International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

Sentimental Analysis of Facial Expression Recognition

Avinash Kumar Jaiswal[1], Ravi Sharma[2]

[1]Computer Science Engineering, Galgotias University

[2]Assistant Professor,Galgotias University

Abstract: - The sentimental analysis is a process of extracting human feelings from data. Which can be seen in the processes such as natural language processing, computational linguistics, text analysis .Its basic idea is to classify human emotions in different moods such as happy, sad, neutral, etc. It uses the pattern of movement of both the lips and the eye shape that it takes during different feelings of humans. It has a huge variety of applications because of its ability to extract insights from data sets and social media. It will be using machine learning algorithms for the detection of emotions and it will be trained on the huge dataset with varying sample size. It will also use facial recognition to perform a specific analysis of a person after identifying their face. This will provide a separate report for each person on their emotional state. This report then will be used for the expression mining in different modern systems such as online streaming, video interviews, etc. This will empower the existing tools to perform multi tasks and gives a lot of data to work with.

Keywords: Sentimental Analysis, Facial Recognition, Machine Learning, Image processing, Artificial Intelligence

Introduction: - We see a lot of images circulate in the world, often on social media and in newspapers as well. We humans can recognize the photographs without their extensive captions, but on the other hand computers require photographs first to obtain sufficient knowledge then only the facial expression of human beings could be analyzed.

Facial expression Recognition has a variety of use cases such as improving people's facial expression by real-time environmental effects by camera capture, increasing social medical awareness by translating facial expression to social feed images, as well as speech signals. Assisting young learners in object identification and Understanding the vocabulary.

Face recognition is already on some of our everyday appliances, such as TV, and mobile devices, where access passwords are being replaced with face- prints, biomedicine, commerce, web searching and military etc. Social networking such as Instagram, Facebook etc. will automatically create captions from the images. The key purpose of this survey paper is to acquire a little knowledge of the strategies of deep learning. For the analysis of photographs, we use two techniques primarily CNN and LSTM.

Basic Terminology:-

1. Face Detection: - Face detection is used to determine whether a given image includes a face. We should be in a position to describe the general face structure. After that we can detect the emotions of humans. Fortunately human faces are no different from each other; we all have nose, eyes, forehead, chin, and mouth; these are the normal structure of the face.

2. Face Identification: - The program compares the given person to all other individuals in the database and, thus, offers a ranked match list.

3. Face Verification: - In this the program compares the given person to who the entity says they are and gives a yes or no judgment.

4. Facial Expressions: - Facial expression is the location of the muscles Beneath the facial tissue. Those gestures remind the observer of the mental condition of the person.

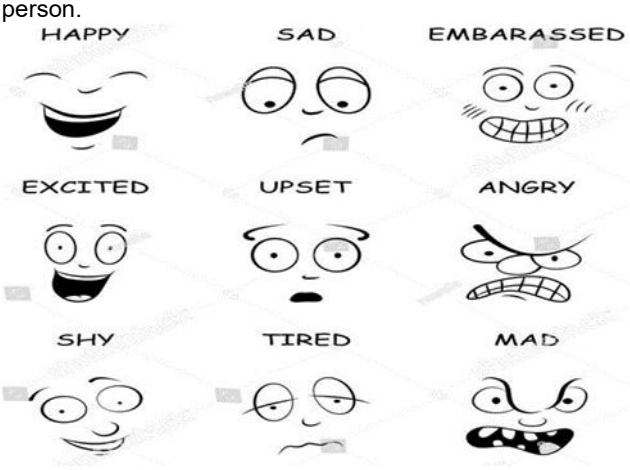


Image Captioning Techniques CNN:-

Convolutional neural systems are specific important neural systems that can produce information that has an information shape, for example, a 2D lattice. Pictures are effortlessly suggested as a 2D lattice and CNN is valuable for working with pictures. It examines pictures from left corner to the right corner and through to extricate significant highlights from the picture, and consolidates the element to characterize pictures. It can deal with interpreted, pivoted, scaled, and modified pictures. The Convolutional neural system is a profound learning calculation that takes in the info picture,

allocates significance to various components/protests in the picture, and recognizes it from each other. The required pre-handling in ConvNet negligible when compared with other order calculations. In spite of the fact that channels are hand-designed in crude strategies, with sufficient preparation, ConNets is fit for learning these channels/highlights. The curved device configuration is like the human mind's neural network architecture, which is inspired by the visual cortex connection. Singular neurons respond to changes in only a confined region of the visual field called the open field. The assortment of such fields covers the whole visual region.

CNN Architecture: - In processing large images and recordings, An architecture where all neurons in one layer makes contact with all its neighboring neurons is governed as the pure vanilla neural network used in analyzing large sector and videos. If we consider a normal picture with pixels and RGB shading, then there will be millions of parameters using a typical neural network that can cause overfitting. To restrict parameters and highlight the neural system to significant pieces of the picture, a CNN uses 3d models arrangement for breaking down into minute element. Inspite making way for the neurons to pass through the channel network. A CNN uses a pictures elements such as the noes ear mouth and hair.

Working of CNN:-

As we have discussed previously, a fully connected neural network where the input in the preceding layers is connected to every input in the following layer is convenient for the task at hand, along such lines, according to CNN, the neurons in a cell may be connected with a specific cell area before it, rather than all the neurons in a totally similar way. This helps in reducing the complexity of the neural network and acquire less computing power. As per new computer under standard image with the use of numbers at each pixels. When we generally compare two images we check the pixel values of each pixel. This technique only helps us to compare two identical images only but when we keep different images to compare the comparison fails. In CNN image comparison takes place piece by piece. The pieces are called feathers or fritters. By finding rough features that match, in the same position in both pictures CNN gets a best at seeing the likeness that the whole picture matches the schema. CNN has four different layers:

- Convolution Layer
- Rectified Linear Unit Layer
- Pooling Layer
- Fully Connected Layer

Convolution layer: in this layer feature/filter is moved to every possible projection on the image. It consists of following steps.

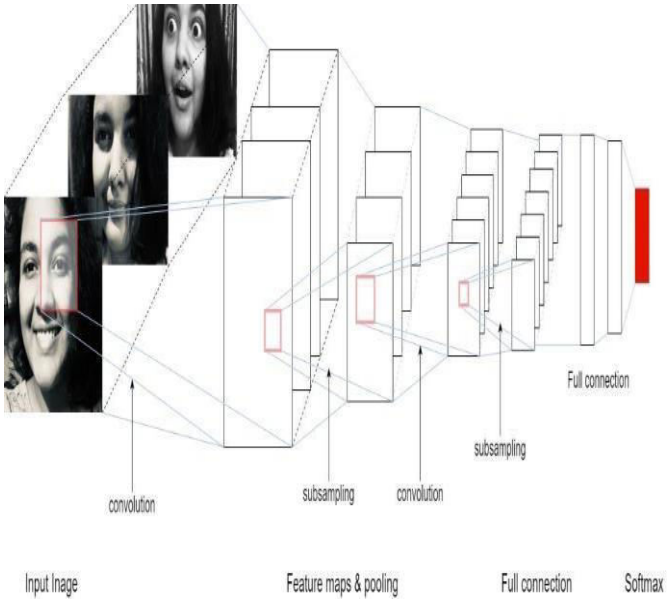
1. Line up the element and the picture
2. Increase each picture pixel by the relating highlight pixel.
3. Include them up
4. Gap by all out number of pixels in the element
5. (to monitor where that component was)we make a guide and put the estimations of the channel at that place.
6. Repeat all above steps for all the features.

ReLU Layer:(Rectified Linear Unit) A node where the function is activated when the input reaches certain value, whereas when input reaches zero output gives 0 and when the input is above threshold, it forms linear relation to vector. This form of transition function is known as Rectified Linear Unit layer. It simply eliminates all negative values and replaces them with null. This is done with all of the data generated by the Convolution sheet.

Pooling Layer: In this layer the picture stack is compressed into minute scale by using following step:

1. We take window size as 2 or 3 during selection
2. Hold a stride
3. The window is moved through the filtered photos.
4. The greatest value is achieved from each window.
5. Repeat all the steps for each output of Rectified Linear Unit layer.

Fully Connected Layer: Few times the output of polling layer is been used to reduce the size of the result. Now the value of the pixels of filtered and shirked are put into a single list/vector. This list will show a pattern of high value in similar sports for matching images after the training process. The process of prediction is done by comparing the list or vector of input image with the trained vectors or lists. The comparison is done by dividing the sum of all values in the same index as the high value of tested data by the sum of the high values of the trained data with all the tested data. The input image is predicted to be similar to the image with highest value after the division.



DATASET:

Exp	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total
Label	0	1	2	3	4	5	6	-
Count	4953	547	5121	8989	6077	4002	6198	35887

To assess the adequacy of our strategy, we utilize the recently discharged [23] Facial Expression Recognition-13 (FER-13) dataset. The dataset was made utilizing Google image search API with feeling related catchphrases. It comprises forty eight*forty eight pixel grayscale snapshots. Brief the dataset has seemed in the Table above. One component to peer is that the human exactness in this dataset is 65±5%

We used an 80% of dataset as a training set and a 20% for validation set. To verify our version, we use three unmistakable execution estimations: 1) the objective paintings; 2)the Top1 precision of both getting prepared and check set; 3) the Top2 Accuracy of both making plans and taking a look at the set.

The consequence of every statement of the entire model appears in Table below.it has been seen with a high (80±%) exactness as "Happiness" and "Shock". These two articulations are additionally the most effortless two for a human to perceive. [23]The general precision of the model on the FER-13 dataset is 65.05%, which are viewed as near human level 65±5%

Expression	precision	recall	F-Score	Accuracy
Angery	0.6017	0.7172	0.6544	62.12 %
Disgust	0.7175	0.5938	0.6554	68.45 %
Fear	0.5959	0.6635	0.6279	58.59 %
Happy	0.8885	0.6784	0.7342	78.86 %
Sad	0.6127	0.4760	0.6365	59.30 %
Surprise	0.9285	0.6868	0.7726	81.50 %
Netural	0.8293	0.8293	0.7096	65.90 %

Predictive Sentimental Analysis:

Predictive analysis is an analysis in which we have some independent and dependent variables in which first we train the model by the help of statistical techniques, data-mining techniques, machine learning and predictive modelling to analyze the current and past historical data like independent variables and predict the targets for the future. There are some benefit to use predictive analysis.1.we can predicts the future values for the target from the past data. 2. It is a less expensive method among all models. For doing predictive analysis we need past data, data modelling techniques, statistics, machine learning, math's, and data mining techniques etc.

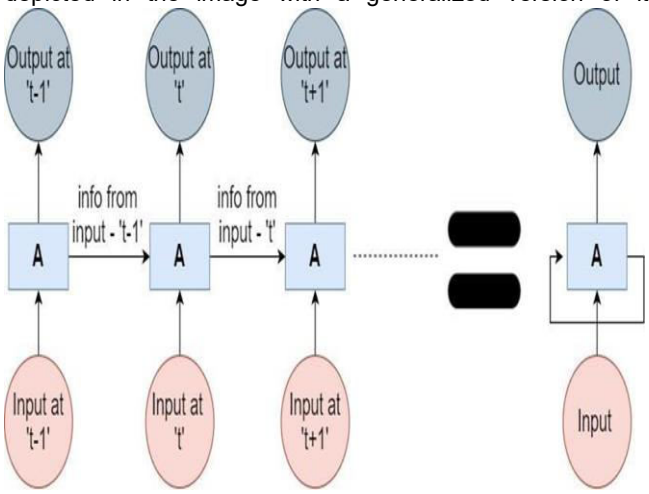
CNN are mostly used in Face recognition, object recognition, computer vision, deep learning, artificial intelligence etc. Using CNN we can get accurate results for face recognition, computer vision etc. CNN has some limitations to use it.1. CNN doesn't encode the position and direction of the item into their forecasts. They totally lose all their inside information about the posture and the direction of the item and they course all the data to similar neurons that will be unable to manage this sort of data. 2. CNN can be used for deep learning techniques for the analysis of image data whereas predictive analytics are used to make predictions on continuous data only. The CNN based model discussed above helps us to identify the extracting human feelings from data with the prediction of any kind of phenomena or data is based on past data and the accumulation of current

scenarios. The success rate of 65.03% (). But cnn does not store the output. In the case of predictive analysis algorithms not only need to store the current output but also perform calculation on the stored data .

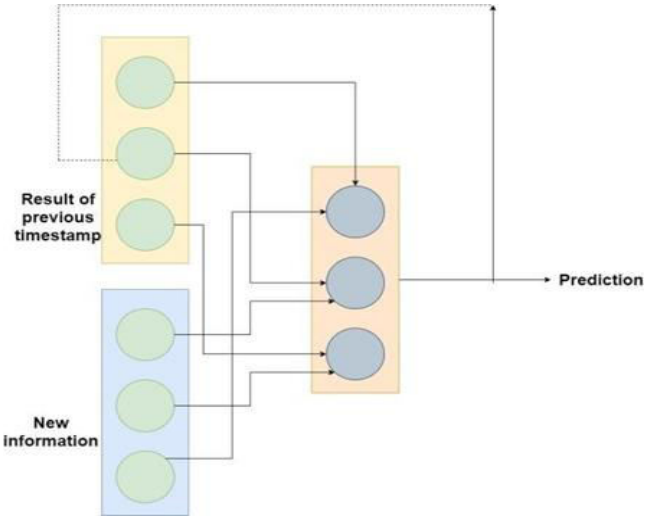
The previous conversation and model work on a feedforward neural system which when presented to any irregular assortment of photos, and the main picture which calculation is presented to won't really change how it arranges the subsequent picture. The output received by the algorithm at time t will not affect the output at time t+1.

Therefore we can conclude the output is independent of each other. But there are certain scenarios where the previous data or result is required for calculating the new output, for example, partial fingerprint or image or for predictive analysis as mentioned above. To overcome the challenges presented by the FNN or CNN we will use the Recurrent Neural Network.

First instance, at time stamp 't-1' we pass our input to our network 'A' and we get the output at 't-1'. And then in the next timestamp which is 't,' a different input will be given to the network along with the information from the previous timestamp i.e. 't-1' which will help the network to get an output at timestamp 't'. Similarly, for output at timestamp 't+1,' the network has two inputs one is the new input provided to the network and another is the information coming from the timestamp't' and it will go on. The same is depicted in the image with a generalized version of it.



The algorithm is provided with the inputs in the form of vectors (list of data). The image description of the process is given below.



The mathematical representation of RNN will be:-

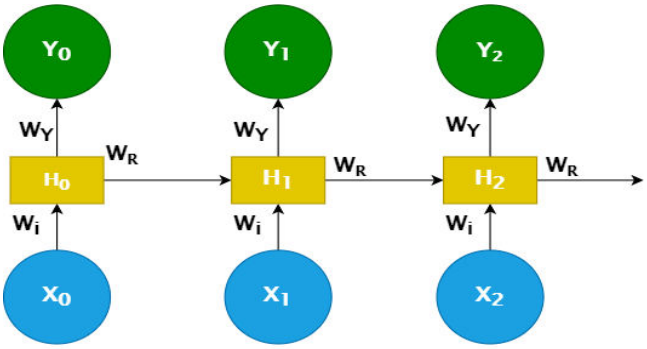
$Ht = gh(wixt + wrh(t-1) + bn)$

$Yt = gy(wyht + by)$

And the graphical representation of the

above equation is shown below.

The equation WRH(t-1) does not apply when time = t, as time cannot be negative.



W_i = Weight vector for hidden layer
W_y = Weight vector for the Output layer
W_R = Same weight vector for the Output layer
$X^{(t)}$ = Image vector for Input image
$Y^{(t)}$ = Image vector for Output image
g_h or g_y = Activation function
b_h or b_y = Bias

Training a recurrent Neural Network: RNN utilizes a backpropagation calculation, yet it applied for each timestamp, generally governed as Backpropagation Through time, But Backpropagation has two major problems:

- Vanishing Gradients
- Exploding Gradients

While using backpropagation we calculate errors. Errors (e) can be calculated by using formula e=(Actual Output - Model Output)²

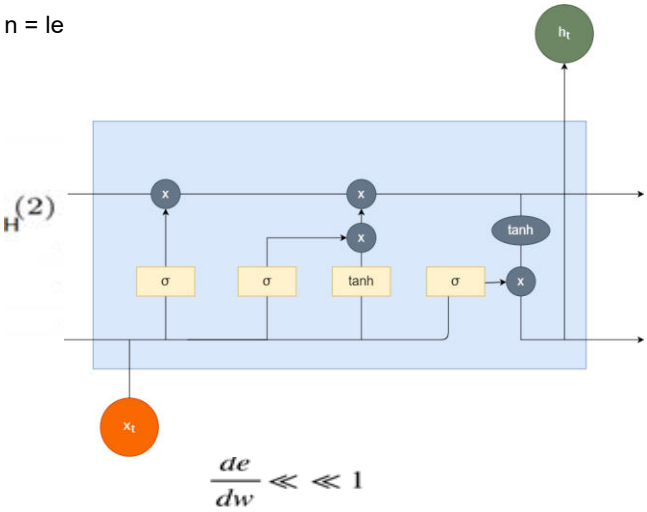
This error participate in the calculation of new weight and reducing the error in the following way:

wη = wσ + Δw

wη = New Weight wσ = Old Weight

Δw = Difference in weight

n = le



$$\frac{de}{dw} \ll \ll 1$$

In many cases which leads to a change in weight to be very small hence making old and new weight almost similar. This effect when our input data is long because then the algorithm needs to store the data from the start or known as long term dependencies. Due to that, the algorithm will become very complex with a lot of iteration.in such conditions there will be no visible change happening in new weight. Which is known as vanishing gradients.

Similarly when the change in weight will be exponential, which will be leading to an unstable neural network. This is known as Exploding Gradients.

Following ways will help in solving exploding and vanishing gradients:

For exploding gradients:

- Truncated BTT- Instead of starting backpropagation at the last timestamp, choosing a smaller timestamp will be appropriate.
- RClipping inclinations at a limit By cutting the angle when it goes higher than an threshold
- RMSprop to adjust the learning rate

For vanishing gradients:

- ReLU activation function- the use of activation function like ReLU, which give yield as one while ascertaining the slope will help
- RMSprop- clipping the gradient when it goes higher than a threshold
- LSMT, GRUs- It's a different network architecture that has been specially designed to be used to combat this problem. We will discuss it in the extent below.

Long short term memory:-

It's an exceptional sort of RNN, which is equipped for learning long haul conditions, for example at the point when the hole between the significant data and where it is should have been extremely huge LSTM has a chain-like structure similar to RNN.

All RNN has a chain of a reaped module of the neural network. The repeating modules would have a basic structure in a regular RNN, with a single tanh function that function act as a squashing function. Squashing function in standard RNN is responsible for converting the input values from -1 to 1. Below, there is a structure of a singular module of RNN with LSTM.

Unlike standard RNN where there is only one neural network, there are four in LSTM interacting in a very special way. The horizontal line at the top of the module

The diagram is called a Cell state. It can be considered as a conveyor belt going across the entire chain with just a few small linear interactions

The step by step working of LSTM

1. The initial phase in the LSTM is to distinguish that data which isn't required and will be discarded from the cell state. This choice is rendered from a sigmoid layer called an overlook entrance layer.

It takes a gander at $ht{-}1$ which is yield from the past timestamp and xt which is the new information. The yield of the layer is ft which is somewhere in the range of zero and one, where zero recommends getting totally freed of the information and one proposes to totally save the information for each number in the cell state $ct{-}1$ .The function use here:

$ft = \sigma(wf[ht{-}1, xt] + bf)$

Where:

wf = Weight

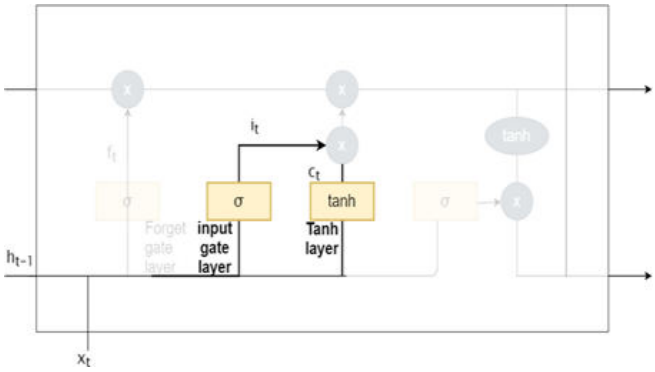
$ht{-}1$=Output from previous timestamp

xt= New input

bf = bias

ft will later combine to cell state

2. The next layer is to select which new data to store in the cell state. It comprises two sections. The initial segment is a sigmoid layer called the "input door layer" which is liable for choosing the values which are subject to Update.



The second part is the tanh layer which creates a vector Of new candidate values of tct~ that could be added to the cell state.

it = $\sigma(wi[ht-1, xt]$ + $bi)$

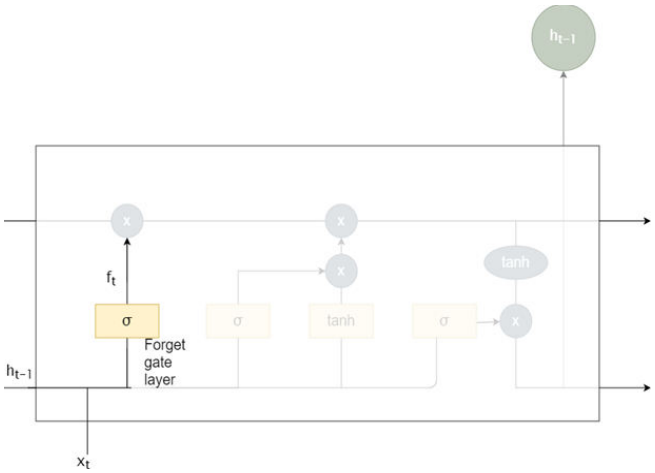
ct~ = tan$h(wc$ $[ht-1, xt]$

+ $bc)$

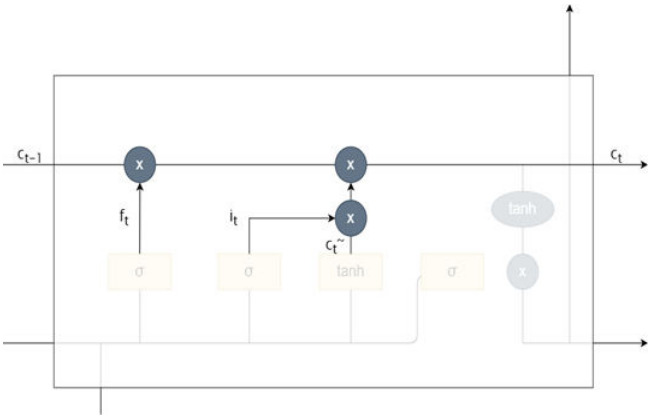
wi & wc= weight matrix

bi & bc = bias

The contribution from the past timestamp (ht-1) and the new info (xt) will go through the σ work which will compute it which will duplicate to ct~. The ct~ is determined by passing the contribution from the past timestamp (ht-1) and the new info (xt) will go through the tanh work. Furthermore, the outcome is later added to the cell state.



3. In this progression, the calculation will refresh the cell state from $ct-1$ into the new cell state ct. Initially, the old state $(ct-1$) is duplicated with $ft=$ which infers forgetting about the information which was chosen isn't helpful in the overlook entryway layer. At that point the outcomes add to the increase of it and ct~= which is the up-and- comer esteem, scaled by how much the calculation chooses to refresh each state esteem. The condition utilized:

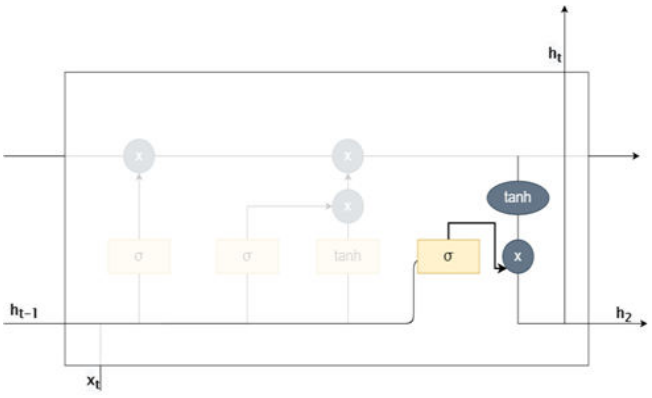


4. In this progression, the calculation runs a sigmoid layer which chooses what part of the cell state will be yielded. At that point we put the cell state through tanh work which pushes the qualities to be between less one and one. At that point it is increased by the yield of the sigmoid gate(ot). This progression just lets the yield to pass which was determined before. The conditions utilized here are:

ot = $\sigma(w0[ht-1, xt]$

+ $b0)$

ht = ot × tan$h(ct)$



The dataset and experimental setup:

To accomplish the errand of the programmed assessment of facial loss of motion, we gather the recordings of the facial analysis activities, which are made by the facial loss of motion patients, and every one of these recordings are utilized as the exploratory information by the patients' authorization. The recordings of 103 facial loss of motion patients are gotten, and 40 ordinary volunteers are included

for video assortment. For every individual, they have to make seven facial activities including raise eyebrow, close eyes, team up nose, stout cheeks, open mouth, grin and grimace, and the succession of these seven facial activities have to be rehash multiple times by every individual. During the time spent in the video assortment, one video section is utilized to record one kind of facial activity for every individual. The seriousness of facial loss of motion is separated into four levels including ordinary, somewhat sick, moderate sick or fundamentally sick, which can be meant with 0,1, 2 and 3 as their marks individually, and three expert specialists in our agreeable clinic help us to evaluate the seriousness of facial loss of motion. On the off chance that the assessment results given by specialists are predictable, at that point the outcomes are utilized as the ground truth. Be that as it may, in the event that their assessment results are conflicting, at that point the conclusive outcomes given by a specialist with broad analytic experience would be the ground truth. Consequently, we separated all these 3003 video tests into four gatherings, which are compared to the four degrees of the seriousness of facial loss of motion. 70% of recordings in each gathering are utilized to prepare the model, and these recordings are dispensed on 5:1 as the preparation set and the approval set. The stayed 30% of recordings in each gathering are utilized as the test information. All the exploratory outcomes are given by utilizing 5-overlay cross-approval. For LSTM, the underlying learning rate is 0.0001, and the cycles are 20 0. In addition, the casing grouping with an arrangement of 30 edges chosen from a video is utilized as one preparing or test, and furthermore as an essential handling unit. Be that as it may, some conventional techniques need to utilize the static facial pictures to assess the seriousness of facial loss of motion, so we select the keyframes from all the video portions as their exploratory information, and these keyframes mirror the facial conditions of having the best scope of facial activities. Besides, exactness, accuracy, review, and F1 score are utilized as the four parameters for assessing the exhibitions of the exploratory techniques. These four assessment parameters are determined by the accompanying equations:

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

$$Precision = \frac{TP}{TP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Where:

TP = an example which is really +ve and is anticipated as 1

FN = example which is really +ve but predicted as a -1

FP = example which is really -ve yet is anticipated as 1

TN = example which is really -ve and is anticipated as -1

Result:

To confirm the viability and predominance of LSTM for the assessment of facial loss of motion, a few existing strategies are utilized as the correlation techniques, what's more, the preliminary outcomes are showed up in Table beneath. There are four sorts of appraisal methods for facial loss of movement, including the customary systems using counterfeit highlights, CNN-based strategies, outward appearance, and smaller scale demeanor acknowledgment techniques and LSTM-based strategies. Moreover, a few CNN-based strategies are likewise applied to our assignment, and tried on the static facial pictures. Contrasted and conventional strategies, the techniques dependent on GoogleNet and VGG-16 have comparative exhibitions as far as exactness and accuracyanyway have better presentations in Recall and F1 score. While the methods reliant on Resnets have remarkable improvements in the four evaluation parameters, and Resnet 50 has a prevalent display than the past ordinary methodologies with a precision of 66.67% and a F1 score of 67.90%.These exploratory results show the force of significant features on the appraisal of facial loss of movement.

Network	Accura cy	Precisi on	Recall	F1
Gabor+SVM	0.6087	0.6731	0.5446	0.5801
LPB+SVM	0.6488	0.5673	0.5423	0.5018
GoogleNet	0.5675	0.5958	0.6014	0.5936
VGG-16	0.6389	0.6580	0.6292	0.6346
Resnet34	0.6286	0.6726	0.6786	0.6419
Resnet50	0.6667	0.6862	0.6814	0.6790
Resnet101	0.6571	0.7279	0.6306	0.6544
CNN-FER	0.3889	0.5786	0.3888	0.4590
MicroExpST CNN	0.5102	0.5307	0.5102	0.5165
CNN-LSTM	0.6364	0.6705	0.6417	0.6475
LSTM	0.7347	0.7396	0.7242	0.7219

Contrasted and fake highlights, CNN- based techniques can catch progressively summed up facial highlights for facial loss of motion assessment. Likewise, Resorts have more convolution and waiting layers with the effect of the troupe, which enables Resnets to keep up logically fruitful features.

In addition, we apply outward appearance and small scale articulation acknowledgment strategies to our assignment. As appeared in Table above. We can see that the demeanor and small scale appearance acknowledgment techniques can apply to facial loss of motion assessment, and MicroExp-ST CNN [21] accomplishes preferred execution over CNN-FER [21] as far as exactness, review, and F1 score individually. In any case, these techniques can't accomplish great execution for facial loss of motion assessment because they primarily center around the general facial movement instead of the asymmetry highlights of facial developments. Interestingly, our strategy LSTM considers the asymmetry highlights of worldwide and nearby facial developments and has 0.2244, 0.2090, 0.215, and 0.0.2053 upgrades over MicroExpSTCNN [21] to the extent exactness, precession, audit and F1 score separately. Finally, the methods reliant on a single stream of LSTM, CNN- LSTM, and our LSTM are performed on the accounts of the patients' facial exercises. LSTM-based strategies have general common displays. We acknowledge that is in light of the fact that these techniques use the dynamic features isolated from the facial muscle advancements. Especially for LSTM, it has 12.58% higher appraisal exactness and 14.20% greater estimation of F1 score than Wang's strategy. Additionally, it moreover has a 6.80% higher exactness and 4.28% greater estimation of F1 score than Resnet50. Among the LSTM-based strategies, LSTM likewise has the best execution. That is because of the consecutive separated highlights extricated by LSTM that can all the more likely mirror the adjustments in the distinctions in facial developments. Besides, LSTM intertwined the worldwide and neighborhood separated highlights for facial loss of motion assessment as opposed to just utilizing the facial movement highlights extricated from the entire countenances. These asymmetry highlights of facial loss of motion assessment are essentially steady with that of the specialists to assess the facial loss of motion.

Conclusion:

My research is not only the research that has been done in this field using this similar algorithms to find how people are feelings by their facial expressions but in every paper research or experiment which I came through did not succeeded in getting in accurate result which will be more than 80% and that too only in some limited cases and this result was obtained after rest training of our system through a large data set which mean this approach is not enough. Various research in human psychology and the way we humans express ourselves have been done. There was much research supporting that at a given moment of time people don't only convey their feelings with their facial expressions but through the tone of their voice too. Therefore for the future work I suggest a separate module that keep track of the voice modulation of the subject like speed and peach in which they are speaking so that it could be collaborated with the psychological research and help detect direct current emotional status which can be combined with the results which we get by our facial expression recognition module and get more accurate result.

References:-

[1] Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In Proceedings of the ThirdWorkshop on Statistical Machine Translation. Association for Computational Linguistics, 115–118.

[2] Ahmet Aker and Robert Gaizauskas 2010. Generating image descriptions using dependency relational patterns. In Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 1250–1258.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998 (2017).

[5] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning. In Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition. 5561–5570.

[6] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, et al. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In the International Conference on Learning Representations (ICLR).

[8] Shuang Bai and Shan An. 2018. A Survey on Automatic Image Caption Generation. Neurocomputing. ACM Computing Surveys, Vol. 0, No. 0, Article 0. Acceptance Date: October 2018. 0:30 Hossain et al.

[9] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Vol. 29. 65–72.

[10] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015.Scheduled sampling for sequence prediction with recurrent neural networks. In Advances in Neural Information Processing Systems. 1171– 1179.

[11] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. Journal of machine learning research 3, Feb, 1137–1155.

[12] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. Computational linguistics 22, 1 (1996), 39–71.

[13] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, Barbara Plank, et al. 2016. Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. Journal of Artificial Intelligence Research David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research 3, Jan (2003), 993–1022.

[14] Cristian Bodnar. 2018. Text to Image Synthesis Using Generative Adversarial Networks. arXiv preprint arXiv:1805.00676.

[15] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data. AcM, 1247– 1250.

[16] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory. ACM, 144–152.

[17] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Troster. 2011. Eye movement analysis for activity recognition using electrooculography. IEEE transactions on pattern analysis and machine intelligence 33, 4 (2011), 741–753.

[18] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In Computer Vision and Pattern Recognition (CVPR), 2011

[20] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, WillCukierski, Yichuan Tang, David Thaler, and Dong HyunLee. Challenges in representation learning: A report on three machine learning contests.Neural Networks,64:117–124, 2014.

[21] Automatic evaluation of facial nerve paralysis by dual-path LSTM with deep differentiated network by (Pengfei Xu, Fei Xie, Tongsheng Su, Zhaoxin Wan, Zhaoyong Zhou, Xiaoyu Xin, Ziyu Guan)

[22] Recognizing Emotion from Speech Based on Age and Gender Using Hierarchical Models by (Ftoon Abu Shaqra, Rehab Duwairi, Mahmoud Al- Ayyoub).

[23] Kuang Liu, Mingmin Zhang, Zhigeng Pan. "Facial Expression Recognition with CNN Ensemble", 2016 International Conference on Cyberworlds (CW), 2016