EmoNet: Facial Expression Recognition using Deep Learning

Lilly Kumari (lkumari@uw.edu)

Practical Introduction to Neural Networks - EE 596B - Spring 2019

UNIVERSITY of WASHINGTON

Introduction

More than 90 % of the human communication is nonverbal. So, understanding and recognising human expressions is crucial and has intensive usage in areas such as human-computer interaction, security & surveillance, robotics and advertising. So, this project focuses on exploring several deep convolutional neural networks (CNNs) such as Inception v2 and ResNet 50, for recognizing seven basic human emotions. It also leverages several deep learning techniques for better convergence and ensemble learning to achieve the best results.

Dataset

We use the FER-2013 (Facial Expression Recognition) Kaggle dataset [1] which is uniformly distributed with respect to gender, race and ethnicity of the subjects. This dataset consists of 35887 grayscale (8 bit) images of size 48 x 48, out of which 28709 images comprise the training set with validation and testing set containing 3589 images each. The images are annotated with 7 key emotions: *happiness, sadness, anger, disgust, fear, surprise & neutrality*, with individuals posing at various angles. FER-2013 has almost an even distribution of emotions across the subjects except for the "disgust" emotion as shown in Figure 4

Figure: Sample images from FER 2013 dataset

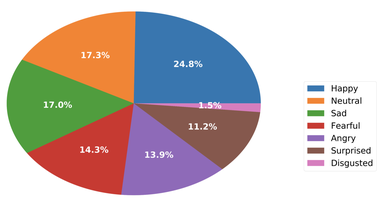


Figure: Distribution of 7 emotions in FER 2013 dataset

Approach & Models

We use two well known state of the art convolutional neural networks, Inception ResNet v2 (Model A) and ResNet-50 (Model B) models to extract feature vectors from raw image pixels data. These ConvNets can successfully capture the spatial and temporal dependencies in an image via the application of relevant filters, along with the benefit of reduced parameters to be learnt. Both networks make use of residual (or skip) connections [2] to make the network more dynamic so that it can optimally tune the number of layers during training instead of treating the number of layers in the network as an hyperparameter to tune.

Along with the CNN based feature vectors, we also feed extra information in the form of bag of visual words which include facial landmarks and Histogram of Gradients (HOG) to the fully connected network described in Figure 7. The entire network with Inception-ResNet v2 model as convolutional feature extractor has 59,231,463 parameters whereas the network using Resnet-50 model as convolutional feature extractor has 28,980,167 parameters.

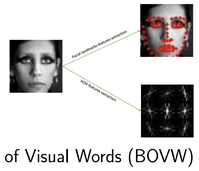


Figure: Bag of Visual Words (BOVW) Feature Extraction

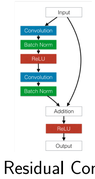


Figure: Residual Connections

Network Architecture

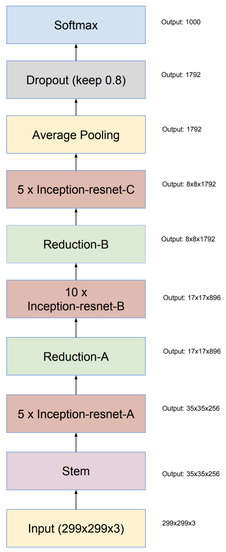


Figure: Inception-ResNet v2 - Model A

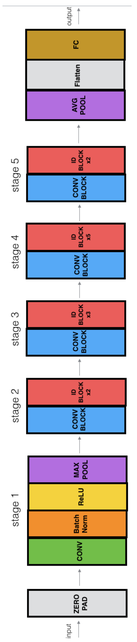


Figure: ResNet-50 - Model B

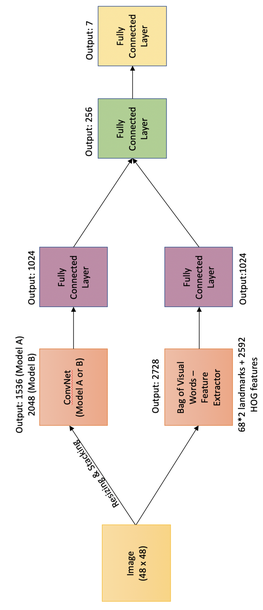
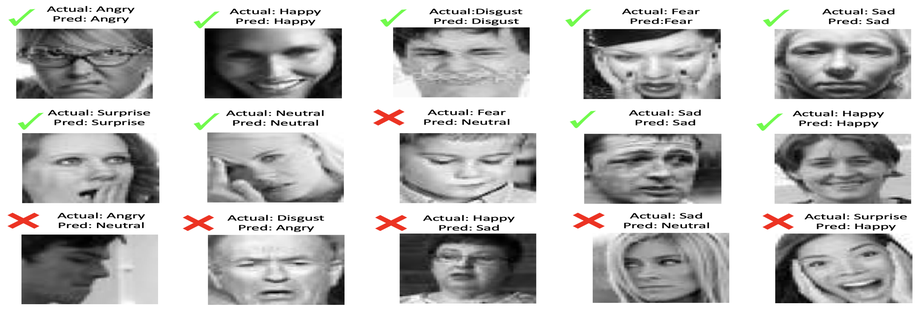


Figure: Overall Network Architecture

Qualitative Results



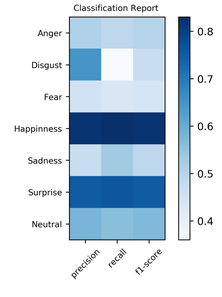


Figure: Network using Inception-ResNet v2 Model (A) for raw pixels + BOVW

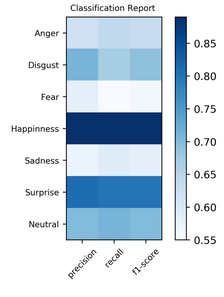


Figure: Network using ResNet-50 Model (B) for raw pixels + BOVW

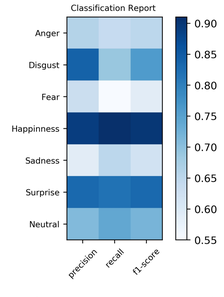


Figure: Ensemble Learning Using Model A and Model B outputs

Implementation Settings

- openCV and dlib libraries used for face detection in order to filter out images containing zero or more than one faces.
- Data processing: Mean centering, standard resizing and stocking, dlibś pretrained facial landmark detector used to identify 68 landmarks, HOG feature extraction with a sliding window size of 24 and a window step of 6
- *Transfer Learning* used for Convolutional Feature Extractor Networks - network initialized from ImageNet weights in Model A and VGGFace weights in Model B
- For first 5 epcohs with a batch size of 64, the pretrained convolutional blocks are frozen and the remaining network is trained using an Adam optimizer with a learning rate of 0.001 with categorical cross entropy loss.
- For next 50 epochs with same batch size, the entire network is fine tuned using a SGD optimizer with nesterov momentum factor of 0.9 and a learning rate of 0.001.
- Learning rate scheduler is used to shrink the learning rate by monitoring validation loss changes and early stopping criterion is used to prevent the model from overfitting on the training set.
- Used differential evolution algorithm for learning weighted ensemble of different model outputs.

Quantitative Results & Conclusions

We report the performance of our models with respect to overall accuracy as well as per-class precision, recall and f1-score metrics. The results show that weighted average ensemble model learnt using both Inception ResNet v2 and ResNet 50 models outputs perform the best.

	Accuracy
Inception-ResNet v2 (pixels only)	62.75%
Inception-ResNet v2 (pixels + BOVW)	64.55%
ResNet-50 (pixels only)	68.92%
ResNet-50 (pixels + BOVW)	70.97%
Weighted Ensemble Learning	73.39%

Table: Performance of different models on FER-2013 dataset

The state of the art performance till now on FER-2013 dataset is 75.1% but the improvement is made via augmenting three datasets together in order to increase the training data size and compensate for the small size of FER images [3].

References

[1] "https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge,"

[2] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[3] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: state of the art," *arXiv preprint arXiv:1612.02903*, 2016.