OMNO AI

ONWARD AND UPWARDS

Emotion Detection

Author:
Ibtihaj Tahir

Team Leads:
Hussam Habib
Kaleem Khan

October 21, 2019

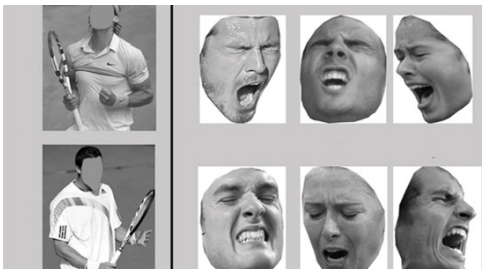
OMNO AI

Contents

1 Problem Statement

Facial expressions are a very powerful way of communication without words. By just looking at someone's expression, we can tell that weather the person is sad, happy or angry and all. A set of muscles are present on human's face that adjust there self with respect to the mood and hence form facial expressions. We, as humans can easily tell the mood by looking at the other person's expressions but for a machine to recognize it is not that simple.

Goal is to device a method by using the modern Deep Learning approaches to recognize human's face expressions.



2 Datasets

2.1 FER2013 [1]

The data consists of 48x48 pixel gray-scale images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. The task is to categorize each face based on the emotion shown in the facial expression in to one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

The cvs contains two columns, "emotion" and "pixels". The "emotion" column contains a numeric code ranging from 0 to 6, inclusive, for the emotion that is present in the image. The "pixels" column contains a string surrounded in quotes for each image. The contents of this string a space-separated pixel values in row major order. The dataset consists of 35,887 images.

The dataset can be obtained from
https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data

Figure 1: Samples from FER2013

2.2 CK+ [2]

The Images (cohn-kanade-images.zip) - there are 593 sequences across 123 subjects which are FACS coded at the peak frame. All sequences are from the neutral face to the peak expression.

The Landmarks (Landmarks.zip) - All sequences are AAM tracked with 68points landmarks for each image.

The FACS coded files (FACS_labels.zip) - for each sequence (593) there is only 1 FACS file, which is the last frame (the peak frame). Each line of the file corresponds to a specific AU and then the intensity. An example is given below.

The Emotion coded files (Emotion_labels.zip) - ONLY 327 of the 593 sequences have emotion sequences. This is because these are the only ones the fit the prototypic definition. Like the FACS files, there is only 1 Emotion file for each sequence which is the last frame (the peak frame). There should be only one entry and the number will range from 0-7 (i.e. 0=neutral, 1=anger, 2=contempt, 3=disgust, 4=fear, 5=happy, 6=sadness, 7=surprise). N.B there is only 327 files- IF THERE IS NO FILE IT MEANS THAT THERE IS NO EMOTION LABEL (sorry to be explicit but this will avoid confusion).

The dataset can be obtained from
http://www.consortium.ri.cmu.edu/ckagree/

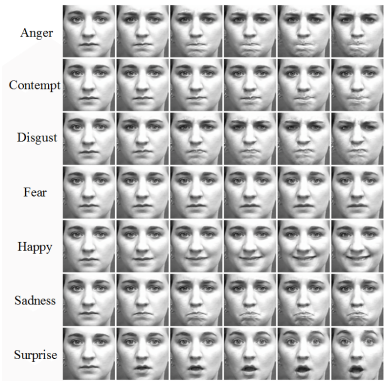


Figure 2: Samples from CK+

3 Deep Learning Models

3.1 VGG16 [3]

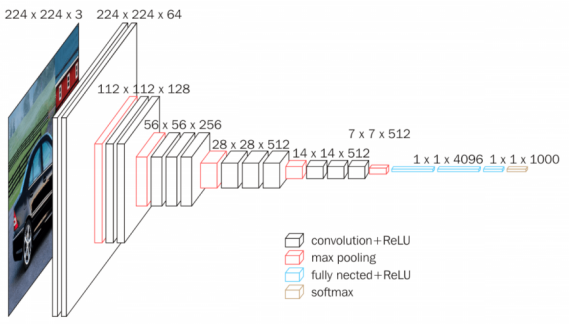


Figure 3: VGG16

VGG16 is a Convolutional Neural Network model proposed by K. Simonyan and A. Zisserman from the University of Oxford. The model achieves 92.7% top5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. It was one of the famous model submitted to ILSVRC-2014. It makes the improvement over AlexNet by replacing large kernelsized filters with multiple 3x3 kernelsized filters one after another.

3.2 Resnet50 [4]

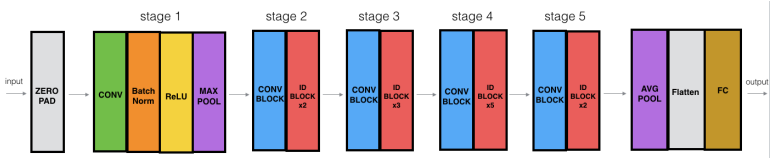


Figure 4: Resnet50

Residual Networks is a classic neural network used as a backbone for many computer vision tasks. This model was the winner of ImageNet challenge in 2015. The fundamental breakthrough with ResNet was it allowed us to train extremely deep neural networks with 150+layers successfully. Prior to ResNet training very deep neural networks was difficult due to the problem of vanishing gradients. AlexNet, the winner of ImageNet 2012

and the model that apparently kick started the focus on deep learning had only 8 convolutional layers, the VGG network had 19 and Inception or GoogleNet had 22 layers and ResNet 152 had 152 layers.

The ResNet-50 model consists of 5 stages each with a convolution and Identity block. Each convolution block has 3 convolution layers and each identity block also has 3 convolution layers. The ResNet-50 has over 23 million trainable parameters.

3.3 Inception V3 [5]

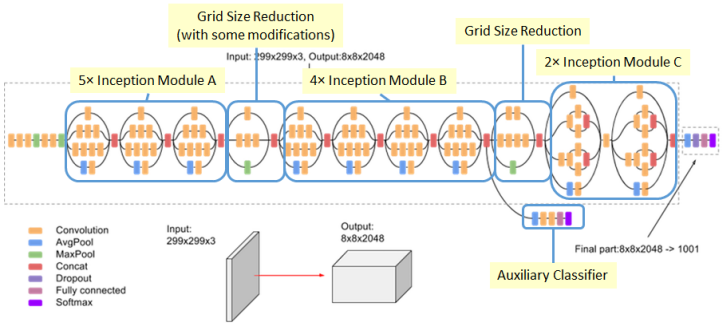


Figure 5: InceptionV3

Inception v3 is a widely-used image recognition model that has been shown to attain greater than 78.1% accuracy on the ImageNet dataset. The model itself is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concats, dropouts, and fully connected layers. Batchnorm is used extensively throughout the model and applied to activation inputs. Loss is computed via Softmax.

3.4 Inception-Resnet-V2 [6]

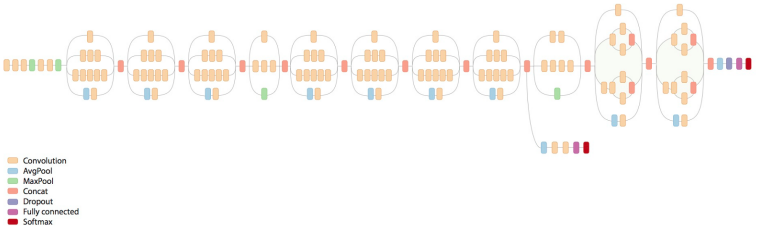


Figure 6: Inception_Resnet_V2

Very deep convolutional networks have been central to the largest advances in image recognition performance in recent years. One example is the Inception architecture that has been shown to achieve very good performance at relatively low computational cost. Recently, the introduction of residual connections in conjunction with a more traditional architecture has yielded state-of-the-art performance in the 2015 ILSVRC challenge; its performance was similar to the latest generation Inception-v3 network. This raises the question of whether there are any benefit in combining the Inception architecture with residual connections. Here we give clear empirical evidence that training with residual connections accelerates the training of Inception networks significantly. There is also some evidence of residual Inception networks outperforming similarly expensive Inception networks without residual connections by a thin margin. We also present several new streamlined architectures for both residual and non-residual Inception networks. These variations improve the single-frame recognition performance on the ILSVRC 2012 classification task significantly. We further demonstrate how proper activation scaling stabilizes the training of very wide residual Inception networks. With an ensemble of three residual and one Inception-v4, we achieve 3.08% top-5 error on the test set of the ImageNet classification (CLS) challenge.

3.5 DeXpression [7]

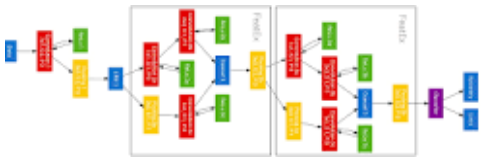


Figure 7: DeXpression

The proposed deep Convolutional Neural Network architecture consists of four parts. The first part automatically preprocesses the data. This begins with Convolution 1, which applies 64 different filters. The next layer is Pooling 1, which down-samples the images and then they are normalized by LRN 1. The next steps are the two FeatEx (Parallel Feature Extraction Block) blocks. They are the core of the proposed architecture. The features extracted by theses blocks are forwarded to a fully connected layer, which uses them to classify the input into the different emotions. The described architecture is compact, which makes it not only fast to train, but also suitable for real-time applications. This is also important as the network was built with resource usage in mind.
The key structure in the architecture is the Parallel Feature Extraction Block (FeatEx). It is inspired by the success of GoogleNet. The block consists of Convolutional, Pooling, and ReLU Layers. The first Convolutional layer in FeatEx reduces the dimension since it convolves with a filter of size 1 x 1. It is enhanced by a ReLU layer, which creates the desired sparseness. The output is then convolved with a filter of size 3 x 3. In the parallel path a Max Pooling layer is used to reduce information before applying a CNN of size 1 x 1. This application of differently sized filters reflects the various scales at which faces can appear. The paths are concatenated for a more diverse representation of the input. Using this block twice yields good results.

4 Implementation

Both the dataset (CK+ and fer2013) are used to train all the above mentioned models. The following hyper-parameters were used.

- Random Seed is set to 7 for reproduction of the results
- A train-test split of 5%
- For CK+, batch size is set to be 8 for train and 4 for test
- For Fer2013, batch size is set to be 32 for train and 16 for test
- **'categorical_crossentropy'** loss is used to train all the algorithms
- **'Adam'** optimizer is used to train all the algorithms
- 200 epochs
- 100 step-size
- 100 validation-step

Note: These parameters can be adjusted to achieve different results.

Implementation can be found at
https://github.com/IT-OMNOAI/Smart-Gandola

5 Results

5.1 CK+

- **VGG16**

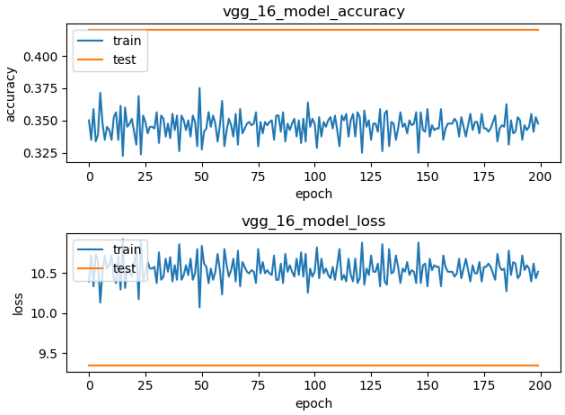


Figure 8: VGG16 on CK+

- **Resnet50**

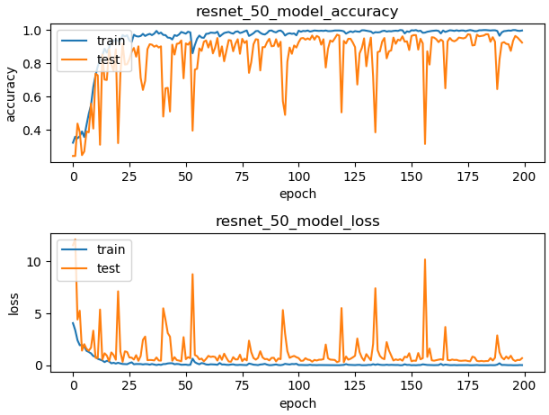


Figure 9: Resnet50 on CK+

- **InceptionV3**

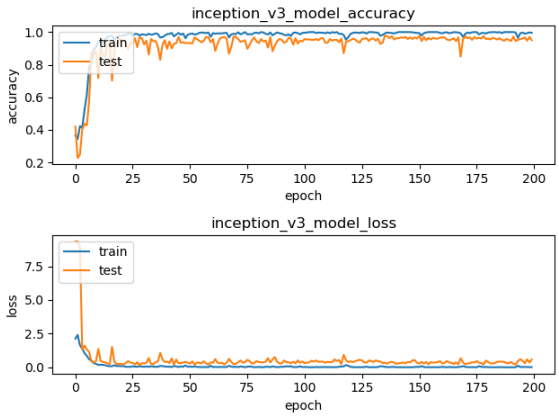


Figure 10: InceptionV3 on CK+

- **InceptionResnetV2**

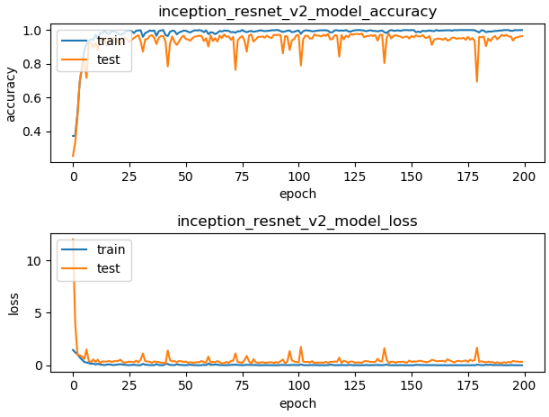


Figure 11: InceptionResnetV2 on CK+

- **DeXpression**

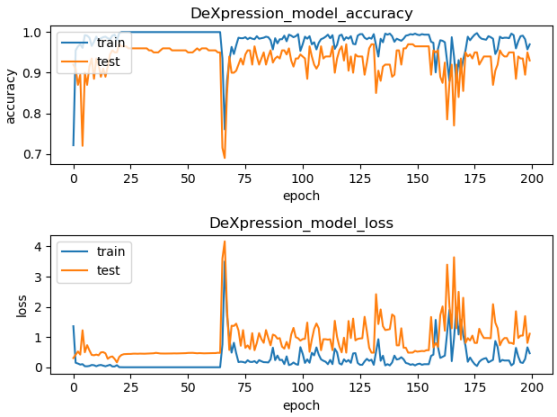


Figure 12: DeXpression on CK+

5.2 FER2013

- **DeXpression**

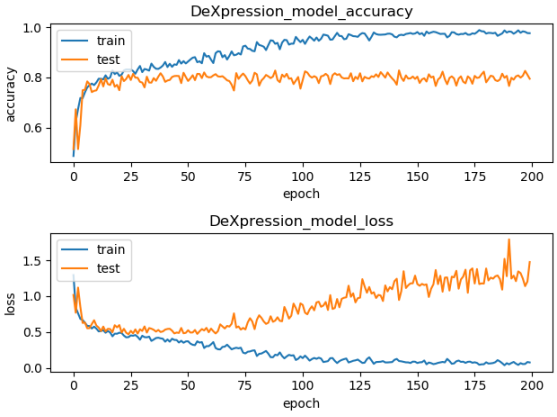


Figure 13: DeXpression on FER2013

InceptionV3 trained on **CK+** and **DeXpression** trained on **FER2013** has been tested and have proved great accuracy on finding the right expression. The stability of models can be seen with the training results of the model.

References

[1] P.-L. Carrier, A. Courville, I. J. Goodfellow, M. Mirza, and Y. Bengio, "Fer-2013 face database," *Technical report*, 2013.

[2] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pp. 94–101, IEEE, 2010.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[6] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[7] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "Dexpression: Deep convolutional neural network for expression recognition," *arXiv preprint arXiv:1509.05371*, 2015.