Machine Learning Engineer Nanodegree

Capstone Proposal

Peter Raafat

23/3/2019

Proposal

Domain Background

Human beings convey emotions verbally and non-verbally. Facial expressions are one of the methods human beings convey emotions non-verbally. In the field of computer vision facial expression detection has been a topic of interest and a challenging problem. It can be used as a method for lie detection [1][2]. In clinical applications it can be used as a method to assess the patient pain regardless of their age [3]. In the field of automotive detection of human facial expressions can prove of high importance for smart cars, in case of high weather temperature and elevated state of non-comfort adjusting the air conditioner temperature might be necessary, or in case of fear, alarming the control system of a possible accident might help avoid the accident or if a presence of a human threat outside the vehicle, locking the doors and closing the windows might be appropriate.

Problem Statement

Detection of human facial expressions can be a challenging task for a computer. In this project I would like to analyze the possibility of using machine learning techniques particularly CNNs in order to solve the problem of finding what emotion the human being is experiencing from his facial patterns. The basic human emotions can be categorized into neutral, happiness, anger, disgust, fear, sadness and surprise.

Datasets and Inputs

For this problem the Japanese Female Facial Expression (JAFFE) database is a great candidate as the input data. The database contains 219 images in total representing 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. The images are all in grey scale and in size of 256x256 pixels. Each model has at least three images for each dominant emotion except “happiness” which consists of 4 additional images, “sadness” with 2 additional images, “disgust” with 1 additional image and “fear” with one additional image. The dataset can be considered as unbalanced. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects. The database was planned and assembled by Michael Lyons, Miyuki Kamachi, and Jiro Gyoba. We thank Reiko Kubota for her help as a research assistant. The photos were taken at the Psychology Department in Kyushu University. The database is available free of charge on the following link http://www.kasrl.org/jaffe.html.

Solution Statement

For the given problem I would incorporate the usage of deep neural network techniques such as CNNs in order to find what emotion best describes the image. The labels are converted from numeric quantities describing each emotion for each image into a vector of zeros and ones with the only a single one for the most dominant emotion. The dataset would be divided into training and testing sets in order to evaluate the performance of the model.

Benchmark Model

As the benchmark model I will compare my approach to the model provided by Lopes, A. T., de Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. [4] where they used CNN on the JAFFE database, CK+ database and BU-3DFE in order to attain facial expression recognition. Their model architecture is consisted of data augmentation, rotation correction, image cropping, down-sampling, intensity normalization followed by training using CNN. They used accuracy as a metric for evaluating their model.

Evaluation Metrics

The evaluation metric used for comparing the result of the model with the benchmark model will be accuracy since the classes can be considered unbalanced. Accuracy can be measured using the following equation:

$$accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions}$$

Or it can be expressed as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP = true positives, TN = true negatives, FP = false positives and FN = false negatives.

For adjusting the weights of the CNN log-loss will be used which can be calculated using the following equation:

$$log - loss = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

Where y are the true labels and $\hat{y}$ is the predicted labels.

Project Design

Architecture

Pre-processing

The labels available in the JAFFE database are numeric values for each image describing the intensity of each of the six basic emotions, happiness, fear, anger, sadness, surprise and disgust plus the neutral emotion. In order to change the problem into a classification problem the labels are converted to a vector of seven elements for every image where a one is positioned at the most dominant emotion for the image.

The JAFFE database contains few samples. CNNs need a lot of samples in order to achieve good training results. In order to resolve the problem artificial images can be obtained using data augmentation from the original database, this increases the number of samples available for training [4].

Face detection and cropping is then applied to the dataset. This decreases the number of irrelevant features available in the images hence decreasing the amount of time needed for training.

The images are then normalized such that all pixels have a value between 0 and 1.

Training

The images are then converted into 4D tensors and divided into training and testing sets. The model CNN layers are constructed then trained using the input 4D tensors for a multiple of epochs till converging. The model hyperparameters are tuned until acceptable performance is achieved. The architecture that I am planning to use will be the following. An input 2D convolutional layer with 64 filters and input of the same size as the cropped image and relu activation function followed by a max pooling layer with pool size of 2 and stride of 2 followed by a second conv2D layer with 128 filters and relu activation function followed by a max pooling layer of size 2 and stride 2 followed by a third conv2D layer with 256 filters and relu activation function followed max pooling layer and global average pooling layer, this is finally connected to a fully connected dense layer with softmax activation function. The loss function used will be categorical cross entropy. After tuning the hyperparameters grid-search could be used in order to

achieve even better hyperparameters without having to search a large space of hyperparameters. Transfer learning could also be used. Models like VGG-16 pre trained can achieve outstanding results since it is trained on a huge dataset.

Post-processing

The model accuracy is obtained from the neural network using the testing set and a confusion matrix is obtained to evaluate which emotions are best predicted by the model and which the model suffers to predict. The model results are then compared next to the benchmark model mentioned above. In case of predicting images that are not available in the JAFFE database the images will have to be converted to grey scale first, and the face extracted and cropped to the same size as the images used for training.

References

[1] Chen, J., Chen, Z., Chi, Z., & Fu, H. (2014, August). Facial expression recognition based on facial components detection and hog features. In *International workshops on electrical and computer engineering subfields* (pp. 884-888).

[2] Gosavi, A. P., & Khot, S. R. (2013). Facial expression recognition using principal component analysis. *International Journal of Soft Computing and Engineering (IJSCE)*, *3*(4), 2231-2307.

[3] Jiang, M., Gia, T. N., Anzanpour, A., Rahmani, A. M., Westerlund, T., Salanterä, S., ... & Tenhunen, H. (2016, April). IoT-based remote facial expression monitoring system with sEMG signal. In *2016 IEEE Sensors Applications Symposium (SAS)* (pp. 1-6). IEEE.

[4] Lopes, A. T., de Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, *61*, 610-628.