Facial expression Recognition Using Convolutional Neural Networks

Hao Chuan Lei
haochuan

Victor Chien
vicchien

Terrence Jo
joterry

Abstract

The ability of Convolutional Neural Networks to perform facial recognition and facial expression classification tasks is well-documented. In this paper, we set out to explore the generalizability of CNNs across different samples of the same dataset. This report describes two experiments, one where we train the network on female faces and test on male facial expressions, and another where we train on peak expressions and test on mid-level expressions in what we call "pre-expression detection". Our findings indicate that the generalizability of CNNs across temporally different samples of the same population (pre-expression detection) is much higher than that of distinct population samples.

1 Introduction

Recent enhancements in Neural Networks have led to a rapid rise in applicability and scope of so-called Deep Neural Networks (DNNs), which use hidden layers to propagate information between nodes in a feedforward nature. These changes have given rise to exciting new applications in areas ranging from natural language processing to image recognitionm (LeCun et al., 2015). In this paper, we explore one such application of Deep Neural Networks, namely analyzing its generalizability across different populations and temporal contexts.

The ability for DNNs to classify images and visual data is a well-documented phenomenon. This ability is especially prevalent in Convolutional Neural Networks (CNNs), which use convolutional filters to preserve spatial information in the learning process. This report focuses on using CNNs for the purpose of facial expression recognition. There exists a wide array of datasets available for public use with respect to expression detection (Gross, 2005; Ng et al. 2015). Similar work has also utilized the CNN architecture for this purpose (Yu et al., 2015).

This paper describes two experiments performed on the same dataset of grayscale faces. Each face represents a frame in a burst from neutral to peak face expression, and is split into a set of male and female faces. We used two networks, AlexNet and a smaller variant, which we have named MiniNet, for our classification task. Our first experiment consisted of training our network on female faces and testing the network's ability to detect facial expressions on male faces. In our second experiment, we trained our networks on peak facial expressions and tested on mid-level expressions (between neutral and peak expression). The goal of both experiments was to better understand the generalizability of CNNs across sample and temporal contexts. This paper outlines the findings of these two experiments.

In Part 2, this paper describes the Cohn Kanade Dataset and the format of the input to the network. It also delves in detail into the pre-processing methodology. In Part 3, we describe the algorithms used in the experiments, beginning with AlexNet and following up with MiniNet. In Part 4, we describe the hyper-parameter tuning done prior to the actual experiments. In parts 5 and 6 we describe Experiment 1: Cross-Sample Experiment and Experiment 2: Pre-Expression Detection, as well as the results and a brief discussion. We end with a conclusion and a discussion of future work.

2 Data Processing

The Cohn Kanade Facial Expression Dataset is a set of grayscale images for facial expression detection. The data came in the form of directories, each of which represented a set of images

for a particular male or female model. Each of the directories had numbered subdirectories, and each subdirectory contained a burst of around 20 images relating to a facial expression, from start to finish. Images were mostly black and white, with the occasional color image, so only the black and white images were considered in our experiment for consistency. In total there were 1075 images of female faces and 202 images of male faces, with the following breakdown by emotion: Men {anger: 15, disgust: 35, happy: 62, sad: 20, surprised: 70}, Women {anger: 105, disgust: 160, happy: 360, sad: 155, surprised: 295}.

First, because the Cohn Kanade data was not separated into females and males, we manually categorized our data into two directories, male and female, for Experiment 1. After categorizing our data, we ended up with 33 male models and 62 female models.

Then, there was preprocessing done to convert all images to numpy arrays, so that the images could easily be entered as inputs into the convolutional neural networks - AlexNet and MiniNet, for the two experiments. After loading and converting the images in the dataset into numpy arrays, the last aspect of data processing that was necessary was assigning labels to each image. In the Cohn Kanade Dataset, each subdirectory that contained image bursts was labeled with a number, which could be interpreted into a score using the Cohn-Kanade Database FACS code spreadsheet. So for each image, we extracted the images AU descriptive score from the Cohn-Kanade Database FACS code spreadsheet and converted each score into an emotion label. We then assigned these labels to their respective images in the numpy array.

Cohn-Kanade Database FACS code:

AU	*Name*	N	AU	*Name*	N	AU	*Name*	N
1	*Inner Brow Raiser*	173	13	*Cheek Puller*	2	25	*Lips Part*	287
2	*Outer Brow Raiser*	116	14	*Dimpler*	29	26	*Jaw Drop*	48
4	*Brow Lowerer*	191	15	*Lip Corner Depressor*	89	27	*Mouth Stretch*	81
5	*Upper Lip Raiser*	102	16	*Lower Lip Depressor*	24	28	*Lip Suck*	1
6	*Cheek Raiser*	122	17	*Chin Raiser*	196	29	*Jaw Thrust*	1
7	*Lid Tightener*	119	18	*Lip Puckerer*	9	31	*Jaw Clencher*	3
9	*Nose Wrinkler*	74	20	*Lip Stretcher*	77	34	*Cheek Puff*	1
10	*Upper Lip Raiser*	21	21	*Neck Tightener*	3	38	*Nostril Dilator*	29
11	*Nasolabial Deepener*	33	23	*Lip Tightener*	59	39	*Nostril Compressor*	16
12	*Lip Corner Puller*	111	24	*Lip Pressor*	57	43	*Eyes Closed*	9

Table 1. *Frequency of the AUs coded by manual FACS coders on the CK+ database for the peak frames.*

Emotion	Criteria
Angry	AU23 and AU24 must be present in the AU combination
Disgust	Either AU9 or AU10 must be present
Fear	AU combination of AU1+2+4 must be present, unless AU5 is of intensity E then AU4 can be absent
Happy	AU12 must be present
Sadness	Either AU1+4+15 or 11 must be present. An exception is AU6+15
Surprise	Either AU1+2 or 5 must be present and the intensity of AU5 must not be stronger than B
Contempt	AU14 must be present (either unilateral or bilateral)

Table 2. *Emotion description in terms of facial action units.*

3 Algorithms

For Experiments 1 and 2, two algorithms, AlexNet and MiniNet were used, both of which are convolutional neural networks. MiniNet was created as a smaller, more compact network. We were motivated to create MiniNet due to the relatively small dataset and the low accuracy exhibited by initial training on AlexNet.

3.1 AlexNet

For AlexNet, we used a version of the widely known convolutional neural network that is trained on more than a million images from the ImageNet database. We modified this AlexNet to fit our project and experiment specifications. In the AlexNet, there was one input channel, due to the black and white nature of the images, and 5 convolutional layers and 3 fully connected layers. For AlexNet, the ReLU activation function was used and Dropout regularization was used to minimize overfitting. The kernel size, padding, and stride for each layer were as follows:

AlexNet Algorithm:

Convolutional Layer #1: `in=1`, `out=64`, `kernel_size=11`, `stride=4`, `padding=2`, ReLU activation, MaxPool: `kernel_size=3`, `stride=2`

Convolutional Layer #2: `in=64`, `out=192`, `kernel_size=5`, `padding=2`, ReLU activation, MaxPool: `kernel_size=3`, `stride=2`

Convolutional Layer #3: `in=192`, `out=384`, `kernel_size=3`, `padding=1`, ReLU activation

Convolutional Layer #4: `in=384`, `out=256`, `kernel_size=3`, `padding=1`, ReLU activation

Convolutional Layer #5: `in=256`, `out=256`, `kernel_size=3`, `padding=1`, ReLU activation, MaxPool: `kernel_size=3`, `stride=2`, Dropout

Fully Connected: `in=1536`, `out=4096`, ReLU, Dropout

Fully Connected: `in=4096`, `out=4096`, ReLU,

Fully Connected: `in=4096`, `out=4096`

3.2 MiniNet

As mentioned previously, the MiniNet algorithm was a more simple convolutional neural network that was made with fewer layers to test overfitting in the more complex AlexNet algorithm. In our MiniNet, there was still an input channel of 1 due to the black and white images, but there were only 3 convolutional layers and 3 fully connected layers. Similar to AlexNet, the ReLU activation function was used and Dropout regularization was used to minimize overfitting. The kernel size, padding, and stride for each layer was as follows:

MiniNet Algorithm:

Convolutional Layer #1: `in=1`, `out=64`, `kernel_size=11`, `stride=4`, `padding=2`, ReLU activation, MaxPool: `kernel_size=3`, `stride=2`

Convolutional Layer #2: `in=64`, `out=200`, `kernel_size=5`, `padding=2`, ReLU activation, MaxPool: `kernel_size=3`, `stride=2`

Convolutional Layer #3: `in=200`, `out=256`, `kernel_size=3`, `padding=1`, ReLU activation, MaxPool: `kernel_size=3`, `stride=2` Dropout

Fully Connected: `in=8960`, `out=1000`, ReLU, Dropout

Fully Connected: `in=1000`, `out=500`, ReLU

Fully Connected: `in=500`, `out=4`

4 Hyper-parameter Tuning

For Hyper-parameter Tuning, we tested two main hyper-parameters: batch size and learning rate. For each of these hyperparameters, we experimented with low, medium and high values. Namely, for batch size, the values we tested were 100, 200, and 300, and the values tested for learning rate were 0.0001, 0.00001, and 0.00005. We ran both AlexNet and MiniNet algorithms on our dataset for each hyper-parameter value, and observed the training and test accuracies.

Hyper-parameter Accuracies:

AlexNet

Batch Size		
Value	Train	Test
100	0.9205	0.3822
200	0.6307	0.3244
300	0.4413	0.3067

Learning Rate		
Value	Train	Test
0.0001	1.0000	0.3911
0.00005	0.9962	0.4178
0.00001	0.4508	0.3200

MiniNet

Batch Size		
Value	Train	Test
100	0.8693	0.3288
200	0.6477	0.3244
300	0.4867	0.3244

Learning Rate		
Value	Train	Test
0.0001	1.0000	0.4044
0.00005	0.9943	0.4
0.00001	0.8220	0.3822

We observed high levels of overfitting despite the use of Dropout. In both networks, we chose the parameters which had the best test accuracy with lower overfitting. For AlexNet, this was a batch size of 100 and a learning rate of 0.00005, and for MiniNet the most optimal hyperparameters were a batch size of 100 and a learning rate of 0.0001.

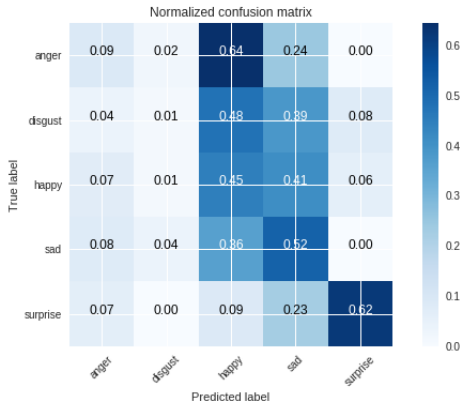
5 Experiment 1

The goal of this experiment is to see how well AlexNet and MiniNet generalize expression classification across women and men. This experiment trains AlexNet and MiniNet on women faces with expressions and tests on men faces. We also trained both models on women and tested on women as a base comparison.

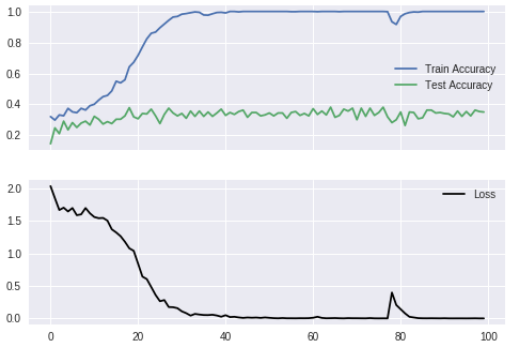
5.1 AlexNet Results

AlexNet train on women and test on women:
Train: 1.0000 **Test**: 0.3478

AlexNet Confusion Matrix (on test data):
Train: 1.0000 **Test**: 0.3509



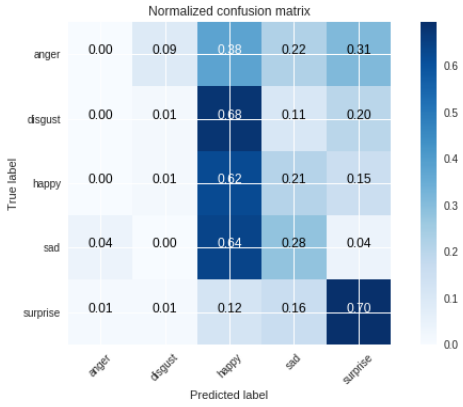
AlexNet Plot over 100 epochs:



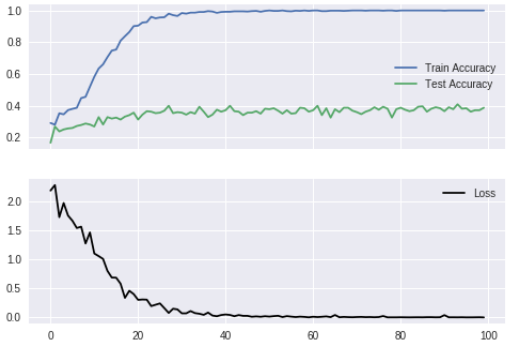
5.2 MiniNet Results

MiniNet train on women and test on women:
Train: 0.9987 **Test**: 0.3665

MiniNet Confusion Matrix (on test data):
Train: 1.0000 **Test**: 0.3882



MiniNet Plot over 100 epochs:



5.3 Discussion

From the confusion matrices, we can observe that AlexNet was able to predict surprise better than any of the other expressions. However, overall, it generalizes poorly from women to men on the other expressions. It tended to classify expressions as either happy or sad. One of the most common misclassifications AlexNet made was classifying anger as happy. MiniNet also had a similar confusion matrix as AlexNet, in that it classified most expressions as either happy or sad and did the best on classifying surprise. However, it did better than AlexNet on correctly classifying happy but worse on correctly classifying sad. Looking at the plots, we can observe that even though both AlexNet and MiniNet were both able to achieve a high train accuracy, they both had around the same low test accuracies. When compared to training on women and testing on women, the accuracies for test and train for both models were very similar to those produced by training on women and testing on men.
The reason for such low test accuracies can be potentially attributed to the fact that both AlexNet and MiniNet had a difficult time generalizing across different faces in general, not from women to men. This is because after training on women faces and testing on women faces, the two models achieved relatively the same test accuracies. Another, more probable, reason is that we did not have enough training data. It is likely that AlexNet did not have sufficient data to fully learn all the necessary CNN parameters.

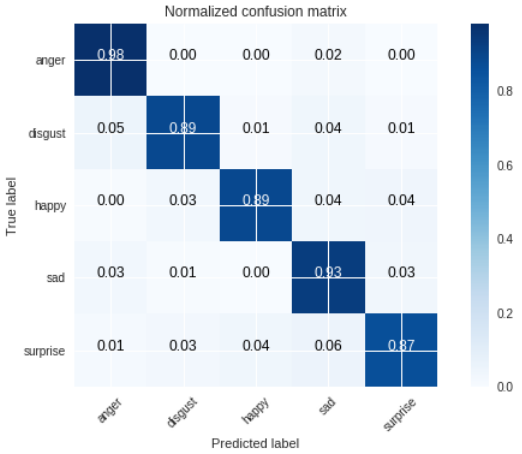
6 Experiment 2

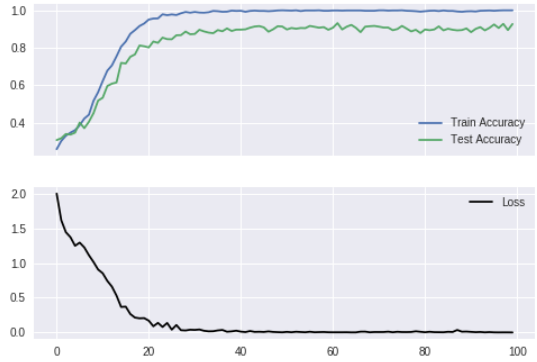
The goal of this experiment is to see how well AlexNet and MiniNet generalize expression classification across across faces in different stages of emotional expression. Both the training and test sets include faces from the same individuals, as we are not interested in generalizing across different faces or genders. AlexNet and MiniNet are trained on the last five images (peak) of a photo burst sequence for an expression and tested on the five images (mid-level) before the last five. We also trained both models on peak expressions and tested on peak expressions as a base comparison.

6.1 AlexNet Results

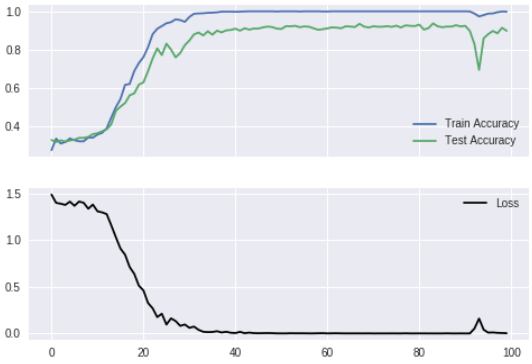
AlexNet train on peak and test on peak:
Train: 1.0000 **Test**: 0.9896

AlexNet Confusion Matrix (on test data):
Train: 0.9984 **Test**: 0.8984





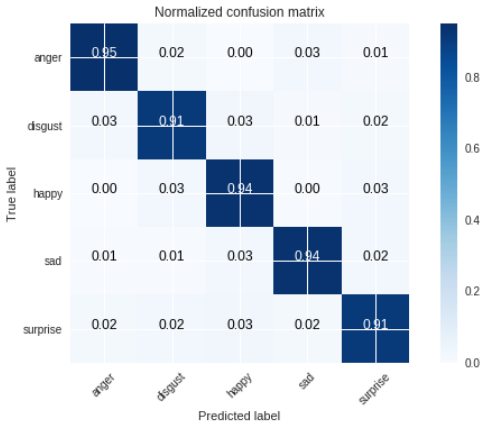
AlexNet Plot over 100 epochs:



6.2 MiniNet Results

MiniNet train on peak and test on peak:
Train: 1.0000 **Test**: 0.9791

MiniNet Confusion Matrix (on test data):
Train: 1.0000 **Test**: 0.9270



MiniNet Plot over 100 epochs:

6.3 Discussion

From the confusion matrices, we can observe that both AlexNet and MiniNet did very well on correctly classifying all five expressions on the test data. Both made relatively few mistakes. From the two plots, we can also see that the test accuracies closely followed the training accuracies. The test accuracies reached the high 80% to 90% range. When compared to training on peak and testing on peak, the accuracies for test and train for both models were very similar to those produced by training on peak and testing on mid-level.
The reason for this higher accuracy, compared to the last experiment is likely due to the faces from the same individuals appearing in both the training and test sets. From the previous experiment, both AlexNet and MiniNet likely have a difficult time classifying faces they have never seen before. Therefore, they would understandably do a better job on the test set for experiment 2.

7 Conclusion and Future Work

After our experimentation, we can conclude that, in general, both Convolutional Neural Networks performed poorly in Experiment 1, as the AlexNet and MiniNet algorithms had difficulty generalizing facial expressions on new faces and new models, while maintaining high accuracies on faces that were trained on. The difference in male and female subjects in our test images had little effect on accuracy levels.
For Experiment 2, or pre-expression detection, the Convolutional Neural Networks performed well, likely because each individual in the test set also appeared in the training set. We were able to detect emotions from early in the burst with the Convo-

lutional Neural Networks, and achieve high accuracies. This shows that the CNN generalizes well over faces from the same sample of individuals.
We conclude that these particular Convolutional Neural Networks, AlexNet and MiniNet, were able to generalize better with the same subject sample with different levels of expression compared to cases where it was forced to generalize across different subject samples.
Some future work regarding these two experiments may include performing pre-expression detection on neutral faces rather than mid-level expressions. We may also want to hold out a portion of the dataset to see if the networks can generalize across different individuals for different levels of expression. With regards to the male/female expression detection, it may be worth the time to try different network architectures such as ResNets in combination with larger, more expressive datasets to see if the lack of generalization is a result of the structure of the network itself.

References

Ng, Hong-Wei, et al... 2015. Deep learning for expression recognition on small datasets using transfer learning. *Proceedings of the 2015 ACM on international conference on multimodal interaction.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E.. 2012. Imagenet classification with deep convolutional neural networks. *In Advances in neural information processing systems* , (pp. 1097-1105).

LeCun, Y., Bengio, Y., & Hinton, G.. 2015. Deep learning.*Nature* 521(7553), 436.

Liu, X., Yang, D., & Gamal, A. E.. 2017. Deep Neural Network Architectures for Modulation Classification. arXiv preprint arXiv:1712.00443.

Ng, Hong-Wei, et al... 2015. Deep learning for expression recognition on small datasets using transfer learning. *Proceedings of the 2015 ACM on international conference on multimodal interaction.*

R. Gross. Face Databases, Handbook of Face Recognition, Stan Z. Li and Anil K. Jain, ed., Springer-Verlag, February 2005, 22 pages. http://www.face-rec.org/databases/

Yu, Zhiding, and Cha Zhang.. 2015. Image based static facial expression recognition with multiple deep network learning. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction.*