FINAL REPORT:
Emotion Classification of Video Clips

Malvika Menon, Stuart Neilson, Jeff Winchell, Michelle Xie

Group 1

Problem Statement

For this project our group would like to classify the emotions of the human face from a dataset of short video files as accurately as possible.

Data Resources

We will be using the BAUM-1 dataset (Bahcesehir University Multimodal Face Database of Spontaneous Affective and Mental States), a sample that includes 1500 video clips from 31 subjects who show an unscripted variety of both emotional and mental states in Turkish. Based on stimuli presented, synchronous facial recordings were recorded to classify eight emotions (happiness, anger, sadness, disgust, fear, surprise, boredom, and contempt) and certain mental states (unsureness, thoughtfulness, concentration, interest, bothered).

We chose this dataset for the following reasons:

1) It provides a relevant challenge that will utilize the skills learned in this class.
2) It is a manageable size.
3) It has a rich variety of data versus comparable datasets only recording the six basic emotions.

Strategy & Overview

First, we explore the data set and select for certain emotions, making other adjustments along the way that will be described. After loading and processing the data, we then construct three models in the interest of finding the most accurate means of classifying emotions:

1. Conv3d

   C3D has been shown to be effective in video analysis tasks, with the ability to process appearance and information about motion at the same time (Fan et al. 16). (A variant also might include a linear classifier, which also performs well according to video analysis benchmarks.)

Most importantly, C3D networks have shown to perform well in emotion recognition, so it is a worthy model structure to look at.

2. TimeDistributed Conv2d plus RNN

   In the same study referenced to above, Conv2d's classification abilities have been shown to be heightened. The hybrid network with C2D and RNN has been shown to improve results for emotional classification, and thus, such hybridization is worth a simulation. (Fan et al. 16)

3. "DeepFace" based on vgg_face_weights

   The Deep Face model is a very large model with 145 million trained parameters. We have tried to extract a lightweight version of it by taking only some of the layers.

There were many interesting considerations and modifications necessitated along the way including: selection of frames, scaling of data, tweaking of model hyperparameters, etc. We will present the results from these three models and discuss their outcomes as well as other avenues we could have taken in terms of strategy and execution. In the end, we found that the Model 1 correctly predicted 42 emotions (29% of the 144 test videos) and that Model 2 correctly predicted 44 (31%). Lastly, Model 3 always classified inputs as one emotion, an issue we could not overcome sufficiently and will further discuss. With these myriad results, we found relative success in the first two models, and the opposite in the third model, a lesson learnt that will be further detailed.

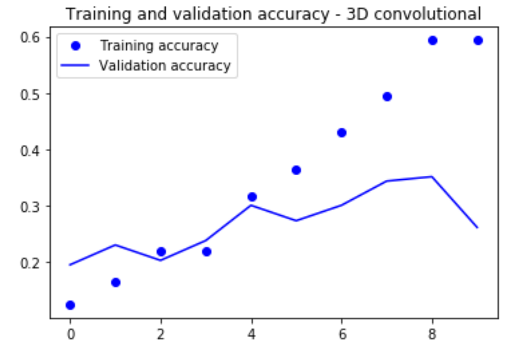
Literature Review & Model Selection Motivation

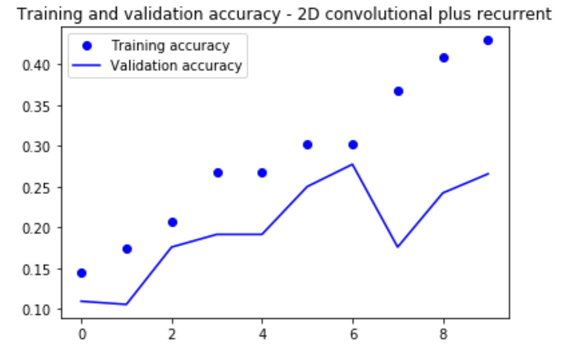
Videos for emotional classification purposes have been shown to have a unique level of power of prediction, looking at properties like feature-point tracking and texture-based features that can be utilized to train classifiers (Bargal et al. 16). While a snapshot is an important indicator of emotion, the dynamism of the sequence of frames, often supplemented by tonal features of audio components, make videos a powerful yet complicated data set for the purposes of emotional classification. We were inspired by the EmotiW challenge, which looks at real world issues that can be solved with affective computing, with the goal being to benchmark algorithms on ‘in the wild’ data. There are two aspects of our dataset that will be pertinent for this literature review 1) emotional classification strategies and models 2) video analysis.

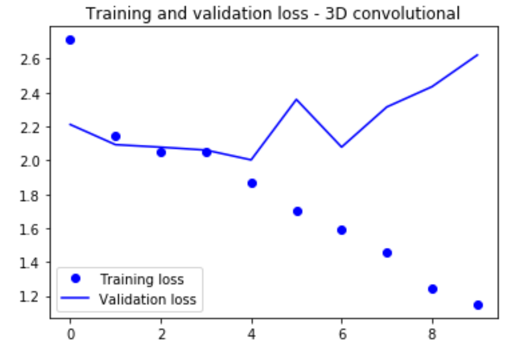
Conventionally, unsupervised learning methods seem to perform less well for this purpose. However, “Emotional classification of Youtube videos” utilizes an ensemble model of both supervised and unsupervised learning methods to classify emotions from Youtube videos with relative success (Chen 17), indicating that such creative strategies might be worth testing in our case. Additionally, an EmotiW Challenge 2016 contender paper also deems that based on reality TV videos, CNN and LSTM-RNN are comparable methods with regards to accuracy (Ding et al. 16). Indeed, based on our exploration, we attempt to use hybridized model structures, as they have been shown to be effective for projects in emotional classification. Conv3d models have been established to be helpful in video analysis, and the incorporation of RNN's has only improved such performance (Fan et al. 16). This research motivates the selection of the first two model structures. A further examination also indicated that the utilization of VGGFace NN and pretrained weights might make our endeavors in this project more streamlined, motivating our third model's selection. (Rassadin et al 17)

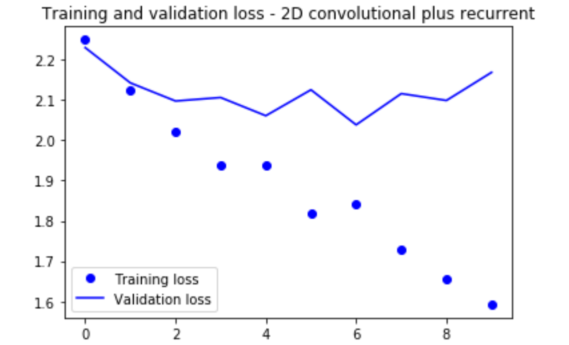
Overall Results & Summary:

Of the 131 videos in the test dataset, Model 1 correctly predicted 36 of them (27%). Model 2 correctly predicted 33 (25%). These stats are better than what is reported by Keras during training, as they haven't gone through the sample rebalancing. On the other hand, we observed that Model 3 tends to misclassify, only classifying to one emotion despite variability of input. We haven't succeeded in setting up a model structure that works with the pretrained model weights to get a useful prediction. Building a workable transfer learning model for faces that can fit within our computing resources turned out to be a difficult task.

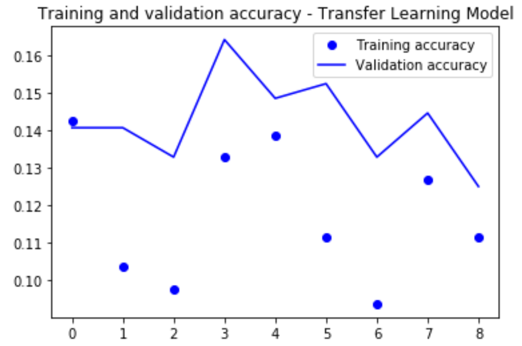
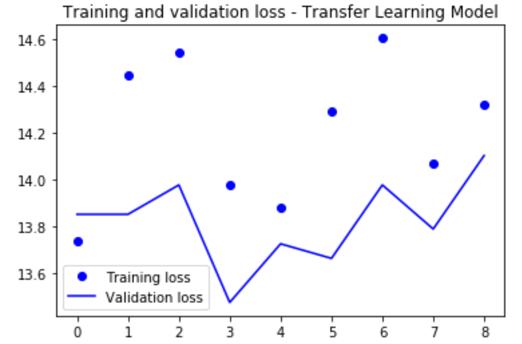
<table>
<tr><th></th><th>Model 1</th><th>Model 2</th></tr>
<tr><td>Accuracy</td><td>


</td><td>


</td></tr>
<tr><td>Loss</td><td>


</td><td>


</td></tr>
</table>

	Model 3
Accuracy	
Loss	

Conclusions

In the EmotiW contestants' papers of finalists, we observed 50-60% accuracy, and we have achieved about 30% accuracy for our own models, a comparable achievement given the time constraints and the learning curve of having to deal with the complexities of a non-standardized and large video dataset. We were able to detect emotions with relative levels of accuracy, and have learned many lessons along the way.

Note that for Model 3, we continually had misclassification issues. We had similar issues with the first two models, before troubleshooting and realizing that we had not scaled the data. Upon further examinations we realized the importance of this normalization. Without scaling input training vectors, the distribution range of feature values would differ for each feature, allowing the learning rate to create variable dimensional corrections, creating problematic differential compensations across inputs. This issue might prove to be problematic with the backpropagation of CNN's especially. Thus, we added these scaling methods and were able to produce the statistics we reported. However, if we had realized the challenge of working with VGG Face for our purposes earlier, we would have spent more time in trying to troubleshoot this issue in order to learn how to achieve such a tranfer learning model within our computing capacities. Alternatively, we could and should have tested comparable methods to VGG

Face to figure out whether this was a VGG specific issue we were facing. These obstacles were definitely a lesson worth learning in dealing with hefty and rich data sets and properly strategizing and reappropriating tools like VGGFace given a probable learning curve. Thus future models would

- further tune hyperparameters of the first two model structures that functioned properly
- work more effectively with structures like VGGFace and
- test out more types of hybridized models (ex. combining a CNN with other RNN's we explored in class)

Additionally, tasks involved in this project that we had not done before are :

- extracting data from video files and setting relevant benchmarks
- implementing creative hybridization of models specific to such data
- working with VGG Face

Emotion classification from video data is a fairly new frontier in machine learning, where not much is yet known about what works the best. We have experimented in this area and learned about the challenges involved, an relatively successful effective endeavor. The high dimensionality of the data was a big challenge given our available computing resources. We dealt with that by training video frames resized to a smaller size and by feeding the training with a generator, although this adds to the training time required. The model setup we have in place runs up against the memory capacity limits of JupyterHub. Moreover, we played around with different hybridized model structures, drawing off the success of such approaches we'd observed during our literature review. Although only two out of our three models performed decently, we still took away many important lessons.