Recognition and Classification of Human Emotions From Facial Expressions

Atharva Dikshit Jong Hoon Park Nischal Suresha Prakrit Tyagi Feiya Zhu
Carnegie Mellon University
Mechanical Engineering Department
5000 Forbes Avenue Pittsburgh, PA 15213
adikshit, jonghoon, ngandige, prakritt, feiyaz@andrew.cmu.edu

Abstract

Human facial expression classification has attracted great attention in the field of both machine learning and computer vision and there has been active research and applications in this fascinating area. In this paper, a facial expression classification algorithm is proposed which uses a shallow neural network architecture for human emotion recognition and classification purposes. The system uses images of a person from the dataset, AffectNet, to classify eight basic emotions: neutral, happiness, sadness, surprise, fear, disgust, anger, and contempt. With our model trained on AffectNet, the algorithm is implemented in real-time for expression classification along with OpenCV facial detection algorithm. The performance of our algorithm has been reported and presented with the comparison between our baseline neural network and the pre-trained residual neural network, ResNet18.

1. Introduction

[1] Facial emotion recognition (FER) is an interesting field, which has several applications such as healthcare, human-human interactions, and human-machine interactions. Furthermore, FER is an important aspect of predicting the psychological states of interlocutors during social interactions. If a machine can recognize the emotion of a person based on its facial expression, there is a huge potential for the industry and market to take advantage of understanding their consumers' mental states and thus promote improved user and customer satisfaction. Identifying the emotion from a person's speech is also a big field of interest, but changes in facial expressions are the first indicators of a person's emotions. There are many instances where there will be no verbal/vocal inputs available in which case vision-based approaches will be useful[7]. Previous studies have established that around 7% of communication is verbal, 38% of the communication is vocal and 55% of the communication is visual[4]. This justifies the interest and importance of using vision-based techniques for identifying emotions.

In this project, the specific objective we would like to accomplish is the classification of human facial expressions with a real-time facial recognition analyzer. As much research has been conducted in this field and many of its applications are released. We believe that by applying our shallow neural network, we can strengthen the knowledge we have learned from the 24-787 (Introduction to AI and Machine Learning) course offered by Carnegie Mellon University and lead us to become an expert in this field within the near future.

1.1. Related Work

Recent developments in convolutional neural network (CNN) has demonstrated great success of as automatic feature detector. The best object classifier YOLOv7 [10] is based on CNN. The convolutional neural network (CNN) has demonstrated its high efficiency in image classification [3], object detection [6], and voiceprint recognition [9]. This makes us very confident in using a convolutional neural network (CNN) for the recognition and classification of human emotions from images. Razavian [8] showed that the features directly extracted from a CNN trained on ImageNet can produce superior results compared to some state-of-the-art systems on a variety of visual recognition tasks such as scene recognition and image retrieval.

Karayev et al. [2] also applied CNN features to recognize image style without any knowledge of the data and task and achieve results that are comparable to human performance. We took inspiration from the work done by Anatas [1] in using pre-trained deep learning models for facial classification of students in an online classroom and do a comparative study between light weight models and also to do a feedback analysis on online teaching and its effectiveness by predicting facial expressions of students.

[1] Git Repo: https://github.com/sjhpark/Facial_Emotion_Classifier

1.2. Data

In this work, we used a benchmark dataset to train our convolutional neural network. We present a comparison between the performance of the shallow neural network and that of a well-trained residual neural network.

1.2.1 Data Collection

Training the neural network with examples is one of the keys to deep learning success. To help researchers with this task, several FER databases are now available. Each one differs from the others in terms of the quantity and size of images and videos, variations in illumination, population, and face pose. The dataset utilized in this report comes from AffectNet, a large benchmark facial expression dataset created by Ali Mollahosseini and Mohammad H. Mahoor [5]. AffectNet is a collection of facial images (RGB; 224x224) attained from the Internet by searching 1,250 emotion related keywords in six different languages in 3 different major search engines. AffectNet is one of the largest databases of facial expressions and about half of the retrieved image annotations were hand-crafted.

There are eleven discrete human emotion categories in AffectNet: neutral, happy, sad, surprise, fear, anger, disgust, contempt, none, uncertain, and non-face.

1.2.2 Data Processing

The full AffectNet dataset was too big in size to train our model per time allotted, so we used AffectNet8, a mini-version of AffectNet database. AffectNet8 contains eight categories of emotions: neutral, happiness, sadness, surprise, fear, disgust, anger and contempt. Originally, the dataset came with four different annotations per image, arousal, expression, landmarks, and valence. However, we only used the annotated facial expressions for the features to train our model. AffectNet8 contains 287,651 compared to the full AffectNet, containing 420,299 images. On the right is a table: 1 showing the distribution of images according to their classes in AffectNet8.

For better training of our neural network model, we divided the training dataset into training dataset (80%), validation dataset (10%), and testing dataset (10%). The training dataset was used to fit our model, the validation dataset was used to provide an evaluation of our model fit on the training dataset, the testing dataset was used to provide an unbiased evaluation of our final model fit on the training dataset.

2. Methods

Expressions	No. of pictures in Full AffectNet
Neutral	80,276
Happy	146,198
Sad	29,487
Surprise	16,288
Fear	8,191
Disgust	5,264
Anger	28,130
contempt	5,135
None	35,322
Uncertain	13,163
Non-face	88,895
Total	420,299

Table 1. Number of Images per Category in Full AffectNet

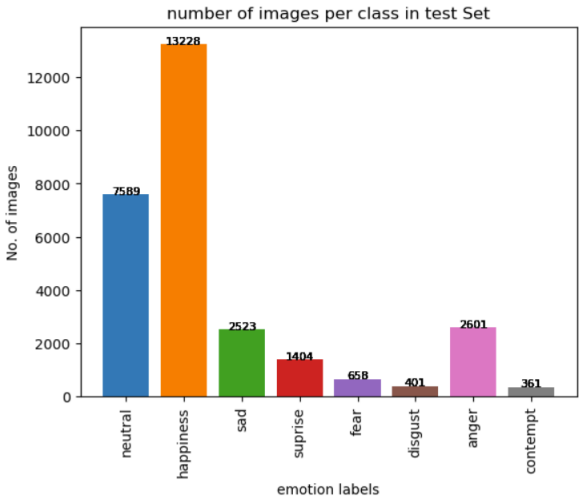

Figure 1. Number of Images in Each Category in Training Dataset After 80/10/10 Split

2.1. Model Architecture

We created a baseline model using the convolutional neural network (CNN) architecture. A CNN is a type of artificial neural network and is widely used in image classification problems due to its ability of feature extraction from images. For instance, a CNN filters features from an input image and maps the features to create a so called feature map, thus the model can be trained with these extracted features.

To keep our model shallow to see its performance compared to one of the State of the Art methods for image classification, ResNet18, we kept our model architecture simple. Our baseline CNN consists of 5 convolutional layers and a ReLU activation with a MaxPooling after every layer except for the output layer. We also implemented a fully connected layer at the end of the architecture. Thus in total there were 5 layers of convolution and one linear layer. The

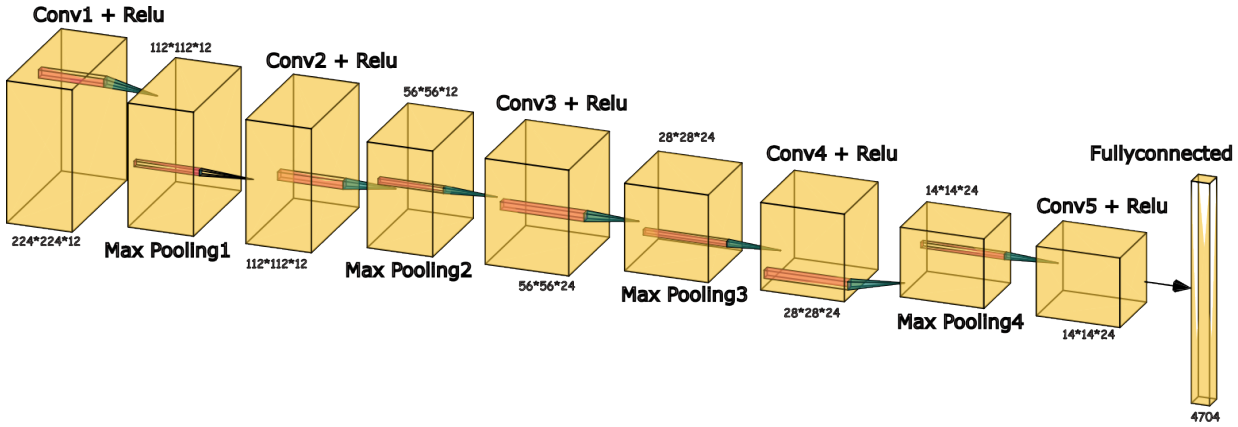

Figure 2. Baseline CNN with 5 Convolutional layer & Fully Connected Layer

choice to use this model is as anywhere between 5-10 layers with 20 to 50 nodes is a good starting position to train on your dataset. The kernel has a size of 3x3 after each convolutional layers with zero-padding applied. Simnple calculations show the feature map after each kernel layer will decrease in size by a factor of 0.5.

2.2. Model Training

We batchlized our CNN model to prevent RAM memory overshoot. We created custom datasets for the training, validation, and test datasets in order to feed a small batch (we used batch size of 64) per iteration throughout the model training process. In each iteration, we performed the forward-pass and back-propagation to update the parameters of the model. We used the Mean Cross Entropy Loss (eq:1 and eq:2) during the back-propagation for updating model parameters.

$$l(x,y) = \Sigma_{n=1}^{N} \frac{1}{\omega_{yn} * 1[y_n \neq ignoreindex]} l_n \quad (1)$$

$$l_n = -\Sigma_{c=1}^{C} log \frac{exp(x_{n,c})}{\Sigma_{i=1}^{C} exp(x_{n,i})} y_{n,c} \quad (2)$$

3. Experiments

In hope to see a performance increase and convergence over epochs, we trained the model over 50 epochs and observed that the prediction accuracy of the model on the validation images converged to around 70%. However, the loss increased and thus the model tended to over-fit over epochs. The cause and analysis of this issue have been discussed in "Confusion Analysis" subsection of this report.

3.1. Comparison to SOTA Method

We also used one of State of the art (SOTA) image classification pre-trained models, ResNet18. We fine-tuned the

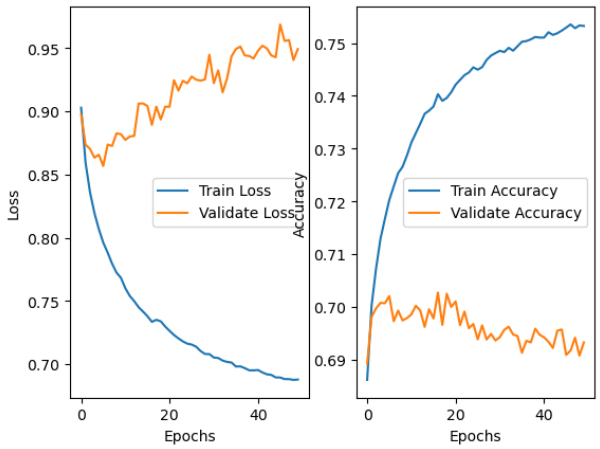

Figure 3. CNN Performance Graph over 50 Epochs

ResNet18 model with our own training dataset. The purpose of this was to compare the emotion classification performance to our trained model's and see how the shallow baseline model can compete against the benchmark model. As shown in the figure below, the ResNet18 model's emotion prediction accuracy to our validation dataset converged to around 72% over 50 epochs of fine-tuning process. Since our baseline CNN achieved around 70% accuracy, we could observe that a baseline model with the sufficient number of training data could be competent among SOTA benchmark models. It is just the matter of increasing a few more percentage of accuracy after that through innovative methods.

3.2. Prediction Evaluation

Our baseline model after 50 epochs of training performed decent against the ResNet18 model. We demonstrated that our trained CNN correctly predicted facial expression annotations 7 out of 10 given human face images. Since our validation and test datasets have a few emotion classes (happy and neutral) dominant, we can't confirm that

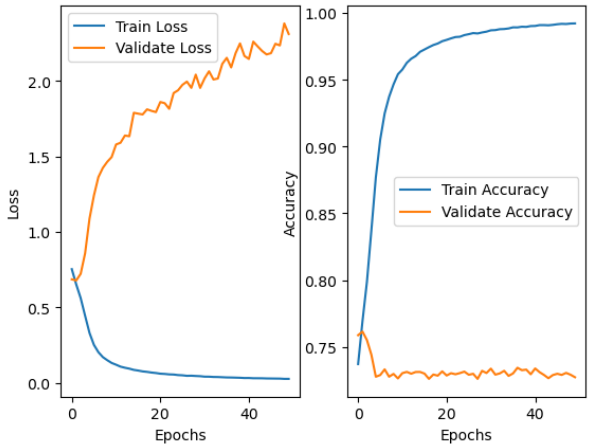

Figure 4. ResNet18 Performance Graph over 50 Epochs

the trained model will perform great for predicting the emotions of a large batch of images of other types of emotions. Training the model with a dataset of equal distribution of different emotion classes can be left for our future work.

3.3. Confusion Analysis

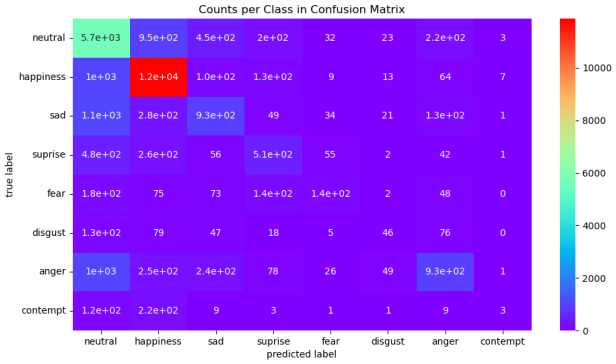

Figure 5. Confusion Matrix - Image Counts per Class

It has caught to our attention that the CNN model was over-fit during the training process. We have analyzed in depth to figure out the cause and concluded that it was due to the unbalanced distribution of emotion classes in our dataset. As shown in figure: 1, some of the emotion classes such as "happy" and "neutral" images were taking over 50% of the whole images in the dataset. The model could have been trained with a bias towards these dominant classes and become over-fit.

Furthermore, figure: 7 shows such a low f-1 score for our baseline CNN model's prediction on "contempt" images. This was because there was a big confusion between "happy" and "contempt" classes during our CNN model's prediction on the validation images as figure: 5 and figure: 6 display. 60% of the time when our model predicted on "contempt" images, it guessed them as "happy" images. This was due to the fact that there were around 30 times more "happy" images than "contempt" images in AffectNet8 as shown in figure: 1. Moreover, it was also due to the subtle different between "happy" and "contempt". When the annotators were labeling these images, there might have been inconsistency in differentiating one from another.

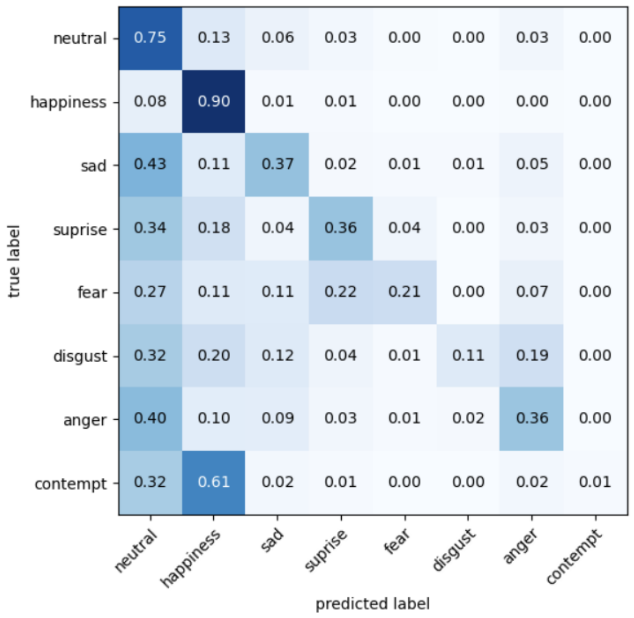

Figure 6. Confusion Matrix - Ratio

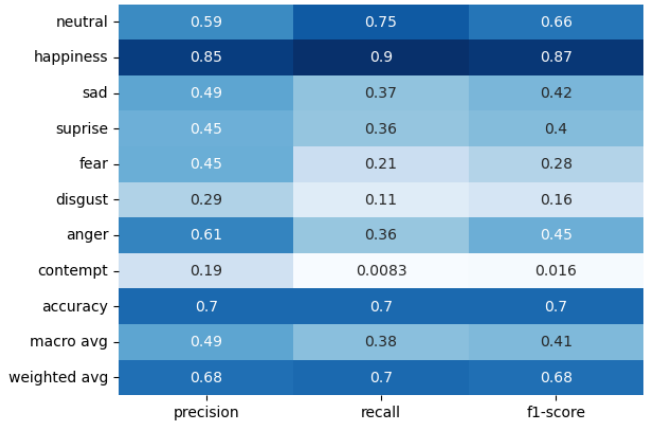

Figure 7. Precision, Recall, F-1 Scores

3.4. Static2Live - Live Emotion Prediction

We have demonstrated our trained baseline CNN's performance on live video-fed images. One with live webcam images of our teammates' faces and the other with faces of multiple characters from a clip of TV series *The Office*. Our demonstrations are displayed in figure: 8 and figure: 9. We

call this live human emotion recognition and classification as *Static2Live* where the image classification problem becomes live as predicting a sequence of static images.

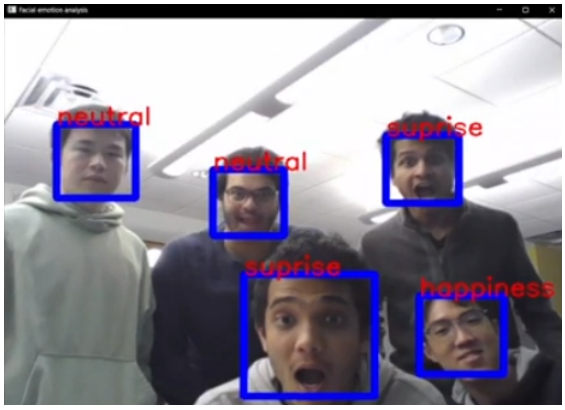

Figure 8. Live Emotion Detection Demonstration - Project Team


Figure 9. Live Emotion Detection Demonstration - TV Series

We further tested our trained CNN model to predict live emotions from animated characters. As shown in the figure: 10 below, our model performed well on classifying appropriate emotions for the detected faces of the characters. This may show the potentials of utilization of the CNN model trained with real human images to fictional characters.

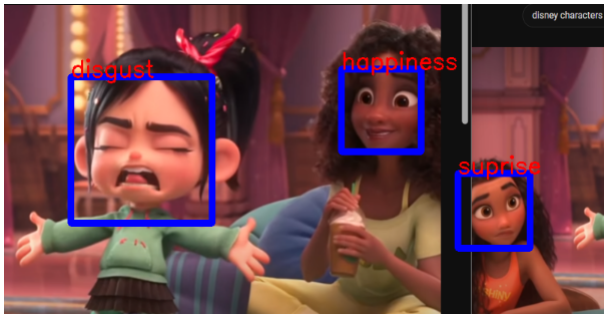

Figure 10. Live Emotion Detection Demonstration - Disney Animation: Wreck it Ralph2

4. Conclusion

In conclusion, we were successfully able to classify and quantify human facial expression with our baseline CNN model. Our model is not only able to classify facial expressions in the images but also in real-time using a webcam. From our testing, we found that our model was predicting "neutral" and "happiness" more often than other expressions. Upon investigation, we saw that we had more number of "neutral" and "happiness" images in our training dataset than other expressions. Thus, our model was biased towards the two expression over other expressions. With a more balanced dataset, we believe that this problem could be avoided.

There are still many places to work on further to sharpen and hone our baseline CNN since it was over-fitting a little bit during the training process. As we discussed in the previous section, retraining our model with a more balanced dataset (equal distribution of different facial expression images) may enhance the accuracy and recall performance of our model. Our future work will be creating a generative model where we can train the model with human face images to generate (or mimic) human faces with a certain facial expression on. We are also looking forward to using the rest of the annotations (arousal, landmarks, and valence) for each face image as extra features while training the generative model.

We hope our work is used as a stepping stone for new researchers in the human emotion recognition and classification field and fosters a more active research field. Studies in this field has a much potential to be leveraged and exploited for the benefits of humanity and business.

5. Acknowledgment

We, Jong Hoon, Nischal, Feiya, Prakrit, and Atharva, would like to express sincere thanks to Prof. Amir Barati Farimani and the teaching assistant, Tong Lin, for informative and supportive assist and supervision throughout this work. We also would like to acknowledge the other teaching assistants Francis, Akshay, Parth, Kazem, Linji, Vishnu, Zefang, Badal, and Achu who helped and advised our team.

References

[1] A. Atanassov and D. Pilev. Pre-trained deep learning models for facial emotions recognition. In *2020 International Conference Automatics and Informatics (ICAI)*, pages 1–6. IEEE, 2020.

[2] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. 2012 advances in neural information processing systems (nips). *Neu-*

ral Information Processing Systems Foundation, La Jolla, CA, 2012.

[4] C. Marechal, D. Mikolajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, and K. Wegrzyn-Wolska. Survey on ai-based multimodal methods for emotion detection. *High-performance modelling and simulation for big data applications*, 11400:307–324, 2019.

[5] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[6] S. Ren, K. He, R. Girshic, and J. Sum. 9faster rycnn: Toward realytime object detection with region proposal networ s, 9 in. NIPS, 2015.

[7] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2014.

[8] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[9] C. Sun, Y. Yang, C. Wen, K. Xie, and F. Wen. Voiceprint identification for limited dataset using the deep migration hybrid model based on transfer learning. *Sensors*, 18(7):2399, 2018.

[10] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.