Facial Emotion Recognition Using Network Surgery

Masih Zakavi, Jasur Okhunov
sz2803@columbia.edu, jo2662@columbia.edu
December 8, 2023

Abstract

Facial Emotion Recognition (FER), the detection of human emotions based on images of faces, has been one of the oldest and most notoriously difficult tasks in computer vision. The primary goal is to enable machines to understand and respond to human emotions, which has various applications, including human-computer interaction, virtual reality, gaming, and emotion-aware systems. Despite breakthrough progress in many aspects of computer vision, state-of-the-art models continue to struggle with FER. Here, we experimented with three relatively simple architectures and achieved an accuracy of 63.7% on the FER2013 test set without any training on additional data.

Introduction

Facial expressions play a crucial role in human communication. However, even for most humans, recognizing emotions can be challenging. As a problem, FER has long been criticized on this basis as an underspecified problem. The rise of deep learning in the past decade has significantly improved Facial Emotion Recognition, surpassing human accuracy in many cases.

Despite criticisms, FER continues to remain relevant for a wide variety of applications in digital advertisement, online gaming, customer feedback assessment, and healthcare [1]–[3]. In this paper, we design and study three simple architectures for FER on the FER2013 dataset. We aim to find the most effective model and leverage model interpretation frameworks to better understand our results and how they were obtained.

Methods

1. Data Cleaning and Processing

In this project, we employed the FER13 dataset from Kaggle, comprising 35,877 images, each 48x48 pixels in grayscale. This dataset was designed by Kaggle in response to the low quality of the datasets used for FER, which relied heavily on the expected human accuracy on it is expected at 65% [4]. FER13 images come in seven categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The categories are not balanced, with 4,953 in Angry, 547 in Disgust, 5,121 in Fear, 8,989 in Happy, 6,077 in Sad, 4,002 in Surprise, and 6,198 in Neutral. Before the training process, a thorough dataset cleaning was conducted to

mitigate noise and enhance the data quality. In particular, five types of images were removed from the dataset:

1) Images that were not human faces. This type includes images of statues, comic book faces, cartoon faces, and drawn faces from other media. Below are a few examples:

Figure 1: From left to right: cartoon, drawing, statue

2) Images that appear in more than one class. Since these form a barrier to effective training, one copy was randomly selected and removed from the data. Below are two examples:

Figure 2: Overlapping images between classes (left image in both *fear* and *sad;* right image in both *angry* and *neutral)*

3) Images in which a portion of the face remains obscured, preventing a full and clear observation of the entire facial features.

Figure 3: Partially visible faces

4) There were a small number of images without a human face.

Figure 4: Image from *fear* category: no human

5) We eliminated the disgust category because its exceptionally small size made classification more challenging.

2. Transfer Learning

Transfer learning and various network surgery techniques have had tremendous success in a diverse set of computer vision tasks. Inspired by the last assignment of the course and successes in such tasks, our first experiments involved transfer learning on ResNet50 [5], VGG16 [6], and EfficientNetB0 [7] architectures, using ImageNet weights provided by TensorFlow's Keras library. In all cases, the classifier head was replaced by a three-layer multilayer perceptron.

Parameters of training are a batch size of 32, 10 epochs of training with the option for early stopping based on validation loss with patience of 3, and Adam for learning rate updates. Training was done by using ImageNet weights, freezing the original network, training the new head, then unfreezing the entire network, and training some more.

To further investigate network surgery, we also trained with gradual unfreezing, where layers were gradually unfrozen moving from the head to the first convolutional layers. The results were nearly identical to basic transfer learning.

2. Designing a Custom Network

As an alternative to the complex and deep networks considered earlier, we designed a simple and much shallower network. As demonstrated in Figure 5, the customized network involved two pairs of subsequent convolutional layers with pooling for feature extraction, each with batch normalization and dropout, to avoid overfitting. Naturally, these convolutional layers are followed by fully connected layers for prediction.

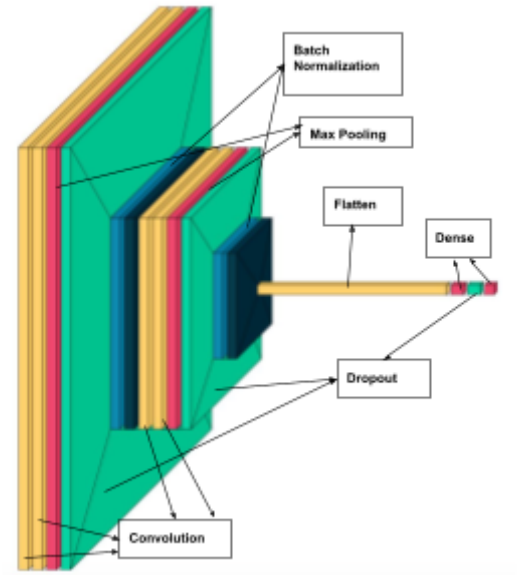

Figure 5: The custom-designed shallow network

The hyper-parameters here are identical to the ones used for transfer learning earlier.

3. Modifying VGG16 Architecture

Another architecture trained was a modified version of the VGG16 architecture. The VGG16 architecture was adopted and modified by removing the head, adding two convolutional layers with max-pooling and batch normalization, and finally, adding two fully connected layers. The figure below outlines the architecture:

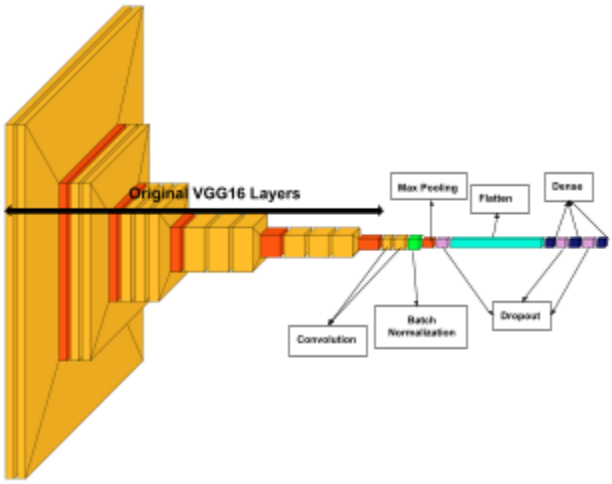

Figure 6: Modified VGG-16 used

The rationale behind this architecture was that if VGG16 can be trained to do precise object recognition, one more set of convolutional and max-pooling layers needs to be added to learn facial expressions from the detected facial feature embeddings. Unlike in the transfer learning section, ImageNet weights were not used in this case, and the network was trained from scratch.

Results

Model Interpretation

A strong FER classifier needs to learn facial features essential to expression recognition, like eyes, eyebrows, mouth, and chin. The ability of our models to learn these features was tested using the model interpretation method, Grad-Cam [8].

For brevity, we omit a complete discussion of the Grad-Cam algorithm here. In summary, gradients with respect to the

last convolutional layer are used to understand the network's predictions with respect to each output class.

Applying Grad-Cam to our predictions, we obtained heatmaps based on the importance of each pixel in the input image for the final prediction and used these results to inform our design and training parameters.

1. Results for Transfer Learning

In all three cases of VGG16, ResNet50, and EfficientNetB0, early stopping based on validation loss was enabled. As a consequence, training stopped in all three cases before 10 epochs of training were reached. The results on the test data were surprisingly weak. The loss and accuracy failed to improve over time and fluctuated around 16-17%, which would make the classifiers as bad as random predictors. (see figure below)

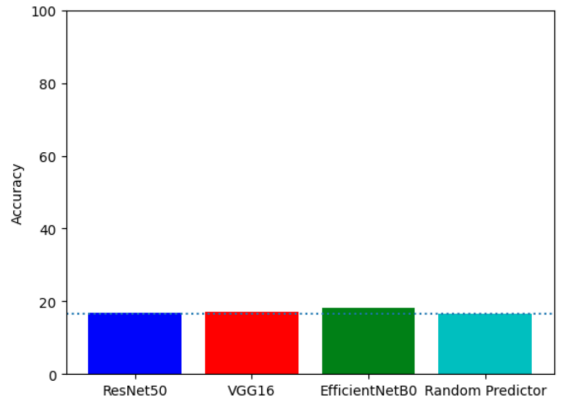

Figure 7: Transfer Learning vs Random Classifiers

As mentioned earlier, Grad-Cam was used to investigate the results. Figure 8 shows Grad-Cam applied to the ResNet50 transfer learning model's predictions.

Figure 8: Heatmap generated by ResNet50 Transfer Learning on a test data point. Actual Label: Happy, Predicted Label: Surprise with probability 17.3%

As expected by the stagnant validation loss and accuracy, the Grad-Cam heatmap illustrates that the transfer learning networks didn't learn facial features, which in turn explains the failure to achieve a high accuracy on the test data.

This failure can be attributed to the pre-trained weights used for the three models. ImageNet weights that were used for the transfer learning networks all use the 1000 categories in ImageNet, which don't include human expressions or any parts of the human face. Consequently, the pre-trained models could learn lots of features in the input images, but none of the ones crucial for FER.

2. Results of Custom Network

The custom-designed network outperformed the transfer learning networks with respect to accuracy on the test data. Accuracy on the test data improved to 34.8%. As reflected in the loss and accuracy curves below, unlike transfer learning, the customized network does, in fact, learn some of the important facial features.

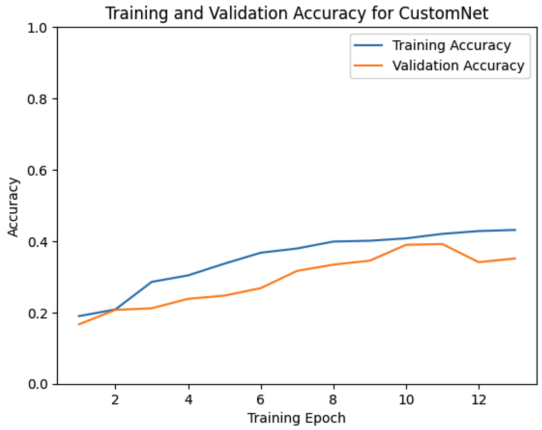

Figure 9: Training and Validation Accuracy for Custom Network

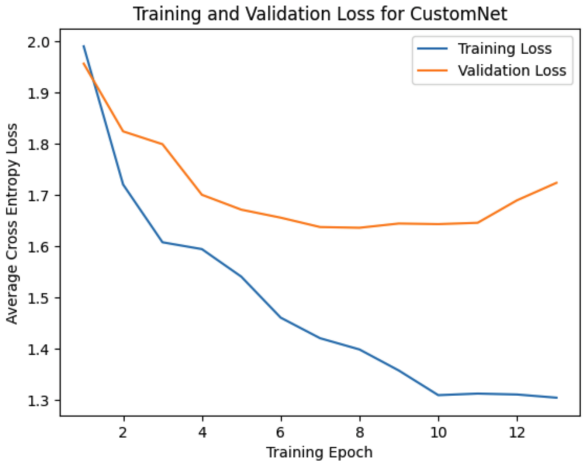

Figure 10: Training and Validation Loss for Custom Network

A major challenge with the custom-designed network is that it overfits to the training data. This can be clearly seen in the sharply increasing validation loss. It's noteworthy that this was despite the presence of both batch normalization and dropout layers. A possible explanation for overfitting is that the network is too simple to capture all facial features, even though it does in fact, learn some of them, unlike the transfer learning models.

3. Results of Modified VGG16 Network

Given the added complexity of the modified VGG16 architecture, we expected this design to learn facial features better and obtain better performance on the test data than either of the two other architectures. This assumption was proven correct by the results. The curves below show the accuracy and loss curves for this architecture:

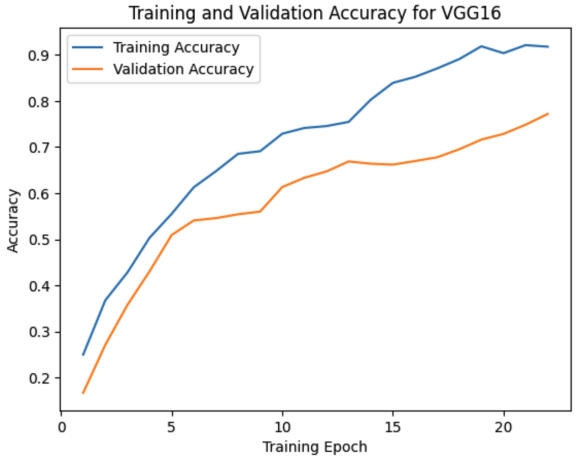

Figure 11: Training and Validation Accuracy for modified VGG16 Network

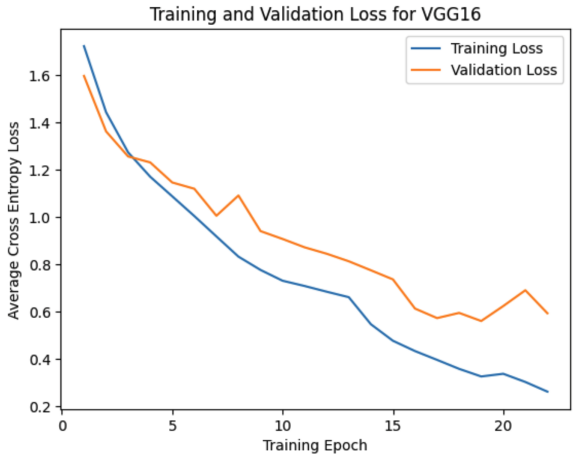

Figure 12: Training and Validation Loss for modified VGG16 Network

Grad-Cam applied to modified VGG16's predictions further demonstrate that important facial features like eyes and

mouth were learned well in the modified VGG16 network.

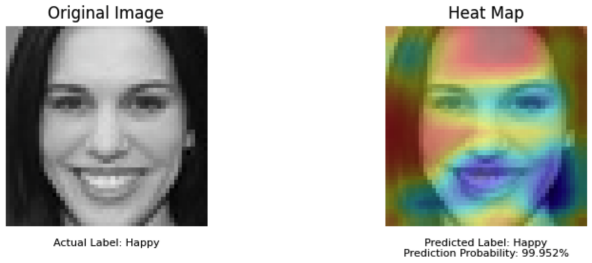

Figure 13: Heatmap generated by modified VGG16 on a test data point. Actual Label: Happy, Predicted Label: Happy with probability 99.952%

Conclusion

While our best model, the modified VGG16 doesn't reach state-of-the-art, it still achieves an accuracy of 63.7%, a performance level comparable to the expected human accuracy of 65%. Notably, our analysis revealed that our model effectively learned to prioritize key facial features for decision-making. The model's performance may be improved by training on external facial feature data and using those weights with transfer learning. Further improvements may be made by the use of ensemble methods or boosting, as was done in the papers mentioned in the introduction. However, this requires computational resources unavailable to us.

Acknowledgment

The code used for Grad-Cam images was adopted from the following repository: https://github.com/wiqaaas/Coursera_Certificates

Author contributions statement

J.O. cleaned and curated the data used in this paper, M.Z. and J.O. designed the experiment. M.Z. designed the architecture. Both authors worked on software development. J.O. did training and testing. M.Z. integrated Grad-Cam and worked on visualization. All authors worked on writing the report and reviewed the manuscript.

References

[1] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2003, vol. 5, doi: 10.1109/CVPRW.2003.10057.

[2] F. Abdat, C. Maaoui, and A. Pruski, "Human-computer interaction using emotion recognition from facial expression," in Proceedings - UKSim 5th European Modelling Symposium on Computer Modelling and Simulation, EMS 2011, 2011, doi: 10.1109/EMS.2011.20.

[3] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," Pattern Recognition, vol. 36, no. 1. 2003, doi: 10.1016/S0031-3203(02)00052-3.

[4] K. He, X. Zhang, J. Sun, S. Ren, "Deep Residual Learning for Image Recognition," arXiv preprint arXiv:1512.03385, 2015

[5] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint arXiv:1409.1556, 2014

[6] M. Tan, Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," arXiv preprint arXiv:1905.11946

[7] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee et al., "Challenges in

representation learning: A report on three machine learning contests," in International Conference on Neural Information Processing. Springer, 2013, pp. 117–124.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.