ISSN: 2586-7652 (Print) Vol. 03, No. 02, September 2020
ISSN: 2635-7607 (Online)
International Journal of Advanced Engineering
Source: http://ictaes.org
Manuscript received: July 14, 2020; Revised: August18, 2020; Accepted: September 02, 2020

Speech to Sign Language Interpreter Using Spectrogram and CNN

Lisa Rajkarnikar[1], Dipinti Manandhar[2], Lujina Maharjan[3], Monika Pokharel[4], Surendra Shrestha[5]

[1,2,3,4]*Advanced College of Engineering and Management, Tribhuvan University, Lalitpur, Nepal*
[5]*Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering, Lalitpur, Nepal*

lisarajkarnikar@gmail.com[1], dipintimanandhar@gmail.com[2], lujinamaharjan@gmail.com[3], monikapokharel1@gmail.com[4], surendra@ioe.edu.np[5]

Abstract

Speech interpreter system can enhance the life of deaf people as Sign Language is the only means of communication for them. The paper proposes the solution of the need for a middle person as an interpreter by translating the voice into text and text into its respective sign language. The paper contributes to the field of Convolution Neural Networks. It begins with signal pre-processing and signal conditioning. Feature extraction process is implemented using Spectrogram. Spectrogram extracts the important features from speech which is necessary for speech recognition. Then it is fed to CNN architecture for feature matching. The recognized word is finally matched with the gif label and sign language gif is displayed. We obtained a training accuracy of 90.64%. With this, obtained interpreter bridges the communication gap between deaf people and normal people. Moreover, the paper aims to eliminate the necessity of attending expensive Sign Language classes, as it can be learnt at home easily using the developed interpreter system.

Keywords: *Deaf Human, Speech interpreter System, Speech Recognition, Spectrogram, CNN*

1. Introduction

About 70 million people around the world face the problem of hard of hearing and deafness. Sign Language is a common form of communication that is widely used by speech and hearing impaired. Nowadays, there is the facility of a middle person as an interpreter for exchanging information. But the professional interpreter may not be always available and may cause a communication barrier and delay in the reach of information. In this context, Speech to sign language interpreter acts as a boon to mankind by providing a better aid for communication with the normal ones. Speech to sign language interpreter is to provide the helping aid for people to learn sign language and eliminate the requirement of a middle person as an interpreter. Furthermore, the GIF

Corresponding Author : Surendra Shrestha
Author's affiliation : Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering
Email: surendra@ioe.edu.np
Copyright © ICT-AES

animation with facial expressions can be seen as its output for better interaction with deaf users. This system has a wide variety of applications in gaming, traffic control, conference presentation, and various other areas.

There are a lot of problems faced by deaf people in their day to day lives. They face difficulties in educational organizations, employment problems due to miscommunication. So, to solve this problem and provide better aid for communication speech to sign language interpreter has been developed. Speech to sign language interpreter system enables deaf people to communicate with normal ones. The converter uses algorithms such as the Spectrogram method and Convolutional Neural Network technique for the process of feature extraction and feature matching and provides a suitable sign language as output for the input speech provided. Speech to Sign language desktop application can be used by hearing disabled people. It can also be used by various organizations for better communication with their employees. There is a serious challenge for deaf people trying to integrate into the educational environment. So, it can also be used by educational organizations which enable the deaf to understand the subject contents. It can also be used to reach more audiences (hearing impaired) during speeches and campaigns. The main objective is to provide educational aid for the hearing disability. The system enables the active participation of deaf people in their day to day professional and personal lives and doesn't limit their participation in any field of works. The model provides us with good accuracy and can also serve as a means to provide basic education to children. It can further be implemented in various fields and serve as a communication aid for deaf people as it is a one-way communication-based system.

2. Materials and Methods

We have used Google Speech Command Dataset for the purpose of training. It consists of 30 labels with a total of 65,000 .way files. Each audio file is of 1-second length. It is recorded by a number of speakers making our system speaker-independent. The speech interpreter system is intended to help the hearing impaired people visualize the sign language of respective spoken words and thus provide ease in communication. The speech interpreter system generally implements two main modules: feature extraction and feature matching. The feature extraction module identifies the audio features which carry important linguistic content. Also, it's capable of eliminating unwanted background noises and room reverberation. Feature matching is executed so that unknown words are identified measuring extracted features from speech input and converts to text [1].

Among various methods for speech interpreter systems like RASTA-PLP, Linear Prediction Coding (LPC), Spectrogram, and Mel-Frequency Cepstral Coefficients (MFCC), we chose spectrogram which provides better visualization for the 3-dimensional audio signal and Convolutional Neural network (CNN) which can accurately analyze the visual aspects of the image.

Initially, the voice of the people is taken as input through the microphone. Then, signal preprocessing and cleaning is done to make it suitable for feature extraction as shown in the block diagram in Figure 1. For feature extraction, the Spectrogram method is implemented which generates the visual representation of input audio in the waveform. The spectrogram is taken as input by the Convolutional Neural Network model. After providing certain learning environment, a CNN model is generated for speech recognition which recognizes the spoken word and converts it to text. Then, the output of the recognized module is fetched to the sign language matched module which matches the converted text with its respective sign language GIF file. The detailed explanation of the methods used can be seen in the next section.

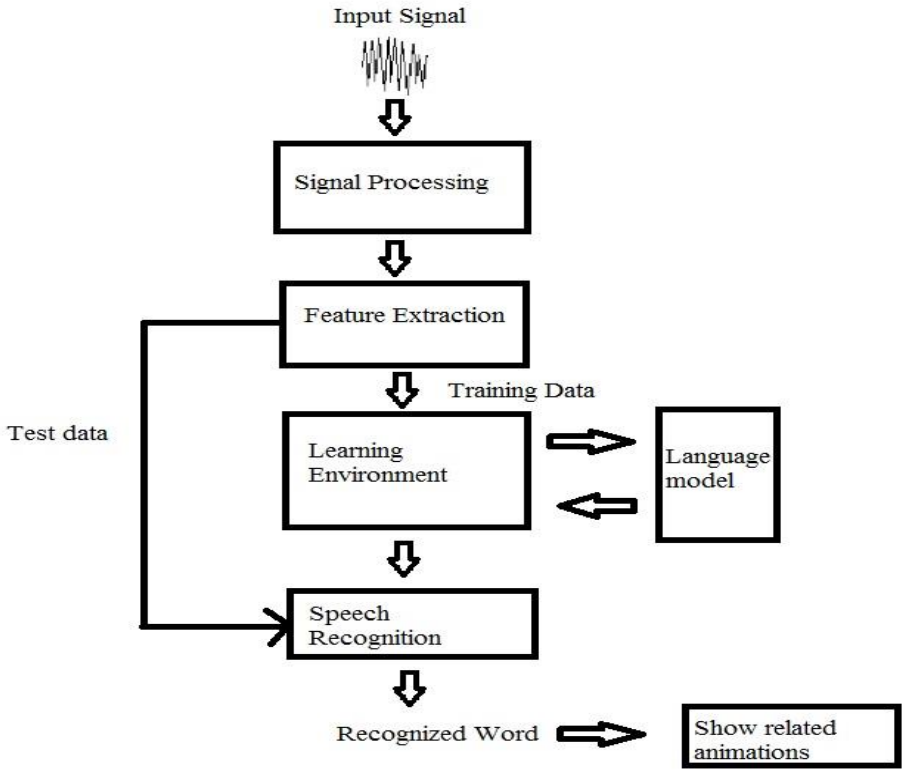


Figure 1. Block Diagram of Speech to Sign Language Interpreter System

2.1. Spectrogram

Spectrogram refers to the visual representation of the signal strength, or "loudness", of a signal over time at various frequencies present in a particular waveform. They have wide applications in fields of music, linguistics, sonar, radar, speech processing, seismology, etc. Short-time Fourier Transform (STFT) is implemented to convert raw data into spectrograms. STFTs divide longer time segments into shorter time segments (having the same length) and perform the Fourier transform on them individually. By windowing and taking the discrete Fourier transform (DFT) of every window, we can obtain STFT of the signal. Specifically, for an input signal x (t) and window w (n), the transform is defined as:

$$STFT\{x(t)\}(\tau,\omega) \equiv X(\tau,\omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)\, e^{-j\omega t}dt \quad (1)$$

The magnitude squared of the STFT yields the spectrogram representation as:

$$Spectrogram\{x(t)\}(\tau,\omega) \equiv |X(\tau,\omega)|^2 \quad (2)$$

Where ω is the window function, commonly known as a Hanning window or Gaussian window centred around zero, and x (t) is the audio signal to be transformed. $X(\tau,\omega)$ is essentially the Fourier Transform of $x(t)w(t-\tau)$ a complex function representing the phase and magnitude of the signal over time and frequency.

Spectrograms can be visualized as the graph of frequency versus time along with varying color intensity as its third dimension, as shown in Figure 2. The Y-axis represents the frequency of the audio signal which can be understood as pitch or tone while X-axis represents time measured in second for particular frequencies. The colors in Spectrogram represent the amplitude or loudness of particular frequency at particular time. The color "Green" corresponds to low amplitude and the color "Yellow" corresponds to high or strong amplitude as shown

in Figure 2. The darkness in the color represents the amplitude at that particular time and frequency.

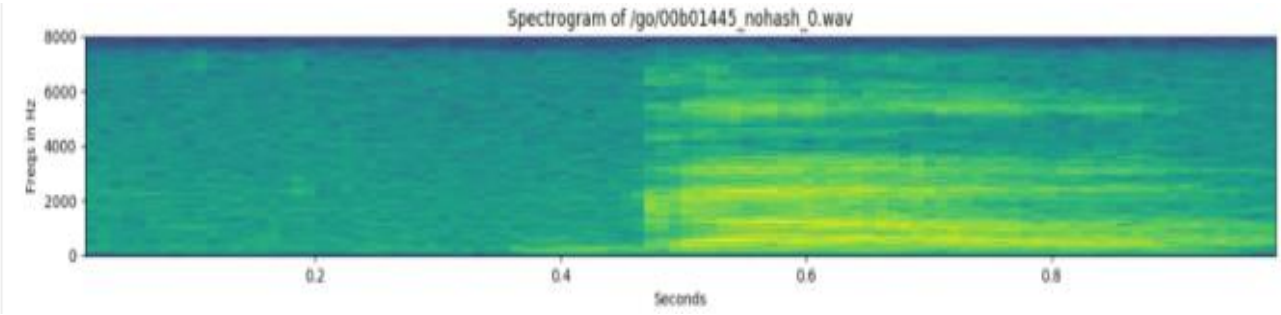


Figure 2. Spectrogram with time in X-axis, frequency in Y-axis and third dimension (amplitude) represented by colors of the spoken word "go"

2.2 Convolutional Neural Networks

In deep learning, a Convolutional neural network is a class of deep neural networks, most ordinarily applied to analyzing visual aspects of images. It can learn highly abstract features and may identify objects efficiently [2]. Due of its remarkable performance ,it has wide implementation within the areas of image classification, object detection, face detection, speech recognition, vehicle recognition, diabetic retinopathy, facial expression recognition and lots of more [3].

CNN use additional layers at the start of the neural network to reduce the size of an image and pre-process those. It uses the concept of weight sharing, due to which the number of parameters that needs training is substantially reduced, leading to improved generalization and prevent over fitting [4]. The diagrammatic view of Convolutional Neural Network can be shown in Figure 3.

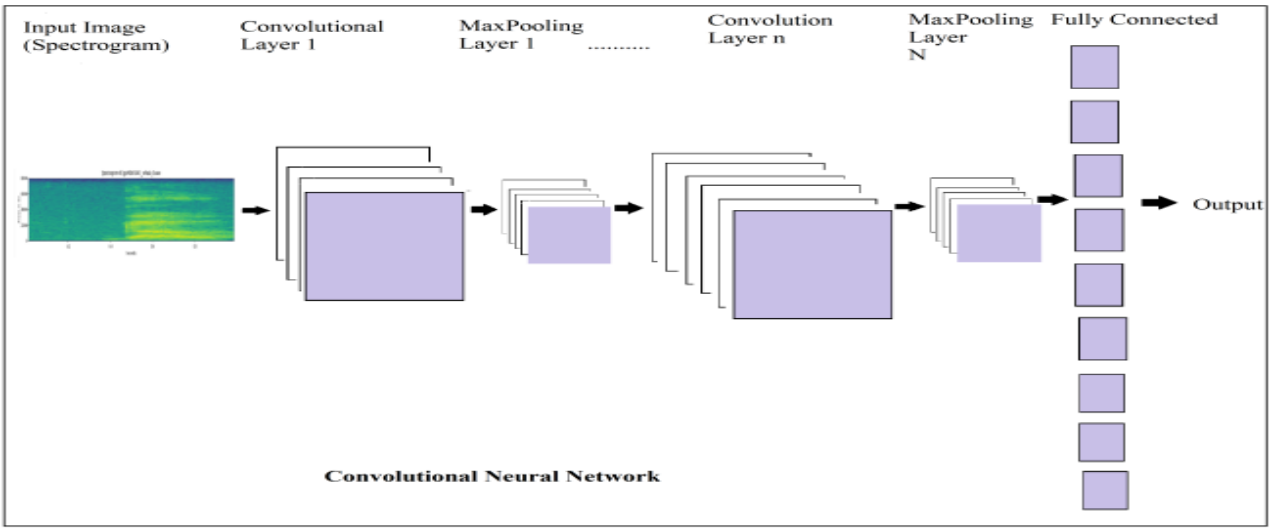


Figure 3. Convolutional Neural Network

2.2.1 Convolutional Layer

The name CNN is derived from the "convolution" operator. Convolutional layer feeds the images into input layers, filters them, extracts necessary image features and predicts the class label for the image using the

convolution operation.

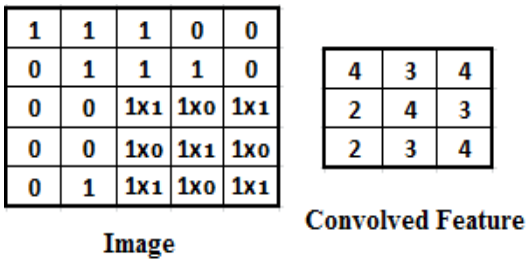


Figure 4. Convolution process performed on input pixel

Each neuron in the previous layer has local correlation with each neuron in the next layer called as Receptive Field [5]. It is capable of extracting local features and hence forms the filter or kernel [6]. These filters slides over the input image by 1 pixel (called as Strides) performing the convolution operation as shown in Figure 4. The 3*3 filter slides over horizontally and vertically over 5*5 matrix and performs the convolution operation. The resulting output is called as Convolved Feature / Feature Map [7]. It depicts the required image features and hence significantly reduces the number of trainable parameters [8].

The output a_{ij} in the next layer for location (i, j), is computed after applying convolution operation using formula given by (3) as shown below where, X is the input provided to the layer, W is filter or kernel which slides over input, b is the bias, * representing the convolution operation, and σ is non-linearity introduced in the network [3].

$$a_{ij} = \sigma\big((W * X)_{ij} + b\big) \tag{3}$$

We have implemented 3 convolutional layers with 32 filters and 3*3 sized window in each layers. We have taken convolutional stride as 1 and padding of type ‘same’ for the purpose of feature extraction of input spectrogram. It enables the model to extract and learn necessary features from the image (Spectrogram). The resulting convolved features are forwarded to Pooling layer for further processing.

2.2.2 Pooling Layer

The Convolution layer is followed by the Pooling layer which reduces the dimensionality of each map. Spatial pooling is capable of preserving significant information and reduces the number of parameters for training. There are three types of Spatial Pooling which are Max Pooling, Average Pooling, and Sum Pooling. Among them, Max Pooling has been shown to perform better as per practice. Max pooling identifies the largest element from the rectified feature map, hence preserving significant information. Pooling operation is performed by sliding the 2*2 filter over the feature map by 2 cells (Stride=2) and fetching to the pooling function as shown in Figure 5. Pooling layers: 3 Filter per Layer: 32, Window: 2×2, Stride: 2

The output from convolutional layer is fed to pooling layer which performs down sampling of the extracted image features .The Speech to Sign Language Interpreter model is developed with 3 pooling layers with 32 filters and 2 strides in each layer. It performs down sampling and prevents the over fitting of the input data. The obtained features are fetched to fully connected layer for classifying the label.

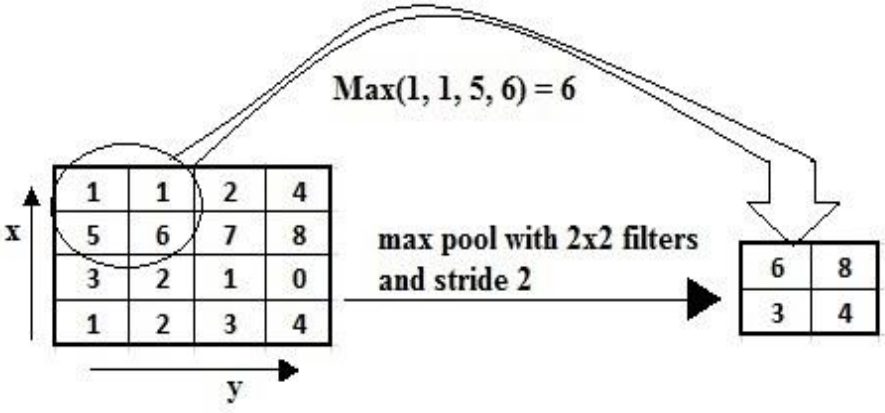


Figure 5. Max-pooling layer with 2*2 filter

2.2.3 Fully Connected Layer

The Fully Connected layer acts as Multi-Layer network in which every neuron in the first layer is connected to each neuron on the subsequent layer. The Convolutional and Pooling layers provide the high level features as output. Fully connected layer takes these features and determines which features most correlate to a specific class hence classifying the input images into various classes. Softmax activation function is implemented in the output layer which flattens the matrix into vector and sends it into fully connected layer. The softmax function ensures that the sum of output probabilities from the Fully Connected Layer is always 1 and provides the model more prediction power to classify the input images efficiently.

The input to fully connected layer is the downsampled result from pooling layers. The model is developed with single dense layer of 64 filters that provides us classified class as output i.e. textual version of the spoken word (input). We have also implemented Batch Normalization along with Softmax activation function to provide the probability distribution of the output .The output (textual format) from dense layer is then matched with the .gif animation tag in the dataset collection and required gif animation is displayed.

3. Results and Discussion

Sign language delivers communication aid between normal people and deaf people. With the assistance of varied deep learning techniques, we've come up with a Speech to Sign Language Interpreter System which could act as a boon to the life of deaf people. The user interface of the interpreter system can be comprehended in Figure 6 where the user can record the voice through a microphone after clicking on the "Live Voice" button. The wave file of the recorded voice is generated and then visualized in a 3D representation graph through the spectrogram method. This method helps to eliminate the unwanted noises and focuses on loud speech (intensity of sound), hence extracting necessary audio features for speech recognition. The resulting sound features presented in the spectrogram format undergoes through CNN architecture to uncover its true label. It includes three Convolutional layers, three Pooling layers, and one dense layer alongside Batch normalization to surge stability and dropout to prevent overfitting. The resulting output predicts the label (of recorded voice) in textual form and then it exhibits its matching sigh language from the database on the screen.

Through the implementation of varied CNN architectures and feature extraction methods, we've acquired the test accuracy of 95.13% for 4 labels and 90.64% for 30 labels. It can be trained for more number of datasets to

enhance accuracy. The performance of the Speech to Sign Language Interpreter can be visualized through the confusion matrix as shown in Figure 7. The X-axis represents the label which is to be predicted and Y-axis represents the true label. Analyzing the confusion matrix, we have obtained that each class in the X-axis has a higher probability that it is matched with its own class in the Y-axis. Hence, the matched label traces its respective sign language, thus it is revealed on the screen. The results clearly illustrate that it is effective in converting the recorded voice into its respective sign language with quite good accuracy. Moreover, the accuracy of the Speech to Sign language model can be uplifted by implementing Recurrent Neural Networks alongside Convolutional Neutral Network.



Figure 6. User Interface of Speech to Sign Language Interpreter

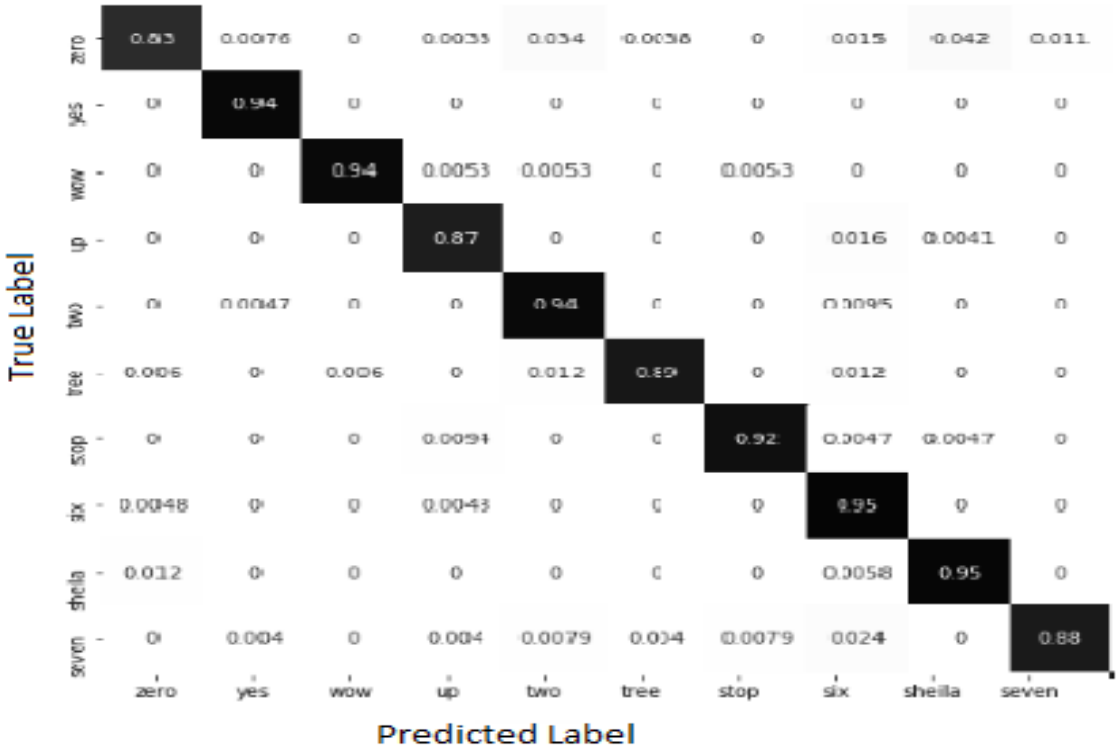


Figure 7. Performance of Speech to Sign Language Interpreter visualized through Confusion Matrix for 10 labels

4. Conclusion

The Speech to Sign Language Interpreter system mainly aims at eliminating the requirement of old fashioned interpreter (real person) which might be costly and might not be available always. They also may cause a delay in the reach of information. The Speech to Sign Language Interpreter system intakes input speech and delivers the respective sign language gif. It assists the active participation of deaf people in their day to day professional and personal lives and doesn't limit their participation in any field of works. Furthermore, it eliminates the necessity of attending expensive Sign Language classes, as it can be learned at home easily using the developed system. The model provides us with good accuracy and can also serve as a means to provide basic education to children. Moreover, it can further be implemented in various fields such as gaming, simulations, and various other sectors.

In the near future, the Speech to Sign Language Interpreter system will be deployed as a mobile application which enables the user to possess quick access to the system. For further improvement, more datasets can be trained and various other deep learning techniques can be implemented.

Acknowledgement

This research has been great learning experience for all of us. We thank our Research Department of Advanced College of Engineering and Management. Also, we would like to thank our colleagues and professors from Advanced College of Engineering and Management who provided us insight and expertise that assisted to progress our research. We are immensely grateful to the Department of Electronics and Communication, Advanced College of Engineering and Management for sharing their wisdom and offering this great opportunity.

References

[1] Kasaei, H. M., & Kasei, M. M. (2014). Development an Automatic Translate Voice to Sign Language Animation Based-on MFCC and Vector Quantization Method. International Journal of Computer and Electrical Engineering

[2] Zhang, Z. (2016). Derivation of Backpropagation in Convolutional Neural Network(CNN)

[3] Indolia, S., Goswami, A. K., Mishra, S. P., & Asopa, P. (2018). Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach. International Conference on Computational Intelligence and Data Science, 132(1), 679-688

[4] Arel, I., Rose, D., & Karnowski, T. (2010). Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]. IEEE computational intelligence magazine, 5(4), 13-18.

[5] Nebauer, C. (1998). Evaluation of convolutional neural networks for visual recognition. IEEE Transactions on Neural Networks, 9(4), 685- 696.

[6] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86 (11), 2278-2324.

[7] Zhou, Y., Wang, H., Xu, F., & Jin, Y. Q. (2016). Polarimetric SAR image classification using deep convolutional neural networks. IEEE Geoscience and Remote Sensing Letter, 13(12), 1935-19.

[8] Dai, D., Yu, L. & Wei, H. (2019) Parameters Sharing in Residual Neural Networks. Neural Process Lett 51, 1393–1410