Real-time Facial Emotion Recognition Using Convolutional Neural Networks

Anmol Singh Suag
Department of Computer Science
University of Massachusetts Amherst

Motivation

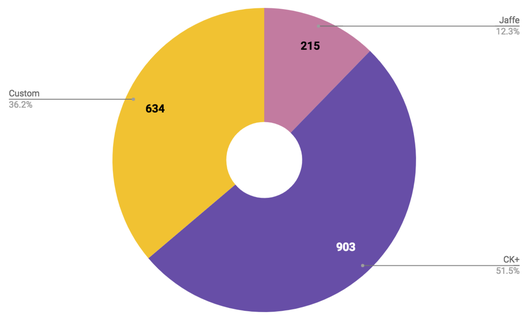
- Recognising Facial Expressions is vital for understanding human emotions
- Myriad applications make it an active area of research
- CNNs models have proved to be very successful in classifying facial expressions
- Difficult because of lack of extensive datasets to train and processing time for real-time classification

Approach

- Collect a dataset using webcam images, apply illumination variations and save as multiple images.
- Convert CK+ and Custom dataset to JAFFE format to have same labels for the 7 emotions : Angry, Disgust, Fear, Neutral, Happy, Sad and Surprise
- Fine-tune a pre-trained VGG_S network trained on CASIA dataset on a collection of JAFFE, CK+ and custom dataset
- All training images are first cropped to face using Multi-task Cascaded CNNs/Haar Cascades.
- Collect realtime video from webcam, pre-process the frame, find faces bounding boxes and feed them into the fine-tuned CNN
- Get classified emotions for all bounding boxes using an ensemble of fine-tuned and pre-trained network (optional for processing time reduction)
- Output frame with bounding boxes and classified expression with probability
- Analyse Convolution layer filters, confusion matrix and optimise for quicker real-time classification

Datasets

- Japanese Female Facial Expressions (JAFFE)
- Extended Cohn-Kanade Dataset (CK+)
- Custom dataset with illumination variants

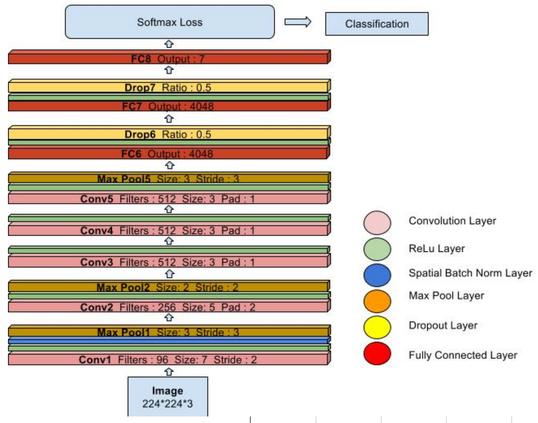


Network Architecture and Training

Network Architecture

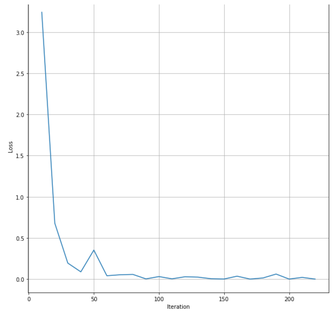
Pre-trained EmotiW Network has the following architecture. Original network has been trained on CASIA-Webface dataset.

The network was Fine-tuned further on our collective data-set starting from the pre-trained weights for about 250 iterations.

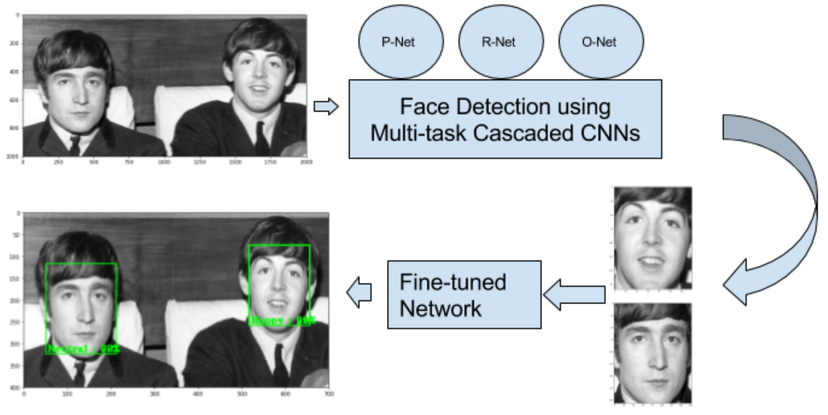


Training

As the weights were already learnt, loss came down quite quickly.

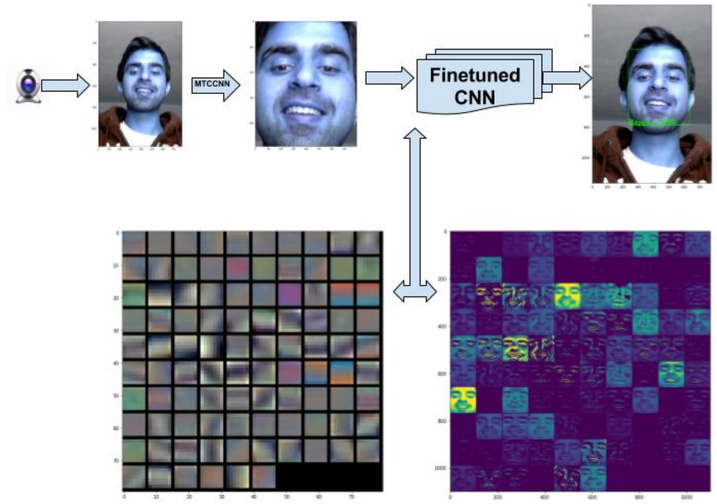


Execution Pipeline



- Multi-task Cascaded CNN is a cascaded architecture with 3 stages of CNNs to predict faces in a coarse to finer manner. Pre-trained MTCCNN was used as is as it provides state-of-the art accuracy. A faster implementation here could have been Haar Cascades
- Faces from the input image are fed into the fine-tuned CNN to classify the expressions
- Rectangles for bounding boxes and classified expression is superimposed on the input frame

Real-time run & Layer Visualisation



1st Conv Layer Filters

1st Conv Layer Output

- A frame from webcam video is processed using MTCCNN to find bounding boxes. These cropped images are classified by the Fine-tuned CNN network
- 96 filters of the first Convolution layer and their outputs are visualised above

Results

Accuracies

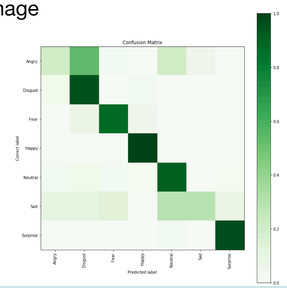
- Original VGG_S Network had a train accuracy 23% on CK+ dataset. Cropping the bounding boxes increased the train accuracy to 86.58%.
- Fine-tuned VGG_S on CK+ and Jaffe received a train accuracy of 90.4% on CK+.
- Test accuracy on custom dataset was 21.3%.

Execution Time

- Time to find Bounding Boxes = 0.143 sec/image
- Time to classify = 0.347 sec/image
- Total Time = 0.490 sec/image

Confusion Matrix for CK+

- Confusion matrix shows the test accuracies as found on CK+ and helps visualise what confuses the network
- Network performs best for Happy and worst for Sad expression



References

- *Levi, G., Hassner, T.: Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In: ACM ICMI (2015)*
- *M. K. M. J. Lyons and J. Gyoba. "japanese female facial expressions (jaffe)," database of digital images, 1997.*
- *T. K. J. S. Z. A. P. Lucey, J.F. Cohn and I. Matthews. "the ex- tended cohn-kanade dataset (ck+):*
- *K. Zhang and Z. Zhang and Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks",IEEE Signal Processing Letters, 2016*
- *S. Ouellet. real-time emotion recognition for gaming using deep convolutional network features, corr, vol. abs/1408.3750, 2014.*