Proceedings of the International Conference on Communication and Electronics Systems (ICCES 2018)
DVD Part Number:CFP18AWO-DVD; ISBN:978-1-5386-4764-6

Emotional Jukebox

DR. BALIKA. J. CHELLIAH [1], S SHRIRAM [2], MEGHNA ANAND [3], HARIHARA SUDHAN [4]

Associate Professor(S.G) [1], [2,3,4] UG Scholers

SRM Institute of Science and Technology, Chennai

balika888@gmail.com[1],

shriram@ankor.us [2], meghna@ankor.us [3], hari@ankor.us[4]

***Abstract* — AI systems have often been developed mostly to understand human thoughts and emotions. This project is also one design that allows an AI system to classify and understand the emotion of a human user. Humanizing of an AI system in a way that can recognize various facial expressions, concludes an emotion with the help of convolutional neural networks and reciprocating the emotion as a song or a genre of music is what we are focusing on. The applicability of the final model portrays that not only can an AI depict profound IQ but also can possess an emotional conscience. Our final model was able score accuracies of ninety-six within the IMDB gender dataset and sixty-six within the FER-2013 feeling dataset. With this our model has radio-controlled backpropagation visual image technique. This uncovers the dynamics of the weight changes and evaluates the learned options. Once the AI is able to classify it starts playing songs from the respective playlist.**

Keywords: CNNs- Convolutional Neural Networks, AI- Artificial Intelligence, Emotional Recognition, Dataset, Modal, Jukebox

I. INTRODUCTION

In human evolution nonverbal communication has been very curial. The same use of this non-verbal information is essential for the growth and evolution of a machine. In the interaction between humans the face provides a vast range of information about the individual, such as sex and age as well as his emotional state. This information allows the computer system worldwide evolve into something more extravagant.

This study of detecting the emotions from an image has been extended in real time as well. Using a webcam, the faces detected in the frames of the video will be analyzed and the features of the face will be utilized to identify the facial emotion. In our project, very large amount of image datasets was collected organized and sorted according to various i.e., 7 emotions. Our final model will try to classify the FER-2013 dataset images in Figure 1 within the following classes {"happy", "disgust", "neutral", "angry", "sad", "surprised", "fear"}.

Face recognition is becoming one of the most interesting aspects in terms of AI and intelligent security, so teaching an AI to learn and detect emotions from a human face in real time would be a huge breakthrough in the field of computer systems and that is what we want to achieve.

The model is real time emotional classifier, that takes in webcam video as input and processes the data and outputs a song or a playlist based on the emotions identified. Currently this model is projected to be able to classify each and every frame input of the webcam into one of seven separate classes. The 7 classes are neutral, surprised, sad, happy, fearful, disgusted, and angry.

The networks are programmed with the use of the TFLearn Library on top of TensorFlow, running on python 3.6. This specific environment lowers the complexity of the code, as only neural layers are created instead of individual neurons

Fig. 1: Samples from FER – 2013 dataset

Also, the model provides real time feedback on training process, so it's easy to save and reuse the model after training. The model is a real- time visualization of the guided-gradient back-propagation proposed by Springenberg [9] in order to validate the features learned by the CNN.

II. RELATED WORK

Our research for this project deals with the main question of - "If an AI can understand human emotions and derive an output based on the facial expression on a human's face." For approaching to the techniques, we will be using, a large amount of data will be collected for experimenting with different emotions portrayed by people with different facial features. We will then come to the part where we show that facial expressions play a vital role in depicting one's emotion. This is the part where a playlist is suggested to the user depending upon the mood they are portraying. The aim of our project is to not design a system that is new but to improve it in a more unique and efficient way. Feature extraction model includes a set of fully connected layers at the end as they tend to contain most parameters in a CNN. Inception V3[10] reduced the number of parameters in their last layers by including a Global Average Pooling operation (GAP). GAP reduces feature maps into scalar value by taking average of all elements in the feature map. This forces the network to extract global feature from the input image. Also, we combined 2 experimental assumptions in our CNN: using

residual modules [4] and depth-wise separable convolutions [10]. With this the model is loaded with FER2-2013 dataset and CNN is trained with square hinged loss which achieves an accuracy of 71% [5] using approximately 5,000,000 parameters. 2nd best is presented in [5] which has accuracy of 66%. It uses ensemble of CNNs

III. EMOTIONAL JUKEBOX

A. Emotional Classifier

Two model are proposed in congruity to their precision and number of parameters. Both the models were created aiming at accuracy over parameters ratio. Reducing the number of parameters facilitates us in overcoming 2 necessary issues: First, it reduces the performance issues caused by the hardware constrained systems. Next, reducing the constraints(parameters) provide better performance under Occam's razor framework. Our primary model focuses on the idea of eliminating fully connected layers. Another model focuses on removal of the fully connected layer and the inclusion of the combined depth-wise separable convolutions and residual modules. Both of the models were trained by ADAM optimizer.

Following the previous design schemas, our initial architecture used Global Average Pooling to fully delete any totally connected layers. This was achieved by having last layers the same number of maps and features and applying an activation function to each of the reduced feature map. Our initial proposed design is a normal fully-convolutional neural network composed of nine convolution layers, ReLUs, Batch normalization and global Average Pooling. It was trained on the IMDB gender dataset, that contained 460,723 RGB pictures where every image belongs to the category "woman" or "man", and it achieved an accuracy of 96% in this dataset. We additionally validated this model within the FER-2013 dataset. This dataset contains 35,887 grayscale images where every image belongs to one of the subsequent categories "angry", "disgust", "fear", "happy", "sad", "surprise", "neutral". Our initial model achieved an accuracy of 66% during this dataset. We are going to sit down with this model as "sequential fully-CNN".

Our second model uses inspiration from Xception architecture. This architecture combines the utilization of residual modules and depth-wise separable convolutions. Residual modules modify the required mapping between two sequent layers, so the learned features become the distinction of the initial feature map and also the desired features.

In order to solve the problem, the desired features H(x) are changed so as to resolve an easier learning problem F (X) such that:

$$H(x) = F(x) + x$$

Since our initially proposed design deleted the last fully connected layer, we tend to reduce any the quantity of parameters by eliminating them currently from the convolutional layers. This was done through the utilization of depth-wise separable convolutions. Depth-wise separable convolutions are composed of 2 different layers: depth-wise convolutions and point-wise convolutions. The primary purpose of those layers is to separate the abstraction cross-correlations from the channel cross-correlations. They do this by initial applying a D D filter on each M input channels so by applying N x 1 x 1 x M convolution filters to mix the M input channels into N output channels. Applying 1 1 M convolutions combines every value within the feature map while not considering their position among the channel. Depth-wise separable convolutions reduces the computation with reference to the quality convolutions by a factor of N1 + D12 . A visualization of the distinction between a traditional convolution layer and a depth-wise separable convolution will be determined.

Our final design is a fully-convolutional neural network that contains 4 residual depth-wise separable convolutions where every convolution is followed by a batch normalization operation and a ReLU activation operation. This design has roughly 60; 000 parameters; that corresponds to a reduction of 10 when put next to our initial naive implementation, and 80 when put next to the initial CNN. The last layer applies a global average pooling and a soft-max activation operation to provide a prediction. Our final design weights can be kept in an 855 kilobytes file. By reducing our architecture's process cost we are currently able annex a part of each models and use them consecutively within the same image without any serious time reduction. Our complete pipeline as well as the openCV face detection module, the gender classification and also the emotion classification takes 0:22 0:0003 ms on a i5-4210M CPU. This corresponds to a speeding of 1:5 compared to the initial design of Tang.

We also added to our own implementation of a real-time guided back-propagation visualization to look at that pixels within the image activate a component of a higher-level feature map. Given a CNN with solely ReLUs as activation functions for the intermediate layers, guided-back propagation takes the by-product of each component (x; y) of the input image I with relation to a component (i; j) of the feature map fl in layer L. The reconstructed image R filters all the negative gradients; consequently, the remaining gradients are chosen such that they solely increase the value of the chosen component of the feature map. Following, a completely ReLU CNN reconstructed image in layer l is given by:

$$Ri;jl = (Ri;jl+1 > 0)\ \ Ri;jl+1$$

B. Music Player

The music player is a web app and this app is connected to the classifier with a local database. The python classifier updates the emotions recorded in a database field using native MySQL Library. After this the Web app when it runs out of song it looks into the database for the current emotional state of the test subject and picks a song from predefined playlist of songs loaded. The Music player also has an option to play songs directly of any popular streaming service like Youtube, Prime music or Spotify.

The Youtube mode lets the app play any song from youtube using the embedded link or play an entire playlist the same way. Using the same method its easy to incorporate any other web streaming service by providing an embedded link.

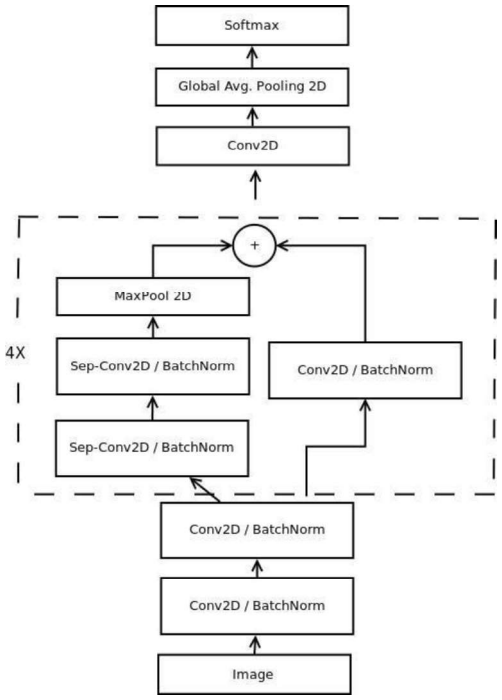

Fig. 2: Proposed model for real-time detection

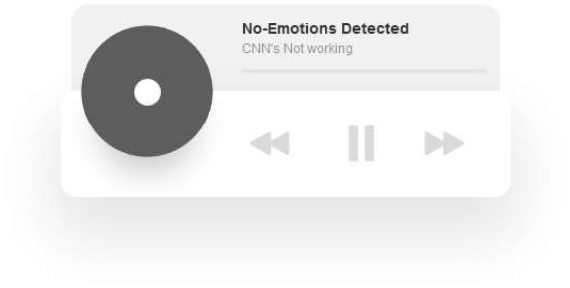

Fig. 3: Music player from the proposed system.

IV. RESULT

Our entire continuous pipeline including: person recognition, feeling and sexual orientation characterization have been completely incorporated in our classifier.

The white regions in figure compare to the pixel esteems that enact a chose neuron in our last convolution layer. The neuron was constantly chosen in agreement to the most astounding initiation. We can see that the CNN figured out how to get initiated by considering highlights, for example, the glare, the teeth, the eyebrows and the enlarging of one's eyes, and that each component stays steady inside a similar class. These outcomes console that the CNN figured out how to decipher comprehend capable human-like highlights, that give generalizable components. These interpretable outcomes have helped us comprehend a few regular misclassification, for example, people with glasses being named "furious". This occurs since the name "anger" is profoundly initiated when it trusts a man is glaring and scowling highlights get mistook for darker glass outlines. Additionally, we can likewise see that the highlights learned in our Xception model shows that there are more interpretable than the ones gained from our consecutive, complete CNN. Subsequently the utilization of more parameters in our innocent executions prompts less powerful highlights.

V. FUTURE WORK

The accuracy of the machine learning model depends on their training data. Our trained CNNs are biased towards western facial features and accessories. We hypothesize that this mismatch occurs due to the trained dataset's facials not being similar to that of the testing data. We will have to find a better way of training the model to make it more humanized. Also the music player could be modified and many other features like track skipping, song queue, custom playlists which the player currently lacks will be an option. Improving the current responsiveness of the player on all platforms will surely be an option too.

Fig. 4: A Guided backpropagation visualization of the FER-2013 dataset images using both proposed models.

VI. CONCLUSIONS

We have proposed and tested a music player that integrates AI capabilities by recognizing human emotions. Our model has been built systematically to reduce parameters using real time CNNs.The idea started by eliminating fully connected layers and by reducing the number of parameters in the remaining layers. This was done using depth wise separable convolutions. Our system recognizes face, gender and emotions all in a single integrated module. There by reducing parameters and still obtaining optimal results. This classified data is then fed to the music player and based on the predictions, a playlist or a song is played.

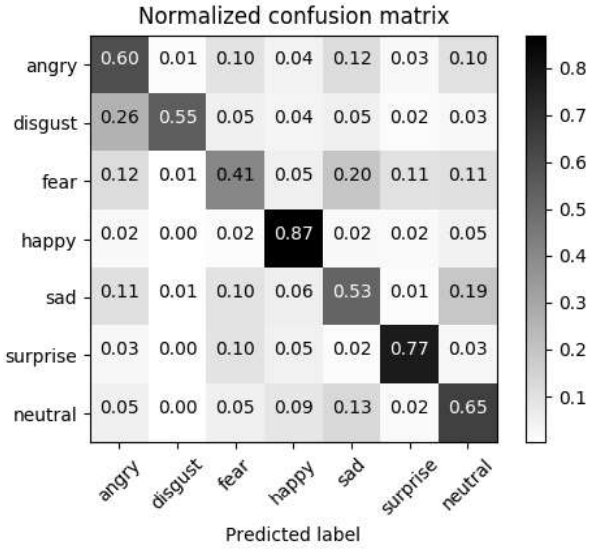

Fig. 5: One of the model's output in a normalized confusion matrix.

REFERENCES

[1] Andrew G. Howard et al. Mobilenets: Efficient convolutional neuralnetworks for mobile vision applications, 2017.

[2] Dario Amodei et al. Deep speech 2: End-to-end speech recognition in english and mandarin, 2015.

[3] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 315–323, 2011.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[5] Ian Goodfellow et al. Challenges in Representation Learning: A report on three machine learning contests, 2013.

[6] Francois Chollet. Xception: Deep learning with depthwise separableconvolutions. CoRR, abs/1610.02357, 2016.

[7] Yichuan Tang. Deep learning using linear support vector machines, 2013.

[8] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[9] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net.

[10] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2818–2826, 2016.

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning, pages 448–456, 2015.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[13] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision (IJCV), July 2016