Journal of Multidisciplinary Engineering Science Studies (JMESS)
ISSN: 2458-925X
Vol. 7 Issue 5, May - 2021

Cloud-Based Human Sign Language Digit Classification Using CNN: A Case Study of King's-Center, Akure, Nigeria.

Ajibike Eunice Akin-Ponnle
The Federal Polytechnic,
Ile-Oluji, Ondo state,
Nigeria.

Abstract— Gesture communication or sign language is very important to the hearing impaired to communicate among themselves. Also, the hearing impaired oftentimes need to relate with other people that are not with hearing difficulty, be it at, official; educational; social; or recreational gatherings. However, it is not everyone that understands how to sign, and this oftentimes leads to communication barrier between the hearing-impaired and the normal people. Therefore, building machines and/or systems to classify gestures, such that the hearing and the hearing-impaired can be enabled to relate and understand themselves becomes very imperative. This led to the interesting subject of gesture communication classification known as 'sign language'. In this study, a method of Machine Language Algorithm, which is deep learning, was adopted to classify and arrive at convenient model, by using a holistic convolutional neural networks (CNNs) to recognise digits shown by human hands by an interpreter in a worship centre in Nigeria. The CNN is known for both extracting a comprehensive feature representation from the input image and learning a classifier for each desired output simultaneously. How to implement a CNN for hand sign language digit classification from scratch in Google Colab and train the model in cloud environment is described in this study. The cloud-based training permits the use of the powerful GPUs for training, which reduces the training time in comparison with training on CPU. Experimental results show 93.5% accuracy on the test set.

Keywords— cloud-based deep model training; communication barrier; convolution neural networks; digit recognition ; hand sign language.

I. INTRODUCTION

Human hand sign known as gesture originates as a manner of natural communication which was used not only for the hearing impaired but even for the normally hearing persons as well. This was based on the situation and premise surrounding the particular kind of communication that was being undertaking. In some situations, people enjoy using gesture communication and they do so intuitively and naturally. In any case, gesture could be generally enjoyable once the natural essence of sending and receiving information effectively with understanding from one person to another is properly established. Gesture has been an adopted manner of communication to the hearing impaired, among themselves. An example of gesture communication is as shown in figure 1.

Fig 1: *Hand Sign Representation of Digits from 0-9 left to right; top-to-down.*

Communication between people is said to be a process of information and idea sharing, either verbally or non-verbally. The non-verbal communication is defined to be silent communication with a person or among people without the use of speech in order to gain the audience of the listeners. [1] It is effective when both the sending and receiving channels are properly established. Hand sign language is considered a non-verbal communication in this study, and a scenario is drawn of 300 people; among whom are 60 hearing impaired, gathering together to worship at King's centre, Akure, Nigeria. This is taken as ratio of 1:4 of hearing impaired to the hearing persons in total audience. The challenges arising among this nature of congregant ranges from but not limited to, inability of the members of congregant who are hearing to be able to understand the hearing impaired and communicate with them freely; inability of the hearing-impaired to express themselves freely before the remaining member of the congregant; and, inability of the hearing-impaired to communicate their needs or worries to the leaders of the congregation.

Human hand sign language recognition is classified into alphabet and digit classification, each of which has critical applications for welfare to people with hearing impairment. A hand sign language digit recognition system should receive a hand image as the input, and yields a number between 0 to 9, corresponding to the given input image, while that of alphabet recognition will take a hand image describing alphabet as input and brings out a letter. In this study, a hand sign

language recognition is studied for the hearing impaired congregants described in this paper, with difficulty in carrying out specific instructions relating to numbers during the time of worship. As a scenario, figure 2 describes a chart of how effective the hearing impaired were able to understand how to open books to certain page numbers. For such a situation, there were 4 categories from A - to - D of the 60 people that could not verbally understand what the leader was saying, and if the leader could not gesture correctly.

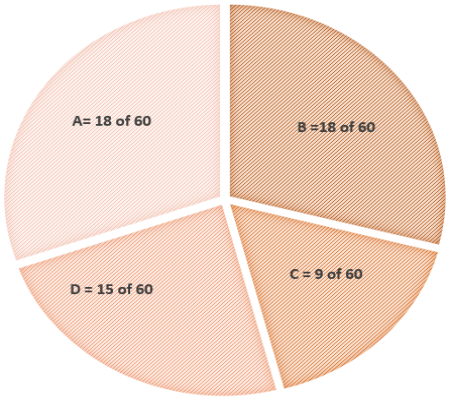

Fig 2: *Different level of Comprehension for Hearing Impairment*

Category 'A', opened their books to wrong page; category 'B', did not open at all because they could not understand and so they did not bother; category 'C', managed to open by following what other people around them did, so some were correct and others wrong depending on whether the examples they followed got it wrong or right; the last category, which is 'D', was able to open correctly. Hence it be can convincingly argue that the leader has successfully communicated with about 85% of the total number of people present effectively. This is as shown in figure 3.

Therefore, building a machine and /or system that could classify hand sign or gesture appropriately in order to enable the hearing and the hearing-impaired relate among themselves and also communicate with others effectively is very important. In this paper, a holistic convolutional neural networks (CNNs), as a type of deep learning, which is a method of Machine Language Algorithm, was adopted to classify and arrive at convenient model, to recognize digits shown by human hands by an interpreter at King's centre in Nigeria. Experimental results show 93.5% accuracy on the test set.

The remaining sections of this paper is grouped as; literature review; data description, visualization and statistical analysis; data pre-processing; methodology; results and discussion; and conclusion.

II. Literature Review

It is not everyone that understands how to sign, or how to correctly recognize gesture communication.

This has brought about many research on the subject of 'sign language', while trying to adopt various methods of Machine Learning Algorithm to classify and arrive at a convenient model.

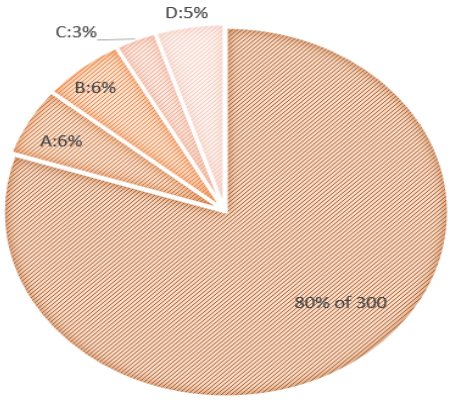

Fig 3: *Levels of Effective Communication at King's Centre.*

Gesture is a form of sign language; which could be presented in various other forms such as hand shapes, facial expressions, and movement of hands [2]. An example is the American Sign Language which has been primarily adopted as a means of communication for the deaf in countries such as Canada, United States of America, and other nations of the World [3]. However, gesture has to be correctly classified.

In an effort to create an effective means of communication between the hearing-impaired and the normal people, there is need to design automated systems that will enable the hearing impaired to communicate among themselves, as well as with other normal people since it is not everyone that understands how to sign.

The sign for digits is a type of 'sign language' as there are 10 digit numbers from figure '0' to '9' as can be seen in Fig 1. The sign digit could also be a form of hand rotation as presented by researchers in [3].

State of the Art. As a result of the quest to overcome the challenges of gesture classification and recognition, many studies have been carried out with the designs of electrical and mechanical machines as well as robots [4], [5]. In this study, there's review of some of the studies carried out to classify digits using CNN, and some other situations where CNN was used for classification generally, thereby, comparing and contrasting the accuracy presented by CNN from different studies. A 10 layer CNN to detect sign language of hand rotation at different angles was proposed by Abdul Kalam et. al. [3], whereby they prepared 7000 rotated images from 700 sign digit images using the proposed model and they were able to arrive at 97.28 percent accuracy for the 7000 rotated images. This is one of the recent related studies in literature. Suharjito et.al [2], reviewed extensively and reported several applications of CNN in computer vision projects to recognise sign language into texts or speech with different results. Certain accuracy as high as 95.5 percent was reported as part of result obtained by a group of researchers who used a 3D CNN for the creation of LipNet. This was a design made to perform lipreading by visual means.

This was used to recognize an end-to-end sentence lipreading. Also, it was reported in [2] that CNN was used for Italian sign language recognition whereby, 20 Italian gestures could be recognised with accuracy of 91.7 percent. The particular design of Hybrid CNN-Hidden Markov Model (CNN-HMM), which was used for continuous sign language recognition system and has sequence modelling capabilities was also presented. Likewise, studies about hand gesture recognition using CNN was carried out by M. Han et.al. in [6], where they propose a convolutional neural network that is biologically inspired, in an attempt to reduce the difficulty of gesture recognition from image of cameras. They adopted Gaussian skin model and background subtraction for CNN data testing and training to filter non-skin colours of image, thereby obtaining 93.8 percent classification rate from experiment. Yawei Hou and Huailin Zhao [7] in an attempt to improve recognition results for handwritten digit recognition, studied a combined depth network for CNN and BP neural network, and reported a more accurate result from the combined network recognition, than when they were separately implemented using the same dataset. Z. Lu et. al. [8] implemented one-shot learning hand gesture recognition with their 13D lightweight network using spatial-temporal separable 3D CNN. This is in an attempt to form a more satisfactory classification while handling very few samples or only one gesture class. Their model was tested using ChaLearn gesture dataset among others, and it was reported that they obtained a satisfactory result. Z. Hu and X. Zhu [9] propose the use of CNN approach for hand gesture recognition from RGB images in order to simulate a real time scenario where depth image for RGB- D (RGB image and depth approach; which has been so much studied in literature), is not available. Their experimental result was reported to have better accuracy in recognition. However, with all the reported studied on CNN in literature, there is yet not enough report on digit number classification in particular. Therefore, in this research, a model of hand digit number classification using CNN is presented.

III. METHODOLOGY

Convolution Neural Network (CNN) is a deep learning algorithm [2], and it is feed-forward neural network that is aided by the visual cortex of human containing convolution and subsampling layers which is found very helpful in the field of computing vision. Complexity of processing is usually reduced by the use of several kernels in each layer [3]

A. Data Description and Processing

In this study, the 'Sign Language Digits Dataset' that is employed comprises of 2180 number of images, collected from 218 students, with 10 images per person, with every Individual in this dataset only to have used their right hand to show digits between 0 to 9. The size of images is 100 X100 with RGB channels. Several samples of this dataset are given in figure 4.

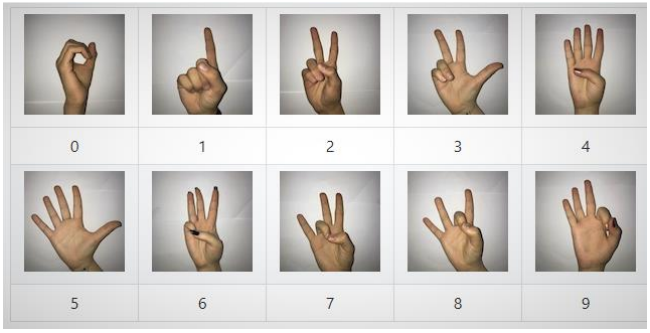

Fig 4: Sample Images from Sign Language Digits Dataset.

The Sign Language Digits Dataset are 100X100 images with consistent background, and the hands are located in the center of the image. The images are normalized from [0, 255] to [0,1] pixels, because CNN can learn better from a normalized dataset, [10].

In addition, to boost the performance of the model, several common data augmentation techniques such as rotation, width-shifting, height-shifting, shearing, zooming, were used. However, horizontal and vertical flipping were avoided as these augmentations deform the usual data, and since all the images on the dataset are taken from right hand, horizontal flipping cannot help the model to learn new features from the existing dataset. Also, while preparing the dataset, 20% of the data for test phase were randomly selected, and the rest were used for learning the model.

B. Recognition Process

The recognition process with a holistic CNN has two main steps:
1) extracting discriminative features from the input image, and 2) learning a classifier based on the image features and its ground truth label. For the first step, which is feature extraction, four convolutional layers followed by a linear activation function, and regularization layers, were used. Given that the problem is a multi-class classification task, an output layer with 10 nodes to predict the probability distribution of an image belonging to each of the 10 classes, was required. To this end, a fully connected layer with 10 neurons with softmax activation function was used. The detailed information of model summary is given in Fig. 5. The analysis is as follows:
1) All the activation functions in the model are Rectified linear unit (Relu) activation functions. The rectified linear activation function is a piece-wise linear function that output the input directly if is positive, otherwise, it output zero. This has become the default activation function for many types of neural networks because a model that uses it is easier to train and often achieves better performance.

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 64, 64, 32)	320
activation_1 (Activation)	(None, 64, 64, 32)	0
conv2d_2 (Conv2D)	(None, 62, 62, 32)	9248
activation_2 (Activation)	(None, 62, 62, 32)	0
max_pooling2d_1 (MaxPooling2	(None, 31, 31, 32)	0
dropout_1 (Dropout)	(None, 31, 31, 32)	0
conv2d_3 (Conv2D)	(None, 31, 31, 64)	18496
activation_3 (Activation)	(None, 31, 31, 64)	0
conv2d_4 (Conv2D)	(None, 29, 29, 64)	36928
activation_4 (Activation)	(None, 29, 29, 64)	0
max_pooling2d_2 (MaxPooling2	(None, 14, 14, 64)	0
dropout_2 (Dropout)	(None, 14, 14, 64)	0
flatten_1 (Flatten)	(None, 12544)	0
dense_1 (Dense)	(None, 512)	6423040
activation_5 (Activation)	(None, 512)	0
dropout_3 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 10)	5130
activation_6 (Activation)	(None, 10)	0

Total params: 6,493,162
Trainable params: 6,493,162
Non-trainable params: 0

Fig. 5: Model summary. Red colored lines indicate the convolutional layers and blue line shows the fully connected layers with 10 output nodes.

2) Several drop out layers were used in the model as regularization layers to prevent the model from early overfitting.
3) A flatten layer to provide the features as a vector to the classifier was used.
4) Max pooling layer was used to compress the features in each level of the model and finally the most important feature representation of the given input image was extracted
5) Parameters of the model are the learning weights, which will be updated in each epoch of training phase. Final weights are used to predict the test set and evaluate the accuracy performance of the model. Google Colab † was adopted to train and test the proposed model for hand sign language digit recognition. Google Colab is a free cloud service based on Jupiter Notebook that supports free GPU to train and test the model (in a 12-hour active session).

After connecting to a run-time, the GPU was selected as run-time type and found it using the Listing 1:

\begin{lstlisting}
[language=Python, caption=Python example]

1. import tensor flow as tf
2. device_name = tf.test.gpu device_name ()
3. if device_name != '/device:GPU:0':
4. raise system error('GPU device not found')
5. print ('Found GPU at: {}'.format(device name))

\end{lstlisting}

Listing 1: Python example

The dataset from the Github was then downloaded using the Listing 2.

\begin{lstlisting}
[language=Python, caption=Python example]

1. !git clone https://github.com/ardamavi/Sign-Language-Digits-Dataset

\end{lstlisting}

Listing 2: Python example

The model was trained for two different settings of augmentation. When basic augmentation like several degrees of rotation, shearing, zooming, random cropping, etc., were applied, it was observed that there was 1:5 percent improvement (from 92:0 to 93:5) in the test performance.
As shown in Fig 6, when the model is trained for more than150 epochs, it starts to be over-fitted to the training subset. Consequently, the validation loss is not improved. To avoid early over-fitting, firstly, 0.5 drop out was used, in which half of the learned values in the layer was dropped. This way, model automatically will focus on more critical features. Secondly, learning phase was stopped in epoch 150. However, based on Fig 6, training can be stopped in epoch 200 also as the difference between validation loss and training loss is negligible.

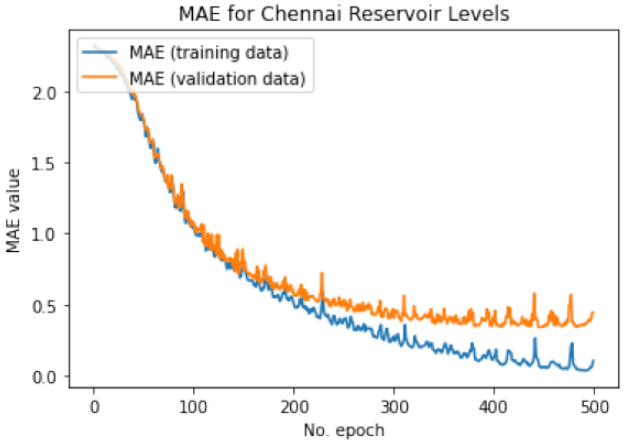

Fig. 6: Loss value visualization in the course of training.

CONCLUSION

In this work, a CNN-based model was implemented to extract and classify a feature representation from the used input hand image. As the dataset is small and easy (i.e., with consistent background), a very shallow model with only 4 convolutional layer was applied. However, if large datasets could be collected, deeper models (such as Inception, ResNet, MobileNet, etc.) can be used, without any concern about over-fitting to the training data. Consequently, one can learn from larger datasets and extend the generalization ability of the model, which is a critical feature in real-world applications as such that has been described by the group of congregant in this paper.

REFERENCES

[1] D. Phutela ‘‘The Importance of Non-Verbal Communications’’, IUP Journal of Soft Skills, vol. 9, no 4, pp 43, 2015

[2] Suharjito, M.C. Ariesta, F. Wiryana and G.P. Kusuma ‘‘A Survey of Hand Gesture Recognition Methods in Sign Language Recognition”, *Pertanika J. Sci. Technol*. vol. 26, no 4, pp 1659 – 1675, 2018.

[3] A. Kalam, N. I. Mondal, and B. Ahmed, “Rotation Independent Digit Recognition in Sign Language”, *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, February 2019.

[4] A. B. Jmaa, W. Mahdi, Y. B. Jemaa, and A. B. Hmadou “A New Approach for Digit Recognition Based on Hand Gesture Analysis”,*International Journal of Computer Science Security*, vol. 2, no 1, 2009.

[5] S.K. Keshari, S.Tyagi, N. Tomar, and S.Goel “Aphonic’s voice: A Hand Gesture Based Approach to convert Sign Language to speech”, *IEEE 2nd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, 2019.

[6] M. Han, J. Chen, L.Li, and Y. Chang, “Visual Hand Gesture Recognition with Convolution Neural Network”, *17th IEEE/ ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing (SNPD)*, 2016.

[7] Y. Hou and H. Zhao, “Handwritten Digit Recognition Based on Depth Neural Network”, *International Conference on Intelligent Informatic and Biomedical Sciences (ICIIBMS)*, 2017.

[8] Z. Lu, S. Qin, L. Li, D. Zhang, K. Xu, and Z. Hu, “One-Shot Learning Hand Gesture Recognition Based on Lightweight 3D Convolutional Neural Networks for Portable Applications on Mobile Systems”, *IEEE Access*, pp 131732 – 131748, September 2019.

[9] Z. Hu and X. Zhu, “Gesture detection from RGB hand image using modified convolutional neural network”, *IEEE 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pp 143-146, 2019.

[10] https://github.com/ardamavi/Sign-Language-Digits-Dataset, visited on 30th June 2020.