

Feeling It: Emotion Classification with Machine Learning

Using Convolutional Neural Networks to Detect and Analyze Human Emotions from Facial Images

Team Members:

Andrew Kroening

Chloe (Ke) Liu

Wafiakmal Miftah

Jenny (Yiran) Shen

Team Number: 7

Abstract

With the onset of the COVID-19 Pandemic, new challenges were introduced to the dynamics of human interaction. As people physically separated or wore face masks, interpreting emotional states became more challenging. Prior to this, computer vision advanced significantly with the arrival of deep learning and convolutional neural network architectures and techniques. This research leverages computer vision models to interrogate the effectiveness of machines at determining emotions from faces and makes a comparison to classical psychological research. The research also experiments with obstructing portions of the face to better approximate the challenges from poor webcam resolution, obstruction, or face masking. We find that computer models perform reasonably well in this space; however, there are complications posed by some similar emotions, such as anger and disgust.

Introduction

The onset of the COVID-19 Pandemic introduced new wrinkles to identifying emotions from observing faces. As much of the world began to socially distance or adopt facemasks as a standard practice, questions about the possible impacts on human interaction and emotional inference naturally emerged. While this challenge has already existed in certain cultures where the wearing of items that obstruct a part of a person's face might be tradition, the COVID-19 pandemic brought this question to a whole new scale.

One of the most notable areas where this question extends today is the more prevalent use of webcams and virtual meetings. Everyone has experienced a virtual meeting where a colleague's face might be partially blocked, poorly positioned in the view pane, or obstructed by other issues. Reading the audience's emotions can be challenging in these environments, a critical task in human communication. With this in mind, this project will seek to gain a greater understanding of where computer vision might be able to augment human interpretation when a portion of the face is obstructed and where a model's limitations are.

Background

In recent years, expanding computational power and advanced methods have brought computers and computer vision to emotion classification. We aim to incorporate and leverage this literature as a baseline understanding for many challenges in this field. To gain further insight into the performance of a model against partially obstructed faces, we explore regions of a face that are significant for predicting emotional states.

In a cursory review of available literature, we found that most were worthy attempts to explore this space. However, some of the research we uncovered needed to be updated and would benefit from more advanced models and computing power developed in the years since publication. We considered numerous pieces for application to our efforts but ultimately determined that the specific problem we sought to address was relatively novel in this space.

Ultimately, the most insightful piece to the project was from Liu, Meng, et al.¹ This paper explores various elements of human emotion, specifically facial activation during different emotional states. Of particular use, discussed in more depth later, was the high-quality mapping of these emotional activation areas. We used these maps when interpreting the saliency maps generated by our models.

¹ Liu, Meng, et al. "Facial Expressions of Emotion Categories are Embedded within a Dimensional Space of Valence-arousal." <https://psyarxiv.com/pw5uh/>.

Data

The facial expression dataset originated from Kaggle². This dataset consisted of 7 facial expressions from various gender, race, and age. In the original state of the dataset, when downloaded, we found over 35,887 images. Our exploratory data analysis began by checking these emotional categories for balance and visual sampling to ensure that the author assigned the categories correctly. The visualization of the category balance shows some skew, while the visual inspection found multiple instances that the team considered either an erroneous assignment or a potentially duplicated image.

A team member conducted an image-by-image review of the dataset to correct the erroneous categorization in the base dataset. This resulted in a slight alteration to the emotional categories, as the overall size of the dataset outweighed the small but concerning number of misclassified images. This process also brought about a discussion about the nuances of emotional interpretation for persons raised in different cultures. While we consider our reassignment to be correct, it is certainly possible that another person examining our resigned labels may find issues.

Once the reassignment was completed, the team reexamined the dataset, inspecting for duplicate images. This was smoother than the manual reassignment process. The image contents were read into a data frame and converted to a hashable data type. The contents were then compared to find duplicates that were subsequently dropped from the dataset. In total, we identified 1,506 images that had at least one exact copy. Anecdotally, we later observed cases where a face was present in more than one image, but the aspect ratio, zoom, or other features were altered very slightly. These cases escaped our inspection because they were not identical copies of one another. Figure 1 (below) presents the dataset before and after these steps were taken.

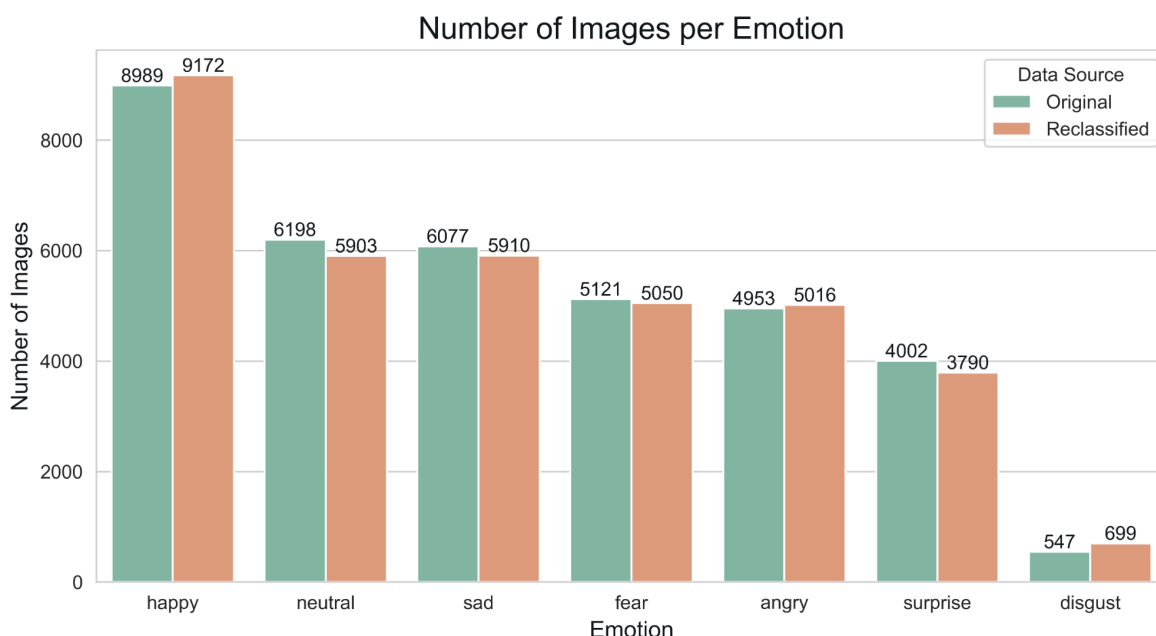


Figure 1. Categories of the Dataset

² Oheix, Jonathan. "Face expression recognition dataset." *Kaggle*, 2019, <https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset>. Accessed 9 March 2023.

The list of facial expressions and the number of images for each train and validation data are stated in Table 1 (below).

Categorical Summary		
Facial Expression	Training Images	Validation Images
Angry	3,977	888
Disgust	551	83
Fear	3,867	916
Happy	7,230	1,752
Neutral	4,580	1,181
Sad	4,696	1,088
Surprise	2,645	580

Table 1. Dataset properties (After relabeling and cleaning)

The images were also resized to a 48x48 pixel size to focus on the person's facial features, greyscaled, and labeled accordingly. Our group chose this dataset because it supports our goal of learning the features most important in facial expression recognition. This robust dataset will also allow us to experiment with data augmentation.

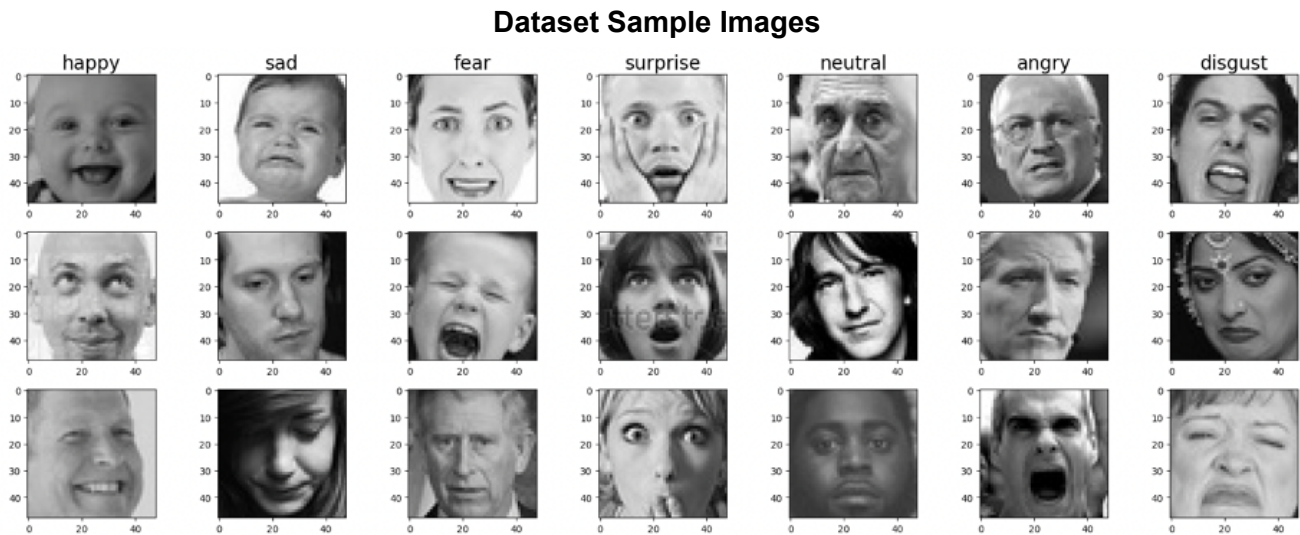


Figure 2. Dataset Excerpt

Above is an excerpt of the dataset, with three samples of each of the seven emotions. As seen from the sample images, there is considerable variation within each class that our model(s) will have to contend with. For instance, we can observe hands-on or covering parts of faces, young people and older adults, angled faces with respect to the photo, and even some watermarks. While the source of the data we derived from Kaggle is unclear, many easily identified personalities suggest these are images collected

from various internet sites and manually cataloged. We further corroborated these suspicions when we observed the incorrectly classified images earlier in data exploration.

Experiments

Methodology

The experiment methodology the team uses is laid out below in the flowchart in Figure 3. In general, we utilized a parallel process utilizing the same cleaned dataset. A VGG-16 model was used in one arm, and the other was based on a ResNet50 model. Both models were fit by comparing the performance of various hyperparameter configurations and validated against a version of the images with no masking and a version with simple masking covering portions of the face. We explored methods that relied on more advanced masking techniques but discovered unreliable performance in several cases, prompting a return to a more straightforward technique that relied on image regions rather than feature recognition.

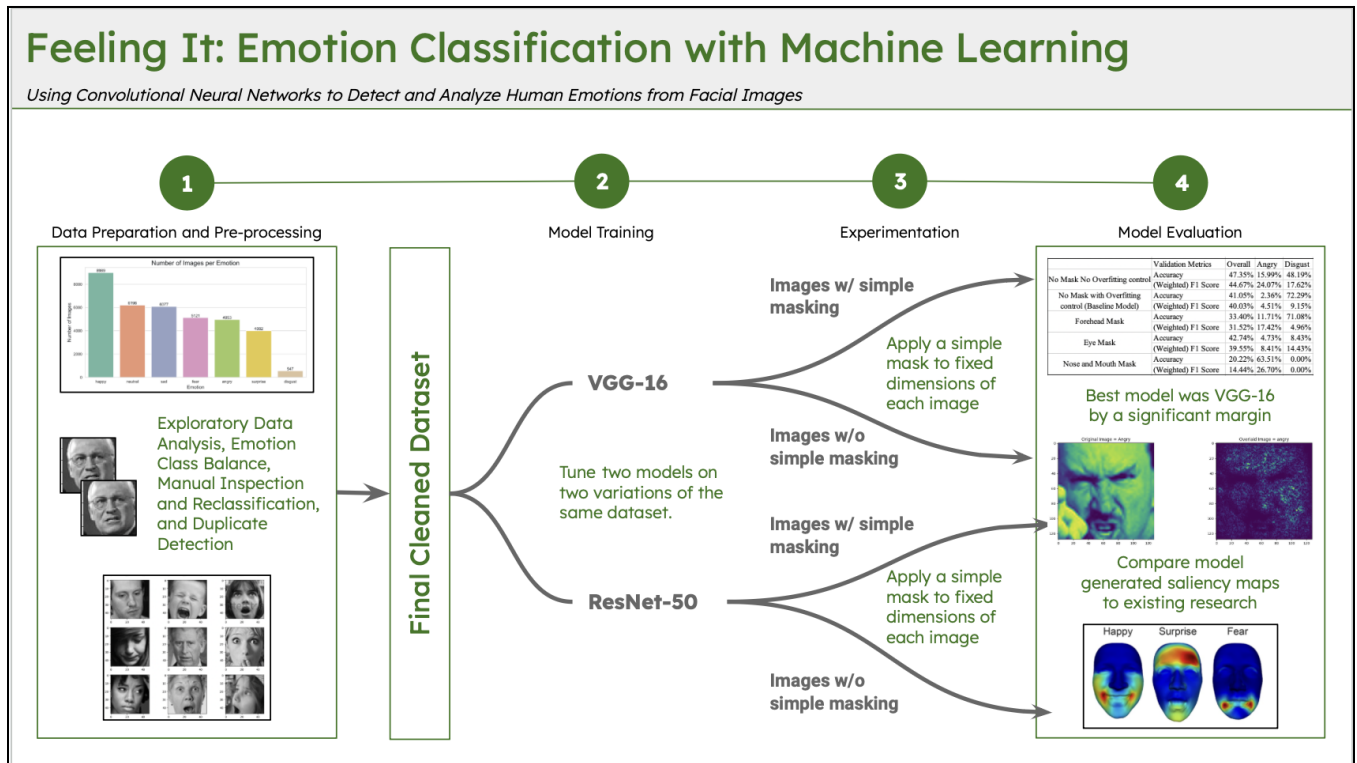


Figure 3. Experiment Flowchart

Model Architecture and Hyperparameters:

We tried two common model architectures for the Face Emotion Classification model. One followed the study by Porcu et al. and used VGG-16 as the model architecture.³ The other followed Li and Lima and utilized ResNet50.⁴

VGG-16 and Hyperparameters

The VGG-16 model architecture was utilized to create a baseline model and three masking layer augmentations. An early stopping callback function was implemented to stop training after three epochs with no improvement in validation loss. We further experimented with a dropout rate of 0.3 and batch normalization layers to account for model overfitting, which increased the overall model accuracy and decreased loss after controlling for overfitting. Therefore, we choose to use the model with batch normalization layers and dropout as the baseline model, and all models with masking layers use the same architecture.

We explored multiple values for learning rate (1e-4, 5e-4, 1e-3) and dropout rate (0.05, 0.1, 0.2, 0.3) on the baseline model. Based on the result of hyperparameter tuning, the final baseline was decided to go with a learning rate of 5e-4 and a dropout rate of 0.3.

ResNet50

The ResNet50 model was utilized to create an alternative baseline model. We also experimented with a dropout rate of 0.3 and batch normalization layers to account for model overfitting, which increased the overall model accuracy and decreased loss after controlling for overfitting. Therefore, we choose to use the model with batch normalization layers and dropout as the baseline model, and all models with masking layers use the same architecture. After exploring multiple values for learning rate and dropout rate, we decided to use the learning rate of 1e-3 and dropout rate of 0.2 as our final hyperparameter, combined with 'Adam' optimizer, Cross-entropy loss, and the same stopping callback logic as VGG-16.

Models with Masking Layers

In many real-life scenarios, facial recognition technology must face various challenges. For example, individuals may not always show their entire faces. They may be wearing glasses or sunglasses, masks in hospitals, hats for fashion, or hijabs due to religious practices. Additionally, cameras may capture only a portion of an individual's face or may be affected by changes in lighting. To address these challenges, we conducted experiments by obstructing regions of the images used by our models.

To compare the effect of different masking techniques on the model's predictive performance, we used a baseline model that was trained on the entire 128x128 image and then compared it against the

³ Porcu, Simone. "(PDF) Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems." ResearchGate, 18 November 2020, https://www.researchgate.net/publication/345940392_Evaluation_of_Data_Augmentation_Techniques_for_Facial_Expression_Recognition_Systems.

⁴ Li, B., & Lima, D. (2021). Facial expression recognition via ResNet50. *International Journal of Cognitive Computing in Engineering*, 2, 57-64.

masking layer augmented models. We explored 3 different masking layers, with each corresponding to real-life scenarios.

- Forehead Mask: corresponds to the covering of the forehead.
- Eyes and Eyebrows Mask: corresponds to the covering of the eyes and eyebrows.
- Lower Face Mask (1/2): corresponds to the covering of the mouth and lower face.

We adopted a sophisticated masking technique to detect facial landmarks for obstruction. We gained specific (x, y)-coordinates of regions surrounding each facial structure and then cropped the region out correspondingly. Below are samples and explanations for the facial landmark masking procedure.

Facial Landmark Masking Samples

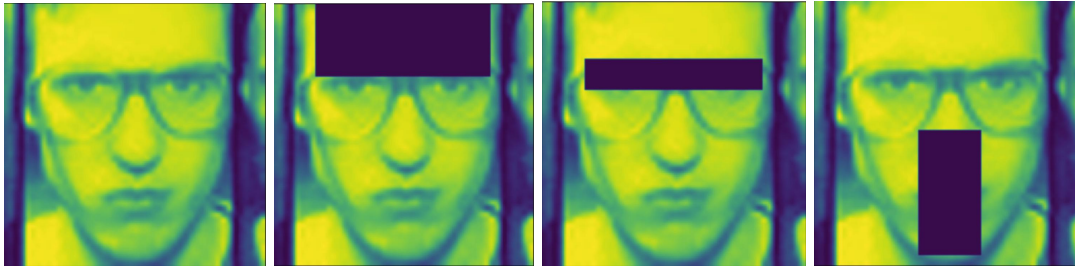


Figure 4. Facial Landmark Masking Samples

Forehead Mask

The Forehead Mask is a physical layer augmentation technique that involves reducing the number of pixels incident on the camera lens, i.e., we block all pixels incident on the camera lens above eyebrows with the width of two eyes. In Figure 4, the left image above represents the original non-masked 128x128 image, and the center-left represents the same image with a forehead mask.

Eyes and Eyebrows Mask

A Middle Mask is a physical layer augmentation technique that reduces the number of pixels incident on the camera lens in the eyes and eyebrows area, i.e., we block all pixels incident on the camera lens below the eyebrows and above eyes with the width of two eyebrows. The center-right image in Figure 4 represents the base image with a eyes and eyebrows mask applied to it.

Lower Face Mask

Lower Face Mask is a physical layer augmentation technique that involves reducing the number of pixels incident on the camera lens in the lower half of the image, i.e., we block all pixels incident on the camera lens from the middle of the nose to mouth with the width of the mouth. The right image in Figure 4 represents the base image with a lower face mask applied to it.

Training and Evaluation

Performance Comparison

Each model was trained on 9-13 epochs with a batch size of 32 images and tested after each epoch with a batch size of 32 images. We had a total of 7,230 happy, 4,580 neutral, 4,696 sad, 3,867 fearful, 3,977 angry, 2,645 surprised, and 551 disgusted images. We split the training data into an 80-20 stratified training validation split that preserved the percentage of samples for each class. Once the early stopping criteria were met or after 30 epochs, the test data was used to evaluate the model's out-of-sample performance and calculate metrics. Tables 2 and 3 (below) depict the outcomes for the VGG-16 model and ResNet50, respectively.

VGG-16 Performance Metrics

	Validation Metrics	Overall	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
No Mask No Overfitting control	Accuracy	56.41%	42.00%	36.14%	40.83%	82.53%	50.97%	44.94%	59.66%
	(Weighted) F1 Score	56.28%	45.38%	33.90%	38.70%	79.89%	51.81%	44.76%	63.31%
No Mask with Overfitting control (Baseline Model)	Accuracy	61.07%	59.23%	38.55%	33.19%	79.28%	57.32%	52.94%	78.97%
	(Weighted) F1 Score	61.01%	45.38%	33.90%	38.70%	79.89%	51.81%	44.76%	63.31%
Forehead Mask	Accuracy	60.07%	42.34%	21.69%	24.78%	89.10%	69.60%	51.38%	57.59%
	(Weighted) F1 Score	58.63%	47.81%	26.87%	34.21%	82.86%	56.95%	48.86%	66.87%
Eye Mask	Accuracy	56.20%	65.20%	22.89%	29.91%	83.39%	44.96%	46.97%	46.72%
	(Weighted) F1 Score	56.10%	46.14%	28.36%	35.06%	82.29%	50.26%	46.86%	58.66%
Nose and Mouth Mask	Accuracy	57.24%	62.95%	8.43%	31.33%	77.74%	58.34%	41.27%	62.24%
	(Weighted) F1 Score	56.85%	51.14%	15.22%	35.50%	78.98%	55.25%	44.04%	65.70%

Table 2. VGG-16 Model Results

ResNet50 Performance Metrics


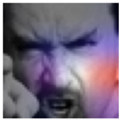





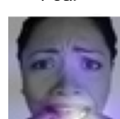
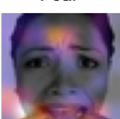

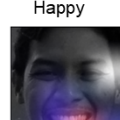

	Validation Metrics	Overall	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
No Mask No Overfitting control	Accuracy	55.90%	55.52%	42.17%	18.01%	83.22%	42.34%	55.42%	64.31%
	(Weighted) F1 Score	54.94%	47.47%	30.44%	26.17%	80.98%	49.58%	45.63%	65.04%
No Mask with Overfitting control (Baseline Model)	Accuracy	58.60%	37.39%	24.10%	29.59%	82.36%	66.05%	49.36%	72.24%
	(Weighted) F1 Score	57.69%	45.67%	27.59%	35.63%	81.55%	56.79%	47.40%	64.26%
Forehead Mask	Accuracy	53.22%	53.60%	24.10%	41.92%	80.99%	48.43%	23.16%	56.90%
	(Weighted) F1 Score	53.01%	41.99%	28.57%	35.87%	80.35%	50.96%	31.94%	61.51%
Eye Mask	Accuracy	49.65%	36.37%	26.51%	50.87%	77.97%	37.26%	35.39%	37.76%
	(Weighted) F1 Score	50.10%	37.56%	29.33%	35.93%	77.88%	42.86%	37.20%	49.72%
Nose and Mouth Mask	Accuracy	35.48%	41.67%	12.05%	42.80%	24.54%	44.45%	38.51%	26.90%
	(Weighted) F1 Score	35.54%	37.56%	12.20%	33.71%	34.97%	39.71%	32.36%	37.91%

Table 3. ResNet50 Model Results

Saliency Mapping

Once the model evaluation was complete, we generated saliency maps of the outcomes to compare the areas of importance for detecting specific emotions in our images. Below is a comparison of the saliency map for predictions created from all facial expressions determined by VGG-16 and ResNet50.

As shown in Figure 5, both models highlighted different facial features for each facial expression. For example, Resnet50 keyed into the cheekbone/nasal region in determining anger, while VGG-16 found additional importance in the central forehead region. Overall, the highlighted facial features by VGG-16 show how robust VGG-16 is in recognizing facial features, yielding a higher accuracy and f1-score than ResNet50, especially when certain facial features were covered in the experiments.

Actual Label	ResNet50	VGG-16
Angry 	Angry 	Angry 
Disgust 	Disgust 	Disgust 
Fear 	Fear 	Fear 
Happy 	Happy 	Happy 


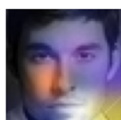


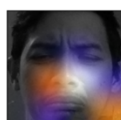
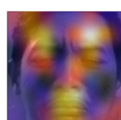

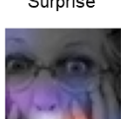

Actual Label	ResNet50	VGG-16
Neutral 	Neutral 	Neutral 
Sad 	Angry 	Sad 
Surprise 	Surprise 	Surprise 

Figure 5. Prediction Results of ResNet50 and VGG-16 Model on All Facial Expressions

With these maps generated, we can compare them with the mappings generated by Liu, Meng, et al. Below, in Figure 6, we show the mapping generated in their paper. A visual comparison shows that both models generally weighted the cheekbone/central face region as humans would when determining anger. The addition of the forehead region by VGG-16 is not expected using a human-focused interpretation. Interestingly, the overlap between disgust and anger is significant in human interpretation, which may contribute to our models' struggles in classifying these emotions. However, the baseline ResNet50 model was very good at detecting disgust (>72%) but atrocious at seeing anger (<3%).

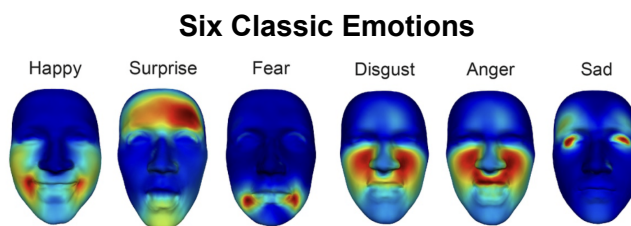


Figure 6. Classic Emotion Mapping⁵

⁵Liu, Meng, et al. "Facial Expressions of Emotion Categories are Embedded within a Dimensional Space of Valence-arousal." <https://psyarxiv.com/pw5uh/>.

Results

The above experiment provides important insights into the performance of two popular deep learning models, VGG-16 and ResNet50, in facial expression recognition tasks. Both models demonstrate significant improvements over random guessing, with VGG-16 achieving a higher accuracy and F1 score than ResNet50, even when applying masks. This suggests that the VGG-16 model, with batch normalization and dropout, may be better suited for facial expression recognition tasks.

A possible reason for the performance difference could be VGG-16's simpler architecture compared to ResNet50, which can be advantageous when dealing with smaller datasets. The VGG-16 architecture consists of 13 convolutional layers, while ResNet50 has 50 layers. The simpler architecture of VGG16 can help prevent overfitting on smaller datasets, which is important when working with facial expression datasets that may be limited in size.

The study also reveals the importance of the nose and mouth regions in recognizing facial expressions. When images with masks are used, occluding these regions leads to the lowest accuracy and F1 score compared to applying forehead or eye masks. This indicates that the nose and mouth regions are critical for correctly identifying facial expressions and that occlusion of these regions can significantly impair the ability of a facial expression recognition model to detect emotions accurately. These findings may help guide the development of more effective facial recognition models that prioritize these critical regions.

Application

We tested our model's ability to detect emotions in real-time using a webcam and tools from the OpenCV library. First, we loaded the model weights that we saved earlier after training it. Then, we used haar-cascade detection from OpenCV to help us detect the front of someone's face.

When we run the script, a new window appears on the screen, showing the webcam feed. It automatically looks for someone's face, draws a box around it, and analyzes the image to classify emotions in real-time. It's like having a virtual assistant that can read people's emotions as they're happening!

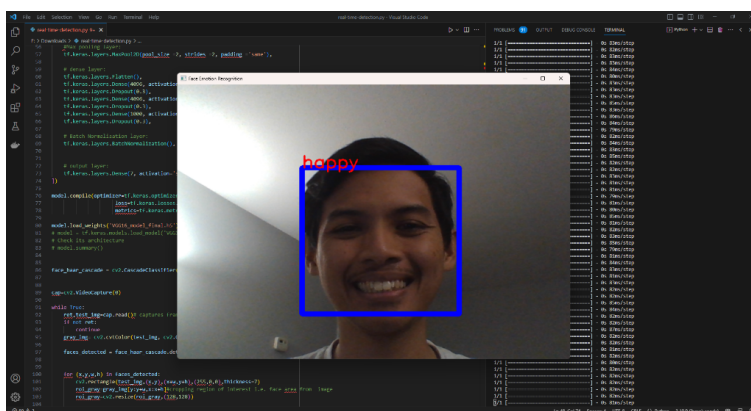


Figure 7. Real life Application of the Model

Conclusions

We find that the VGG-16 model architecture outperforms ResNet50 in detecting the emotional state of persons in images. When the dataset is augmented to obstruct portions of the face, we see alterations in the performance of both models to detect specific emotions. This is not surprising, although we find additional insights into the cause of this drop-off when we examine the saliency maps of both models and compare them to existing psychology research in the area.

There were some surprises when examining the saliency maps, such as VGG-16's addition of the forehead region for detecting anger. However, these maps generally followed our expectations. The facial obstruction and the model's performance provide insights into how the human brain might adapt to detecting emotions when other people in the environment are wearing face masks, for instance. We observed model performance increase in some cases, such as VGG-16 at detecting neutral or sad emotions. This suggests that the human brain might theoretically be able to adapt to detecting emotions in scenarios where a full face is not visible.

One limitation of the project is that the base data lacks a detailed description; we do not have aggregated information on features such as gender, race, face views, and age. As a result, it is impossible to determine if the dataset is balanced regarding these features, which may lead to potential biases in the model. These biases could negatively impact the model's generalization and limit its applicability to real-world scenarios. Therefore, we believe further investigation in this area is needed.

Roles

Andrew Kroening:

- Exploratory Data Analysis
- Report Compilation and Editing
- CVV Facial Masking
- Task Management

Chloe (Ke) Liu:

- VGG-16 Model Training and Evaluation
- Facial landmarks detection

Wafiakmal Miftah:

- Baseline ResNet50 Model Building and Training
- Saliency Maps
- CVV Facial Masking

Jenny (Yiran) Shen:

- Baseline ResNet50 Model Building and Training
- ResNet50 Model Building and Training with Simple Mask
- ResNet50 Model Performance Evaluation

References

- Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.
- Dajose, Lorinda. "Facial Expressions: How Brains Process Emotion." *Caltech*, 24 April 2017, <https://www.caltech.edu/about/news/facial-expressions-how-brains-process-emotion-54800>. Accessed 9 March 2023.
- Dores, Artemisa, et al. "Recognizing Emotions through Facial Expressions: A Largescale Experimental Study." *National Library of Medicine*, vol. 20, 2020, p. 7420. *National Library of Medicine*, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7599941/>.
- Goodfellow, Ian J., et al. "Challenges in representation learning: A Report on Three Machine Learning Contests." *Neural Information Processing: 20th International Conference*, vol. 8228, 2013, pp. 117-124. Springer, https://link.springer.com/chapter/10.1007/978-3-642-42051-1_16.
- Hase, Peter, Chaofan Chen, Oscar Li, and Cynthia Rudin. "Interpretable image recognition with hierarchical prototypes." In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, pp. 32-40. 2019.
- Heaven, Douglas. "Why faces don't always tell the truth about feelings." *Nature*, 2020, <https://www.nature.com/articles/d41586-020-00507-5>. Accessed 08 March 2023.
- Jeon, J., et al. "A Real-time Facial Expression Recognizer using Deep Neural Network." *International Conference on Ubiquitous Information Management and Communication*, 2016, pp. 1-4.
- Kim, B. K., et al. "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition." *Journal on Multimodal User Interfaces*, vol. 10, 2016, pp. 173-189.
- Liu, Meng, et al. "Facial Expressions of Emotion Categories are Embedded within a Dimensional Space of Valence-arousal." <https://psyarxiv.com/pw5uh/>.
- Minaee, S., et al. "Deep-emotion: Facial expression recognition using attentional convolutional network." *Sensors*, vol. 21, no. 9, 2021, p. 3046.
- Minaee, Shervin, and Amirali Abdolrashidi. "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Networks." 2019, <https://arxiv.org/pdf/1902.01019v1.pdf>.

- Oheix, Jonathan. "Face expression recognition dataset." *Kaggle*, 2019, <https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset>. Accessed 9 March 2023.
- O'Toole, A.J., and H. Abdi. "Face Recognition Models." *International Encyclopedia of the Social & Behavioral Sciences*, 2001, pp. 5223-5226. *Science Direct*, <https://www.sciencedirect.com/science/article/pii/B0080430767006902>.
- Pei, Zhao, et al. "Face Recognition via Deep Learning Using Data Augmentation Based on Orthogonal Experiments." *Electronics*, 2009. *Semantic Scholar*, <https://www.semanticscholar.org/paper/Face-Recognition-via-Deep-Learning-Using-Data-Based-Pei-Xu/8449af7f2950e1beacdb5d759ca743815bb59748>.
- Porcu, Simone. "(PDF) Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems." *ResearchGate*, 18 November 2020, https://www.researchgate.net/publication/345940392_Evaluation_of_Data_Augmentation_Techniques_for_Facial_Expression_Recognition_Systems. Accessed 5 April 2023.
- Ranganathan, G. "A study to find facts behind preprocessing on deep learning algorithms." *Journal of Innovative Image Processing (JIIP)*, vol. 3, no. 01, pp. 66-74.
- Shin, M., et al. "Baseline CNN structure analysis for facial expression recognition." *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 724-729.
- Vepuri, Ksheeraj Sai. "Improving Facial Emotion Recognition with Image processing and Deep Learning." *SJSU ScholarWorks*, 2021. *San Jose State University*, https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=2029&context=etd_projects.
- Zeiler, Matthew, and Rob Fergus. "Visualizing and Understanding Convolutional Networks." 2013. *Visualizing and Understanding Convolutional Networks*, <https://arxiv.org/pdf/1311.2901.pdf>.