Capstone Project: Deep Learning Methods for Facial Emotion Recognition

Monica Palacios Boyce
MIT ADSP - 2022 E cohort
FINAL REPORT

Executive Summary

This project aims to construct a best-fit Convolutional Neural Network (CNN) model that accurately performs multi-class classification for facial emotion recognition. Specifically, the model must accurately detect four specific emotions in images of people, including: 'happy', 'sad', 'neutral', and 'surprise' from the FER 2013 dataset (>30,000 images [1]). To this end, a series of CNN models were designed, optimized and evaluated. Transfer learning strategies were also employed, utilizing the pre-trained feature-extraction layers of well tested high performance CNNs, namely, VGG16 [2], ResNet V2 [3], and EfficientNet [4] . Finally, a deeper complex CNN was designed and assessed for accuracy of multi-classification of facial characteristic - emotional state encoding or recognition.

Of the models assessed, the final model was identified as the best performance for the stated purpose, with a overall ~93% accuracy of discriminating correct emotions on a test set of facial images. This model was able to generalize well from its training and validation performance to testing performance. In terms of key takeaways, factors that impact model performance on test images include suboptimal training due to dataset issues that include a certain amount of labeling error in the FER 2013 dataset (as is the case with any dataset), inherent ambiguity of facial characteristics for certain more subtle emotions, the potential for culture specific nuances in surface emotional expression, and data sparsity, to name several. Another important dataset related issue is well known demographic bias with respect to imbalances in gender, race, age and other factors that lead to reduced accuracy of model training.

Key next steps to mitigate these known issues could involve: increasing the size of the dataset, increasing the accuracy of dataset labeling, correction of bias by balancing demographic factors (equal representation of genders, ages, and racial phenotypes), and using transfer learning. Within the constraints of this project, the use of transfer learning did not yield improved performance counter to what would be expected. Pre-trained feature descrimination convolution layers of candidate CNN models such as VGG16, ResNet V2 and EfficientNet have been trained on much larger and more diverse datasets than the current models 1, 2 and final model which have been trained solely on the FER 2013 dataset.

The final model has a sufficient level of performance (**~93%**) that can be further optimized during model training by using other larger image datasets for training, such as ImageNet (>14 million annotated images [5]), CelebA (>202,000 annotated images [6]), FFHQ (Flickr-Faces-HQ, 70,000 high resolution diverse image set [7]), to name a few.

Problem Summary

Beneath the surface of modern culture, mediated through our many devices and consumption habits, lay vast amounts of both obvious data collections as well as data that are untapped but of great interest to countries, research groups and companies seeking to quantize and monetize it. Dual impulses to efficiently capture and control big data have driven forward advances in deep learning. Considering that humans have evolved to communicate not just with the spoken word but with changes in our facial expressions, capturing and interpreting accurate emotions from images of facial expressions will yield a deeper, richer context of the user experience.

The evolution of computer vision to process images into meaningful, accurately encoded data led to the creation of neural networks and more specifically Convolutional Neural Network models (CNNs) that are able to process image data in such a way as to learn higher order patterns (features) that can yield predictions of value on new images. Examples of this might include a facial emotion recognition model used on live video in a public space to detect people displaying emotions that suggest that they intend to do harm to others (threat assessment), remote telemetry of at-risk patients in their homes for emotions that indicate pain or some other medical emergency, and perhaps even to gauge a consumer's level of interest as they browse through a store's inventory online. This last example illustrates the potential for ethical risks that could arise from latent, intrusive use of this technology.

Solution Design

The core objective of this project is to design, train, test, and optimize a CNN model that performs well at predicting emotions in images of people displaying various facial characteristics or features. The CNN model presented herein is capable of correctly labeling emotional states in test images with sufficient accuracy for use in most settings where emotion detection is of interest.

Candidate model designs, including those employing transfer learning, were trained on an industry standard dataset called FER 2013 that contains thousands of grayscale facial images that were culled from online sources by Google image search and then validated by human encoders. The prediction performance of the models was then assessed against similar grayscale images to determine the ability of the model to learn meaningful encodings (features as filters) that correspond to the correct emotion.

Figure 1 (**See Appendix**) shows the network architecture for the best-fit model in this project, referred to as the 'final model'. Key features include 5 convolutional blocks, each with layers that minimize data loss and overfitting so that the model is optimized for rich feature encoding and good generalized performance on test data.

Analysis and Key Insights

The goal for all models studied in this project is accurate classification (prediction) of an emotion from an image showing a facial expression. The key metrics used to assess performance was the accuracy of the model during training, validation and then during testing on novel images.

Additionally, the best-fit candidate was assessed for how it failed to properly classify images so as to gain insight into potential optimization pathways.

Six different model architectures were assessed during this project, three of which were based on transfer learning (vggmodel, resnetmodel, efficientmodel). Model architecture diagrams have been included in the **Appendix** of this report. Assessment data is shown in **Table 1** below.

Model	Training Accuracy	Validation Accuracy	Testing Accuracy
Model 1	65.4%	66.1%	69.5%
Model 2	73.5%	69.4%	68.8%
vggmodel	53.8%	52.1%	51%
resnetmodel	30.6%	33.4%	30.5%
efficientmodel	26.3%	24.4%	25%
final model	92.4%	93.6%	93.3%

Table 1: Performance measures for all models assessed in this project

Final model accuracy during training and validation are shown in **Figure 2** below. This chart illustrates, visually, the concept of “generalization” of performance over the duration of model training. The blue line shows training performance and the orange line shows performance during validation. As these two lines follow very similar paths, the generalization can be considered sufficient. If the model had been underfit or overfit (both of which would indicate a model not capable of making accurate and meaningful predictions on novel images) then the orange validation line would be substantially out of register or shifted with respect to the training data line.

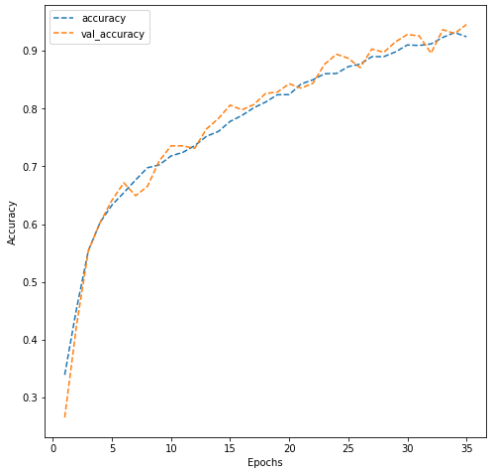


Figure 2: Plot of training and validation accuracy against epochs during model training.

The model that performed the best was the final model, with an accuracy of prediction on novel images (test images) of approximately **93%** and good generalization.

Thus, this final model is the CNN model that is supported as the best-fit CNN model in this project.

Key Limitations

In terms of limitations of facial emotion recognition, of the four emotion classes in the FER 2013 dataset, the 'Neutral' face represents a challenge for humans and algorithms alike. (Note: While only four classes were used in this project, there are actually six classes in the FER 2013 dataset). A neutral face can be easily misinterpreted by humans and can fall into other predicted emotion classes when those emotions are subtle.

To garner insights on the performance of the final model with respect to discriminating the different classes, a confusion matrix may be used. **Figure 3**, below, shows one such confusion matrix. It illustrates the incidence of prediction errors by the model on 128 test images, 32 in each class (happy, sad, neutral, surprised). Along the left side is the actual labeled set which is arrayed against the predicted set along the bottom. If a number is shown in a box, it represents the number of times that the model did not predict an emotion correctly. The color intensity indicates incidence of error (0 - 8 in this case). The diagonal line indicates the number of times that the predicted value was the same as the actual value for each class.

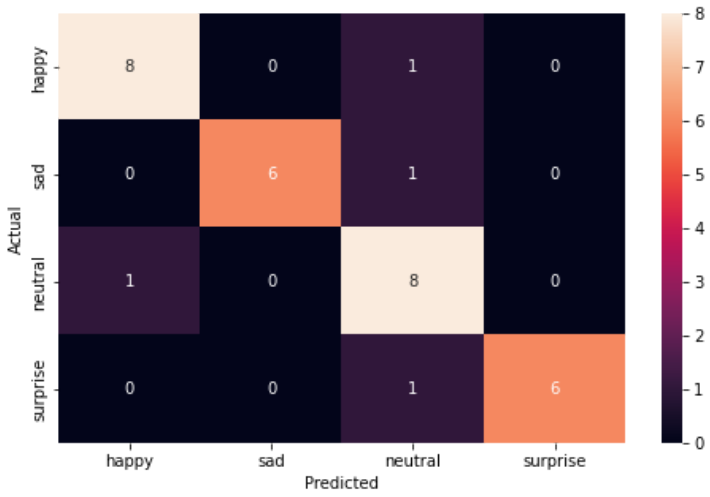


Figure 3: Confusion Matrix for the final model

Figure 3 visually demonstrates how the number of errors by the final model are quite low. This matrix also shows how errors are occurring for a specific class - 'neutral'. Those boxes showing an error (1 error, dark purple) are associated with the 'neutral' emotion, in agreement with the intuition mentioned above regarding the difficulty of interpreting a neutral expression.

Key Recommendations for Further Analysis and Implementation

Results shown in this project suggest that there are several key actionables for achieving the goal of creating and deploying a robust facial emotion recognition model.

One action is further optimization of the final model to improve informative feature extraction by training on larger datasets with higher label fidelity and demographic diversity. The benefits that would be realized from this potentially costly endeavor would be higher emotion recognition accuracy across a more diverse demographic spectrum. This would certainly make any product using this optimized model fit for use in global markets.

As mentioned previously, the main risk in further training of the convolutional (feature extraction) layers of the final model is the computational cost that is required to train on very large datasets.

As such, it is worthwhile to revisit the use of transfer learning to take advantage of the feature extraction layers of pre-trained CNN models, which only continue to grow ever more powerful.

Recent advances in CNN model design for the purpose of robust facial emotion recognition include the use of a dual-channel CNN architecture that first identifies a region of interest (ROI) and then applies higher resolution feature extraction to the “pre-qualified” ROIs. The intent of this design is to reduce model confusion on more subtle facial expressions [8].

Business use-cases for facial emotion recognition are expanding significantly to areas such as capturing metrics of student engagement in online education, psychological analysis of job applicants by human resource groups during hiring to optimizing personalized learning milieu through the analysis of not only visual facial features but EEG data as a neurological emotion-ground-truth reference [9].

Opportunities for entry into this market are considerable. Innovations in model design, growth of high quality image datasets, the application of facial emotion recognition to novel use-cases, and the continued trend in lower computational costs will enable profound breakthroughs in areas such as medicine, education, social science, public policy and other unforeseen settings.

Ethical risks regarding privacy and ownership issues will require an open societal level discourse that should be considered a necessary component of any development plan (whether that is research or corporate).

Bibliography

1. Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, 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.

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

3. Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence.

4. Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

5. Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

6. Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision (pp. 3730-3738).

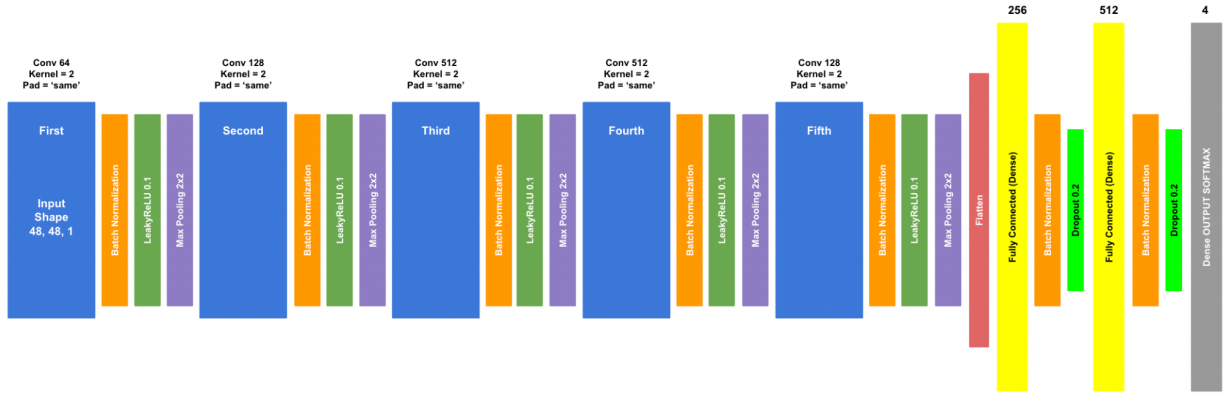
7. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

8. Song, Z. (2021). Facial expression emotion recognition model integrating philosophy and machine learning theory. Frontiers in Psychology, 12.

9. Zhu, X., Rong, W., Zhao, L., He, Z., Yang, Q., Sun, J., & Liu, G. (2022). EEG Emotion Classification Network Based on Attention Fusion of Multi-Channel Band Features. Sensors, 22(14), 5252.

Appendix

Figure 1: Final Model Architecture



Legend

= Convolutional layer

= Batch Normalization layer

= LeakyReLU layer

= Max Pooling layer

= Flatten layer

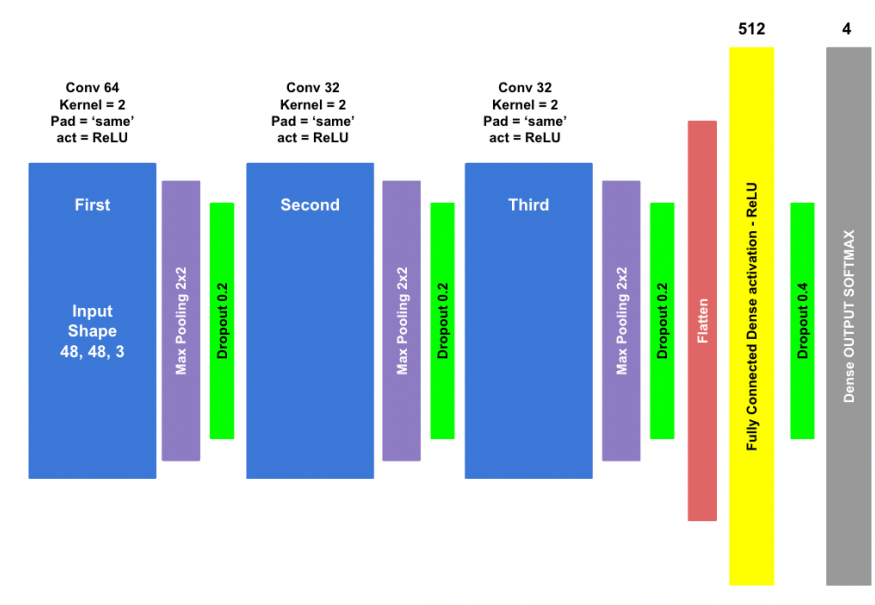
= Dropout layer

= Fully Connected (Dense) layer

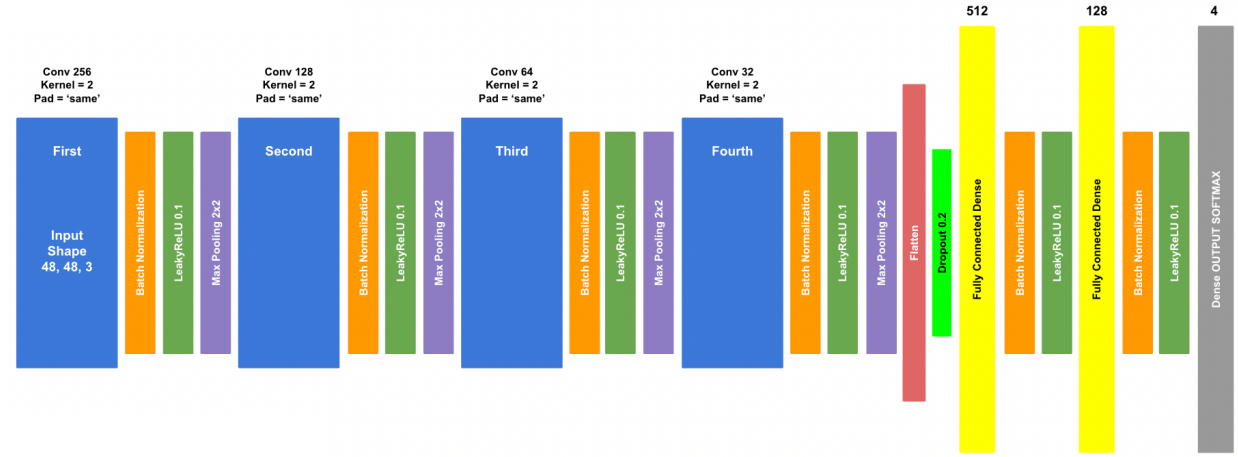
= Fully Connected (Dense) OUTPUT layer

Other models assessed in this project

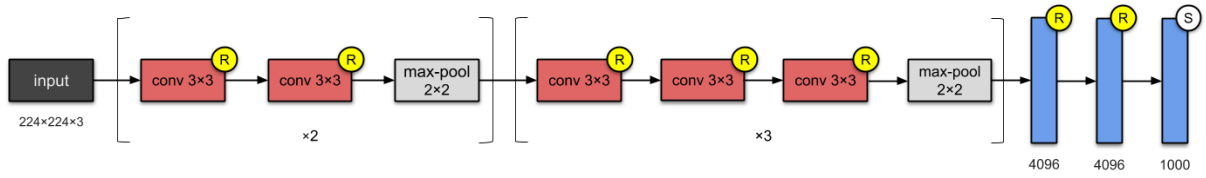
Model 1 (Milestone 1)



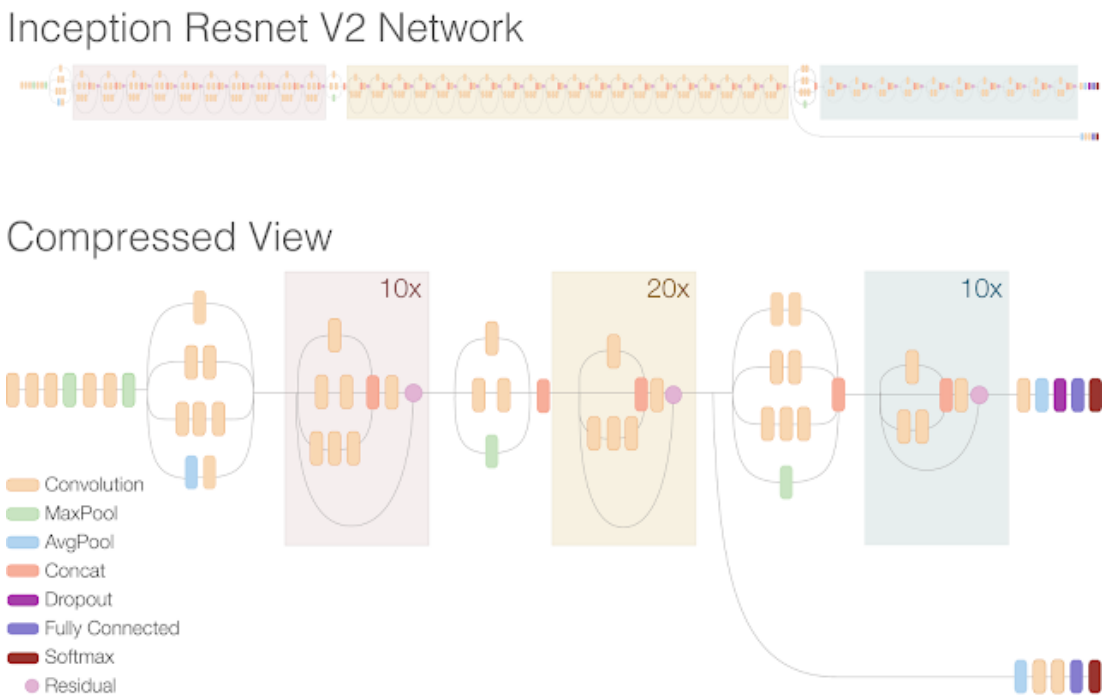
Model 2 (Milestone 1)



VGG16 model



ResNet V2 Model



EfficientNet mode