DeXpression: Deep Convolutional Neural Network for Expression Recognition

Jan Samuel C. Matuba
Energy Engineering Program
College of Engineering
University of the Philippines Diliman

ABSTRACT

DeXpression is a convolutional neural network (CNN) architecture proposed by Burkert et al. for use in facial expression recognition. The model is implemented in Google Tensorflow and is evaluated on the Extended Cohn-Kanade (CK+) dataset. Four different models are explored: standard model, with dropout, with batch normalization, and with dropout and batch normalization. Data augmentation techniques are used to increase the number of training samples and 5-fold cross validation was used to evaluate the models. Results showed that the standard model has the highest test accuracy having a mean test accuracy over 5folds of 86.76%. The model that used batch normalization and dropout achieved lower test accuracies.

1. INTRODUCTION

Human facial expression recognition is an emerging field in relation to the related field of robotics and automated image processing and encoding. Being able to understand ones emotions and the encoded feelings is an important factor for an appropriate and correct understanding. Integrating these capability into machines will be able to allow for a more diverse and natural way of communication.

In this work, we present an approach of detecting emotions from images using Convolutional Neural Networks (CNN) which is a form of Artificial Neural Networks (ANN). CNNs have seen large popularity over the past years due to its classification power in images. Aside from high predictive performance capability, CNN models are also real-time capable. This allows for the usage of the raw input images without any pre- or postprocessing which could translate to a lot of potential applications especially on automatic systems. The CNN architecture proposed is inspired by the GoogLeNet architecture and has been evaluated on the Extended Cohn-Kanade Dataset.

2. METHODOLOGY

2.1 Dataset

The dataset used is the Extended Cohn-Kanade (CK+) dataset retrieved from http://www.consortium.ri.cmu.edu/ckagree/. The CK+ database contains 593 sequences across 123 subjects which are FACS coded at the peak frame. Only 327 of the 593 sequences have emotion sequences where those are the only ones the fit the prototypic definition for the classes of emotion which are labeled as classes 0-7 (i.e. 0=neutral, 1=anger, 2=contempt, 3=disgust, 4=fear, 5=happy, 6=sadness, 7=surprise). All sequences are from the neutral face to the peak expression where the last image in the sequence. The images are of size 640490 px as well 640480 px with mixed colored and grayscale.

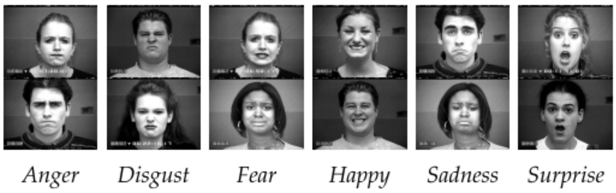

Figure 1. Sample images from the Extended Cohn-Kanade Dataset

Fig. 3: This Figure shows the differences within the Cohn-Kanade Plus (CKP) dataset. The emotion Contempt is not shown since there is no annotated image with the emotion being depicted, which is allowed to be displayed.

2.2 Data Preprocessing

From the raw dataset, a total of 327 labeled images corresponding to the peak frame in the sequence are extracted. The images are first standardized to a size of 224x224 pixels. Then they are converted from colored to grayscale images in order to remove the color bias in classification since skin color is not expected to be a factor in emotion scoring. The frequency of the classes are listed from largest to smallest number of samples: Surprise(7.0):83, Happy(5.0), 69, Disgust(3.0):59, Anger(1.0):45, Sadness(6.0):28, Fear(4.0): 25, Contempt(2.0):18. With a good frequency distribution of classes, the dataset is balanced enough. However, the number of samples itself is too small to be trained in a CNN. To resolve this, image augmentation techniques including blurring and rotation are applied to increase the number of samples using a Python package OpenCV. The Gaussian blurring kernel size is set to 5 and the factor is randomly selected within the range of 20 to 100. For the rotation, the images are rotated randomly from -15 degrees to 15 degrees with respect to the center of the image. For every image, augmentation was done 3 times resulting in a total of 6 additional images for every sample. The final dataset to be used contains a total of 2289 images where each class simply increased by a factor of 7.

2.3 CNN Architecture

The DeXpression convolutional neural network architecture takes an input of 224x224 images and then the sequence of convolutional, pooling, and normalization layers are listed in Table 1. All the convolutional layers are followed by the ReLU activation function. There are two parallel feature extraction blocks as visualized in Figure 1.

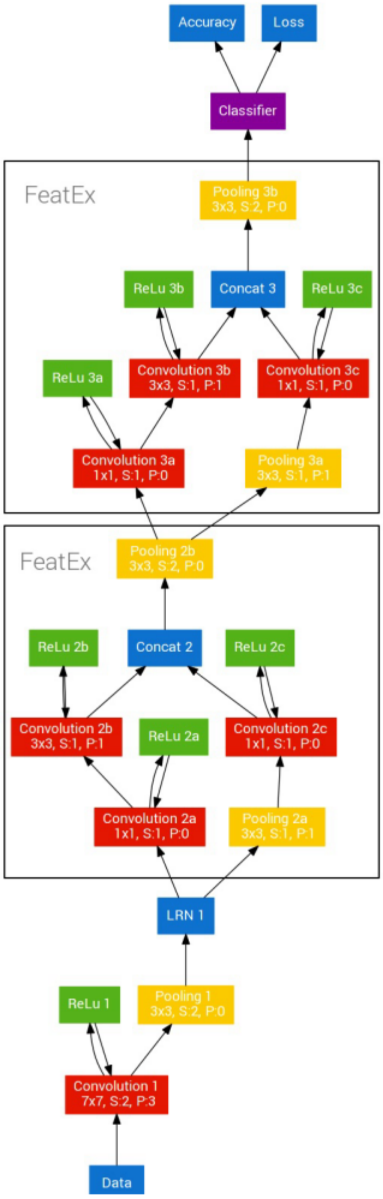

Figure 1. DeXpression Architecture

3. RESULTS AND DISCUSSION

Four different models are trained using the DeXpression architecture: standard, with dropout, with batch normalization, and with dropout and batch normalization. Using 5-fold cross validation and training over 1000 epochs, the accuracy levels and losses with respect to test sets are evaluated. In the following sections the visualizations of the results per model are shown.

Table 1: DeXpression Architecture

1	Convolutional 1
2	Pooling 1
3	LRN 1
4	Convolutional 2a
5	Pooling 2a
6	Convolutional 2b
7	Convolutional 2c
8	Concatenate 2
9	Pooling 2b
10	Convolutional 3a
11	Pooling 3a
12	Convolutional 3b
13	Convolutional 3c
14	Concatenate 3
15	Pooling 3b
16	Fully Connected
16	Softmax Classifier

3.1 Standard

In the standard model, the CNN architecture detailed in Figure 1 is implemented without any changes. At the last epoch, the average training and testing accuracy over 5 folds are 94.84% and 86.76% respectively while the average loss is 0.138. The total runtime is 8334 seconds.

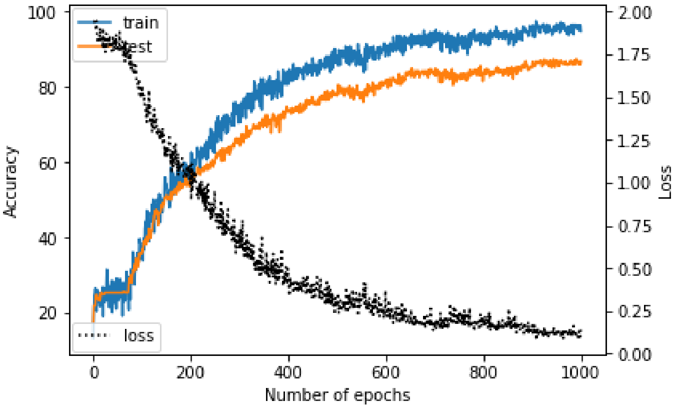

Figure 2. Accuracy and loss plot for standard model

3.2 With Dropout

In this model, the dropout method is applied. At the last epoch, the average training and testing accuracy over 5 folds are 46.87% and 56.06% respectively while the average loss is 1.28. The total runtime is 8316 seconds.

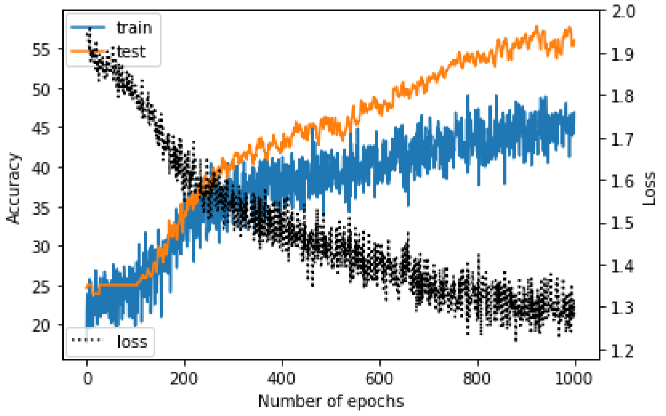

Figure 3. Accuracy and loss plot for model with dropout

3.3 With Batch Normalization

In this model, the batch normalization is applied for each convolutional layer in the standard model. At the last epoch, the average training and testing accuracy over 5 folds are 88.91% and 10.14% respectively while the average loss is 0.312. The total runtime is 10433 seconds.

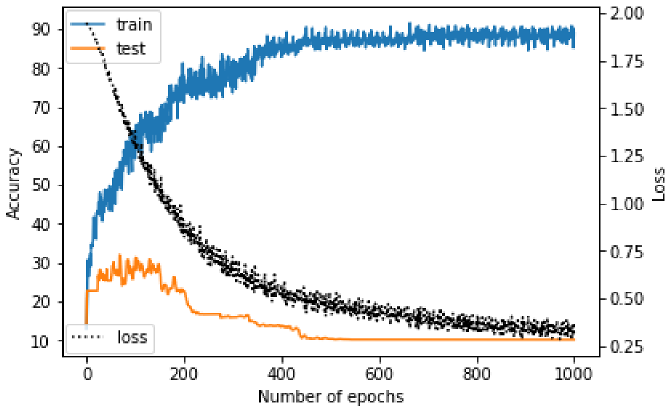

Figure 4. Accuracy and loss plot for model with batch normalization

3.4 With Dropout and Batch Normalization

In this model, dropout is used and batch normalization is applied for each convolutional layer in the standard model. At the last epoch, the average training and testing accuracy over 5 folds are 71.56% and 18.31% respectively while the average loss is 0.736. The total runtime is 10433 seconds.

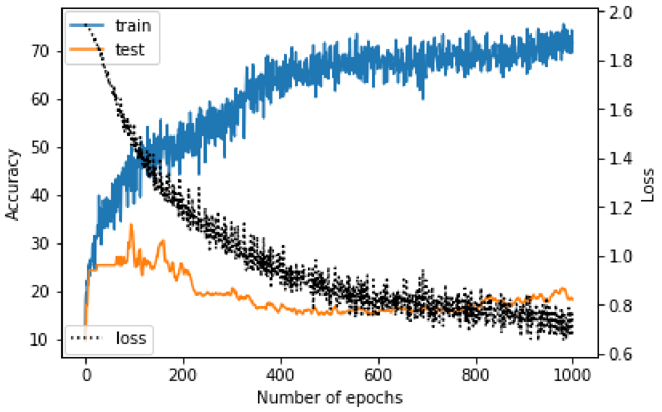

Figure 5. Accuracy and loss plot for model with dropout and batch normalization

3.5 Summary of Results

Out of all the models, only the standard model produced acceptable results with a good enough test accuracy of 86.76%. All the other models are very likely to have been implemented in the wrong way causing very low train and test accuracies.

4. CONCLUSION

In this work the standard DeXpression CNN model for facial expression recognition is implemented in Tensorflow. Aside from the standard model, batch normalization and dropout techniques are also implemented. The CNN models are evaluated on the Extended Cohn-Kanada (CK+) dataset in which the standard model gives the best test accuracy of 86.76%. All the other models were likely to have been implemented poorly resulting in also poor performance results.

5. REFERENCES

[1] Burkert, P. et al., 2015. DeXpression: Deep Convolutional Neural Network for Expression Recognition. , pp.18. Available at: http://arxiv.org/abs/1509.05371.

[2] Lucey, P. et al., 2010. The extended cohn-kande dataset (CK+): A complete facial expression dataset for action unit and emotionspecied expression. Cvprw, (July), pp.94101.