead of a single tag.

Overview

The overview of the approach in this paper is presented in Fig. 4. We reproduced the original study using the FER 2013 dataset and then introduced FERPlus labels to compare the performance of the model on these labels.

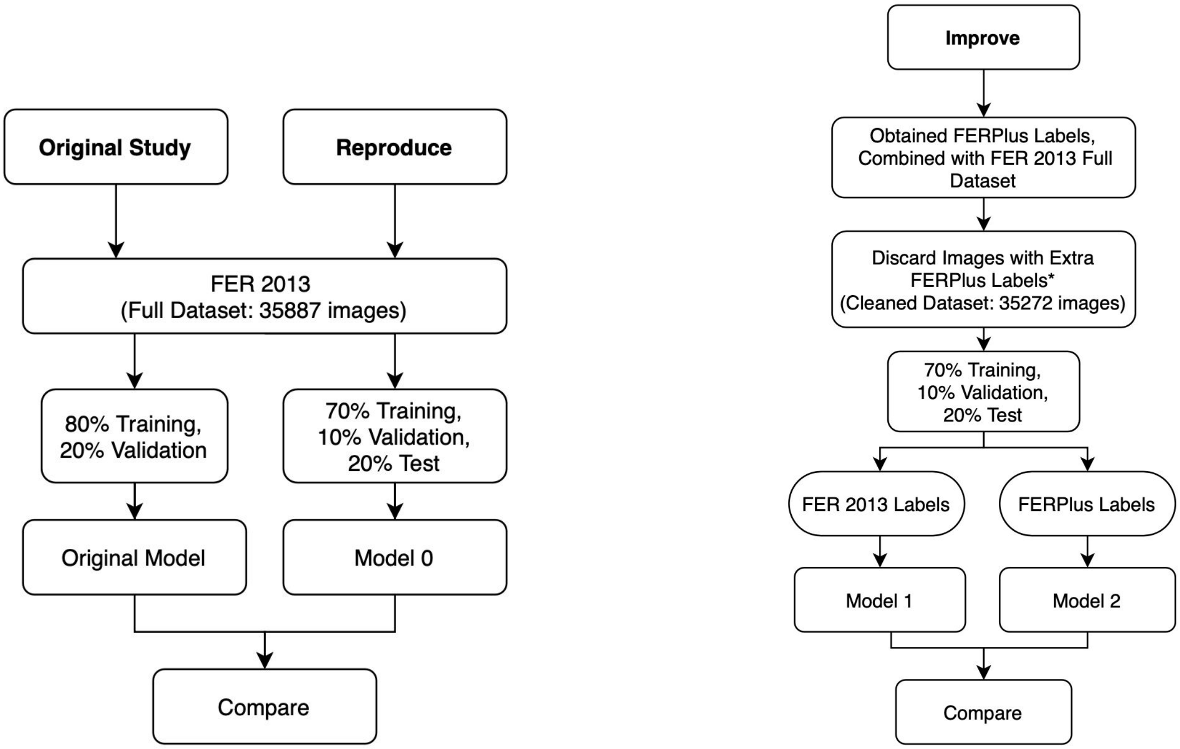


Figure 4: Flow chart of methodology. First we reproduced the original study using FER 2013 full dataset the data was split into training, validation, and test sets in a 70-10-20 split. *removed not-a-face, unknown, contempt labeled images.

Data Preprocessing

Following the approach in the original study (Arriaga et al., 2017), all images were resized to 64x64 in order to help find additional features that may be underrepresented in a 48x48 size (Hasemi, 2019; Topor, 2020). The pixel values were normalized to a range of (-1, 1) to employ equal weightage to every pixel and for the positive invariant nature of ReLU activation function (Meng et al., 2018).

In the original study, the FER 2013 dataset was split into 80% training and 20% validation sets. The training data was used for model building and performance was reported using validation set without a test set, which may not give an accurate estimate of the generalization performance. To overcome this, we split the data into 70% training, 10% validation, and 20% test set, while maintaining similar class distribution between the original study and our reproduction for comparison purposes (Fig. 5, Table 1, Table 2).

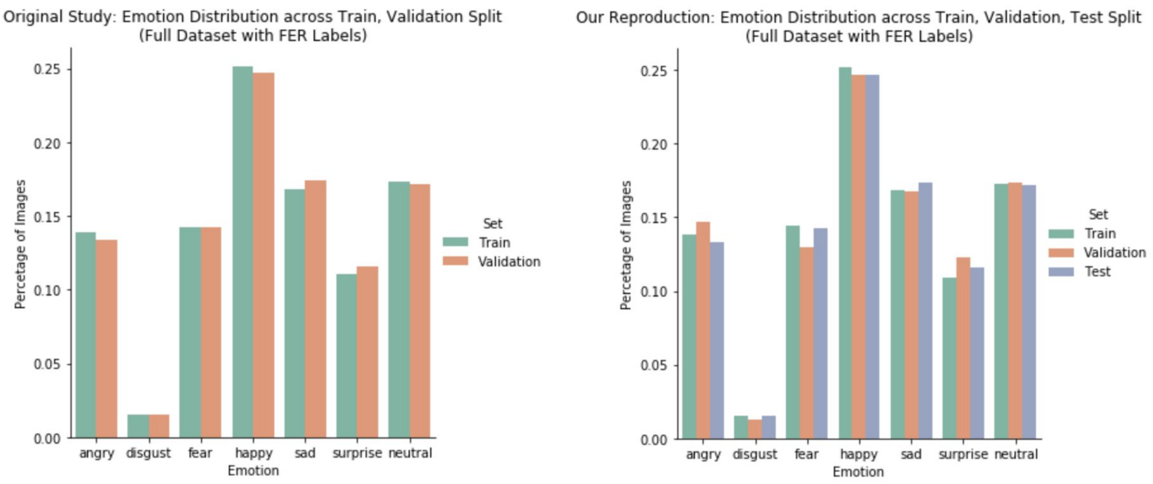


Figure 5: Normalized class distribution of the original study (left) and our reproduction (right)

Emotion	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Train	3995	436	4097	7125	4830	3171	4965
Validation	958	111	1024	1774	1247	831	1233

Table 1: Class distribution of image counts in the original study

Emotion	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Train	3468	391	3632	6328	4229	2731	4341
Validation	527	45	465	887	601	440	624
Test	958	111	1024	1774	1247	831	1233

Table 2: Class distribution of image counts in our reproduction

After combining FERPlus labels to FER 2013 dataset, 615 erroneous images with contempt, unknown, not-a-face FERPlus labels were discarded. For both FER and FERPlus labels, the resulting split on the cleaned dataset maintained similar class distribution across training, validation, and test sets (Fig. 6).

Data augmentation was implemented in real-time for the model fitting process, including 10-degree range for random rotation, 0.1 separate horizontal and vertical random shifting, random zooming of 0.1, and random horizontal flipping (Wang & Perez, 2017; Simard et al., 2003). The purpose of this is to create additional images from the original data to build a more robust dataset to train on (Li & Deng, 2020).

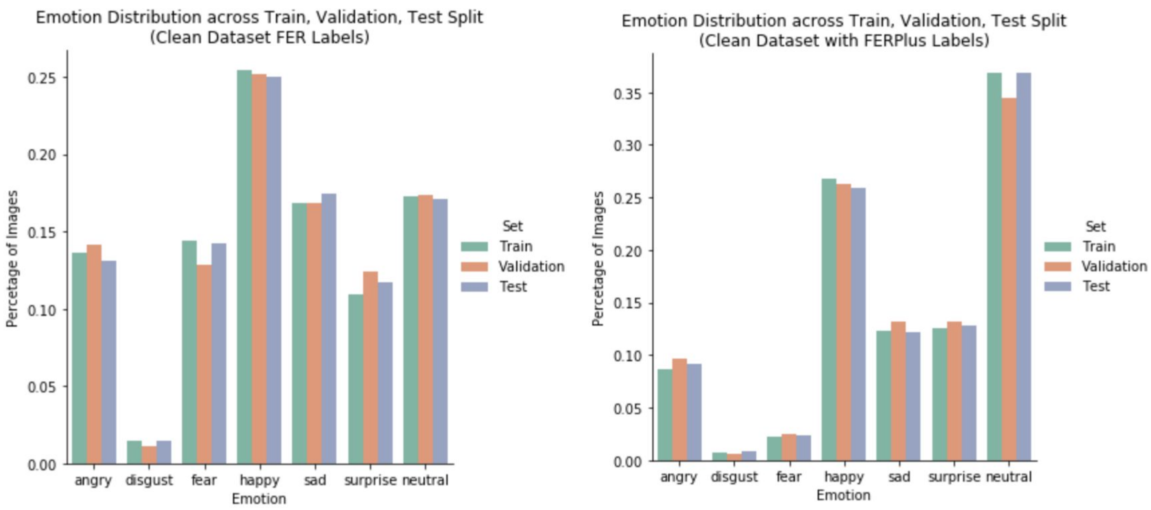


Figure 6: Normalized class distribution after removing images with extra labels (contempt, unknown, not-a-face) with FER labels (left) and with FERPlus labels (right)

Model

For frontal face detection, the pre-trained classifier with Haar-like features were used from the OpenCV module. This algorithm moves a window systematically over an image and detects the face by applying Haar-like features. On top of the facial detection, an emotion classifier was built.

A custom network architecture similar to Xception is used with each convolution layer followed by residual layer and depthwise convolution, ReLU activation, and batch normalization. The residual layer takes the difference of two consecutive layers to help the model learn the difference of features in preceding layers (He et al., 2016). ReLU activation was used as it enables sparse and more efficient activations than tanh or sigmoid would. The depthwise convolutional layer (Sifre & Mallat, 2014) assists in reducing the total number of network parameters by first convoluting using a Pointwise Convolution and later applying spatial convolution on each dimension. This resulted in reduced dimensions after the Pointwise Convolution and hence the number of model parameters are reduced. The illustration of depthwise convolution is represented in figure 7. All fully connected layers were removed by adding a global average pooling layer and the averaging was done by using the same number of output classes and a softmax activation function at the end. The complete architecture for the model, called mini-Xception, is represented in figure 8 (Arriaga et al., 2017).

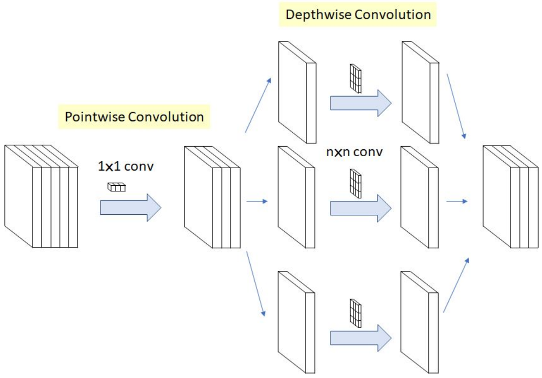


Figure 7: Depthwise convolution diagram (Sifre & Mallat, 2014)

Model Parameter Selection

The multiclass cross-entropy loss was used as the cost function. ADAM was chosen as an optimizer for backpropagation as dynamic learning rates can be applied for different parameters from the first and second moments of the gradients (Szegedy et al., 2015). As per the original study, the stopping criteria was set where if there was no improvement on validation loss after 50 epochs, then the training would terminate (Arriaga et al., 2017).

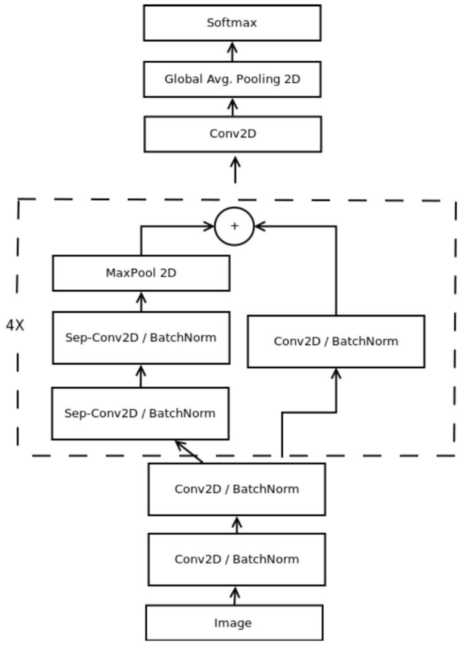


Figure 8: Model Architecture for Real-Time Application (Arriaga et al., 2017)

Model Evaluation Metrics

A confusion matrix for the multi-class classification problem was created to measure the generalization performance of the model. A list of F1-score, Precision and Recall are presented to analyze accuracy for each individual emotion as well as an overall prediction accuracy. To further assess the qualitative performance of our model, misclassified images were manually analyzed after passing through the final mini Xception classifier.

Results

We were able to reproduce the validation accuracy of 66% claimed by the original paper, with test accuracy of 65% of Model 0. Our test set performance shares similar prediction accuracy for each emotion compared with those from the original study (Fig. 9). The minor difference between the two was likely due to the difference in splitting criteria as described above.

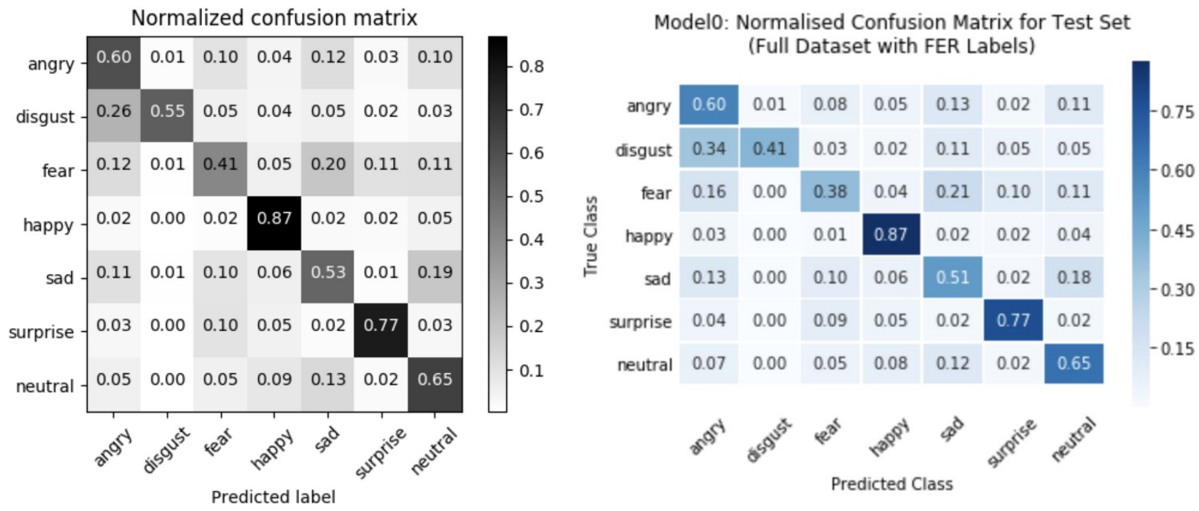


Figure 9: Normalized confusion matrix on FER 2013 dataset (left) (Arriaga et al., 2017) and our replication of the project (right, Model 0: with full dataset and same labels)

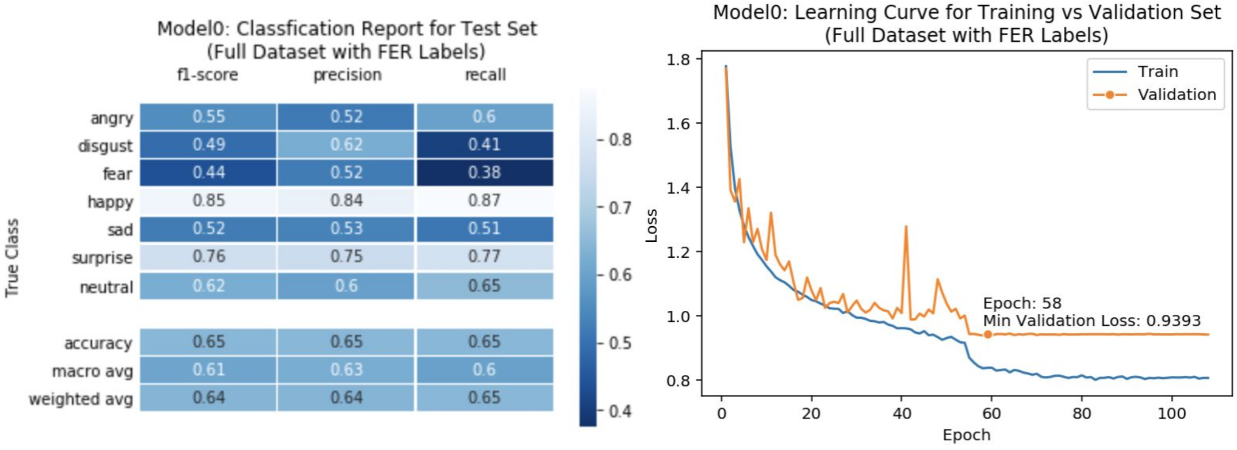


Figure 10: Classification report for test set (left), and learning curve for training vs. validation set (right) from Model 0: with full dataset and same labels

After successfully replicating the results from Arriaga et al., 2017, we compared model performance trained on FER 2013 labels (Model 1) and FERPlus labels (Model 2) from the cleaned dataset. The results obtained from both models suggest that overall Model 2 is performing better, in terms of class-specific accuracy, F1-score, precision and recall, and overall accuracy.

From the confusion matrices and the classification reports of both models, ‘disgust’, ‘fear’, and ‘sad’ have rather low prediction accuracy, while ‘angry’, ‘happy’, ‘surprise’, and ‘neutral’ can be predicted with fairly high accuracy. This is likely resulting from unbalanced emotion classes in the training data where both ‘happy’ and ‘neutral’ make up 60% of the dataset while ‘disgust’ and ‘fear’ make up less than 5%.

On further analyzing specific classes, we observe that the accuracy for correct predictions has improved among all the classes except for ‘disgust’, which has very limited training images (Fig. 11). The number of class pairs with more than 10% misclassified images has been greatly reduced by updating to FERPlus labels. Similarly, we observe that the class-specific F1-score, precision and recall has improved across all the classes except for ‘disgust’ (Fig. 13).

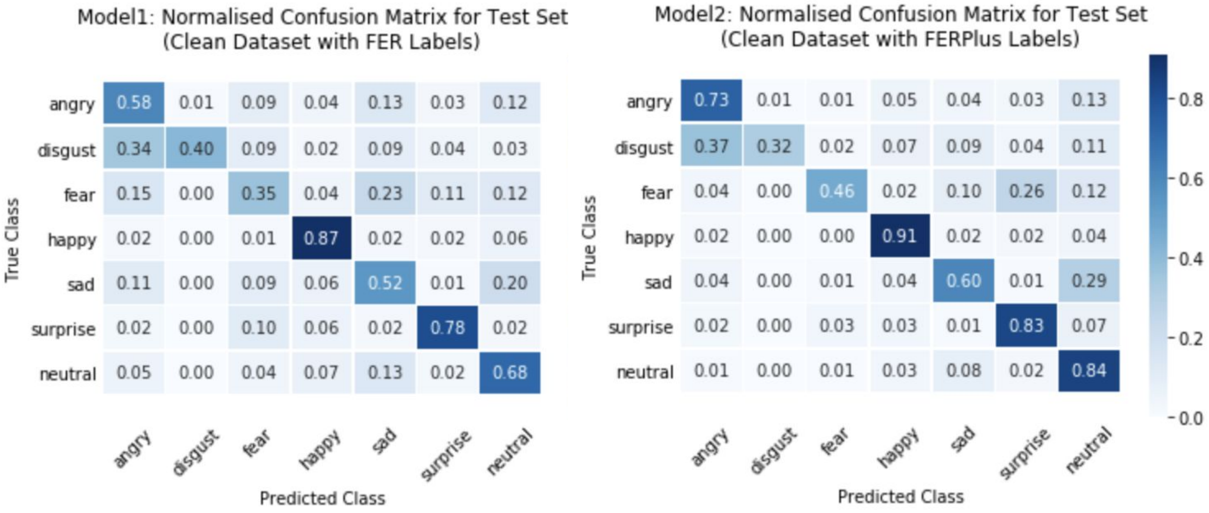


Figure 11: Normalized confusion matrix from Model 1 trained on FER labels (left), normalized confusion matrix from Model 2 trained on FERPlus labels (right), both trained on cleaned dataset discarding extra labeled images

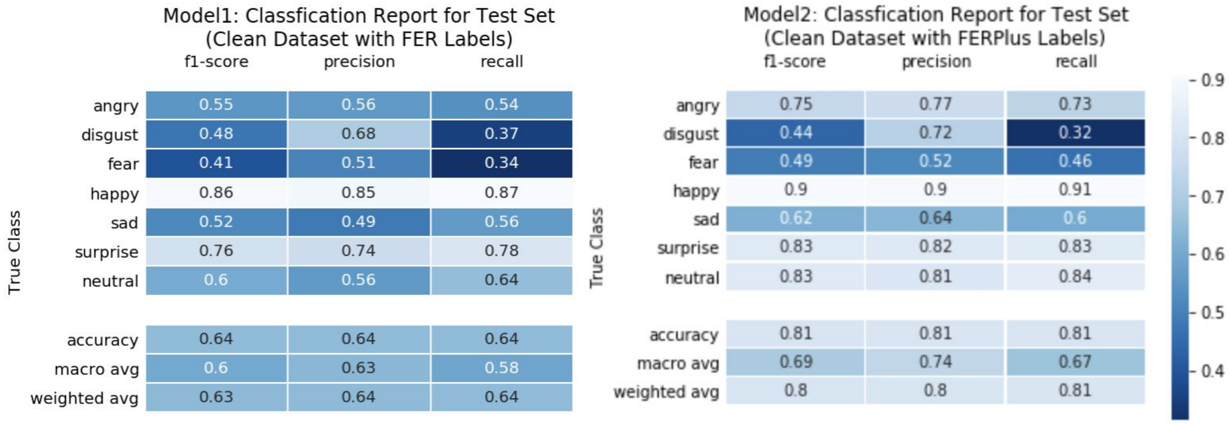
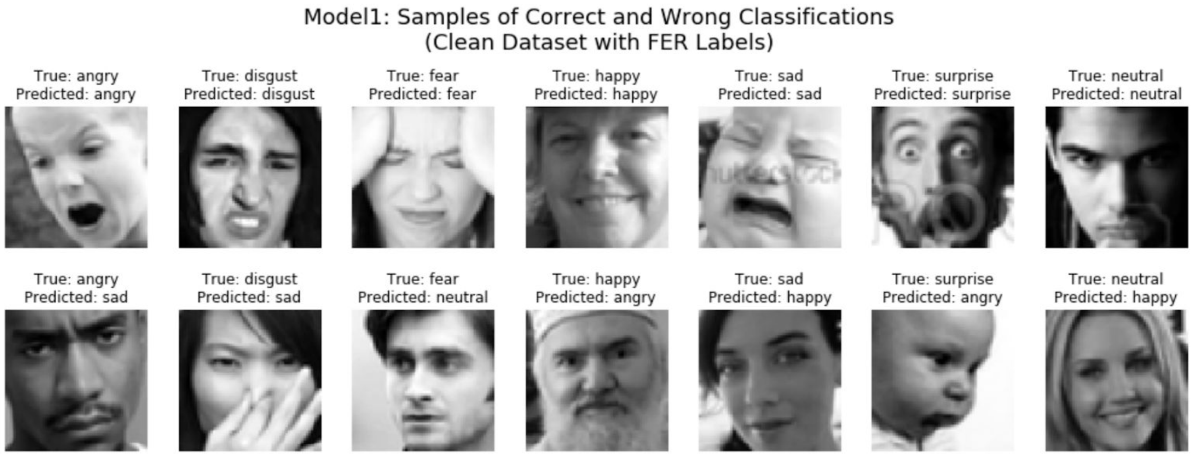


Figure 12: Classification report from Model 1 trained on FER 2013 labels (left); Model 2 classification report trained on FERPlus labels (right), both trained on cleaned dataset discarding extra labeled images

However, there are still three major misclassifications that remain in the model with FERPlus labels, i.e. 'disgust' as 'angry', 'fear' as 'surprise', and 'sad' as 'neutral'. This could be contributed by the fact that the above-mentioned misclassified pairs have similar facial features as shown in figure 13, and the classifier is biased towards the latter class of the pair with more training images. Hence, the model tends to predict 'surprise' over 'fear' for open mouth and wide-open eyes, 'angry' over 'disgust' for frowns and squinted eye', 'neutral' over 'sad' for more relaxed features.

From figure 13, the model was able to correctly predict images with obvious defined features such as teeth being displayed in combination with cheek curvature due to smiling, as happy, or common features of wide eyes and raising eyebrows for surprise. Some misclassified images appear to be a combination of emotions. When a person covers their face with a hand or there is a watermark, it tends to predict neutral as important emotion features are covered up.


Model1: Samples of Correct and Wrong Classifications
(Clean Dataset with FER Labels)
True: angry Predicted: angry
True: disgust Predicted: disgust
True: fear Predicted: fear
True: happy Predicted: happy
True: sad Predicted: sad
True: surprise Predicted: surprise
True: neutral Predicted: neutral
True: angry Predicted: sad
True: disgust Predicted: sad
True: fear Predicted: neutral
True: happy Predicted: angry
True: sad Predicted: happy
True: surprise Predicted: angry
True: neutral Predicted: happy

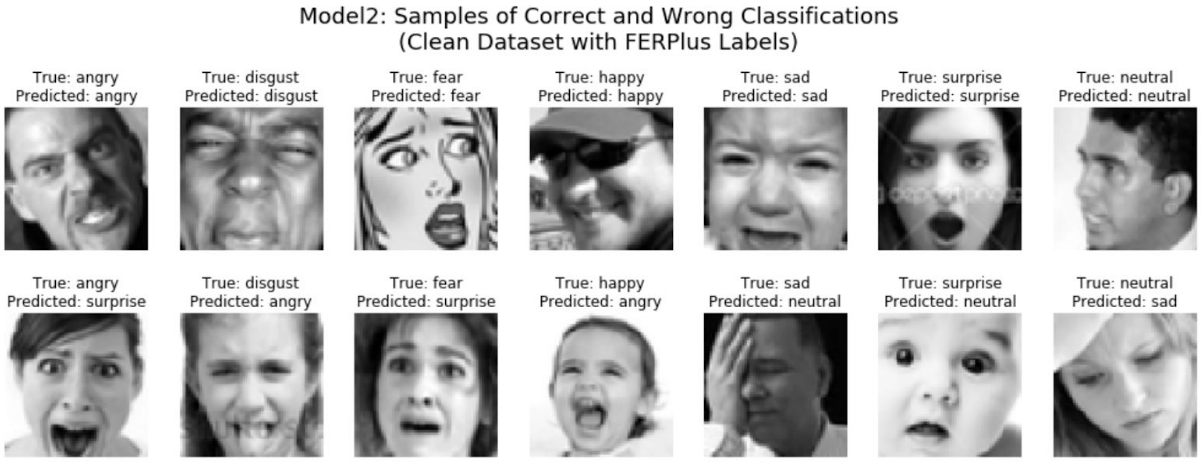


Figure 13: Sample correct and incorrectly classified images from Model 1 (top), and from Model 2 (bottom)

Conclusion

In this paper, we successfully reproduced a real-time emotion classifier based on facial expressions achieving 65% overall accuracy vs. 66% from the original study. By utilizing crowd-sourced labels, we were able to significantly improve the overall accuracy up to 81%. The prediction accuracy for each individual emotion was improved at the same time. This provides valuable insights for experimental design in psychological studies where the ground-truth labels could be more accurately collected by gathering more opinions.

The major challenge faced was the imbalanced class distribution and a lack of information of demographic or cultural variance. The problem could potentially be improved by incorporating other substantial datasets such as Audio/Visual Emotion and Depression Recognition (AVEC), the MMI Facial Expression, The Japanese Female Facial Expression (JAFFE), etc. The light-weight CNN model adopted by the original study allows for real-time emotion tracking in small-sized devices that can be further applied to help monitor individuals suffering from mental illnesses.

One way to apply this research would be able to develop a mobile application that could observe an individual and record their expressions throughout the day. From there a summary score can be calculated where the individual is classified by the emotion that is most prevalent for that day. This would remove quick and sudden changes in emotion and would provide a more accurate representation of how the individual felt that day, which can serve as reference information for one to evaluate his or her mental status.

Roles

Preprocessing of the data was done by all members of the group: Nathen Warren (NW), Yiran Chen (YC), Jiayue Xu (JX), Ravitashaw Bathla (RB), and Abhishek Baral (AB). Each of the members tried different preprocessing techniques and reported their results as to what method of preprocessing provided the best results for classification. Multiple combinations of classifiers were attempted by everyone. Lastly, for the manuscript all parts were shared equally.

References

Alom, Md Zahangir, et al. “The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches.” ArXiv:1803.01164 [Cs], Sept. 2018. arXiv.org, http://arxiv.org/abs/1803.01164.

Arriaga, O., Valdenegro-Toro, M., & Plöger, P. (2017). Real-time Convolutional Neural Networks for Emotion and Gender Classification. arXiv preprint arXiv:1710.07557.

Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016, October). Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 279-283).

Bourke, C., Douglas, K., & Porter, R. (2010). Processing of Facial Emotion Expression in Major Depression: A Review. Australian & New Zealand Journal Of Psychiatry, 44(8), 681-696. doi: 10.3109/00048674.2010.496359

Calistra, Cole. “The Universally Recognized Facial Expressions of Emotion.” *Kairos*, 15 Mar. 2015, www.kairos.com/the-universally-recognized-facial-expressions-of-emotion.

Chollet, F. (2017). Xception: Deep learning with depth wise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).

Daros, A.R., Ruocco, A.C. & Rule, N.O. Identifying Mental Disorder from the Faces of Women with Borderline Personality Disorder. *J Nonverbal Behav* 40, 255–281 (2016). https://doi.org/10.1007/s10919-016-0237-9

Georgescu, M. I., Ionescu, R. T., & Popescu, M. (2019). Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, *7*, 64827-64836.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Zhou, Y. (2013, November). Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing* (pp. 117-124). Springer, Berlin, Heidelberg.

Gruttadaro, D., & Crudo, D. (2012). College students speak: A survey on mental health. *Arlington, VA: National Alliance on Mental Illness (NAMI)*.

Hashemi, M. (2019). Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. *Journal of Big Data*, *6*(1), 98.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Kennedy, D. P., & Adolphs, R. (2012). Perception of emotions from facial expressions in high-functioning adults with autism. *Neuropsychologia*, *50*(14), 3313-3319.

Krebs, P., & Duncan, D. (2015). Health App Use Among US Mobile Phone Owners: A National Survey. *JMIR Mhealth And Uhealth*, *3*(4), e101. doi: 10.2196/mhealth.4924

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Li, S., & Deng, W. (2020). Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, 1–1. doi: 10.1109/taffc.2020.2981446

Meng, Q., Zheng, S., Zhang, H., Chen, W., Ye, Q., Ma, Z. M., ... & Liu, T. Y. (2018). G-sgd: Optimizing relu neural networks in its positively scale-invariant space.

Mental Health and COVID-19 – Information and Resources. (2020). Retrieved 18 April 2020, from https://mhanational.org/covid19

Mollahosseini, A., Hassani, B., Salvador, M. J., Abdollahi, H., Chan, D., & Mahoor, M. H. (2016). Facial Expression Recognition from World Wild Web. 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). doi: 10.1109/cvprw.2016.188

NAMI: National Alliance on Mental Illness, Mental Health By the Numbers. (2019). Retrieved 18 April 2020, from https://www.nami.org/mhstats

Ranganathan, H., Venkateswara, H., Chakraborty, S., and Panchanathan, S. (2017). "Deep active learning for image classification," 2017 IEEE International Conference on Image Processing (ICIP), Beijing, 2017, pp. 3934-3938.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV). doi: 10.1109/iccv.2017.74

Sifre, L., & Mallat, S. (2014). Rigid-motion scattering for texture classification. *arXiv preprint arXiv:1403.1687*.

Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003, August). Best practices for convolutional neural networks applied to visual document analysis. In *Icdar* (Vol. 3, No. 2003).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Siqueira, H., Magg, S., & Wermter, S. (2020). Efficient Facial Feature Learning with Wide Ensemble-based Convolutional Neural Networks. *arXiv preprint arXiv:2001.06338*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

Tang, Y. (2013). Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*.

Topor, J. (2020). Implementing Convolutional Neural Networks for Image Classification and Facial Recognition Using Tensorflow v1.0: An Independent Study. URL https://rstudio-pubs-static.s3.amazonaws.com/279500_57cad8c546724b40ad3e90692f716ae4.html

The State of Mental Health in America – Information and Resources. (2020). Retrieved 18 April 2020, from https://www.mhanational.org/issues/state-mental-health-america

Wang, J., & Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11.