Modified Convolutional Neural Network Architecture Analysis for Facial Emotion Recognition

Abhishek Verma,[1] Piyush Singh, [2] John Sahaya Rani Alex [3]
[1, 2, 3] School of Electronics Engineering, Vellore Institute of Technology, Chennai, India 600127
svermaan@gmail.com

***Abstract* - Facial expressions are one of the key features of a human being and it can be used to speculate the emotional state at a particular moment. This paper employs the Convolutional Neural Network and Deep Neural Network to develop a facial emotion recognition model that categorizes a facial expression into seven different emotions categorized as Afraid, Angry, Disgusted, Happy, Neutral, Sad and Surprised. This paper compares the performance of two existing deep neural network architectures with our proposed architecture, namely the Venturi Architecture in terms of training accuracy, training loss, testing accuracy and testing loss. This paper uses the Karolinska Directed Emotional Faces dataset which is a set of 4900 pictures of human facial expressions. Two layers of feature maps were used to convolute the features from the images, and then it was passed on to the deep neural network with up to 6 hidden layers. The proposed Venturi architecture shows significant accuracy improvement compared to the modified triangular architecture and the rectangular architecture.**

***Keywords* - accuracy, convolutional neural network, deep neural network, facial emotion recognition**

I. INTRODUCTION

Facial emotions play an important role in communication among humans and help us to understand the intentions of others and how they feel. Humans have a strong tendency to express emotions. They play an essential role in our daily lives. Human spend great amount of time in understanding the emotions of others, decoding what these signals mean and then determine how to respond and deal with them. Facial Emotion Recognition [1] is getting into our lifestyle and impacting us more rapidly than we have predicted a few years back. Apple released a new feature on iPhone X called Animoji [2] where the user can get a computer simulated emoji to mimic facial expressions. It is now hard for us to ignore the potential capabilities of such features. Facial Emotion Recognition has a wide range of applications. It can be applied in smart cars where it can detect the emotions of the driver and alerts him if he feels sleepy or drowsy [3]. Facial Emotion Recognition(FER) can be helpful in detecting whether the experience of the gamer was enjoyable by analysing his facial expressions. It can be employed in emotion detection of old age people in old age homes and to monitor the level of stress and anxiety in day to day life. It can help people recognize the expressions of people suffering from autism [4] or speech-impaired people. Moreover, investigation agencies can apply Facial Emotion Recognition (FER) to pre-determine their actions before they are carry out interrogation. This paper proposes a new architecture in the convolutional neural network framework and compares it with different architecture on parameters like the training accuracy of the network, testing accuracy of the model, training loss, testing or validation loss etc.

II. RELATED WORK

One of the early works on facial recognition [5] uses the Nearest Feature Line (NFL) to find out the two feature points on a person's face through which a particular feature line passes but NFL only gave slightly insignificant improvement in the error rate than Convolutional Neural Network (CNN). The Facial Emotion Recognition system in [6] uses the auto-encoders to provide uniqueness to different emotion but the image pixels were vertically fed into the auto-encoders and the structural integrity of image was lost. The emotion recognizer in [7] uses the Facial Animation Parameters (FAP) to create a robust analysis system and then develops a neurofuzzy based on the rules defined by the analysis of the FAP. Gabor filter based feature extraction and the learning vector quantization was used in [8] for facial recognition but in order to improve the accuracy more than 40 images of erratic expressers were removed from the dataset.

The speech recognition system in [9] introduces different architectures of Deep Neural Networks (DNN) and trains the neural network on different Mel-frequency cepstral coefficients (MFCC) values of different speakers. This paper proposes a new Venturi architecture of the CNN for the images of facial emotion and analyse its performance with the Modified Triangular architecture which was used in [9] to classify audio files but in this paper this architecture is being used on image data set. Three different architectures are being analysed on different parameters including training accuracy, testing accuracy, training loss and testing loss when trained on the same dataset of Karolinska Directed Emotional Faces (KDEF) [10] for facial emotion recognition.

III. CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Network (CNN) is a type of Artificial Neural Network that is specifically designed to process the image pixels and extract useful features from the provided images. CNNs are being used in fields like image and video recognition [11], natural language processing (NLP) [12] and artificial intelligence [13]. CNN consists of different layers that include Image layer, Convolution layer, Pooling layer, Flatten layer, Fully connected layer (Input layer & Hidden layers) and Output layer.

978-1-7281-3227-3/19/$31.00 ©2019 IEEE

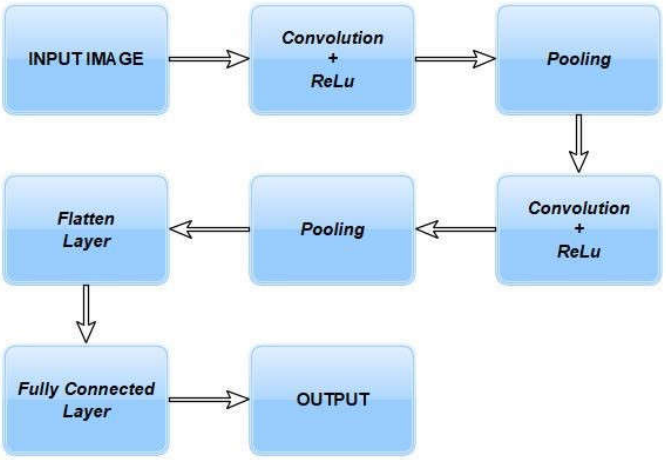

Figure 1. Overview of our CNN

The input layer is usually an image full of pixels and a feature map is created which is then slid over these pixels that result in a convolution layer. In order to reduce the number of features and to form more correlation between the adjacent pixels we perform the pooling step.

This paper uses max pooling method for down sampling and to extract the most important features from images like edges. This paper uses two convolution layers with two max pooling methods after each convolution layer. In the convolution 2D layer the provided input images is scaled down to 256*256 pixels from 562*762 of the original image. Activation function used in each convolution layer is Rectified Linear Unit or ReLu. In the first convolution layer 32 feature maps or filters were used along with 3x3 feature detector matrix. In the second convolution layer the number of feature map or filter was increased to 64 with the same size of 3x3 feature detector matrix. After the first convolution layer the max pooling layer uses a 4x4 feature extraction matrix whereas in the max pooling layer after the second convolution layer the size of feature extraction matrix was reduced to 2x2. The resultant matrices after the 2 convolution layer and 2 max pooling layer is broken down in to a single layer or in to a single column matrix containing all the pixel values from these matrices in one single column also known as the flattening layer. This flattening layer is then used to feed the input layer of the next artificial neural network. The artificial neural network with large number of hidden layers results in a deep neural network [14].

This paper compares and analyze different architecture of the Convolutional Neural Network (CNN). The CNN is then compiled using the stochastic gradient descent algorithm [15] also known as the Adam optimizer [16]. The loss type is the categorical cross-entropy with the accuracy performance metrics.

IV. ARCHITECTURE

In this section, the detailed design of three different architectures, namely Rectangular Architecture, Modified Triangular Architecture [9] and Venturi Architecture are discussed.

A. Rectangular Architecture

The rectangular architecture that is used for the designing of the hidden layer of the deep neural network in the convolutional neural network consists of 6 hidden layers. The architecture is named as rectangular architecture because of its shape that consist of 6 layers of equal number of nodes in each layer which gives it the shape of a rectangle as shown in figure 2 . In this paper the number of hidden layer that is being used for the rectangular architecture is 6 and each layer has a constant number of 256 nodes. The number of output nodes is 7 based on the 7 different emotions by which the results are classified. The activation function used for each hidden layer is ReLu activation function except for the output layer which has the Softmax activation function [17].

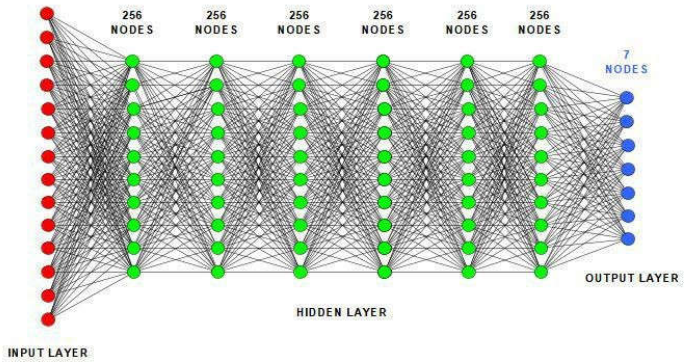

Figure 2. Rectangular Architecture

B. Modified Triangular Architecture

The modified triangular architecture [9] uses a 7 hidden layer architecture in the deep neural network of the convolutional neural network.The modified triangular architecture has 256 nodes in the first hidden layer, 512 nodes in the second hidden layer and from the 3rd layer till the 7th layer the number of nodes decrease in such way that it takes the shape of a triangle and in total it looks like a modified triangle.

The numbers of nodes in the hidden layers are 256, 512, 256, 128, 64, 32 and 16. It also consists of a 7 node output layer based on the 7 different categories in which the 7 different emotions are classified. The activation function used for all the 7 hidden layers is the 'ReLu' activation function. The output layer uses a Softmax activation function [17] as it makes easy to model the probability distributions.

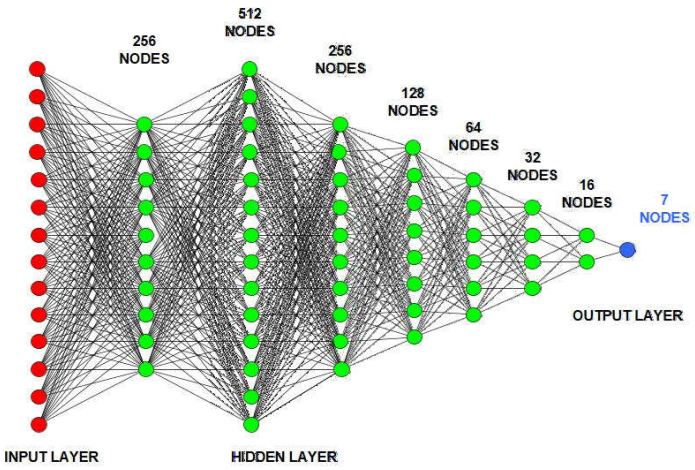

Figure 3. Modified Triangular Architecture

C. *Venturi Architecture*

Venturi Architecture is our proposed architecture for the hidden layer of the deep neural network in the convolutional neural network. The architecture consist of 6 layers in the hidden layer with one output layer consisting of 7 nodes based on the 7 different categories in which the facial emotions are classified. Venturi architecture gets its name from the shape of its hidden layers that looks like a Venturi Tube in figure 4.

After the input layer the first hidden layer consist of 256 nodes which then decreases to 192 nodes in the second layer then for the third and fourth layer the number of nodes remains constant to 128 and then for the fifth and sixth layer the number of nodes are the mirror of the second and first layer i.e. 192 nodes in the fifth layer and 256 nodes in the final and sixth layer. In the overall structure of the hidden layer the number of nodes first decreases then remains constant and then increases till the output layer. The Rectified Linear Unit activation function is being used for all the 6 hidden layers. The output layer uses a Softmax activation function in order to get an accurate model of the probability distribution.

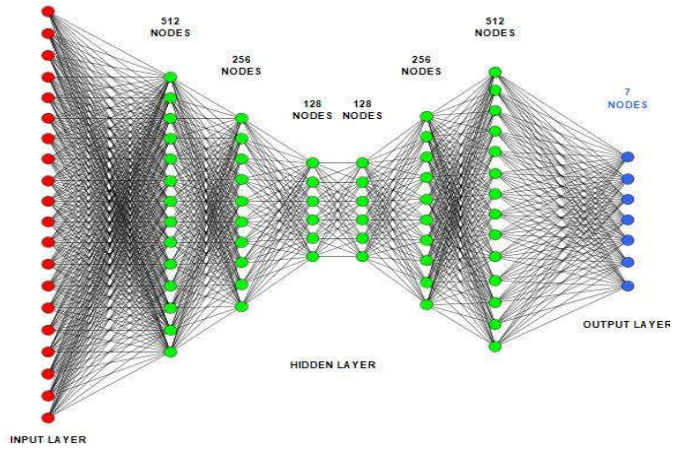

Figure. 4. Venturi Architecture

V. EXPERIMENTAL SETUP

The dataset used to train and test the different architecture of hidden layer in the deep neural network of the convolutional neural network is the Karolinska Directed Emotional Faces (KDEF) [10] developed by the Emotion Lab at Karolinska Institutet Sweden. The dataset consist a total of 4900 images of 562*762 categorized in to seven different emotions i.e. Afraid, Angry, Disgusted, Happy, Neutral, Sad and Surprised. The dataset is divided into test and train dataset in 80% - 20% split. The train dataset consist of 3920 images divided into seven categories and the test dataset consist of 980 images divided into seven categories. The images are scaled down from 562*762 pixels to 256*256 pixels before feeding it into the CNN model.

The Convolution Neural Network model is trained on NVIDIA GEFORCE GTX 960M Graphics Processing Unit (GPU) with 4 GB of dedicated graphic memory and Intel Core i7 6700 HQ CPU 2.60GHz with 8 GB of RAM on Asus ROG GL552VW. The working environment is Spyder editor for Python 3.5. Keras and NumPy libraries were used with TensorFlow as the backend. The models were trained for 25 epochs with 3920 steps per epochs for training set and 980 steps for validation set.

Figure 5. Images in Dataset

VI. IMAGE PROCESSING

The images present in the dataset were preprocessed by using the ImageDataGenerator [18] class which generates batches of tensor images. In this method the images were re-scaled by a factor up to 1/255 . The images were randomly flipped in horizontal direction in order to generate randomness in the input image while training the model. Images were sheared in counter clockwise direction up to 0.2 degrees and the zoom range for the images were set to be about 0.2 to provide random zoom.

VII. COMPARISON AND DISCUSSION

The three different DNN architecture that has been discussed in this paper are The Rectangular Architecture, The Modified Triangular Architecture [9] and our proposed architecture, The Venturi Architecture (in shape of an venturi tube). All the three architectures were trained and tested on the same dataset consisting of total 4900 images depicting seven different human facial emotions. The Rectangular architecture consisting of 6 hidden layers has a training accuracy of 98.64% with a trainingloss of 0.0914 whereas the Modified Triangular architecture consisting of 7 hidden layer in its architecture has a training accuracy of 98.29% which is 0.35% less than the Rectangular architecture but it has a training loss of 0.0591 which is 0.0323 more than the training loss of Rectangular architecture.

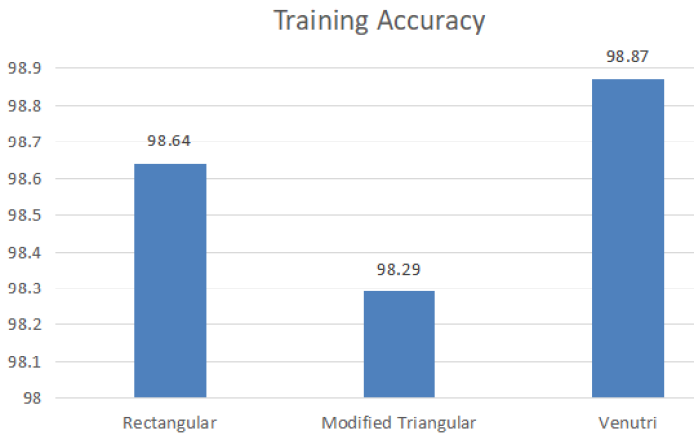

Fig. 6. Training Accuracy Comparison

Our proposed Venturi Architecture which has 6 hidden layers has a training accuracy of 98.87% which is 0.23% more than the Rectangular architecture and 0.58% more than the Modi-fied Triangular architecture. The training loss of the Venture architecture is 0.0224 which is 0.069 less than the Rectangular architecture and it is 0.0367 less than the Modified Triangular architecture.

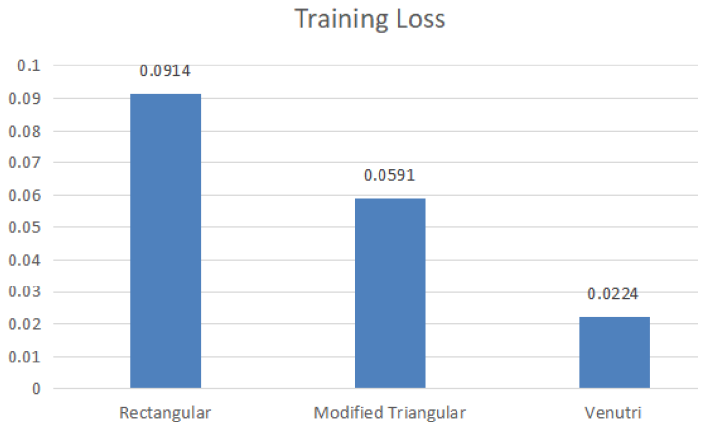

Fig. 7. Training loss Comparison

The Rectangular architecture has a validity accuracy of 79.61% with 1.0522 validity loss whereas the Modified Rectangular architecture has a validity accuracy of 82.70% which is 3.09% more than the Rectangular architecture and its validity loss is 0.9859 which is 0.0663 less than the Rectangular architecture.

Our proposed Venturi Architecture shows a validity or testing accuracy of 86.78% which is 7.17% more than the 6 layered Rectangular Architecture and 4.08% more than the testing accuracy of the 7 layered Modified Triangular architecture. The Venturi Architecture has a testing loss of 0.9693 which is 0.0829 less than the Rectangular Architectures testing loss and it is 0.0166 less than the 7 layered Modified Triangular architecture [9].

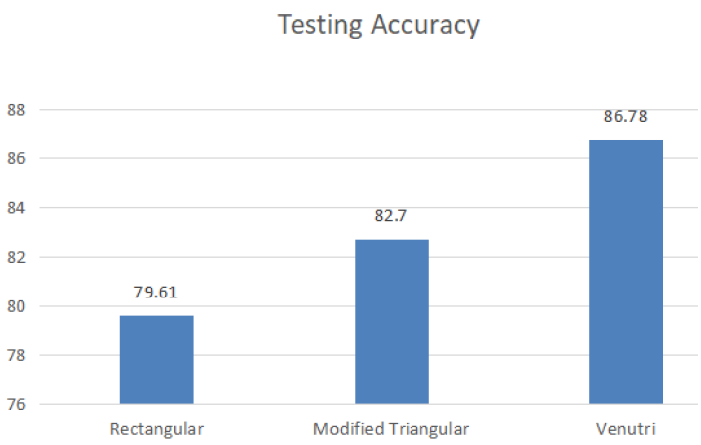

Fig. 8. Testing Accuracy Comparison

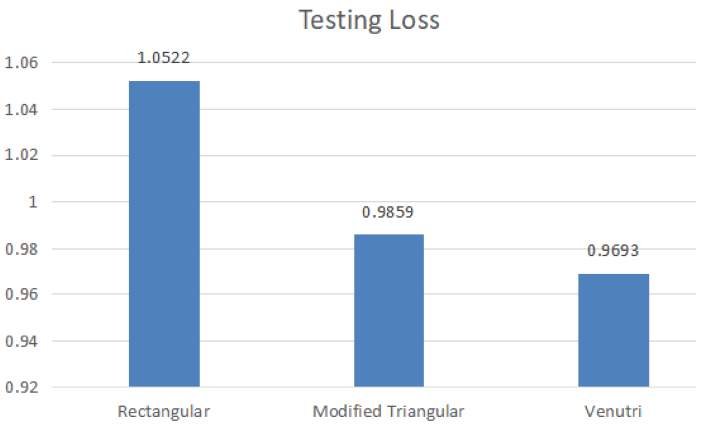

Figure 9. Testing Loss Comparison

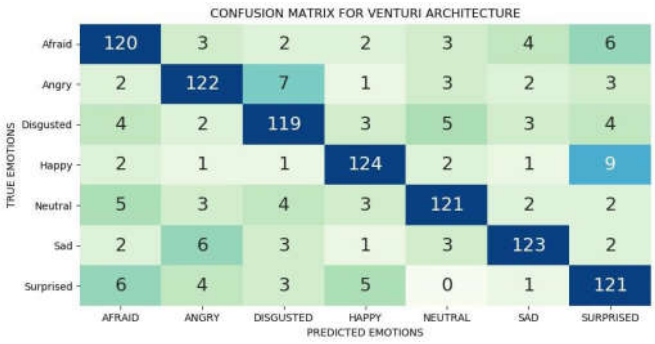

Figure 10. Venturi Architecture Confusion Matrix

Fig. 10 represents the confusion matrix for the proposed Venturi Architecture, which was tested on a total of 980 test images of different emotions categorised as Afraid, Angry, Disgusted, Happy, Neutral, Sad and Surprised. The trained

model was test on 980 images with each emotion category having 140 test images. The Confusion matrix shows the accuracy for different emotions when Venturi Architecture was tested with 980 test images. Happy emotion had the best accuracy of 88.57% whereas the Disguated emotion had the worst accuracy of 85.00%. Afraid had an accuracy of 85.71%, Angry had an accuracy of 87.14%, Neutral had an accuracy of 86.42%, Sad had an accuracy of 87.85% and the last emotion category Surprised had an accuracy of 86.42%. Whereas the overall accuracy of the whole model was at 86.73%.

These comparison shows that the proposed Venturi Architecture has a better training as well as better testing accuracy than the other two discussed architecture. It also shows better improvement in the both areas of training and testing loss during the 50 epochs which was used to train each the three architecture.

VIII. CONCLUSION

In this paper, three different architecture for the deep neural network of the convolutional neural networks i.e. The Rectangular Architecture, The Modified Triangular Architecture [9] and the newly proposed Venturi Architecture were analyzed for their training accuracy, testing or validation accuracy, training loss, testing or validation loss. It was found out that in terms of training accuracy The Venturi Architecture showed the highest training accuracy whereas The Modified Triangular Architecture showed the worst training accuracy at 98.29%. The Venturi Architecture also shows the best testing or validity accuracy as compared to other 2 Architecture at 86.78% whereas The Rectangular Architecture was with the worst validity accuracy at 79.61%. The proposed venturi architecture shows a 4.08% accuracy improvement than the modified triangular architecture and 7.17% accuracy improvement than the rectangular architecture. In future, plan to evaluate the performance of the Venturi architecture on other facial emotion database and to use this architecture to evaluate a multi modal deep neural network with both the facial emotion images and emotion audio samples for better efficacy.

REFERENCES

[1] I. Fasel, J.R. Movellan, M.S. Bartlett, and G. Littlewort, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," in 2003 Conference on Computer Vision and Pattern Recognition Workshop, IEEE, 2003.

[2] N.V. Scapel, G.P. Barlier, A. Guzman, B. M. Sommer, N. Damasky, T. Weisse, T. Goossens, H. Pham, B. Amberg, J. D. Stoyles, A.R. Moha, "Emojicon puppeting," in US20180336714A1 United States of America, Patent application publication, 2018.

[3] K. Kozuka, T. Nakano, S. Yamamoto, T. Ito, and S. Mita, "Driver blink measurement by the motion picture processing and its application to drowsiness detection," in The IEEE 5th International Conference on Intelligent Transportation Systems. IEEE, 2002.

[4] B.-H. Lee, and J.-G. Lee, "Therapeutic behavior of robot for treating autistic child using artificial neural network," in Fuzzy Systems and Data Mining IV: Proceedings of FSDM 2018, pages 358–364, 2018.

[5] J. Lu, and S. Z. Li, "Face recognition using the nearest feature line method," in IEEE Transactions on Neural Networks, IEEE, 1999.

[6] P. R. Dachapally, "Facial emotion detection using convolutional neural networks and representational autoencoder units," arXiv: 1706.01509, 2017.

[7] V. A. Tzouvaras, T. P. Mailis, K. C. Karpouzis, S. D. Kollias, S. V. Ioannou, and A. T. Raouzaiou, "Emotion recognition through facial expression analysis based on a neurofuzzy network," in Neural Networks, Elsevier, 2005.

[8] G. K. Venayagamoorthy, S. Bashyal, "Recognition of facial expressions using gabor wavelets and learning vector quantization," in Engineering Applications of Artificial Intelligence, p.p. 1056–1064, Elsevier, 2008.

[9] N. Venkatesan, Md. A. Haque, and J. Sahaya Rani Alex, "Evaluation of modified deep neural network architecture performance for speech recognition," in 2018 International Conference on Intelligent and Advanced System (ICIAS), IEEE, 2018.

[10] A. Flykt, A. Öhman, and D. Lundqvist, "The Karolinska directed emotional faces," Department of Clinical Neuroscience, Psychology section, Karolinska Institute, 1998.

[11] R. Chellappa, S. Zhou, and V. Krueger, "Probabilistic recognition of human faces from video," in Computer Vision and Image Understanding, pages 214–245, Elsevier, 2003.

[12] M. Yu, H. Schutze, W. Yin, and K. Kann, "Comparative study of CNN and RNN for natural language processing," in Computer Vision and Image Understanding, arXiv: 1702.01923, 2017.

[13] T. P. Karnowski, I. Arel, and D. C. Rose, "Deep machine learning a new frontier in artificial intelligence research," in IEEE Computational Intelligence Magazine, IEEE, 2010.

[14] J. Schmidhuber, "Deep learning in neural networks: An overview, in Neural Networks," Elsevier, 2014.

[15] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in Y. Lechevallier, G. Saporta (eds) Proceedings of COMPSTAT'2010, pages 177–186, Physica-Verlag HD, 2010.

[16] J. Lei Ba, and D. P. Kingma, "Adam: A method for stochastic optimization," in 3rd International Conference for Learning Representations, San Diego, 2015.

[17] W. Chu, S. K. Shevade, A.N. Poo, K. Duan, and S. S. Keerthi, "Multicategory classification by soft-max combination of binary classifiers," in T. Windeatt, F. Roli (eds), Multiple Classifier Systems, MCS 2003, Lecture Notes in Computer Science, vol. 2709, Springer, Berlin, Heidelberg, 2003.

[18] S. Pal, and A. Gulli, "Deep learning with Keras," in Deep Learning with Keras, Packt Publishing Ltd., 2017.