International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.134.18

ResNet-50 and VGG-16 for recognizing Facial Emotions

Poonam Dhankhar[1]
[1]*Department of Computer Science and Engineering, MSIT, New Delhi, India*

Abstract- This paper discusses the application of feature extraction of facial expressions with combination of neural network for the recognition of different facial emotions (happy, sad, angry, fear, surprised, neutral etc.). Facial expression plays a major role in expressing what a person feels. It expresses human perspective or inner feeling& his or her mental situation. A human brain can have lot of emotions but this paper deals with the main 7 emotions. This paper deals with the highly efficient model of the combined models of VGG 16 and ResNet50 increasing the efficiency to 92.4%. The earlier baseline models used were support vector machine. We have named the combine model as Assemble model. These papers leveraged assemble and transfer learning to achieve the best results.

Keywords – End to End learning, emotion recognize, deep learning, facial emotion recognition, conventional FER, deep learning-based FER, convolutional neural networks, long short term memory, facial action coding system, facial action unit .

I. INTRODUCTION

Understanding human emotions [1] is key area of research, as the ability to recognize one’s emotions can give one access to a plethora of opportunities and applications, ranging from friendlier human-computer interactions, to better targeted advertising campaigns, and culminating with an improved communication among humans, by improving the emotional intelligence (“EQ”) of each of us. While there are multiple ways one can investigate the recognition of human emotions, ranging from facial expressions, and posture of the body, speed and tone of the voice, in this paper we shall focus on only one area of this field - visual recognition of emotions. One of the reasons we chose to focus on the area of facial expressions is because certain facial expressions have universal meaning, and these emotions have been documented for tens and even hundreds of years. “Constants across cultures in the face and emotion”[2]. That paper identified the following six key emotions: anger, disgust, fear, happiness, sadness and surprise. These are the same emotions that are being use d by current researchers to identify facial expression in computer vision, or in competitions such as Kaggle’s Facial Expression Recognition Challenge, along with the addition of a seventh, neutral emotion, for classification.

Thus, our research is about using deep learning (a VGG- 16 convolution network and a ResNet50 convolution network) to identify these seven main human emotions [3]. To us this problem is extremely relevant because of its broad spectrum of applicability in a variety of fields, such as systematic recruiting, while being also able to be integrated with a variety of technologies (i.e. smartglasses, VR, wearables, etc.). Emotions and facial responses can also serve as a new dimension of user information (i.e. imagine Facebook or Google analyzing your emotions and reactions to learn more about the user and serve better recommendations and ads). To achieve our goals, we used a support vector machine (SVM) classifier baseline model and develop a convolution neural network (CNN) to classify these emotions. In particular, we will use some of the current state of the art architectures - VGG-16 and ResNet50, while making some adjustments which include applications of various deep learning techniques, and ensemble and transfer learning [5]. We chose to go with VGG-16 and ResNet50 because they won in the past the ImageNet challenge, achieved near state-of-the-art results in terms of prediction accuracy, and follow a relatively standard CNN architecture. The two datasets we will leverage in our research are the Kaggle’s Facial Expression Recognition Challenge and Karolinska Directed Emotional Faces (KDEF) datasets. We found these datasets to be representative because of their size, unstructured nature of faces (in terms of facial orientation, ethnicity, age, and gender of the subjects) and relatively uniform distribution of the data across the seven main human emotions (disgust being the only underrepresented one within the Kaggle dataset, at~1.5%). To evaluate the performance of our models, we will primarily be looking at the accuracy on the training, validation, and test sets. The processes, we will be leveraging other standard statistics such as precision and recall providing further insights on the efficacy of the models. We expect our best model to achieve at least 60% test set valuation because the winner of the Kaggle challenge achieved 71.2% accuracy and the top ten contestants achieved at least 60% accuracy. Expression Recognition Challenge, along with the addition of a seventh, neutral emotion, for classification.

II. LITERATURE SURVEY

Previous works are focused on eliciting results from uni-modal systems. Machines used to predict emotion by only facial expressions [1] or only vocal sounds [2]. After a while, multimodal systems that use more than one features to predict emotion has more effective and gives more accurate results. So that, the combination of features such as audio-visual expressions, EEG, body gestures have been used since. More than one intelligent machine and neural networks are used to implement the emotion recognition system. Multimodal recognition method has proven more effective than uni-modal systems by Shiqing et al. [3]. Research has demonstrated that deep neural networks can effectively generate discriminative features that approximate the complex non-linear dependencies between features in the original set. These deep generative models have been applied to speech and language processing, as well as emotion recognition tasks [4-6]. Martin et al. [7] showed that bidirectional Long Short-Term Memory (BLSTM) network is more effective that conventional SVM approach. In speech processing, Ngiam et al. [8] proposed and evaluated deep networks to learn audio-visual features from spoken letters. In emotion recognition, Brueckner et al. [9] found that the use of a Restricted Boltzmann Machine (RBM) prior to a two-layer neural network with fine-tuning could significantly improve classification accuracy in the Interspeech automatic likability classification challenge [10]. The work by Stuhlsatz et al. [11] took a different approach for learning acoustic features in speech emotion recognition using Generalized Discriminant Analysis (GerDA) based on Deep Neural Networks (DNNs). In the deep neural approaches, the already used approaches are baseline SVM model which is followed by CNN model with 43.8% efficiency. One paper implements ResNet50 for vague training data and gets the efficiency about 72.7%.

2.1 VGG-16

The input to VGG-16 layer is of fixed size 224 x 224 RGB image. The image is passed through a stack of convolutional (conv.) layers, where the filters were used with a very small receptive field: 3×3 (which is the smallest size to capture the notion of left/right, up/down, center). In one of the configurations, it also utilizes 1×1 convolutional filter, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolutional stride is fixed to 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1-pixel for 3×3 conv. Layers (Figure 1). Spatial pooling is carried out by five max-pooling layers, which follow some of the convolutional layers (not all the conv. layers are followed by max pooling). Max-pooling is performed over a 2×2-pixel window, with stride 2. VGG16 was trained for weeks and was using NVIDIA Titan Black GPU’s.

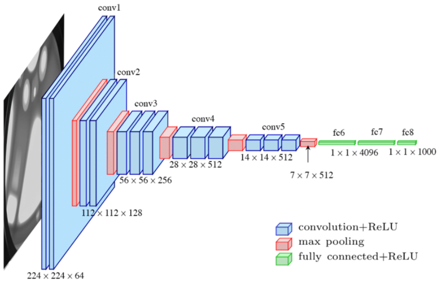

Figure 1: VGG-16 architecture diagram.

The input to our VGG-16 is a 48x48 RGB image. The only preprocessing we do is subtracting the mean RGB from each pixel. The image is passed through a stack of convolution layers ,where weuse3x3filters.Inoneofthe configurations we also utilize 1 × 1 convolution filters, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to 1 pixel; the spatial padding of convolutional layer input is such that the spatial resolution is preserved after convolution(i.e.thepaddingis1pixelfor 3 × 3 conv. layers). Spatial pooling is carried out by five max-pooling layers, which follow some of the convolutional layers (not all the convolutional layers are followed by max-pooling). Max-pooling is performed over a 2 × 2-pixel window, with stride2.

A stack of convolutional layers is followed by three Fully Connected (FC) layers: the first two have 4096 channels each, the third performs 7-way ILSVRC classification and thus contains seven channels (one for each class). The final layer is the so ft max layer. The configuration of the fully connected layers is the same in all networks. All hidden layers are equipped with the rectification (ReLU) non linearity.

To conclude, VGG-16 consists of 16 weight layers that include13convolutional layers with filter size of 3x3 and 3 fully connected layers. The stride and padding of all convolutional layers are fixed to 1 pixel. All convolutional layers are divided into 5 groups and each group isfollowed by a max-pooling layer (Figure 1). Max-pooling is carried out over a 2x2 window with stride 2. The number of filters of convolutional layer group starts from 64in the first group and then increases by a factor of 2 after each max-pooling layer, until it reaches 512. We leveraged the keras implementation of VGG-16.

III .RESNET50

ResNet 50 is current state of the art convolutional neural network architecture. It is similar in architecture to networks such as VGG-16 but with the additional identity mapping capability (Figure2).

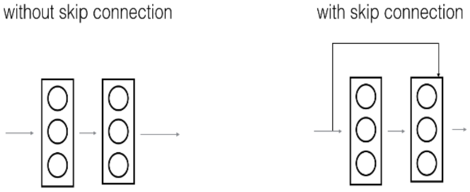

Figure 2: ResNet residual block diagram with skip connection

ResNet short for Residual Networks is a classic neural network used as a backbone for many computer visions 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. ResNet first introduced the concept of skip connection. The diagram below illustrates skip connection. The figure on the left is stacking convolution layers together one after the other. On the right we still stack convolution layers as before, but we now also add the original input to the output of the convolution block. This is called skip connection. There are two reasons why Skip connections work here: They mitigate the problem of vanishing gradient by allowing this alternate shortcut path for gradient to flow.

They allow the model to learn an identity function which ensures that the higher layer will perform at least as good as the lower layer, and not worse.

	Accuracy	Precision	Recall
SVM (baseline)	31.8%	43.7%	54.2%
VGG-16	59.2%	70.1%	69.5%
ResNet50	65.1%	76.5%	74.8%
Ensemble	67.2%	79.4%	78.2%

3.1 Assemble Learning

While VGG-16 and ResNet50 are currently two of the state-of-the-art deep learning architectures, we attempt to combine these two models by leveraging an ensemble approach .From the second to last layers ,we obtain a vector of weights which can be treated as feature vectors. These feature vectors represent the latent representation of the input image which each model learned. We combine these latent representations by concatenating the feature vectors to form an overall feature vector which is inputted into logistic regression models for the final emotion prediction (Figure 3). We train one logistic regression for each emotion, for a total of seven models, and taking the model with the highest score as the prediction. So, for each image we compute nine predictions: one from VGG-16, one from ResNet50.

IV. RESULTS

The precision shows us the positive predictive value, and recall captures the sensitivity or true positive rate of the models. To compute the overall precision and recall, we use micro-averages to combine the results across all seven emotions. For both the Kaggle and KDEF datasets, we used a 80-10-10 split for the train, validation, and test sets. To further understand and assess our models, we examined the metrics for each emotion as well as the confusion matrix.

In Table 1 below, we see the results of the SVM (baseline), VGG-16, ResNet50 and ensemble learning models on the Kaggle dataset. Our baseline SVM accuracy was 31.8% while VGG-16 and ResNet50 had accuracies of 59.2% and 65.1%. Because ResNet50 contains identity bypass layers, it is possible that this is helping the model achieve better performance in terms of accuracy, precision, and recall compared to VGG-16. The ensemble learning model which effectively combines VGG-16 and ResNet50, achieved an accuracy of 67.2%, 2.1% greater than either VGG-16 or ResNet50individually.

The overall accuracies along with precision and recall on the KDEF dataset are greater than those on the Kaggle dataset. SVM achieved an accuracy of 37.9% while VGG- 16 and ResNet50 achieved accuracies of 71.4% and

73.8%, respectively (Table2).The ensemble approach achieved an accuracy of 75.8% and continued to perform better than the individual deep learning models. The ranking of the four models is the same for KDEF as it is for Kaggle. Wefound it surprising that all four models performed better on the KDEF, a significantly smaller dataset than Kaggle. We conjecture that this may be a result of the structure and uniformity of the KDEF dataset in terms of the subjects' postures and number of examples for each subject and each emotion.The images in the KDEF data set are also of higher quality. Aside from better image resolution, there were examples in the Kaggle dataset where there was, for example, text overlay in the background of the image.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP= True Positive, TN= True Negative
FP= False Positive, FN= False Negative

	Accuracy	Precision	Recall
SVM (baseline)	37.9%	50.1%	54.9%
VGG-16	71.4%	81.9%	79.4%
ResNet50	73.8%	83.3%	80.7%
Ensemble	75.8%	85.0%	82.3%

Table 1: KDEF dataset performance (accuracy, precision, and recall) for SVM, VGG-16, ResNet50, and ensemble learning models.

Applying transfer learning further improved the results. After training the VGG-16 and ResNet50 models on the Kaggle dataset, we fixed the layer weights aside from the last few layers of these models and retrained on the KDEF dataset. This led to a 2.5% accuracy improvement in our ensemble model which was our best performing model (Table 3). Precision and recall were similarly improved. This shows that the model was able to leverage the learnings from the faces of the Kaggle dataset which contained a wider and more abundant distribution of data and transfer those learnings to the smaller KDEF dataset.

	Accuracy	Precision	Recall
VGG-16	73.6%	84.2%	81.1%
ResNet50	76.0%	86.1%	82.5%
Ensemble	78.3%	87.3%	84.3%

Table 2: KDEF dataset performance (accuracy, precision, and recall) with transfer learning from Kaggle models.

To help assess the model performance on each individual emotion, we summarize the findings in Figure 7. The minimum accuracy, precision, and recall are 56.1% (neutral), 48.2% (sad) and 56.1% (neutral). Sadness and neutrality, as we further discuss later on, possess similar facial features as each other and a couple other emotions. We also note that we performed the best on happiness, which may be due to having the most data coverage for this emotion. While it is surprising, due to the lack of data coverage, we achieved 81.8% accuracy on disgust, the low precision indicates that the model may not have learned to distinguish disgust amongst other emotions and is predicting disgust more often than it should.

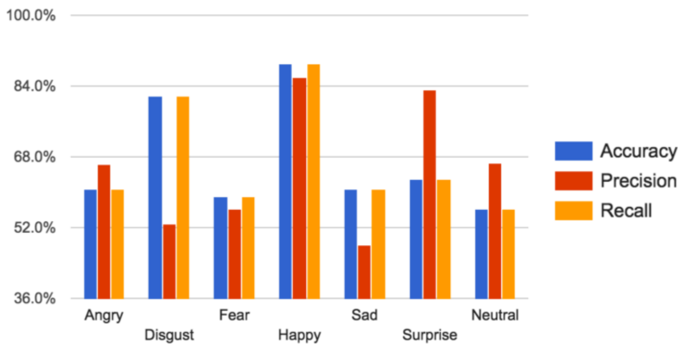

Figure 3 shows the confusion matrix for our best performing model on the Kaggle dataset. The correlations between actual and predicted emotion hold for the other three models we experimented with. The matrix reveals that anger, disgust, fear, and neutrality tend to get miss categorized with sadness. Conversely, sadness tends to be miss categorized with the same set of emotions. Looking at the raw images, we can qualitatively see that the facial expressions for sadness have commonalities with that for those emotions, especially the aspects of the mouth area (aside from anger).Since we did not add additional features aside from the processed image pixels; it isn't surprising that these emotions are confused with one another. Lastly, surprise is confused with both fear and happiness.

	Disgust	Happy	Surprised	Neutral	Angry	Fear	1Sad
Disgust	97.925	0.01	0.039	0.168	0.085	0.01	1.763
Happy	0.005	99.979	0.007	0.0	0.0	0.007	0.0
Surprised	0.005	0.0	99.378	0.0	0.331	0.03	0.256
Neutral	0.004	0.363	23.765	75.48	0.004	0.004	0.378
Angry	0.0	0.001	0.008	0.0	99.99	0.0	0.0
Fear	0.03	0.0	0.256	0.0	0.333	99.375	0.005
Sad	0.004	0.006	0.001	0.001	0.001	1.224	98.763

Figure 3: Confusion matrix with actual (true) emotion rows and predicted emotion columns (Kaggle, ensemble learning)

V. FUTURE WORK SCOPE

We are working towards a machine with emotions. A machine or a system, which can think like humans, can feel warmness of heart; can judge on events, prioritized between choices and with many more emotional epithets. To make the dream reality first we need the machine or system to understand human emotions, ape the emotion and master it. We just started to do that. Though there is some real example exists these days. Some features and services are getting popularity like Microsoft Cognitive Services but still there is a lot works required in the terms of efficiency, accuracy and usability. Therefore, in future Emotion Recognition is an area requires a great intentness

VI. CONCLUSION

We explored the VGG-16 and ResNet50 architectures for recognizing facial emotions using deep learning. The results demonstrated that we were able to achieve acceptable results in comparison to other Kaggle contestants and researchers leveraging the KDEF dataset. We further improved these models by developing an ensemble model to combine the outputs from the two neural networks. Coupled with transfer learning, we achieved 67.2% accuracy on the Kaggle dataset and 78.3% accuracy on the KDEF dataset .For context, the winner of the Kaggle Facial Expression Recognition Challenge achieved an accuracy of 71.2% and the top 10 finalists achieved accuracies of at least 60%.

VII. REFERENCE

[1] Albert Mehrabian. Silent Messages, University of California Los Angeles,2015.

[2] P. Ekman and W. V. Friesen. Emotional facial action coding system. Unpublished manuscript, University of California at San Francisco,2017.

[3] Yaniv Taigman, Ming Yang, Marc'AurelioRanzato, Lior Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. The IEEE Conference on ComputerVisionandPatternRecognition(CVPR),2014,pp. 1701-1708.

[4] Very deep convolutional networks for large-scale image recognition. Visual Geometry Group, Department of Engineering Science, University of Oxford,2018.

[5] Ruiz-Garcia A., Elshaw M., Altahhan A., Palade V. (2016) Deep Learning for Emotion Recognition in Faces. In: Villa A., Masulli P., Pons Rivero A. (eds) Artificial Neural Networks and Machine Learning – ICANN 2016. ICANN 2016. Lecture Notes in Computer Science, vol 9887. Springer, Cham.

[6] A.Mehrabian, "Communication without Words" Psychology Today, Vol.2, no.4, pp 53- 56, 1968

[7] Ekman P, Friesen WV. Constants across cultures in the face and emotion Journal of personality and social psychology 1971; 17:124

[8] Bharati A.Dixit and Dr. A.N.Gaikwad "Statistical Moments Based Facial Expression Analysis" IEEE International Advance Computing Conference (IACC), 2015

[9] S.Ashok Kumar and K.K.Thyaghrajan "Facial Expression Recognition with Auto-Illumination Correction" International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), 2013

[10] Mateusz Zarkowski "Identification-deiven Emotion Recognition System for a Social Robot" IEEE, 2013

[11] Shuai Liu and Wansen Wang "The application study of learner's face detection and location in the teaching network system based on emotion recognition" IEEE, 2010