Convolutional Neural Networks: Real Time Emotion Recognition

Bruce Nguyen, William Truong, Harsha Yeddanapudy

Motivation:

Machine emotion recognition has long been a challenge and popular topic in the realm of research. Our project garnered inspiration from two research papers in specific, “Emotion recognition through facial expression analysis based on a neurofuzzy network,” by Spiros V. Ioannou, Amaryllis T. Raouzaiou, Vasilis A. Tzouvaras, Theofilos P. Mailis, Kostas C. Karpouzis, and Stefanos D. Kollias and “Real Time Facial Expression Recognition in Video using Support Vector Machines,” by Philipp Michael and Rana El Kaliouby. We wanted to take a different approach towards emotion recognition and use a convolutional neural network, granted these networks are specially designed to classify images.

Convolutional Neural Network Architecture:

The architecture of the network is relatively simple. The convolution neural network we designed consists of three convolutional layers. Each layer utilizes filters of 5x5 patches and 2D convolution with a stride size of 1, 0-padded. The first convolutional layer extracts 32 features (from 1 input channel because the images are gray-scaled), the second, extracts 64 features, and the third, 128 features. Between convolutional layers, we used 2x2 max-pooling in order to reduce the spatial size, reducing the amount of parameters and computation in the network, which helps combat overfitting. Lastly, we used a fully-connected layer with 2048 neurons leading into a softmax classifier to decode the image’s class. To further prevent overfitting, we used the regularization technique of dropout with a keep probability of 50% just before the fully-connected layer.

Data:

The convolutional neural network was trained on the Chicago Face Database(CFD), developed at the University of Chicago by Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. This dataset includes standardized, high-resolution photos of faces representing five categories of expressions: neutral, angry, happy(with open mouth), happy(with closed mouth), fearful, giving our network the capability to predict these five emotions. The dataset contained pictures distinguished by specific demographics: First by gender: Male, Female and further by race: Latino, Asian, Black, White. The breakdown of the images by emotion class are as is: 153 happy(with closed mouth), 154 happy(with open mouth), 149(fearful), 154(anger), and 597(neutral), which totals to 1207 images. More information on the dataset can be found here: http://faculty.chicagobooth.edu/bernd.wittenbrink/cfd/index.html. See Figure 1 for an example.

Fig 1, Example Chicago Face Data

Training Methodology:

The network trained on images of the CFD that were preprocessed. All images were cropped to only include just the portion of the image that entailed the subjects face, as facial expressions are indicators of emotion. Then, the images were reduced to a size of 31x32, where they were grayscaled (see Figure 2). This allowed the network to train efficiently and quickly permitting us to run many different trials. The network trained on 90% of these images and tested on the other 10%. After 20 epochs, we were able to attain a test accuracy of ~80%.

Fig 2, Image pre-processing step

Notably, further epochs passed 20 did not result increased performance and the model converged at about 80% test accuracy. We noticed that letting the model run for more than 20 epochs only resulted in an increase in training accuracy and not testing accuracy. This led us to believe that the network was not overfitting per se, but rather learning features that were not relevant. Additionally, we started with a very simple baseline CNN that is similar to our current model except it had two layers instead of three and the fully connected layer only had 1024 neurons. Increasing the layer size by one and adding more neurons in the fully-connected layer resulted in minor improvements, improving our test accuracy by about three percentage points or so, prompting us to use this is our final architecture. In effort to try and increase our testing accuracy, we introduced more complexity to the model by increasing the number of layers, features, and neurons in the fully connected layer. However, these changes only resulted in either no change or a decrease in test accuracy.

Results:

The convolutional neural network performed well in practice when integrated into our web-application. Our web-application utilized a webcam to take a picture every 2.5 seconds and sent the image to the trained CNN in the back end (see Figure 3).

Figure 3, webcam display followed by cropping process on server

Once the image was decoded by the CNN, a response was sent back to the application which was then decoded to represent an emotion that was displayed on the application's web page (see Figure 4). The convolutional neural network did an outstanding job, all on faces that it had never seen before.

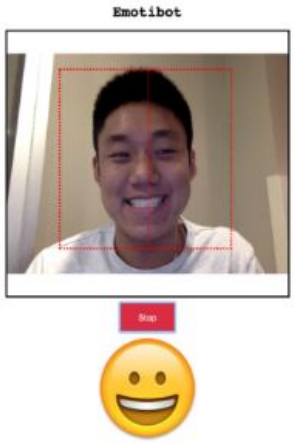

Figure 4, web-application front end

Although we did not gather any form of statistics, on a general basis, we found the application to be usable and effective in recognizing emotions of users. We can safely say that the CNN did, in effect, perform as though its test accuracy were 80%+. The results were not state of the art, but indeed, eerily accurate given our limited dataset and simple network architecture. (see Figure 5)

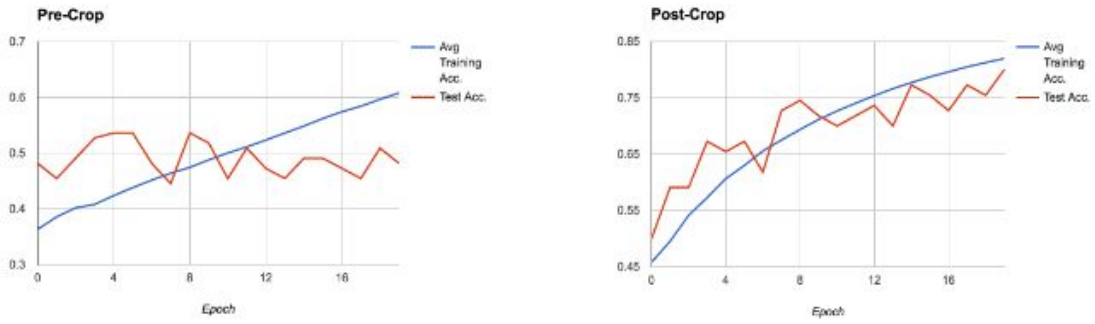

Figure 5, results pre-crop versus post-crop preprocessing

The drawbacks of our application were not so much the CNN, but rather the manner in which the application took photos for the CNN to classify. As noted earlier, our CNN is effective, only if it is given an image of just the face of a subject; incorporating entire images creates too much noise for the network to accurately identify an emotion as not cropping the photos only resulted in about 55% test accuracy on the exact same network architecture. Since our network was trained in such a manner, images must be in the same format for accurate classification. Our method of approaching this nuance was to simply just crop a fixed section of the photo that the web application takes. The web application very roughly specifies which portion of the photo the face should be placed in. With this method of cropping, this means that although the network can identify emotions in real time, there are some very strict guidelines as to where the user's face should be placed. Deviating from these guidelines drastically detriments the applications capability of prediction.

Conclusions:

Convolutional neural networks are ground breakingly effective in the task of image classification. Although we only obtained a test accuracy of approximately 80%, we argue that is level of accuracy is stunningly robust, given our network architecture and dataset. With more time to fine tune the many parameters of our network, we are highly confident that we could find a more tailored CNN architecture that suits our specific goal of emotion recognition. In addition to its simple architecture, we trained the model on an inadequate dataset that did not nearly have enough images per class to constitute the effective learning of five classes of images. Our project serves as an example as to how resilient even a simple convolutional network could be under adverse and inconvenient circumstances.

In practice, we found that the neural network had great predictive power if we were to place our faces in a specified location of the camera. Hence, this leads us to believe that a big part of incorrect predictions were due to the poor quality of pictures being sent to the network. Looking forward, we would look to first, use a better camera to get images similar to the ones that the network trained on, which were of much higher resolution. Secondly, we could easily incorporate the use of a facial recognition algorithm that would crop the image accordingly for us, instead of manually cropping specified coordinates of the image, which may or may not contain the user’s face. Lastly, we would continually train our network and include more types of

images to train on. The images that our network trained on, which were the images in the CFD database, were drastically different than webcam photos that our web application took.

The implications of our research will be potentially impactful. Creating a highly accurate real time recognition system has many applications and could change the realm of HCI. Specifically, the landscape of gaming and virtual reality could be vastly altered if games were to recognize their players emotions and respond accordingly. Moreover, it could lead to new emotion based gestures, which may revolutionize the way consumers can interact with their electronic devices. Real time emotion recognition may also help shape the realm of data collection. For example, businesses can gather data on the emotional responses their viewers exhibit when viewing their content, which would help quantify how effective their content is in eliciting desired reactions. In summation, this new form of emotional response data adds another dimension to data analytics.