Facial Emotion Detection Using Convolutional Neural Networks and Representational Autoencoder Units

- Prudhvi Raj Dachapally, Indiana University

Abstract

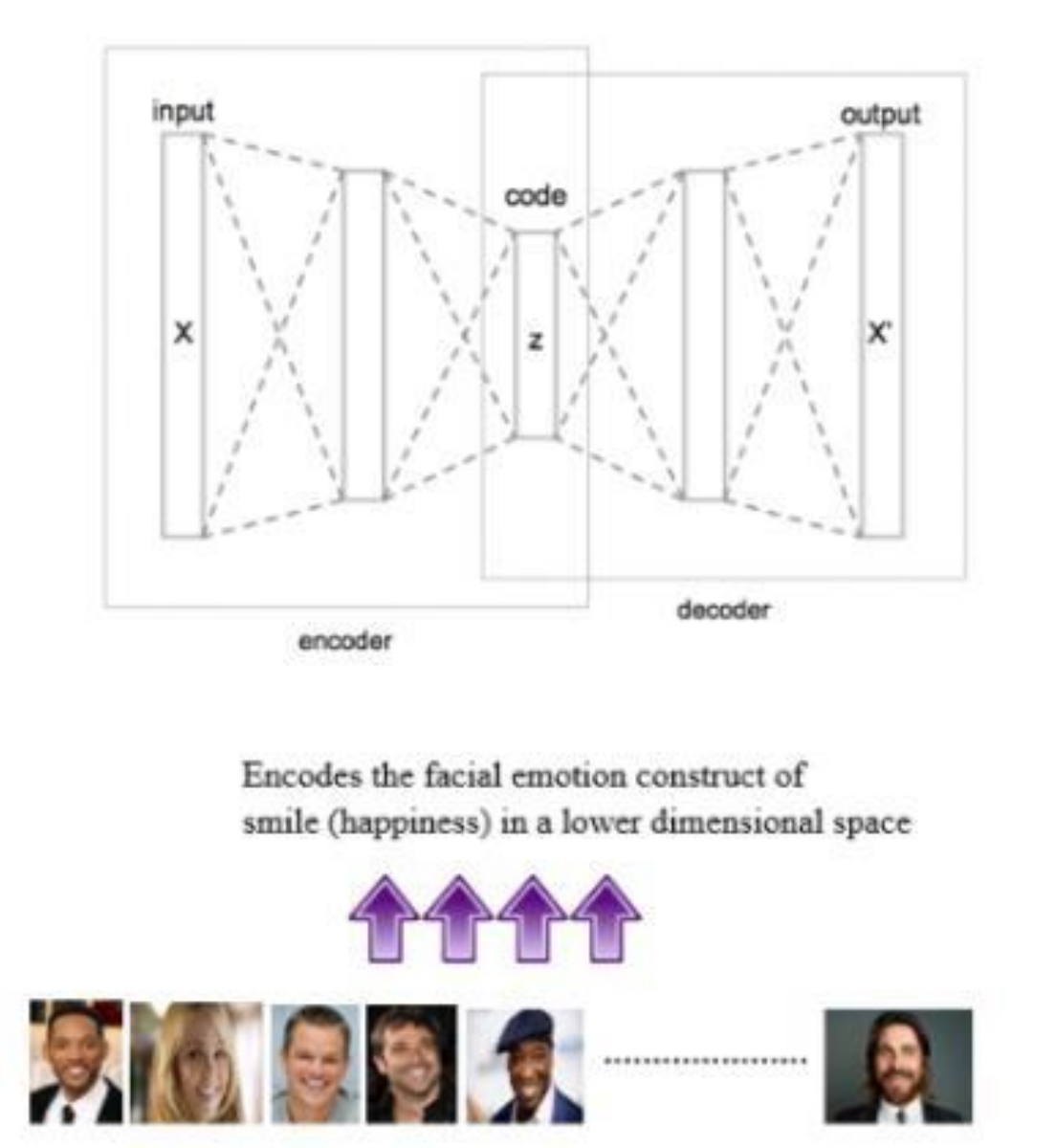
Emotion being a subjective thing, leveraging knowledge and science behind labeled data and extracting the components that constitute it, has been a challenging problem in the industry for many years. With the evolution of deep learning in computer vision, emotion recognition has become a widely-tackled research problem. In this work, two independent methods have been proposed for this very task. The first method uses autoencoders to construct a unique representation of each emotion, while the second method is an 8-layer convolutional neural network (CNN). These methods were trained on the posed-emotion dataset (JAFFE), and to test their robustness, both the models were also tested on 100 random images from the Labeled Faces in the Wild (LFW) dataset, which consists of images that are candid than posed. The results show that with more fine-tuning and depth, the CNN model can outperform the state-of-the-art methods for emotion recognition.

Background and Related Works

- Idea of using Representational Autoencoder Units (RAUs) first came from a paper by Hadi Amiri et al. wherein context-sensitive autoencoders were used to find similarities between two sentences.
- Lopes et al. used a 5-layer CNN on CK+ database for classifying six different classes of emotions.
- Hamester et al. used a 2-channel CNN where the upper channel used convolutional filters while the lower used Gabor-like filters in the first layer.
- Xie and Hu proposed a CNN structure using convolutional modules to reduce redundancy of same features learned and process the best set of features for the next layer.

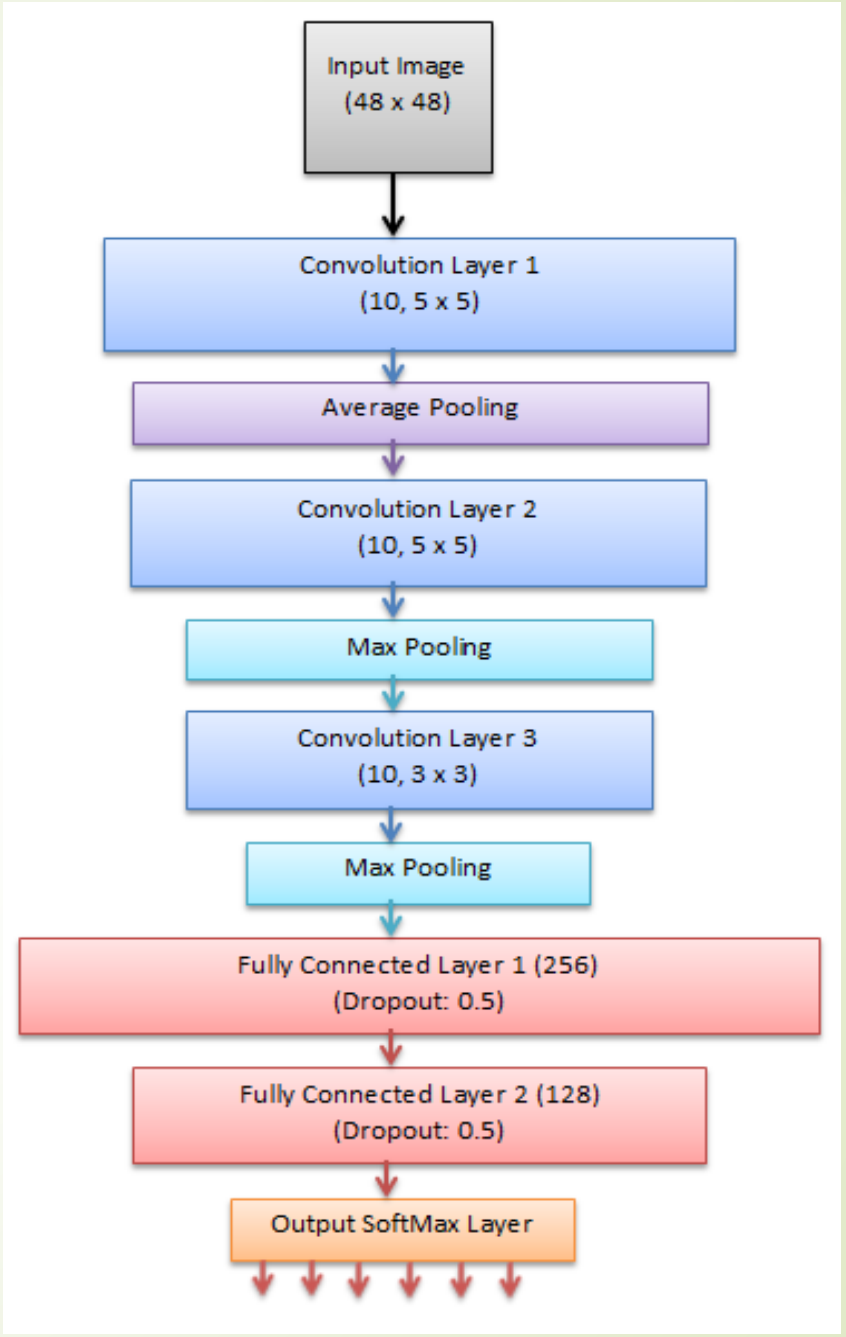
Methods Employed

- Representational Autoencoder Units (RAUs)
  - What are Autoencoders?
    1. A neural network model that can learn a compressed representation of raw data. They can reconstruct their own input in some lower dimensional space.
    2. Composed of encoder and decoder sub-models.
  - The intuition here is that given training examples of a certain kind of emotion, the RAU should be able to construct a representation of the features unique to that emotion.



Methods Employed (Contd.)

- Convolutional Neural Networks (CNNs)
  - 8-layer CNN with 3 convolutional layers, 3 pooling layers and 2 fully connected layers. Structure is as shown:



Experiments/Process of Training

- RAUs
  - 4 different autoencoder networks were designed. The 2 shallow ones have 1 hidden layer with 300 and 500 nodes. The other 2 are deeper networks with 2 layers of 2800 nodes each attached before and after the hidden layer.
  - The JAFFE dataset was used which has 215 images of 7 different emotions. The images were resized to 64 x 64 dimensions and then grouped based on the emotion class such that all images with the same labelled emotion were in the same group.
  - These groups (75% of total) were then trained on the autoencoder architectures from which 7 distinct representational units for each emotion were obtained.
  - For testing, the remaining (25%) of the data was used. Each test image was compared with each representational unit using a simple cosine distance function. Images were labelled the representational unit with which the lowest distance was obtained.

Experiments/Process of Training (Contd.)

- CNNs
  - Due to the small no. of training images available (159), data augmentation in the form of cropping of size 48 x 48 was employed on each 64 x 64-dimensional image.
  - However, all the obtained 48 x 48 patches were chosen for training and each 64 x 64 was condensed into 16 - 48 x 48 sized images. Since the original image size was small, more than 95 % of all the facial features were conserved in the patches.
  - The network gave the best validation accuracy after training for 360 iterations. The model was then tested on a separate set of 852 images.
  - To assess the robustness of the model and to perform cross-database evaluation, the model was also evaluated on the LFW dataset which comprises of candid images (or) real time data.

Results & Discussion

- RAUs
  - Since 4096-dimensional values are being condensed to 300/500 values, the top 2 closest distances were considered.
  - The shallow model got 25 out of 54 correct on the JAFFE test set while the deeper one did a little better, getting 29 correct.
  - Considering the Autoencoders were trained on a training set of Japanese women (same ethnicity and gender), the results obtained on the LFW test set of candid images were decent, with 53 out of 105 correct for the 300-node model.

Representational Autoencoder Units (105 LFW test images)
Baseline (Random Guessing: 14.28%)
(x – 300/500) (Top 2)

Structure	300 Nodes	500 Nodes
4096 – x – 4096	41.90%	44.76%
4096 – 2800 – x – 2800 - 4096	50.48%	48.57%

Representational Autoencoder Units (54 JAFFE test images)
Baseline (Random Guessing: 14.28%)
(x – 300/500)(Top 2)

Structure	300 Nodes	500 Nodes
4096 – x – 4096	46.29%	48.19%
4096 – 2800 – x – 2800 – 4096	53.70%	59.25%

Results & Discussion (Contd.)

- CNNs
  - Here are the results obtained using CNN model:

Convolutional Neural Network (Trained on 2556 images) Baseline (Random Guessing: 14.28%)		
Dataset	Images	Accuracy
JAFFE Test Set	852	86.38%
LFW (Top 2)	105	67.62%

  - The LFW test set predicted 71 out of 105 images correctly. The confusion matrix is as follows:

Confusion Matrix for LFW Test Set (105 Images)							
	AN	SA	SU	HA	DI	FE	NE
AN (15)	4	2	0	1	3	1	4
SA (21)	0	20	0	0	1	0	0
SU (8)	0	1	6	0	0	1	0
HA (29)	1	3	0	24	1	0	0
DI (9)	1	0	1	0	7	0	0
FE (6)	1	0	0	0	1	4	0
NE (17)	1	4	1	2	0	3	6

Results & Discussion (Contd.)

- CNNs (Contd.)
  - This is the confusion matrix for the JAFFE test set:
  - Through the obtained confusion matrices, it is observed that the boundary between happiness, neutral and sadness is quite thin in the facial structure. Most misclassifications for happiness were neutral and sadness and vice-versa.
  - The model developed is fairly robust as it gives decent accuracy on the LFW dataset with candid images. This implies that the model can leverage the subtleties of facial expressions in the images.

Confusion Matrix for JAFFE Test Set (852 Images)							
	AN	SA	SU	HA	DI	FE	NE
AN (110)	102	2	0	0	6	0	0
SA (130)	2	104	3	8	2	7	4
SU (120)	0	1	104	0	0	9	6
HA (131)	2	3	1	110	0	3	12
DI (117)	8	5	0	2	100	2	0
FE (130)	1	9	2	1	1	113	3
NE (114)	0	4	2	5	0	0	103

Results & Discussion (Contd.)

- CNNs (Contd.)
  - Another thing with the CNN model is that if the input image does not have a positive emotion, then the top predictions tend to be negative.
  - The first image has the ground truth label of Sadness, and the model gives the correct prediction of Sadness. Also, the other most probable prediction was Fear, a negative emotion.
  - In the second image, while the ground truth is Angry, the model makes an inaccurate prediction of Fear. However, again, the top 3 predictions happen to be negative emotions.

Original	Sadness
Sadness	99.98%
Fear	0.02%
Surprise	3.27E-04

Original	Angry
Fear	41.50%
Sadness	30.78%
Disgust	27.73%

Conclusion & Future Work

- In this paper, two different methods of solving the problem of emotion detection have been introduced.
- The first using autoencoders was introduced with the right kind of intuition in mind. However, the results obtained were not up to the mark. A possible reason for this could be loss of structural integrity of the image upon concatenating all the pixels vertically for the purpose of feeding into the autoencoder. In the future, replacing the normal hidden units with convolutional layers in the encoder part and deconvolutional layers in the decoder part is something which can be explored.
- The second method was using CNNs with three convolutional layers, three pooling layers and two fully connected layers. Good accuracies were achieved on the JAFFE and LFW test sets.

THANK YOU