Emotion Detection with Ensemble-based CNN

Yun Cheng (yuncheng@andrew.cmu.edu), Yuxin Pei (yuxinp@andrew.cmu.edu), Zhiyi Kuang (zkuang@andrew.cmu.edu)

Carnegie Mellon University

Motivation & Objectives

Facial expression is one of the most powerful, natural and universal signals for human beings to convey their emotional states and intentions. [1]

- Our goal is to utilize the FER-2013 train data to maximize classification accuracy on the test data provided.
- We applied the state-of-the-art models in the field.

Datasets

FER-2013: The data consists of 48x48 pixel gray-scale images of faces classified into seven categories (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral). The training set consists of 28,709 examples and the public test set consists of 3,589 examples.

Figure 1:FER-2013 Dataset Samples

Models

- **Baseline:** consists of 3x3x32 same-padding, ReLU filters, 2 2x2 MaxPool layers, batchnorm, 50% dropout, a FC layer of size 1024 and softmax layer.
- **ResNet50/SeNet50:** pre-trained models with the output layers removed. All layers except last 5 are frozen, 2 FC layers of size 4096 and 1024 with 50% dropout and a softmax output layer are added.
- **VGG16:** pre-trained models with the output layers removed. The feature extraction layers are frozen. 2 FC layers of size 4096 and 1024 with 50% dropout and a softmax classification layer are added.
- **Ensemble:** We ensembled four models(baseline, VGG16, ResNet50, SeNet50) to improve significantly from individual models and achieve **72.2**% accuracy.

Methods

Data Preprocessing

We combined the typical pipeline for preprocessing data with face normalization methods [4]:

- **Facial Landmarks Registration + Affine Transformation**: based on extracted facial landmarks, a canonical alignment is obtained for each face image using affine transformations.
- **Illumination Normalization**: every image is normalized to have a mean of 0 and a norm of 100.
- **Pose Normalization**: frontalization synthesizes frontal facing views of faces appearing in single unconstrained photos. Local facial features (using face detector) are projected to the 3D reference coordinates and soft symmetry is used to produce front facing views of photos [2].

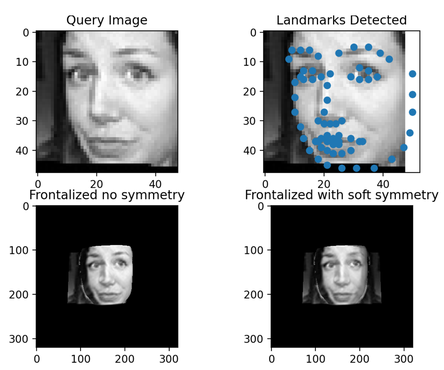

Figure 2:Frontalization Process Example[2]

Multi-class SVM Classification in CNN

We'll attempt to replace the softmax layer objective with the SVM objective

$$\ell(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\max(1-\mathbf{w}^T\mathbf{x}_i, 0)$$

Differentiate the SVM objective with respect to the penultimate layer activation $\mathbf{h}$ [6],

$$\frac{\partial \ell(\mathbf{w}}{\partial \mathbf{h}_i} = -C\mathbf{w}\cdot\mathbb{I}_{\{1>\mathbf{w}^T\mathbf{h}_i\}}$$

Transfer Learning and Ensemble Method

We ensemble baseline, ResNet50, SeNet50, and VGG16, and take a majority voting combination of the classification outputs for each test instance without class weights.

Results

We achieved 72.2% testing accuracy on our ensemble model of baseline, VGG16, ResNet50, and SeNet50. The highest reported state of the art accuracy is 75.2%[5].

Model	Depth	Test Accuracy
Baseline	5	67.5%
VGG16	16	68.3%
ResNet18	18	43.4%
ResNet50	50	68.5%
SeNet50	50	62.4%
Ensemble	-	**72.2%**

Table 1:Summary of Model Accuracy

Error Analysis

We mainly rely on confusion matrix to measure and analyze the classification performance of our models.

1. Among all the models, classifying *Happy* achieves the highest accuracy score;
2. The models generally perform poorly on *Fear* and *Sad* categories;
3. Resnet50 seems to complement on some categories that Senet50 achieves less desirable outcomes and vice versa. Adding this pair of models into the ensembling model boosts classification accuracy.

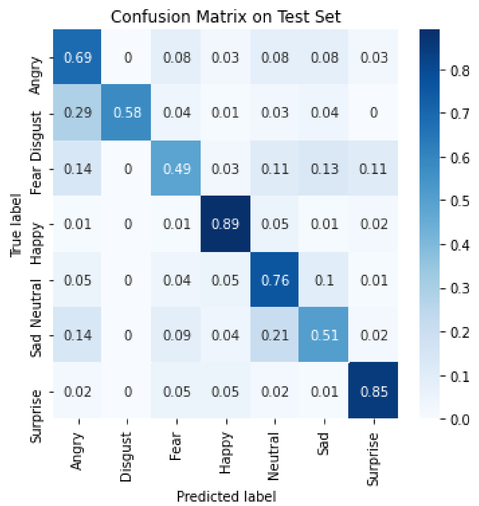

Figure 3:Confusion Matrix of the Ensemble Model

Interpretability

We intend to improve the model interpretability using occlusion-based saliency map, which highlights specific facial features that the models are sensitive to. In particular, we want to focus on those misclassified examples and investigate if

1. the model focuses on less significant features, or
2. the misclassified image contain mixed emotions that are also confusing to human

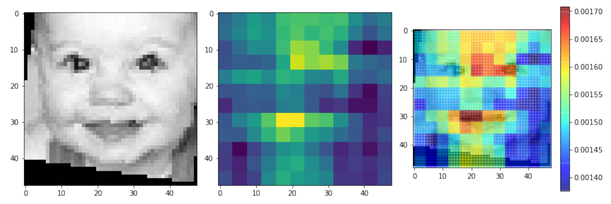

Figure 4:Saliency Maps Example

Future Work

- Improve model accuracy by performing data augmentation and additional robust facial frontalization methods, training on auxiliary dataset
- Experiment with other types of networks like Siamese Net and losses like Triplet loss (a type of contrastive loss that aims to maximize intraclass difference)

References

[1] C. Darwin and P. Prodger, *The expression of the emotions in man and animals.*, (1998).

[2] T. Hassner, S. Harel, E. Paz, and R. Enbar, *Effective face frontalization in unconstrained images*, CoRR, abs/1411.7964 (2014).

[3] A. Khanzada, C. Bai, and F. T. Celepcikay, *Facial expression recognition with deep learning*, 2020.

[4] S. Li and W. Deng, *Deep facial expression recognition: A survey*, IEEE Transactions on Affective Computing, (2020), p. 1–1.

[5] C. Pramerdorfer and M. Kampel, *Facial expression recognition using convolutional neural networks: State of the art*, 2016.

[6] Y. Tang, *Deep learning using linear support vector machines*, 2015.

Acknowledgements

This work is inspired by the awesome project by Khanzada *et al.* [3].

Carnegie Mellon University