Characterization of Facial Expression using Deep Neural Networks

Neha Sharma , Charvi Jain
Department of Computer Science and Engineering
Indian Institute of Information Technology
Una, Himachal Pradesh, India
neha724@gmail.com, jaindirect@gmail.com

***Abstract*—Deep learning plays a significant role in the advancement of computer vision by improving the speed and accuracy to the assigned tasks. It is opening opportunities for improvement and enhancement of processes and to initiate the human-driven tasks in an automated manner. On the basis of this growth, deep-learning algorithms are finding applications in the field CNN and RNN. The key advantage of Deep Learning algorithm is that manually extraction of features from the image is not required. The network extracts the features while training. The only input required is to provide the image to the network. The CNN's and RNN's have given state-of- the art results on numerous classification tasks. The Deep learning algorithm are designed for feature detection / extraction, classification and recognition of the object. The key advantage of a CNN is to remove or reduce the reliance on physics-based models, other processing methods by enabling complete learning directly from the input images of the object. The CNN and RNN together has given effective results in the area of face recognition, object recognition, scene understanding and facial expression recognition.**

Keywords*— *Deep Learning, Computer Vision, CNN, RNN

I. Introduction

With great progress in facial image recognition in the recent past, intelligent emotion recognition system are opening new avenues for analysis and evaluation of human interactions. Emotion recognition has to undertake the psychological findings, facial expressions plays the key role as it exhibits the emotions. The emotions are dynamic happenings in the real life, so recognition of emotion between human and computers is all together is a exciting task. Convolutional Neural Network (CNN), has given breakthrough in several computer vision areas. The image descriptors formed by CNN gives remarkable results for the image classification. The only limitation of CNN is just that it handles the spatial information. It has been also observed that the beginnings from convolutional layers can be analyzed as local features showing the particular image areas. The local features can be aggregated using aggregation techniques established for local features, which results to a global descriptor. Therefore, the image features of each of the video results into a feature vector, which ignores the useful temporal video inputs. The effective results motivate us to move deeper to bring further experiments, which also includes the experiments of hybridizing to the LSTM models. Thus effectively, such hybrid network produces superior outcomes for emotion classification. The key role of this paper is the characterization of facial expression using Deep Neural Networks.

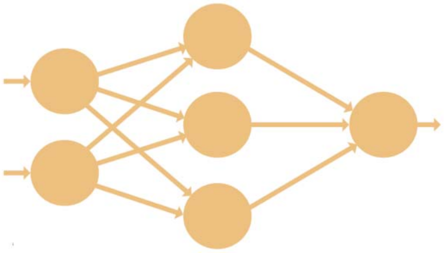

Fig. 1. Feed forward neural network [9]

Recurrent Neural Networks are a commanding and strong type of neural network and provide the most capable algorithms, the reason being that they have internal memory. The Long Short Term Memory [LSTM] are widely accepted [1]. The memory ability of the LSTM helps in processing sequences with the contexts as well. RNN's internal memory features, enable them to remember the important aspects about the input received by them, which eventually helps them to very precise in forecasting the forthcoming events. The precise results are helpful and preferred for sequential data viz: financial data, time series, speech recognition, audio, video, weather , stock market prediction, image captioning, next word prediction and so on [2,4].

The encoder LSTM are designed to address an input sequence into a fixed vector representation. This can be decoded using a single or multiple LSTM. Also LSTM have given efficient output in analyzing the sequence models.

978-1-5386-9533-3/19/$31.00 ©2019 IEEE

II. Methodology

Conventional Neural Network (CNN): It consist of a feed-forward artificial neural network. CNN considers arrays of several pixel values as an input to the network. The hidden layers consist of different layers which are used for feature extraction [3]. There is a connected layer which finally identifies the objects in the image. CNN has four types of heterogenous layers. Convolutional layers are key to the CNN. The convolutional layer practices a filter matrix on the array of image pixels and does convolutional operation to finally result in a convolved feature map [6]. The below example depicts the convolution operation on the input array images.

Convolutional Layers:

Image Pixels

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Filter

1	0	1
0	1	0
1	0	1

Fig.2. Image Matrix

Convolution Pixels:

Image Pixels

1	1	1	0	0
0	1	1_{x1}	1_{x0}	0_{x1}
0	0	1_{x0}	1_{x1}	1_{x0}
0	0	1_{x1}	1_{x0}	0_{x1}

Resulting matrix is called Convolved feature map

4	3	4
2	4	3

4	3	4
2	4	3
2	3	4

Fig.3. Output Matrix

ReLU Layers: It gives non-linearity to the network by setting all negative pixels to 0 and completes element wise operations. The input image is captured in several Convolution Layers and ReLU layers for identifying the hidden features and patterns in the image.

Pooling Layers: It helps in reduction of parameters of the feature map. The result is a pooled feature map. The images are too large, so they are to be reduced to trainable parameters. The pooling layers are introduced between the convolution layers. The purpose behind pooling is to condense the size of the object image. Pooling is done separately on each dimension, thus the depth of the image is not changed.

Featured Map

1	4	2	7
2	6	8	5
3	4	0	7
1	2	3	1

The pooled feature map is converted to a continuous layer vector. This feature is called flattening. The flattened matrix is directed through a fully coupled layer for the classification of the images.

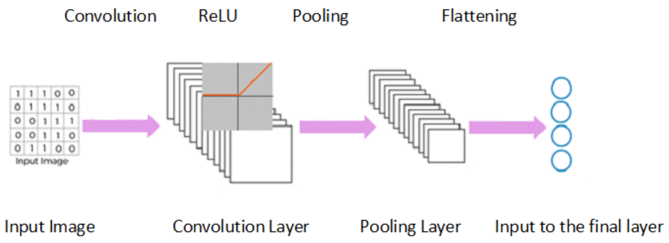

Fig.4. Flattening Matrix [12]

CNN recognition of an image:

Below is an example of the entire flow of how CNN recognizes an image:

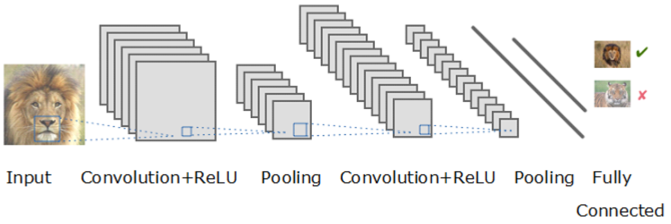

Fig.5. Feature Extraction and Output Layer [11]

Recurrent Neural Networks (RNN): They are strong and powerful type of neural networks and provides promising algorithms as they have an internal memory. The internal memory enable them to remember the important things, which again enable them to be very specific in predicting the forthcoming activities. The RNN is the preferred for sequential data like text, speech, time series, financial data, video, weather, as they perform a deeper understanding of a sequence and the relevant context. The information flows through a loop. While taking decision, it considers the current input and considers the inputs which were received previously. The two images shown below demonstrate the difference between a RNN and a Feed Forward Neural Network.

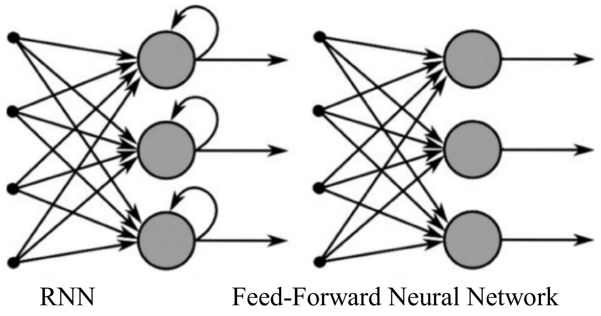
RNN Feed-Forward Neural Network

Fig.6. Information Flow [9]

The image below demonstrates an unrolled RNN. This demonstrates that RNN can be seen as a sequence of Neural Networks.

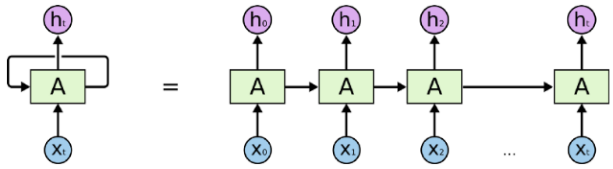

Fig.7. Loop Structure [8]

Long Short Term Memory:

The LSTM network are basically enhancement of RNN features, which extend their memory. LSTM has added benefits over the Feed-forward neural networks and RNN in several ways. They have the property of specifically remembering memory for the lengthy duration of time [7, 10].

The LSTM consist of three gates viz. input, forget and the output gate. The input gate decides whether or not to allow new input in. Forget gate removes the information, if the information is not important. Output gate determines to let it influence the output at the current time step. The illustration of a RNN along with the three gates is shown below:

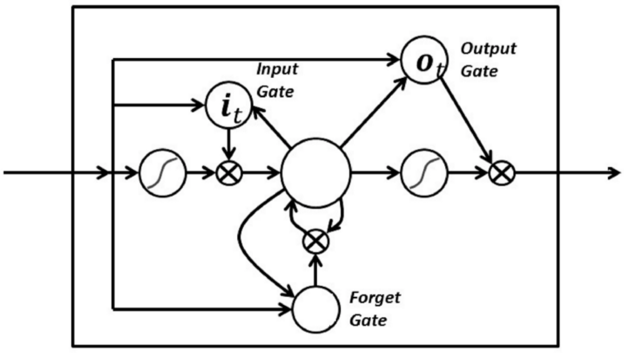

Fig.8. RNN with Input, Output and Forget Gate [9]

III. EXPERIMENT

The model was evaluated on the Cohn-Kanade database. The classification was done for the each of the video sequences into one of the expression classes viz: anger, happiness, sadness, disgust, surprise and fear. The accuracy was checked by using deep neural networks like CNN, RNN, LSTM and hybrid models i.e. CNN-LSTM.

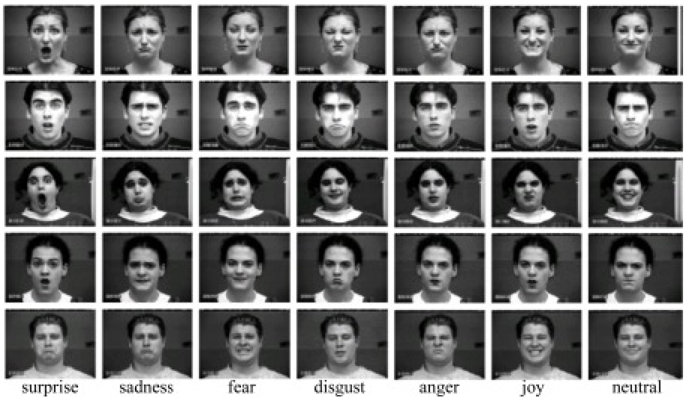

Fig.9. Cohn Kanade Database [3]

IV. EXPERIMENT EVALUATION

The Cohn-Kanade facial expression recognition database was taken for evaluation of the system. The facial expressions portray the emotional demonstrations and represents the characteristics. The expressions have concrete impact when doing the analysis of the emotional state of a human. The facial expressions have universality which unites people together throughout the world. The persons must be speaking different language, but things can be inferred through facial expressions and emotions. The facial expressions have universal, social, cultural and racial invariant emotion characteristics. The objective was to distinguish the video sequences into one of the six established universal emotions viz: “Happiness”, “Surprise”, “Sadness”, “Anger”, “Disgust” and “Fear”. The experiment used 5-fold cross validation to arrive at the classification accuracy [5].

The Cohn-Kanade is publically available with the sole purpose of encouraging research into automatically detecting the individual facial expressions. The Cohn-Kanade database is most widely used for algorithm development and evaluation. The short image sequences were taken for 30 persons of male and female genders, and the age group was between 20 to 35 years. The target expression for each sequence involves specifications for the activation of AUs and the depiction of facial expressions.

The paper suggests a Bidirectional Long Short Term Memory Networks [7] with 80 cells in the forward and backward hidden layers. It has a unidirectional LSTM network with 120 cells, which are in the hidden layers. The bidirectional Long Short-Term Memory Networks and unidirectional LSTM network had thirty inputs and six output units, on the basis of each target class. The hidden layers were connected to each other and also to the input and output.

The Bidirectional Long Short-Term Memory Networks had 90,292 trainable parameters, whereas the unidirectional LSTM had 94,606 trainable parameters [7]. The neural network was trained to classify discrete outputs by affixing a softmax layer and also trained with the cross entropy function for classification. For the training parameters, each frame for each video were analyzed in isolation in accordance to the overall expressions in the video. After the completion of the training, the accuracy of the network was evaluated by adding up the frame classifications, after that normalizing to conclude a probabilistic classification for all the sequences.

Error Rate:

Classifier	Mean Error Rate
Bidirectional LSTM	13.2 +/- .8%
Unidirectional LSTM	14.6 +/- .7%

The Bidirectional Long Short-Term Memory Network gave significantly good results, whereas Uni-directional network could be used for real-time recognition. In the scenarios, the training parameters were easily understood by the classifier and eventually all the sequences were correctly classified.

V. FUTURE WORK

The paper highlighted the facial expression estimation with the combination of model-based image interpretation and sequence labelling. The dataset provides various type of algorithms based on historical data, which can predict the data for the new users. The predictive models induced by the algorithm gives us robust classifiers. Further, the multi classifiers are a better option, as they result after combining the various individual classifiers. The significance of the multi classifier may be further improved by including the applicability on real time basis.

VI. CONCLUSION

The outcome of the paper envisages that the characterization of facial expression can be benefitted massively from features learnt using Deep Neural Networks. In regard to the performance, using even a comparatively simple framework around these features results in performance much better than the current state of the algorithms. This study gave us proper understanding on Recurrent Neural Networks and to select right algorithm for the given machine learning problem. It also enhanced knowledge in distinguishing the Feed-Forward Neural Network and a RNN.

REFERENCES

[1] Liu, M., Wang, R., Li, S., Shan, S., Huang Z. and Chen, X.2014. Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild. *ACM ICMI.*

[2] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)* .4489-4497. IEEE.

[3] http://www.eecs.qmul.ac.uk/~sgg/papers/ShanGongMcOwan_IVC09.pdf

[4] Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., and Pal, C. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 467- 474. ACM.

[5] Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç, Memisevic, R. and Mirza, M. 2013. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 543-550. ACM.

[6] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In ACM MM.

[7] Sak, H., Senior, A. W. and Beaufays, F. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH*. 338-342.

[8] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[9] https://towardsdatascience.com/recurrent-neural-networks-and-lstm-4b601dd822a5

[10] Sepp Hochreiter and J¨urgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[11] https://shafeentejani.github.io/2016-12-20/convolutional-neural-nets/

[12] https://www.quora.com/What-is-max-pooling-in-convolutional-neural-networks