Neural Network Approach

We decided to test the accuracy of a neural network in distinguishing eye contact. Our labelled examples consisted of feature data from OpenFace for different frames in several Youtube videos of infants followed by a 1 or 0, with 1 indicating eye contact in that frame and 0 indicating no eye contact.

To begin with, we normalized feature data from OpenFace so that all inputs are between -1 and 1. To do this, we divided each feature by the absolute value of the largest value for that feature. For example, if we found the largest value of the xgaze feature in all frames of the video to be 50, then every xgaze value would be divided by 50. This removes any inherent bias toward larger value features, since these features would otherwise appear to have heavier weights when evaluated by the NN.

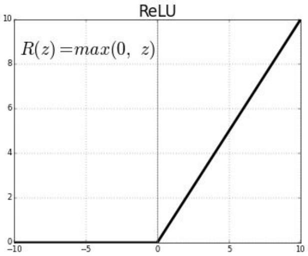
We then divided the labeled examples into training data and testing data. We designated certain videos to be used for training and for testing with an approximately 3 to 1 ratio (3 training videos for every 1 testing video). We then randomized the array of training examples to ensure the training of the model is not influenced by the order of the videos.

Lastly, we designed the model and used gradient descent with a stochastic variable learning rate. The learning rate is the rate at which perceptron weights are updated in each iteration, and depends on how much the total loss decreases in response to that update. Using stochastic gradient descent prevents overfitting by forcing the weights to occasionally update “against the gradient” or in a direction that increases loss. However, the probability of this occurring decreases at higher iterations. Additionally the stepping rate is considered variable because the magnitude of change to weights decreases in later iterations. These attributes ensure that our gradient descent algorithm will eventually converge on a given value for each weight.

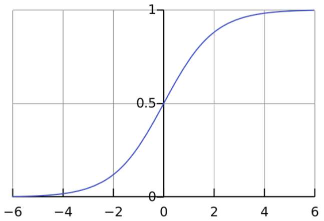
The neural network itself consisted of 3 layers: input, hidden, and output. Each perceptron in the input layer received all feature data for a given frame, and applied a ReLU (Rectified Linear Units) evaluation to determine output. Each perceptron of the much larger hidden layer used the outputs of all input layer perceptrons and applied a softmax evaluation to produce output. Finally the output layer perceptrons used the outputs from the hidden layer and applied a unit step to produce a final 1 or 0.

We found this model worked because because the ReLU mapped inputs linearly between zero and 1, preserving the proportions in the feature data, whereas the hidden and output layers used steeper evaluations, eventually producing either a 1 or 0 with no intermediate values.

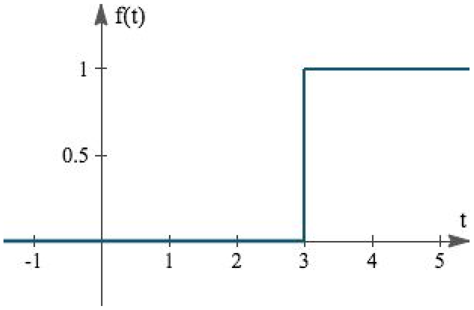
ReLU activation



Softmax activation

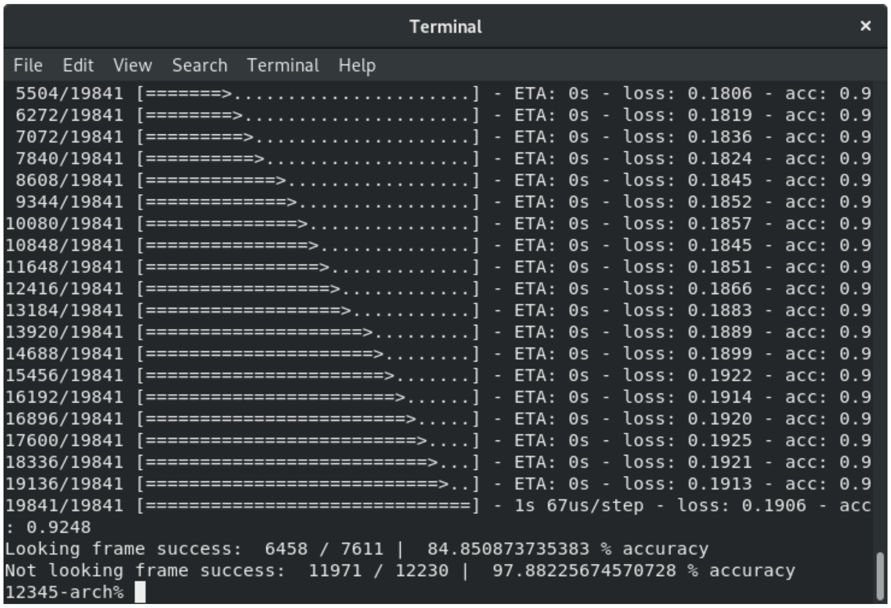


Unit step activation



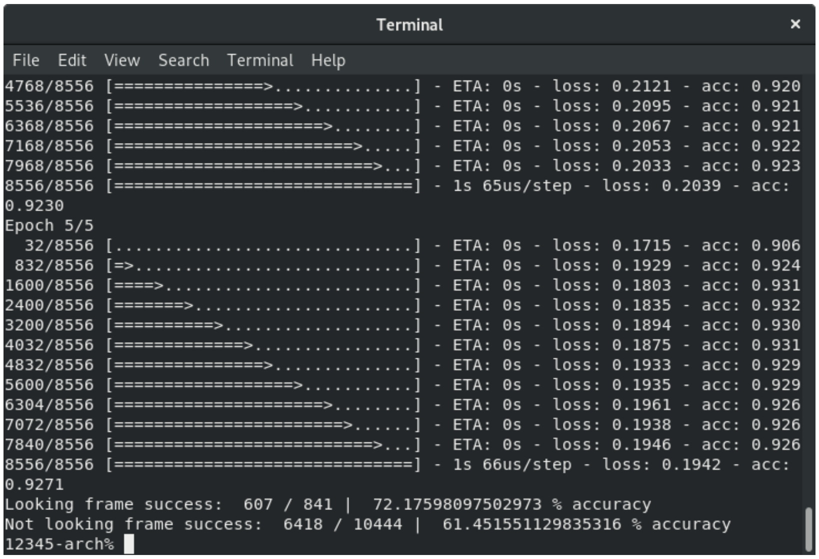
Results:

Unfortunately, we found the accuracy of our NN to depend highly on our choice of testing and training videos. When our testing videos were similar in format to our training videos, we found the resulting model to be highly accurate. However, when the training videos differed from the testing videos (for example, when the infant was closer/further from camera, was off center of the screen, or was accompanied by other individuals) the accuracy was markedly worse. We also found that even in videos which did have similar formats, our model was generally more accurate in frames in which the infant was not making eye contact. This is likely because a higher percent of the frame in these videos (which came from Youtube) show no eye contact than eye contact, so our model is better trained to identify the former than the later. Both of these issues can be corrected by training the model on videos which are more similar to the ones our NN will be used on in the lab. Since the videos capture by the robot in the lab will be all have roughly similar formats, our model can be optimized for this use.



```
 5504/19841 [=======>......................] - ETA: 0s - loss: 0.1806 - acc: 0.9
 6272/19841 [========>.....................] - ETA: 0s - loss: 0.1819 - acc: 0.9
 7072/19841 [=========>....................] - ETA: 0s - loss: 0.1836 - acc: 0.9
 7840/19841 [==========>...................] - ETA: 0s - loss: 0.1824 - acc: 0.9
 8608/19841 [============>.................] - ETA: 0s - loss: 0.1845 - acc: 0.9
 9344/19841 [=============>................] - ETA: 0s - loss: 0.1852 - acc: 0.9
10080/19841 [==============>...............] - ETA: 0s - loss: 0.1857 - acc: 0.9
10848/19841 [===============>..............] - ETA: 0s - loss: 0.1845 - acc: 0.9
11648/19841 [================>.............] - ETA: 0s - loss: 0.1851 - acc: 0.9
12416/19841 [=================>............] - ETA: 0s - loss: 0.1866 - acc: 0.9
13184/19841 [==================>...........] - ETA: 0s - loss: 0.1883 - acc: 0.9
13920/19841 [====================>.........] - ETA: 0s - loss: 0.1889 - acc: 0.9
14688/19841 [=====================>........] - ETA: 0s - loss: 0.1899 - acc: 0.9
15456/19841 [======================>.......] - ETA: 0s - loss: 0.1922 - acc: 0.9
16192/19841 [=======================>......] - ETA: 0s - loss: 0.1914 - acc: 0.9
16896/19841 [========================>.....] - ETA: 0s - loss: 0.1920 - acc: 0.9
17600/19841 [=========================>....] - ETA: 0s - loss: 0.1925 - acc: 0.9
18336/19841 [==========================>...] - ETA: 0s - loss: 0.1921 - acc: 0.9
19136/19841 [===========================>..] - ETA: 0s - loss: 0.1913 - acc: 0.9
19841/19841 [==============================] - 1s 67us/step - loss: 0.1906 - acc
: 0.9248
Looking frame success:  6458 / 7611 |  84.850873735383 % accuracy
Not looking frame success:  11971 / 12230 |  97.88225674570728 % accuracy
12345-arch% 
```

Testing and training videos both feature infant in center screen
Accuracy is high, but our NN is considerably better trained on not-looking-frames, which are more numerous, than on looking-frames



```
4768/8556 [===============>..............] - ETA: 0s - loss: 0.2121 - acc: 0.920
5536/8556 [==================>...........] - ETA: 0s - loss: 0.2095 - acc: 0.921
6368/8556 [=====================>........] - ETA: 0s - loss: 0.2067 - acc: 0.921
7168/8556 [========================>.....] - ETA: 0s - loss: 0.2053 - acc: 0.922
7968/8556 [==========================>...] - ETA: 0s - loss: 0.2033 - acc: 0.923
8556/8556 [==============================] - 1s 65us/step - loss: 0.2039 - acc:
0.9230
Epoch 5/5
  32/8556 [..............................] - ETA: 0s - loss: 0.1715 - acc: 0.906
 832/8556 [=>............................] - ETA: 0s - loss: 0.1929 - acc: 0.924
1600/8556 [====>.........................] - ETA: 0s - loss: 0.1803 - acc: 0.931
2400/8556 [=======>......................] - ETA: 0s - loss: 0.1835 - acc: 0.932
3200/8556 [==========>...................] - ETA: 0s - loss: 0.1894 - acc: 0.930
4032/8556 [=============>................] - ETA: 0s - loss: 0.1875 - acc: 0.931
4832/8556 [===============>..............] - ETA: 0s - loss: 0.1933 - acc: 0.929
5600/8556 [==================>...........] - ETA: 0s - loss: 0.1935 - acc: 0.929
6304/8556 [=====================>........] - ETA: 0s - loss: 0.1961 - acc: 0.926
7072/8556 [=======================>......] - ETA: 0s - loss: 0.1938 - acc: 0.926
7840/8556 [==========================>...] - ETA: 0s - loss: 0.1946 - acc: 0.926
8556/8556 [==============================] - 1s 66us/step - loss: 0.1942 - acc:
0.9271
Looking frame success:  607 / 841 |  72.17598097502973 % accuracy
Not looking frame success:  6418 / 10444 |  61.451551129835316 % accuracy
12345-arch% 
```

Testing videos feature infant off center screen, causes reduced accuracy in both looking and not looking frames