Facial Expression Recognition with Deep Learning

Amil Khanzada (amilkh@stanford.edu), Charles Bai (cbai@stanford.edu), Ferhat Turker Celepcikay (turker@stanford.edu)

CS230 FINAL PROJECT PRESENTATION, STANFORD UNIVERSITY https://github.com/amilkh/cs230-fer

Motivation & Objectives

- Facial expressions are a universal way for people to communicate.
- Our first goal is to maximize accuracy on the test of the FER2013 dataset.
- Our second goal is to showcase a mobile web app which runs our FER models on-device in real time.

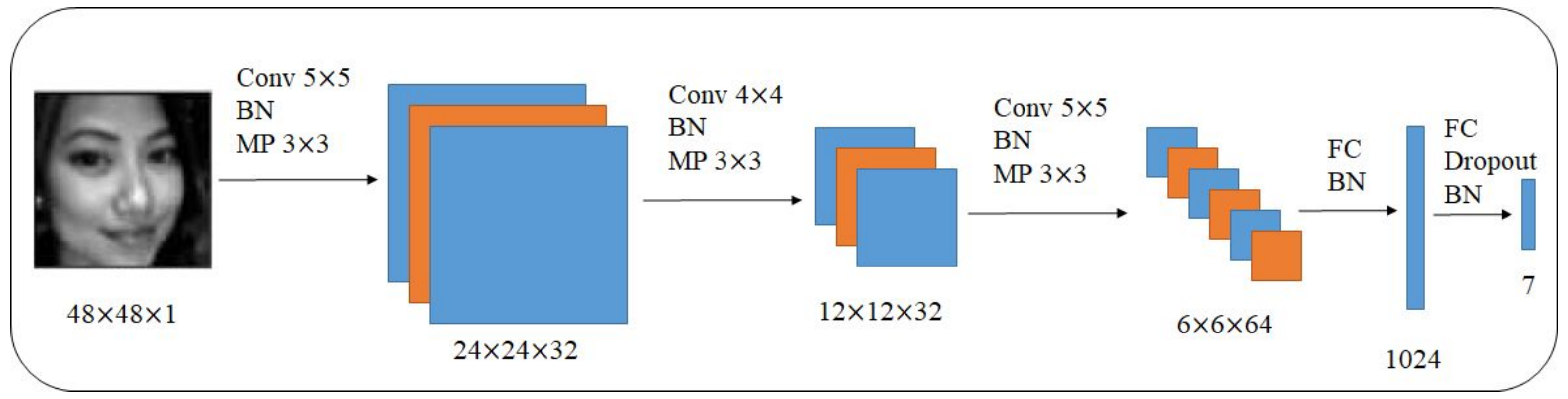
Datasets

- **FER2013:** Contains 35,887 normalized 48x48 grayscale labeled images of 7 classes, including "angry", "disgust", "fear", "happy", "sad", "surprise", and "neutral". Human-level accuracy is 65±5%.

- **Auxiliary datasets:** CK+ and JAFFE.
- **Web app dataset:** We gathered 258 labeled images from 12 people. Although ethnically imbalanced, our dataset was sufficient to meet our web app's evaluation metrics.

Models

- **Baseline:** Consists of four 3x3x32 same-padding, ReLU filters, interleaved with two 2x2 MaxPool layers, batchnorm, and 50% dropout, followed by a FC layer of size 1024 and softmax layer.
- **Five-layer:** Consists of three stages of convolutional and max-pooling layers, followed by an FC layer of size 1024 and softmax layer. The convolutional layers use 32, 32, and 64 filters of size 5x5, 4x4, and 5x5, respectively. The max-pooling layers have 3x3 kernels with strides of 2. It has batchnorm at every layer and 30% dropout after the FC layer.



- **Transfer Learning:** ResNet50, SeNet50 and VGG16 are used as the pre-trained models. The original output layers are removed and 50% dropout is applied. For ResNet50 and SeNet50, all but the last 5 layers are frozen; two FC layers of size 4096 and 1024 with 50% dropout and a softmax output layer are added. For VGG16, the model is entirely frozen and an FC layer of size 1024 with 50% dropout and a softmax output layer are added.
- **Ensemble:** We were able to achieve our highest accuracy of 75.8% by ensembling seven models: our five-layer model, ResNet50, and SeNet50, with/without class weights, and VGG16 without class weights.

Methods

- **Class Weighting:** Applied to alleviate class imbalance. We were able to drop misclassification rate from 61% to 34% for "disgust."
- **Data Augmentation:** Horizontal mirroring, ±10 degree rotations, ±10% image zooms, and ±10% horizontal/vertical shifting.
- **Test-Time Augmentation:** TTA with horizontal flip and seven augmented images improved test accuracy by 1.7% on the five-layer model.

Results

We achieved **75.8%** on our best model, outperforming the highest reported 75.2% test accuracy in a published work [2].

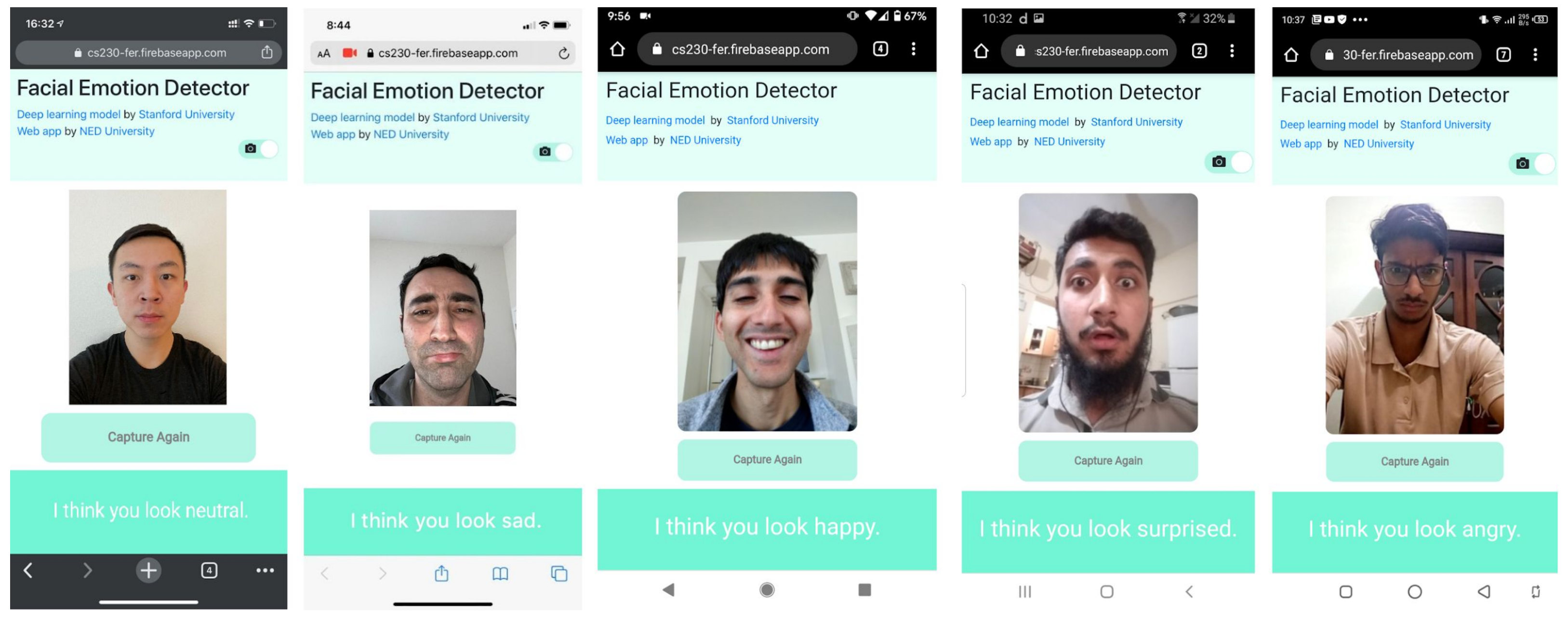
Model	Depth	Parameters	Accuracy
(Human-level)	-	-	65±5%
Tang [1]	4	12 (m)	71.2%
Pramerdorfer et al. [2]	10/16/33	1.8/1.2/5.3(m)	75.2%
Baseline	5	37.8(m)	64.0%
Five-Layer Model	5	2.4(m)	66.3%
VGG16	16	128(m)	70.2%
SeNet50	50	27(m)	72.5%
ResNet50	50	25(m)	73.2%
Ensemble	-	-	**75.8%**

Adding auxiliary data improved accuracy on most of our models.

Dataset	**ResNet50**		**SeNET50**		**VGG16**		**Ensemble**
	NCW	*WCW*	*NCW*	*WCW*	*NCW*	*WCW*	
FER2013	73.2%	67.7%	70.0%	68.9%	69.5%	70.0%	74.8%
Auxiliary	72.7%	72.4%	72.5%	71.6%	70.2%	69.6%	75.8%

NCW = no class weights / WCW = with class weights

Mobile Web App



Facial Emotion Detector

Deep learning model by Stanford University

Web app by NED University

I think you look neutral.

I think you look sad.

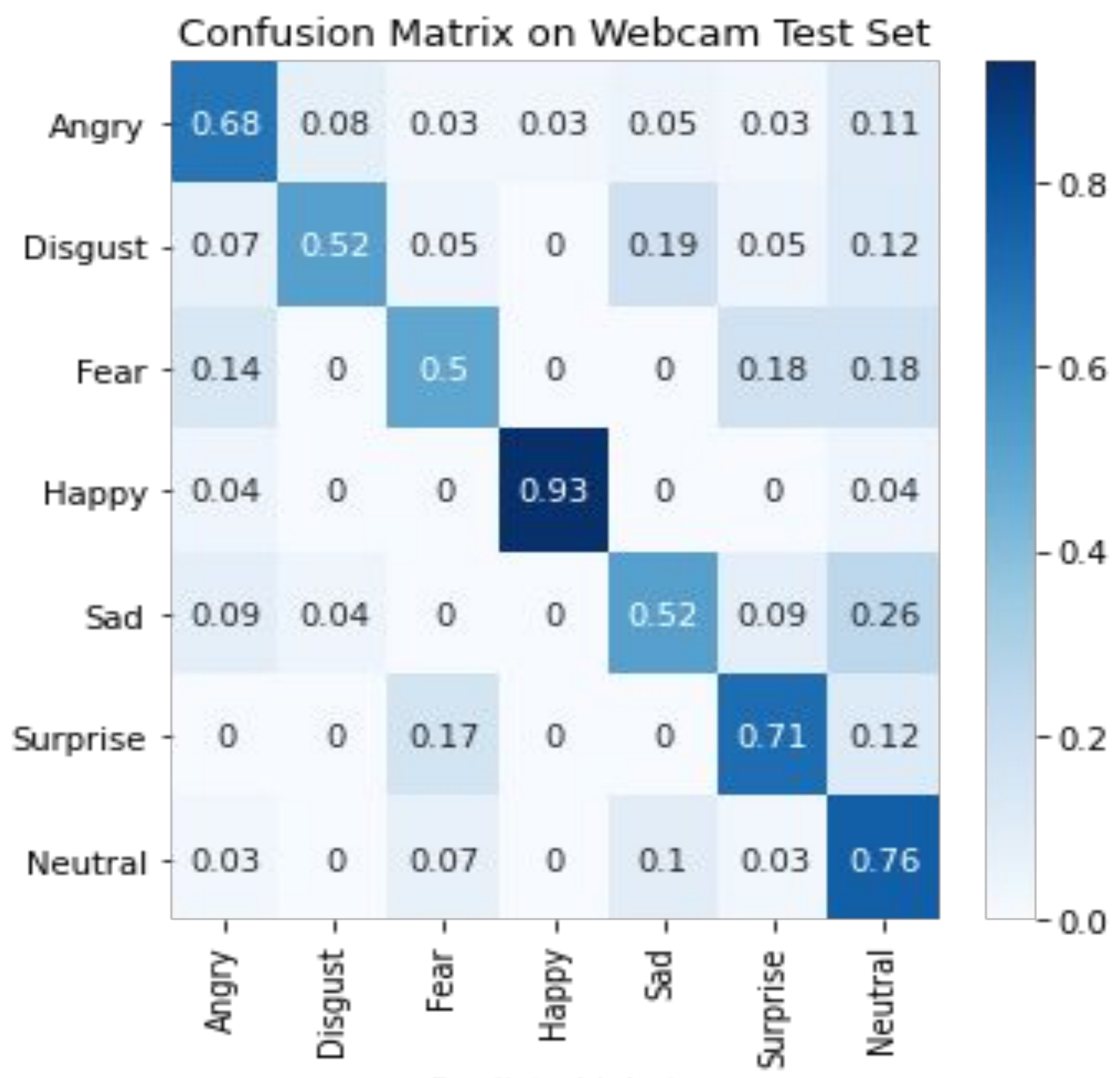
I think you look happy.

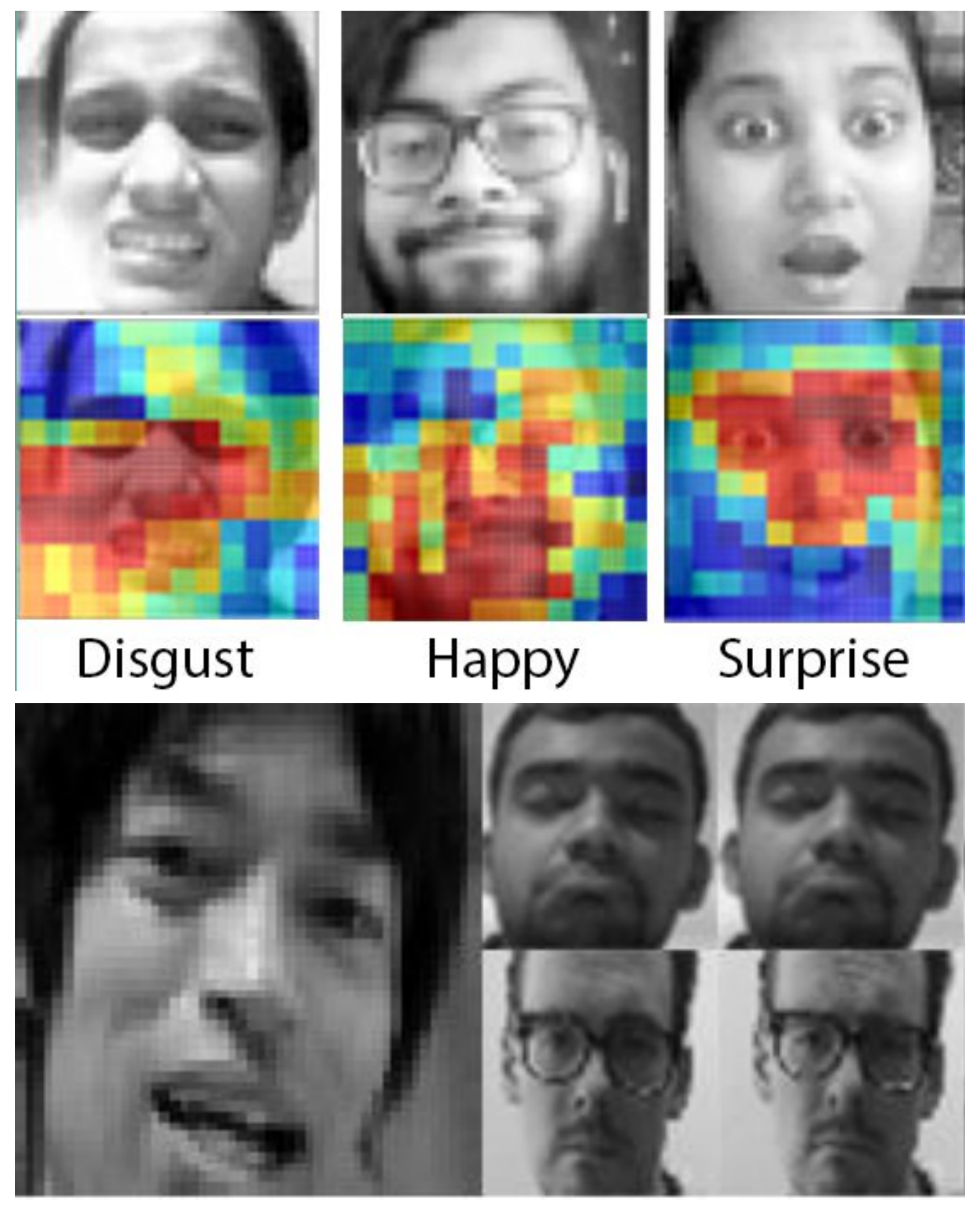
I think you look surprised.

I think you look angry.

- **Frameworks:** Firebase, TensorFlow.js, React.js, face-api.js.
- **Satisficing Metric:** 100ms recognition speed on-device.
- **Optimizing Metric:** accuracy on the webcam test set.
- **Dataset Mismatch:** Overcame poor illumination and tilted angles with 80%/20% train/test split on our web app dataset.
- **Results:** Achieved 40ms recognition speed and 69.8% accuracy after training for 120 epochs on the five-layer model.
- **Web Link:** http://cs230-fer.firebaseapp.com/

Discussion and Error Analysis





- **Ambiguity:** One image was classified angry (29%), fear (28%), and sad (26%), similar to mispredictions by humans on the same image.
- **Misclassifications:** Sad images were often predicted neutral with one of the subjects misclassified on all images. We addressed this by augmenting our dataset with more sad images from the web app.
- **Interpretability:** The network learned to focus on the mouth and nose to make predictions for disgust, mouth for happiness, and eyes and nose for surprise. For neutral images, it focused on all parts of the face except for the nose, which made sense given that small changes in non-nose regions tend to correspond to emotion changes.

Future Work

- Improve accuracy with facial landmark alignment, attentional CNN, occlusion of irrelevant facial features, more auxiliary data, balancing dataset, pipelining models, and additional data augmentation.
- Enhance real world applicability by investigating valence/arousal emotional models and investigating social good use cases.
- Addressing ethnicity bias issue in existing facial datasets by starting the Pakistani Female Facial Expression dataset project (PKFFE.org).

Acknowledgements

- Pawan Nandakishore, data scientist at Colaberry, for guidance in research of related works, interpretability, and model tuning.
- NED University for web app development, auxiliary data, and dataset preparation.

References

1. S. Li and W. Deng, "Deep facial expression recognition: A survey," arXiv preprint arXiv:1804.08348, 2018.
2. Pramerdorfer, C., Kampel, M.: Facial expression recognition using convolutional neural networks: state of the art. Preprint arXiv:1612.02903v1, 2016.
3. 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.