IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 9, Issue 11, November 2020

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 11, November 2020||

# Brain Intelligence: A Multitasking Model using Convolutional Neural Network and Artificial Neural Network

**Anuja Shirke**[1]

IT Student, Department of Information Technology, B.K. Birla College of Arts, Science and Commerce (Autonomous), Kalyan, Maharashtra, India[1]

**ABSTRACT:** Currently Artificial Intelligence has marked a tremendous growth in the technology and business sector. Without Artificial Intelligence, technology has no future. Artificial Intelligence is used in daily social life. Most Information Communication Technology (ICT) models are dependent on big data, lack self-idea, and are complicated. So, Lu et. al. [1] came up with a model named Brain Intelligence (BI). The researchers planned to generate new ideas about events without experiencing them by using Artificial life with an imagine function. Recently Artificial Intelligence is been developed for extracting patterns. Some limitations has been stated by Li. et. al. [1]: Artificial Intelligence concentrates on individual areas, i.e. voice recognition, visual recognition, text analytics and Natural Language Processing, and so on. The other limitation of Artificial Intelligence is: the system cannot make connections between two or more things. They are incapable of linking between things. Also, machine learning results are vulnerable and can be obtained using numerical only it is incapable to function as a human brain i.e. a single part can't function as a whole brain. To solve these problems the Brain Intelligence model has been introduced. The portrayed model represents the working of the human brain. BI is a form of Artificial Intelligence fused with Artificial Life models. This paper focus on the multitask limitation of AI which can be overcome using Convolutional Neural Network (CNN) and Artificial Neural Network (ANN). It is briefly discussed below.

**KEYWORDS:** Artificial Intelligence, Brain Intelligence, Deep Learning models, Multitasking model, Artificial Life.

## I. INTRODUCTION

Nowadays AI has achieved great success in the world of technology. Using AI, we can perform a human specific task (voice recognition, image processing, face recognition, etc.). AI is the future of the current world as it outperforms human tasks. Alexa, google assistant, Cortana are some of the Automatic Speech Recognition (ASR) system which carries out the specified task as commanded. These systems made our lives easier. AI-based games such as Alpha GO [19], Dark forest, Deep blue, etc. are also available which outplays the human role. Real-time robots are invented to carry out human tasks like Amazon Prime Air which is a delivery system designed to safely hand-in the packages to customers using unmanned aerial vehicles [16]. Basically, AI is such a technology that can perform all human tasks. Implementing recurrent neural networks (RNN), Automatic Speech Recognition (ASR) system can be improved using n-gram language models [4]. Initially, Convolutional Neural Network (CNN) was used for image processing purpose. In recent years, CNN can be used for speech recognition, language processing [13] [15]. But AI cannot perform multiple tasks at the same time. A research paper on brain intelligence (BI) [1] proposes a BI model by fusing Artificial Intelligence and Artificial life models to develop new intelligent learning with memory and idea function. The researchers state that the BI model can solve the multitasking limitation of AI. The purpose of this paper is to show, using CNN and ANN we can build multitasking system which can perform image processing and speech recognition at the same time or language processing and image processing at the same time.

In the following, we discuss the related work of Convolutional neural networks and recurrent neural networks, a methodology that has been used, experimental results, and a conclusion drawn based on results.

## II. RELATED WORK

1. Image Processing: Face recognition has been improved in recent years. According to Taigman et. al. [3] coupling the 3D model-based alignment with large capacity feedforward models can effectively overcome the drawbacks of the traditional method. The researchers proposed a system "DeepFace" which closes the remaining gap in unconstrained face recognition and is on the verge of human-level accuracy. The authors used the LFW dataset

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 11, November 2020||

for training and testing purposes without frontalisation obtaining an accuracy of 94.3% (with 2D alignment) while with frontalisation and naïve LBP/SVM combination accuracy obtained was 91.4%. Tio et.al [18]. used Inception V3 model to classify the face shape, using a human face dataset. The accuracy of the retrained Inception V3 model ranges from 84.8% to 97.8%. Retrained Inception V3 model outperforms the classifiers based on linear discriminant analysis, support vector machines, multilayer perceptron, and k-nearest neighbors [18].

2. Voice Recognition: We all are acquainted with Alexa, Google Assistant, Cortana, Siri, etc ASR systems. No doubt these systems have marked tremendous growth. In [17] Saon et. al. suggested, best results can be obtained using CNN and RNN together. Recently it has been found that a small convolution window with many layers yields better results. In comparison between CNN and RNN, long short-term memory (LSTMs) is better at taking advantage of more data than the VGG network while in sequence training VGG network performs well than LSTM [17]. So, the researchers combined both networks for better results. Picheny et. al. [17] used three types of language models: class-based exponential model, feed-forward neural network [9], and six recurrent neural networks yielding WER of 5.5% which is the lowest among all. According to Kombrink et. al. [4] RNN language model is more suitable for mass application in common large vocabulary continuous speech recognition (LVCSR) systems. Recent development has also been seen in noise-robust speech recognition [15]. The model proposed by Qian et. al. [15] obtains a 10.0% relative reduction over traditional CNN. The portrayed model achieves a WER of 8.81% i.e. 17% of relative improvement over LSTM-RNN.
3. Language modeling: Deep neural networks (DNN) performs excellently on difficult problems such as speech recognition or object visualization. In [9], researchers developed an architecture ‘straightforward LSTM’ which can solve general sequence to sequence problem. Sutskever et. al. [9] portrayed a model that had two different LSTMs: one for input sequence another for output sequence and chose LSTM with 4 layers also the researchers found reversing of words at input sequence useful. [10] shows a tremendous work in grapheme to phoneme conversion by comparing seq2seq architecture and sequitur G2P. Milde et. al. observed that the combination of seq2seq with sequitur G2P without multitask learning yields lower WER. A multi-channel convolution neural network that is popular among image processing can be used on language models [14]. The multi-channel model can build a multilingual dialog state tracker with large-scale cross-language corpora.

## III. METHODOLOGY

Three datasets are used to perform experiments using CNN and ANN models.

1. Image processing

CNN is more popular in image processing but in recent years CNN did show some great results in speech recognition and language modeling [13]. So, the CNN model has been chosen for building a multi-tasking model. In image processing, a Face expression recognition dataset [22] was used, consisting of 7 directories i.e. different facial expressions (angry, happy, sad, fear, disgust, neutral, surprise. See fig.1). Four convolutional layers were added with 2 fully connected layers with an Adam optimizer. One convolutional layer consists of Conv2D, batch normalization, ReLU as activation function, max pooling with pool size (2,2), and Dropout. Batch normalization improves the performance and stability of neural networks. Pooling layers are used to reduce the dimensions of the feature maps. Max pooling layer selects the maximum element from the region of the feature map generated by a convolution layer. The output of the max-pooling layer contains the most prominent feature map than the previous feature map. This makes the model more robust to variations in the position of the features in the input image. Dropout is used to reduce the overfitting of the model. ReLU activation function was used in our model as it is more efficient and popular. The advantage of using this function is it does not activate all neurons at the same time. A flattening layer was also used to obtain a long vector of input data that can be passed through neural networks.

Figure 1

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 11, November 2020||

The second dataset used is Intel image classification [23], consisting of natural scenes around the world categorized as buildings, forests, glaciers, mountains, sea, streets (fig.2). Here InceptionV3 model was used to train and test the data. This model is a pre-trained classifier. It has been trained using the ImageNet 2012 dataset that contains about 14 million images and 1000 classes. In [18], the Inception V3 model showed 97% accuracy and the authors confirmed that retraining the Inception v3 image classifier can outperform the traditional classifiers. The model developed for this dataset consists of the Inception V3 model with flattening layer, dropout layer, and two dense layers with ReLU and softmax activation function. A dense layer is used to feed all outputs obtained from the previous neuron to the next neuron.

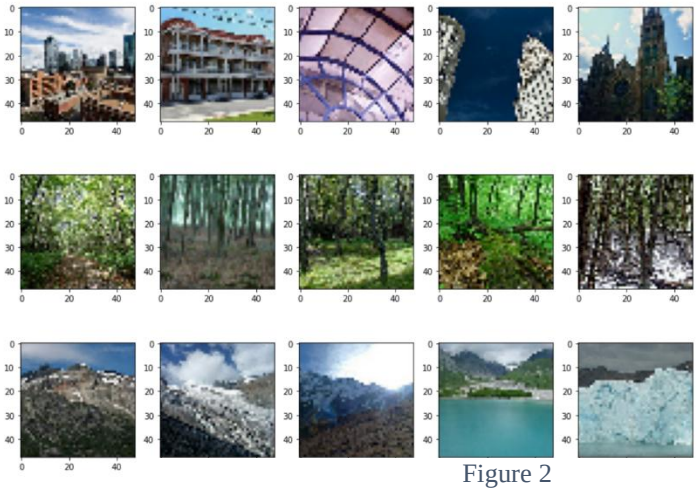

Figure 2

2. Speech recognition

For speech recognition, gender recognition by voice dataset [21] was used available online. The dataset consists of 3,168 recorded voice samples collected from male and female speakers. Here Artificial neural network was used for training and testing of the dataset. Three artificial neural network layers: forward propagation, backward propagation, and cost function were applied. Forward propagation, to feed the input data in forwarding direction through the network. It is used to yield some output, and to ensure that data should not flow in the reverse direction during output generation. The backward propagation algorithm is to calculate the gradient of the error function w.r.t the neural network's weight. The cost function is used to measure how wrong is our model in terms of its ability to find a relationship between X and Y. Parameters are updated gradually to feed forward the data. These weights help in reconciling the differences between the actual and predicted outcomes.

3. Natural Language Processing

The third dataset used is a deep NLP dataset consisting of two .csv files [20]. One of the .csv files contained 80 user responses of chatbot therapy (See fig.3 representing a word cloud of chatbot responses) while the other CSV file contained 125 resumes. A neural network with a sequential model which is a common architecture of ConvNets was developed. To this sequential model Embedding layer, LSTM layer, Dense layer, and Dropout layer were added. The embedding layer turns positive integers to dense vectors of fixed size [14] as the data is in CSV format, it is important to convert them into dense vectors for obtaining better results. Here LSTM is used as the first hidden layer. This layer selectively remembers or forgets things i.e. LSTMs have an edge over conventional feed-forward neural networks. A dense layer is used to feed forward the output of the previous neuron to the next neuron as input. With the dense layer, the sigmoid activation function was been used to fire the next neuron. A sigmoid function exists between 0 to 1, it is used to predict the probability of our model. The dropout layer was been used to reduce overfitting, overfitting of the model means- production of analysis that corresponds too closely or exactly to a particular set of data. Overfitting is caused due to reading details or noise of training data by neurons. So, the Dropout layer was used to prevent random fluctuations or noise been picked from training data. 'Rmsprop' optimizer was used to change the attributes of the neural network such as weights and learning rate to reduce the losses.


Figure 3

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 11, November 2020||

## IV. EXPERIMENTAL RESULTS

1. Image Processing

Two datasets were used: Face expression recognition dataset and intel image classification dataset, available online. Two different algorithms were used on these datasets. The algorithm used for the facial expression recognition dataset was a simple CNN model consisting of 4 convolutional layers and 2 fully connected layers. Convolution will extract relevant features from the input image and a fully connected layer will focus on these features. Using a softmax activation function to the dense layer and an Adam optimizer with learning rate(LR)=0.0001 and epochs =50, the obtained accuracy was 0.649 i.e. 64.9%. The confusion matrix (Fig.4) constitutes the obtained result wherein we can see that the model is good at predicting surprise and sad faces. But this model gets confused between happy, disgusted faces and fear, angry faces. For the second dataset, the Inception v3 model was used which is good at classification [18]. The dataset consists of natural scenes around the world. In the trained and tested model, with the InceptionV3 layer, Flatten, Dense, and Dropout layers were used with categorical cross-entropy as loss function, Adam as optimizer with epochs=10 and lr=0.000003. A flattening layer is used to convert the input data into a long vector which will be passed further in neural networks. The accuracy obtained through this model was 0.903 i.e. 90.3% which is a pretty good accuracy rate. Fig.5 reveals an analysis of the obtained result. As we can see, our model stabilizes after 4 epochs giving an accuracy rate greater than 90%. Therefore, the InceptionV3 model shows more accuracy rate compare to the basic CNN model. According to the obtained results, InceptionV3 is more recommendable for image processing.

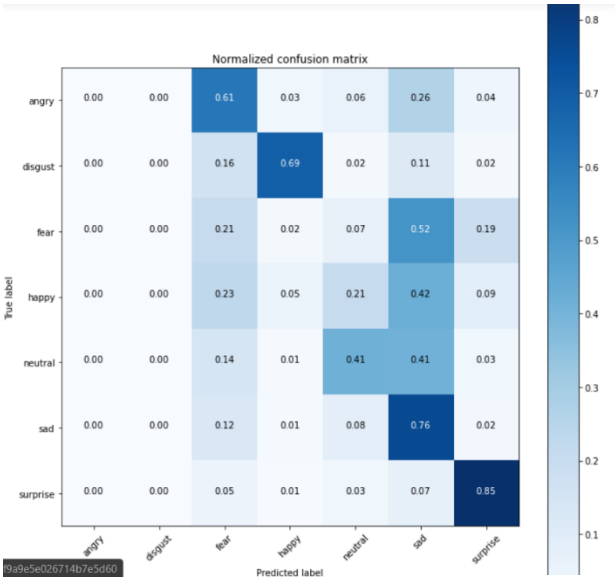


Figure 4

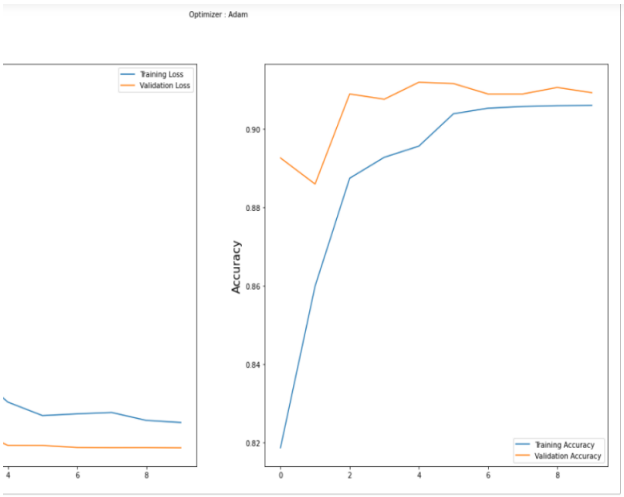


Figure 5

2. Speech Recognition

The dataset used for speech recognition is Gender recognition by voice available online. The dataset consists of 3,168 recorded voice samples of male and female speakers. The voice samples are pre-processed by acoustic analysis in R using the see wave and tuneR packages, with an analyzed frequency range of 0hz-280hz. This acoustic file is in CSV format. Three neural network layers were applied to the dataset: Forward propagation, backward propagation with gradient descent, and cost function. As discussed above, forward propagation is used to pass the input data further, Backward propagation is used to calculate the gradient of the error function and the cost function expresses the difference between the predicted value and actual value. Here, in fig.6 accuracy of the model is shown w.r.t the cost function and number of iterations. The LR given to this model is 0.001 and the number of iterations=5000. Accuracy found using this model is of 98.106%.

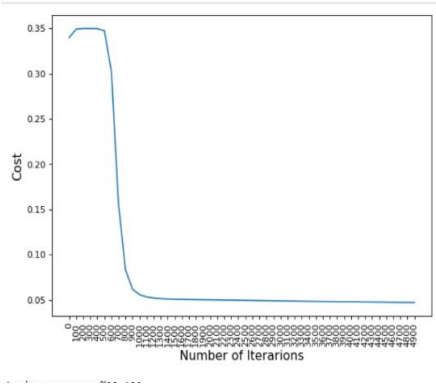


Figure 6

3. Natural Language Processing

The dataset used consists of 80 user responses to a therapy chatbot and 125 resumes queried from indeed.com [20]. In the chatbot dataset, if a response is not flagged then the user can continue to speak and if a response is flagged it is referred to help. In the resumes dataset, if a resume is not flagged then the applicant can submit a modified resume if flagged then the applicant is invited to interview. The sequential model was used with dense, embedding, LSTM, and Dropout layers, rmsprop as optimizer, binary cross-entropy as loss function, and sigmoid as the activation function. Fig.7 shows loss and accuracy of our model for therapy chatbot CSV files. Our model gains greater than 90% accuracy after 6$^{th}$ epoch and after 8$^{th}$ epoch our loss is less than 30%. The predicted model for this dataset gained an accuracy of 81%. Similarly, for 125 resumes CSV files same model was built that gained an accuracy of 98.87%(fig.8) and the predicted model gained 70.58% of accuracy. This model is used to classify the data.

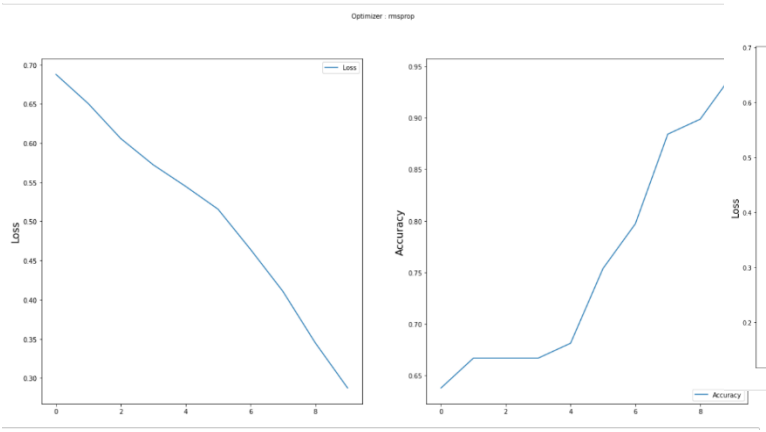


Figure 7

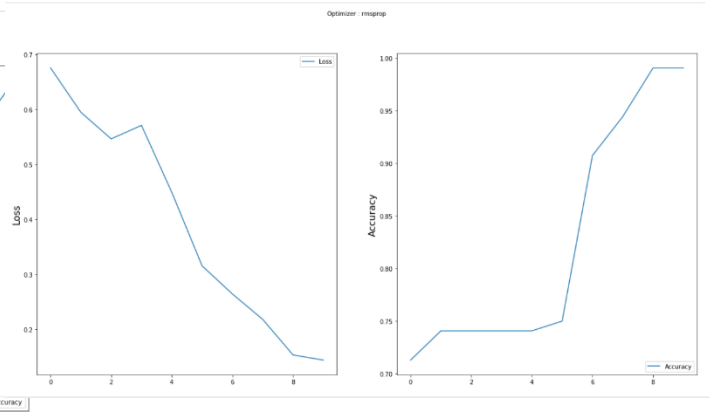


Figure 8

## V. CONCLUSION

For image processing, we can conclude that using a simple four layers convolutional neural network model with 2 fully connected layers isn't that beneficial as the accuracy rate is pretty low (64%). While using an Inception V3 layer with the CNN model is quite beneficial as Inception V3 is good at classification [18], gaining an accuracy rate of 90%. For voice recognition, an Artificial neural network (ANN) was used with forward propagation, backward propagation with gradient descent, and cost function which gains a good rate of accuracy i.e. 98.10% while keeping the number of iterations=5000. For NLP, the common architecture of ConvNet was used 'Sequential model' with dense, embedding, LSTM, and Dropout layers yielding an accuracy rate of 81% and 98.87%.

So, we can conclude that using CNN and ANN we can build a multitasking system that can perform image classification and voice recognition at the same time. While natural language processing i.e. computational techniques for the automatic analysis and representation of human language can also be performed using CNN and LSTM layer which gives pretty good results.

According to my knowledge future scope in this field can be the linking of things. For example, if we know the meaning of the word 'horse' and the meaning of the word 'stripes'. So, we can understand, 'A zebra is a horse with stripes' but the computer cannot make such a connection [1]. Also, predicted model of NLP can be improved using different deep learning models.

## VI. ACKNOWLEDGMENT

I would like to thank Prof. Swapna Augustine Nikale, Department of Information Technology, B.K. Birla College Kalyan for guiding me throughout my research work.

## REFERENCES

[1] Lu, H., Li, Y., Chen, M., Kim, H., &Serikawa, S. (2017). Brain Intelligence: Go beyond Artificial Intelligence. Mobile Networks and Applications, 23(2), 368–375. https://doi.org/10.1007/s11036-017-0932-8

[2]Parikshit Gupta. (2020). Reverse Engineering of Human Brain for the field of Artificial Intelligence. International Journal of Engineering Research And, V9(05), 252–257. https://doi.org/10.17577/ijertv9is050169

[3]Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 1–8. https://doi.org/10.1109/cvpr.2014.220

[4] Kombrink, S., Mikolov, T., Karafiat, M., &Burget, L. (2011). Recurrent Neural Network based Language Modelling in Meeting Recognition. Interspeech, 2877–2880. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=recurrent+neural+network+based+language+model&oq=recurrent+neural+network#d=gs_qabs&u=%23p%3DLW3DvYKBFAMJ

[5]Bengio, Y., Ducharme, R., & Vincent, P. (2003). A Neural probabilistic Language Model. Journal of Machine Learning, 1–7. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=a+neural+probabilistic+language+model&oq=#d=gs_qabs&u=%23p%3D2MIMJitPj6UJ

[6] van Gerven, M., &Bohte, S. (2017). Editorial: Artificial Neural Networks as Models of Neural Information Processing. Frontiers in Computational Neuroscience, 11, 1–2. https://doi.org/10.3389/fncom.2017.00114

[7]Özdemir, V., &Hekim, N. (2018). Birth of Industry 5.0: Making Sense of Big Data with Artificial Intelligence, "The Internet of Things" and Next-Generation Technology Policy. OMICS: A Journal of Integrative Biology, 22(1), 65–76. https://doi.org/10.1089/omi.2017.0194

[8] Mei, H., Bansal, M., & Walter, M. R. (2016). What to talk about and how? Selective Generation using LSTMs with Coarse-to-Fine Alignment. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 720–730. https://doi.org/10.18653/v1/n16-1086

[9] Sutskever, I., Vinyals, O., &V.Le, Q. (2014). Sequence to Sequence Learning with Neural Networks. Neural Information Processing Systems(Nips), 1–9. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-

[10] Milde, B., Schmidt, C., & Köhler, J. (2017). Multitask Sequence-to-Sequence Models for Grapheme-to-Phoneme Conversion. Interspeech 2017, 2536–2540. https://doi.org/10.21437/interspeech.2017-1436

[11] Belk, R., Humayun, M., &Gopaldas, A. (2020). Artificial Life. Journal of Macromarketing, 40(2), 221–236. https://doi.org/10.1177/0276146719897361

[12] Vieira, S., Lopez Pinaya, W. H., Garcia-Dias, R., &Mechelli, A. (2020). Deep neural networks. Machine Learning, 157–172. https://doi.org/10.1016/b978-0-12-815739-8.00009-2

[13]Gheisari, M., Wang, G., & Bhuiyan, M. Z. A. (2017). A Survey on Deep Learning in Big Data. 22017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 146–157. https://doi.org/10.1109/cse-euc.2017.215

[14] Shi, H., Ushio, T., Endo, M., Yamagami, K., & Horii, N. (2016). A multichannel convolutional neural network for cross-language dialog state tracking. 2016 IEEE Spoken Language Technology Workshop (SLT), 559–564. https://doi.org/10.1109/slt.2016.7846318

[15]Yuksel, M. E. (2018). Agent-based evacuation modeling with multiple exits using NeuroEvolution of Augmenting Topologies. Advanced Engineering Informatics, 35, 30–55. https://doi.org/10.1016/j.aei.2017.11.003

[16] Amazon Prime Air Amazon.com: Prime Air

[17]Saon, G., & Picheny, M. (2017). Recent advances in conversational speech recognition using convolutional and recurrent neural networks. IBM Journal of Research and Development, 61(4/5), 1:1-1:10. https://doi.org/10.1147/jrd.2017.2701178

[18] Tio, A. (2019). Face shape classification using Inception v3. *Cornell University*, 1–6. https://arxiv.org/abs/1911.07916

[19] https://deepmind.com/research/case-studies/alphago-the-story-so-far#what_is_go_

[20] https://www.kaggle.com/samdeeplearning/deepnlp

[21] https://www.kaggle.com/primaryobjects/voicegender

[22] https://www.kaggle.com/jonathanoheix/face-expression-recognition-dataset

[23] https://www.kaggle.com/puneet6060/intel-image-classification

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details