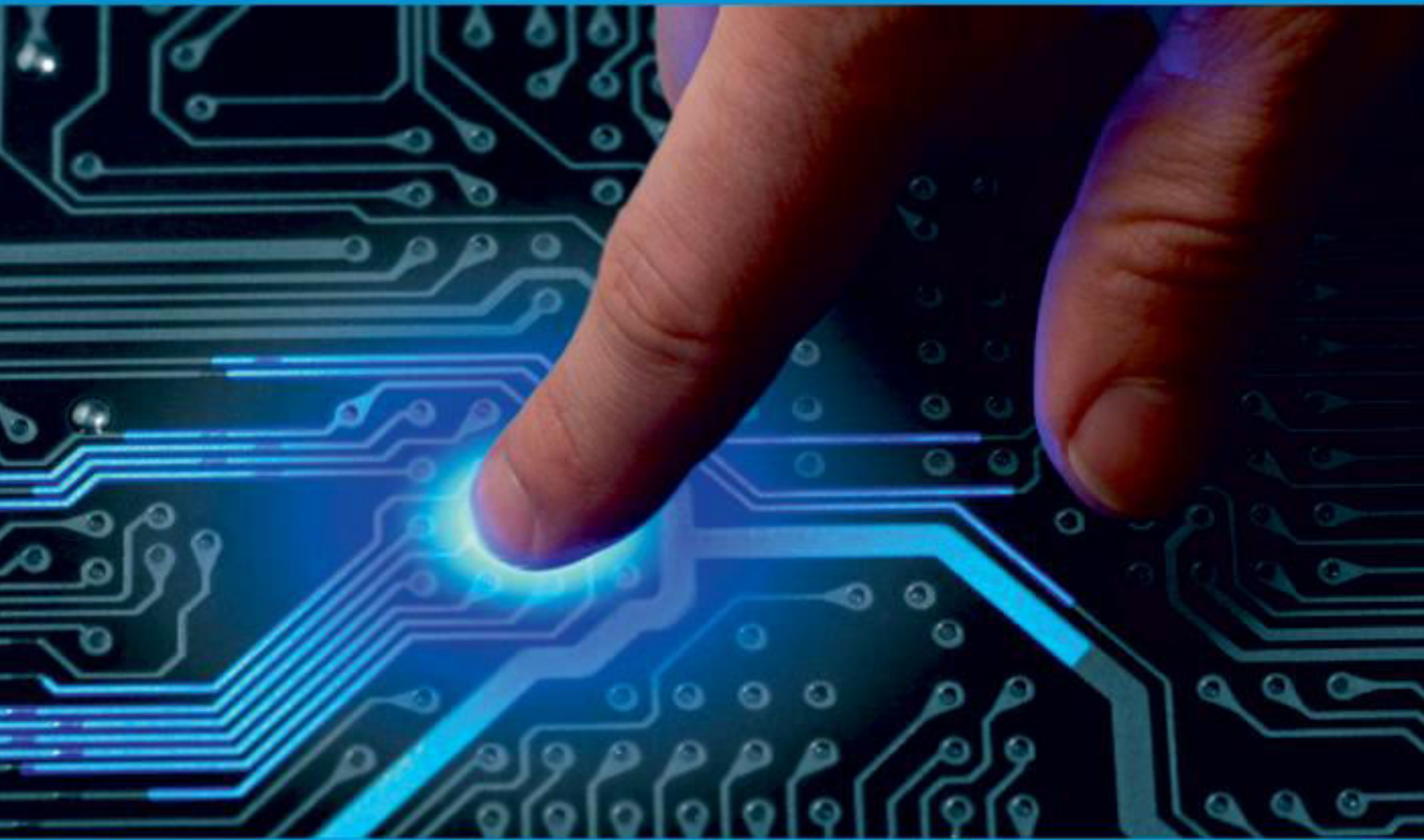
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005086|

Sign Language Translator using Deep Learning

B. Evans Nikith Royan, Darryl Andrade, Davin S. Thomas, Gyandwip Das, Debjit Das, Prof. Rashmi Mothkur

UG Student, Dept. of CS&E., Dayananda Sagar University, Bangalore, India

UG Student, Dept. of CS&E., Dayananda Sagar University, Bangalore, India

UG Student, Dept. of CS&E., Dayananda Sagar University, Bangalore, India

UG Student, Dept. of CS&E., Dayananda Sagar University, Bangalore, India

UG Student, Dept. of CS&E., Dayananda Sagar University, Bangalore, India

Assistant Professor, Dept. of CS&E., Dayananda Sagar University, Bangalore, India.

ABSTRACT: Human beings need to communicate with one another for a variety of reasons to coexist and survive. Unfortunately, millions of people worldwide lack the ability to speak or hear, either from birth or as a result of some accident. For people who suffer from these disabilities, it is difficult to communicate with the rest of the population. Hence, Sign language was created, which relies on physical gestures to represent letters, words or even sentences. A combination of these signs/gestures help a person to convey their message to another person. These signs/gestures, however, are not known by most of the general public, who will most definitely find it impossible to understand what another person is saying in sign language. The general public most often has no immediate need to learn sign language or has no motivation to do so. Hence there is a need for a Sign Language Translator, that would convert these gestures and movements into words and alphabets that are understandable to people who do not know sign language. 5 models, namely a basic convolutional network, ResNet50, ResNet50V2, Inceptionv3 and VGG16 were developed with CNN and Transfer Learning methodologies to translate sign language gestures in real time.

KEYWORDS: Sign Language, Convolutional Neural Networks, Transfer Learning, ResNet50, ResNet50V2, InceptionV3, VGG16.

I. Introduction

A person who is incapable of speaking can communicate by writing, typing or sign language. Regular speech takes place spontaneously, in real time. Writing and typing require one person to complete their communication before the other can access it. Therefore, Sign language is the most spontaneous. Sign Language is how people who are hearing and speech impaired would express their feelings, contribute to a conversation and communicate in general. It is still a language of its own, and should be viewed as so. It is a form of communication, and is just as important as learning Spanish or French or any other language.

The problem here is that not many people know sign language. It is mainly confined to the affected and the people who interact with them on a daily basis. So, a translator is required. Who better to translate than a computer? The goal of this implementation is to allow a person to convey their signs into a camera and the computer will translate it to text for the recipient to understand in real time. In the recent past, research has found that deep learning techniques are very effective in image processing applications. The images of each sign can be translated using an appropriate model. This implementation develops various deep learning models and evaluates their performance in the same task.

II. RELATED WORK

Literature review of the problem statement shows that there have been many attempts at sign language translation and for different sign languages as well. In [1] the authors scaled all images to 50x50 and trained it on a multi layered CNN with 5 layers and leaky ReLU activation. This proposed method used both static(0-9 and A-Z) and dynamic(alone, afraid, anger, etc) gestures in training, validation and blind testing to make the model more robust. The blind test accuracy obtained in this proposed methodology was found to be 99.89%.

Md. Jahangir Hossein et al [2] published a paper on Recognition of Bengali Sign Language by first developing a CNN architecture consisting of 22 layers. Then proposed a vision-based method for classification, featuring deep learning methods and adopted data augmentation. Finally, the authors used SIFT based methods with binary SVM classifiers. The dataset used in this implementation covers 50 sets, 36 necessary Bengali signs and was obtained with the cooperation of deaf volunteers of multiple organizations. The multilayer model achieved a 94.88% of accuracy, the adopted data augmentation achieved an accuracy of 92% and the SIFT based model accomplished an accuracy of about 98%.

Transfer learning is beneficial and can drastically reduce training periods and does not need the manual creation of models from scratch. Manual Eugenio Morocho Cayamcela et al. [3] used transfer learning and fine tuning methodologies applied to pre-trained networks AlexNet and GoogleNet that were trained on ImageNet dataset. The discriminative filters from the pre-trained models were reused on a custom dataset.

Aditya Das et al. [4] proposed a model that used InceptionV3 which stacks various layers of a CNN model in parallel instead of one on top of the other. This model achieved an average accuracy of 90% on static sign language.

Sign Language Recognition using deep learning and computer vision carried out by Kshitij Bantupali et al. [5] used an ensemble of CNN and LSTM. The CNN model, Inception was used to extract spatial features and LSTM whereas a RNN model was used to extract temporal features from the video sequences. Gesture segments identified and processed by CNN and classified by LSTM(sequence data). The results for varying sample sizes 10, 50, 100, 150 the accuracy of softmax were 90%, 92%, 93% and 91% and for Pool layer, accuracy obtained were 55%, 58%, 58% and 55% respectively.

In [6] Murat Taskiran et al. developed a model on 80:20 split and used skin colour and convex hull algorithm to determine the position of the hand. The model made use of a sequential network. Real time algorithm used in this implementation consists of extracting hand bound convex hull points and classifying hand images with CNN. This model was made to operate in real time and each character, 36 in total (10 numbers and 26 alphabets), were tested 10 times giving an accuracy of 98.05%.

Saleh Ahmad Khan et at. [7] made use of Scikit learn label encoding on their custom handmade dataset which provided less Root Mean Square error than One Hot Encoding method. They also used customized ROI segmentation to determine the region of interest for the hand to be captured. By using Scikit learn encoding, their model was able to achieve an accuracy of 86.4% to 97.54%. The customized ROI segmentation provided lesser computation work for locating the hand.

A multilayer Recurrent Neural Network for sign detection was proposed by Mark Borg et al. [8]. Features from a two stream Convolutional Neural Network taking video image data and motion data was taken as input. This proposed system consisting of a two stream RNN compares against the baseline research and was found to give at least a 18% increase against baseline accuracy of 78%.

In [9] Jinalee Jayeshkumar Raval et al. proposed a two segmented approach. First, the image processing is done by applying Gaussian blur to the region of interest, this forms a skin colour mask. Then a series of dilation and erosion were applied to enhance and smoothen the required features. CNN was used to identify the vivid features of the hand and ReLu was used to remove the negatives, Softmax was used to provide an accurate position. After training this model on an 85:15 split the model was tested on different illuminations and backgrounds in real time and provided an 83% accuracy.

Siming He et al. [10] made use of methods like Faster R-CNN used to recognise and locate the hand from the video input and a 3D CNN feature extraction network and a LSTM network based on RNN were constructed to improve the accuracy by learning the context of sign language. To improve the colourtone (RGB signs) problem, all the methods were combined to build a recognition model. Upon testing and training this fusion method in a real time environment the recognition rate is a high 99%.

The paper [11] chose four different models with convolutional neural networks as their basis. A simple CNN and three transfer learning models (VGG16, VGG19, and InceptionV3) pre-trained on the ImageNet dataset. This model obtained an accuracy of 97% and above on all four models, which were trained on 100 epochs.

A paper on Deep Residual Networks by Kaiming He et al. [12] adopts batch normalization after each convolution and before each activation which were done after scale augmentation, with per pixel mean subtracted and standard colour augmentation. Resnet reduced error rate by 3.57% in top-5 error on test set compared to VGG V5 (6.8%) and GoogleNet (6.66%).

A three network pipeline was adopted by Shruti Mohanty et al. [13] which utilizes a dataset annotated with ground truth 3D key points for the signs. The networks employed in this implementation are HandSegnet, Posnet and Poseprior. An error rate of 0.58 with SVM classifier was observed, on extracting additional features with SVM a reduced error rate of 0.39 was obtained. Transfer learning approach reduced error rate even further to 0.37 and on using InceptionV3 a final error rate of 0.36 was obtained.

Gautham Jayadeep et al. [14] implemented a vision based classification system having dynamic ISL signs for bank purposes. This implementation made use of a self-customized dataset from ISL dictionary. CNN was used for feature

| DOI: 10.15680/IJIRCCE.2022.1005086|

extraction and passed on to an LSTM model and finally output the signs to text form. The dataset resulted in higher accuracy compared to other categories. It was observed that the model gave accuracy of 100% during training.

An Inception V3 network model with TensorFlow to retrain the final layers of the Inception model was adopted by Xiaoling Xia et al. [15] to classify 17 species of flowers. Transfer learning was used to keep parameters of the previous layer and remove the last layer of the Inception V3 model, then retrain the last layer. This methodology achieved an accuracy of 95% on Oxford-17 flower dataset and 94% on Oxford-102 dataset

III. METHODOLOGY

3.1 *Dataset*

The data set is a collection of images of alphabets from the American Sign Language, separated in 29 folders which represent the various classes. The training data set contains 87,000 images which are 200x200 pixels. There are 29 classes, of which 26 are for the letters A-Z and 3 classes for SPACE, DELETE and NOTHING. The 3 classes are very helpful in real-time applications and classification. Each class has 3000 image samples. The test data set contains a mere 29 images, to encourage the use of real-world test images. The dataset includes image samples with variations in lighting conditions and skin tones as well to help make the model more robust.

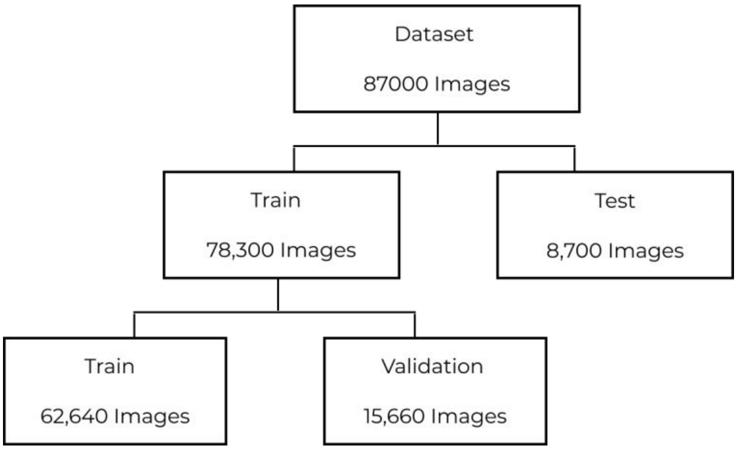


Fig 1. Dataset Split

3.2 *Deep Learning Models:*

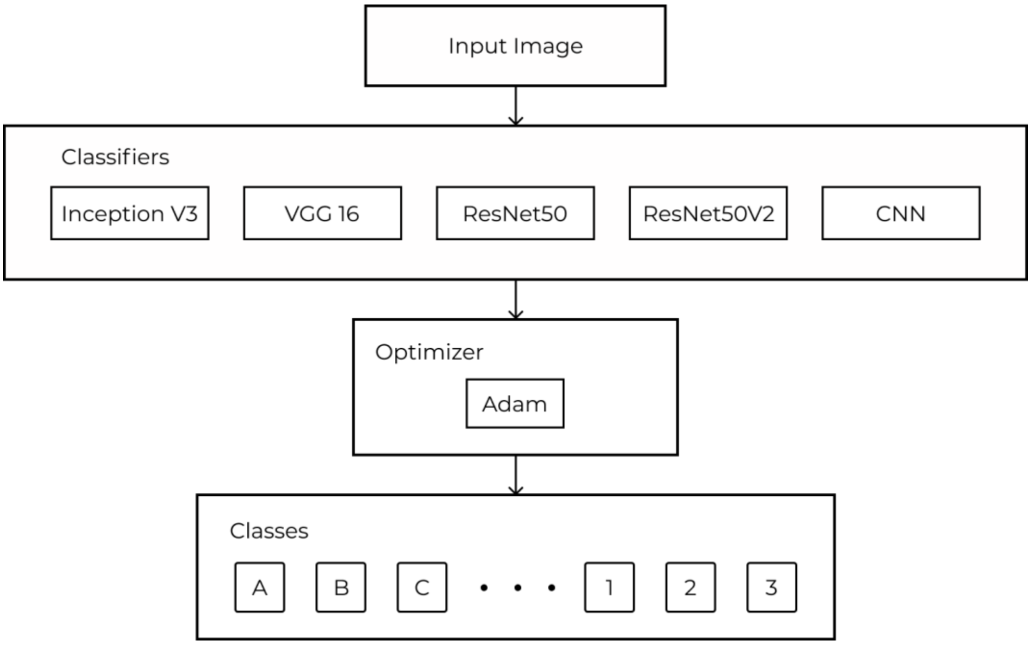


Fig 2. Model Architecture

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005086|

3.2.1 Convolutional Neural Networks

Convolutional Neural Networks popularly known as CNNs are one of the oldest deep neural networks that are widely used in computer vision tasks. They are composed of multiple building blocks such as, convolutional layers, pooling layers and fully connected layers. CNNs take in an input image, assign importance to different aspects and features within the image and are able to distinguish such images from others. CNNs form the basis of the various deep learning models that are implemented for this application.

3.2.2 Transfer Learning

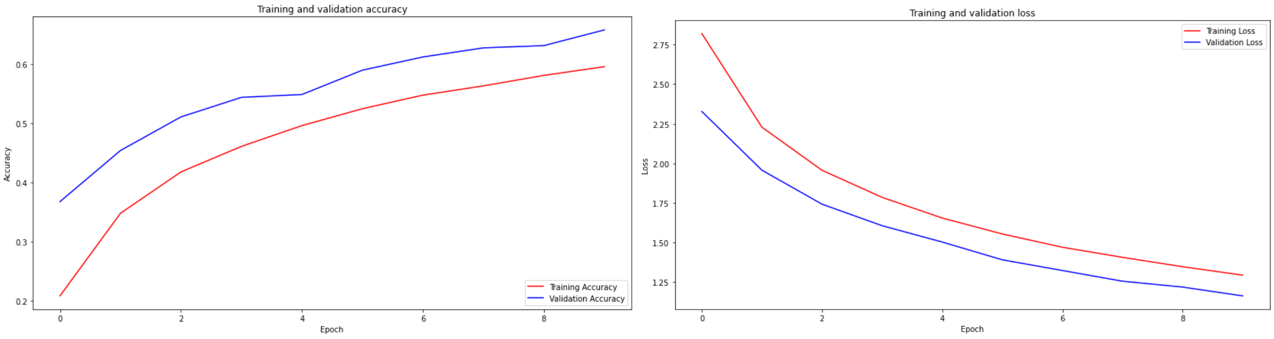
Transfer Learning is the reuse of a model on a problem that is different from the original one that it was previously developed and trained for. Transfer Learning allows us to use previously gained knowledge to newer problems without having to start from scratch. The benefits include shorter training times to develop models that work well in most cases.

- ResNet50
  ResNet-50 is a deep residual network. The “50” refers to the number of layers it has. It’s a subclass of convolutional neural networks, with ResNet most popularly used for image classification.
- ResNet50V2
  ResNet50V2 is a modified version of ResNet50 that performs better than ResNet50 and ResNet101 on the ImageNet dataset. In ResNet50V2, a modification was made in the propagation formulation of the connections between blocks. ResNet50V2 also achieves a good result on the ImageNet dataset.
- Inception V3
  The Inception V3 is a deep learning model based on Convolutional Neural Networks, which is used for image classification. The inception V3 is a superior version of the basic model Inception V1 which was introduced as GoogleNet in 2014. As the name suggests it was developed by a team at Google.
- VGG16
  VGG16 is a simple and widely used Convolutional Neural Network (CNN) Architecture used for ImageNet, a large visual database project used in visual object recognition software research. It won ILSVR (ImageNet) in 2014.

The transfer learning models are developed using the above pre-trained models and the weights obtained on training them on the ImageNet dataset. The proposed approach involves removing the existing fully connected layers from the above pre-trained models and replacing them with another set that would cater to this specific application. The base models would be frozen and hence would not be able to update weights. The models are trained on images from the specified dataset.Fine-tuning these pre-existing models in this fashion allows us to classify images for our specific application.

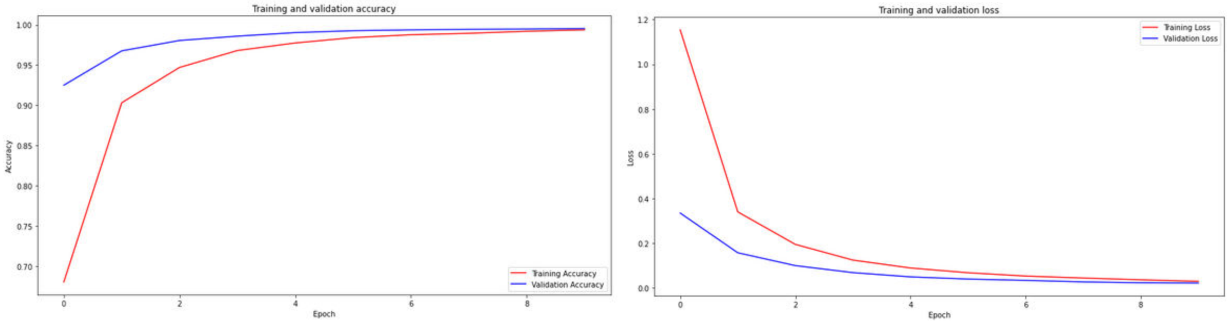
IV. **RESULTS**

The dataset consists of 87,000 images divided into 29 classes. The were 26 classes, one for each letter of the alphabet and 3 additional classes for “space”, “nothing” and “delete”. The models were trained on 62640 images. 15660 images were used for validation and 8700 images of the dataset were used for testing. Kaggle Notebooks were used for training, validating and testing the models. They were trained for 10 epochs each. The graphs depicting the validation accuracy againsttraining accuracy and validation loss against trainingloss of the models are shown in Fig.3, Fig.4, Fig.5, Fig.6 and Fig.7.



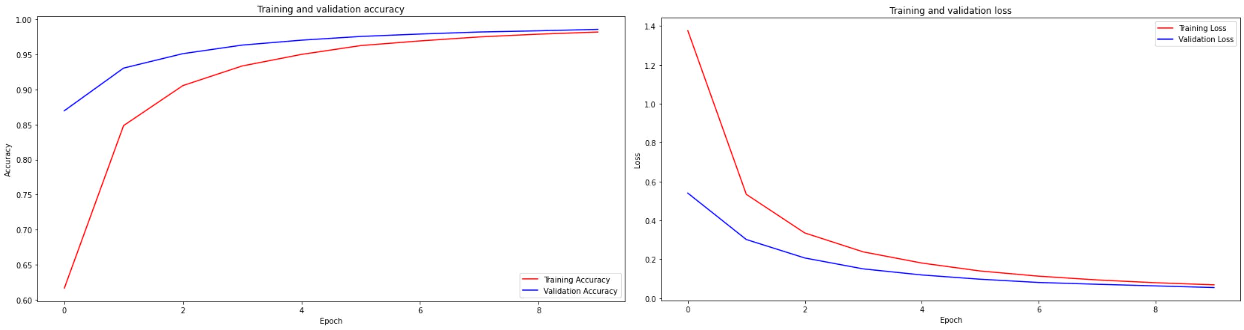
(a) (b)

Fig. 3 Graph of (a) Validation Accuracy against Training Accuracy (b) Validation Loss against Training Loss for ResNet50



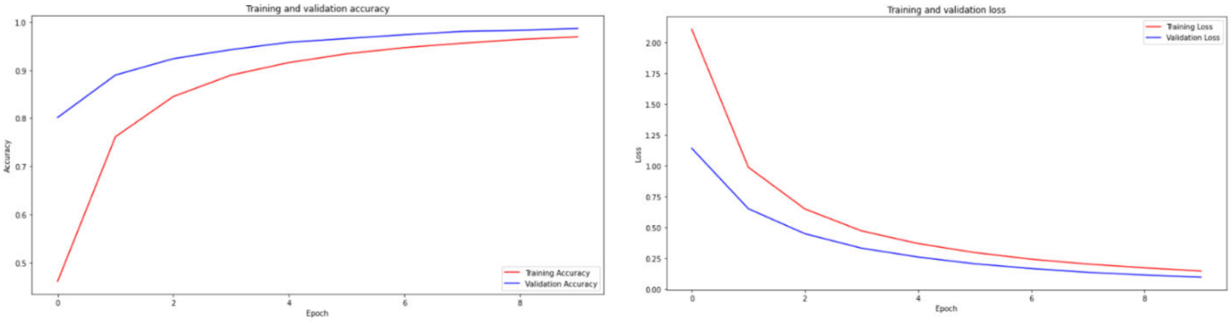
(a) (b)

Fig. 4 Graph of (a) Validation Accuracy against Training Accuracy (b) Validation Loss against Training Loss for ResNet50V2



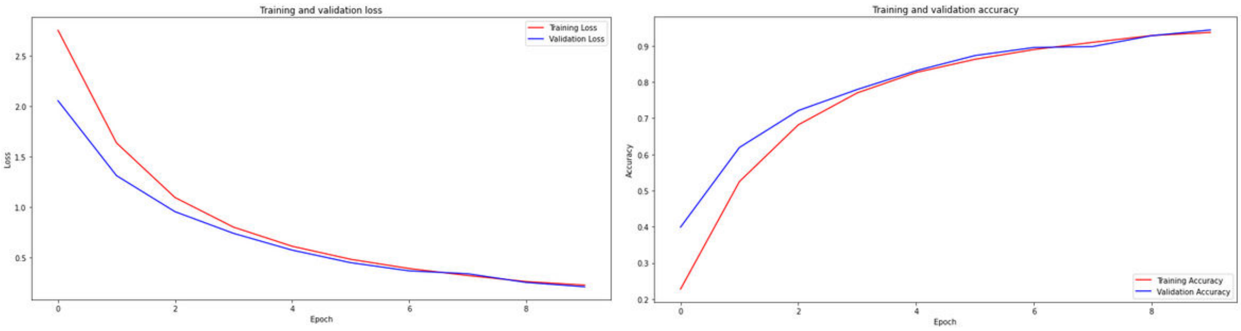
(a) (b)

Fig. 5 Graph of (a) Validation Accuracy against Training Accuracy (b) Validation Loss against Training Loss for InceptionV3



(a) (b)

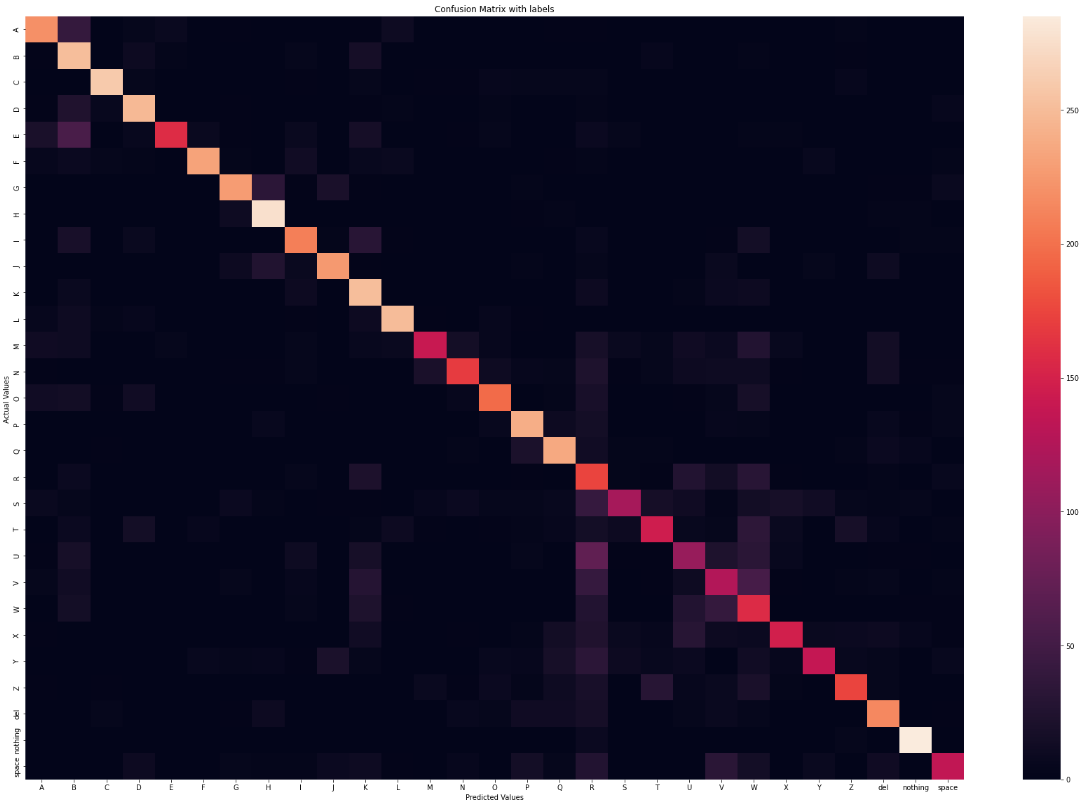
Fig. 6 Graph of (a) Validation Accuracy against Training Accuracy (b) Validation Loss against Training Loss for VGG16



(a) (b)

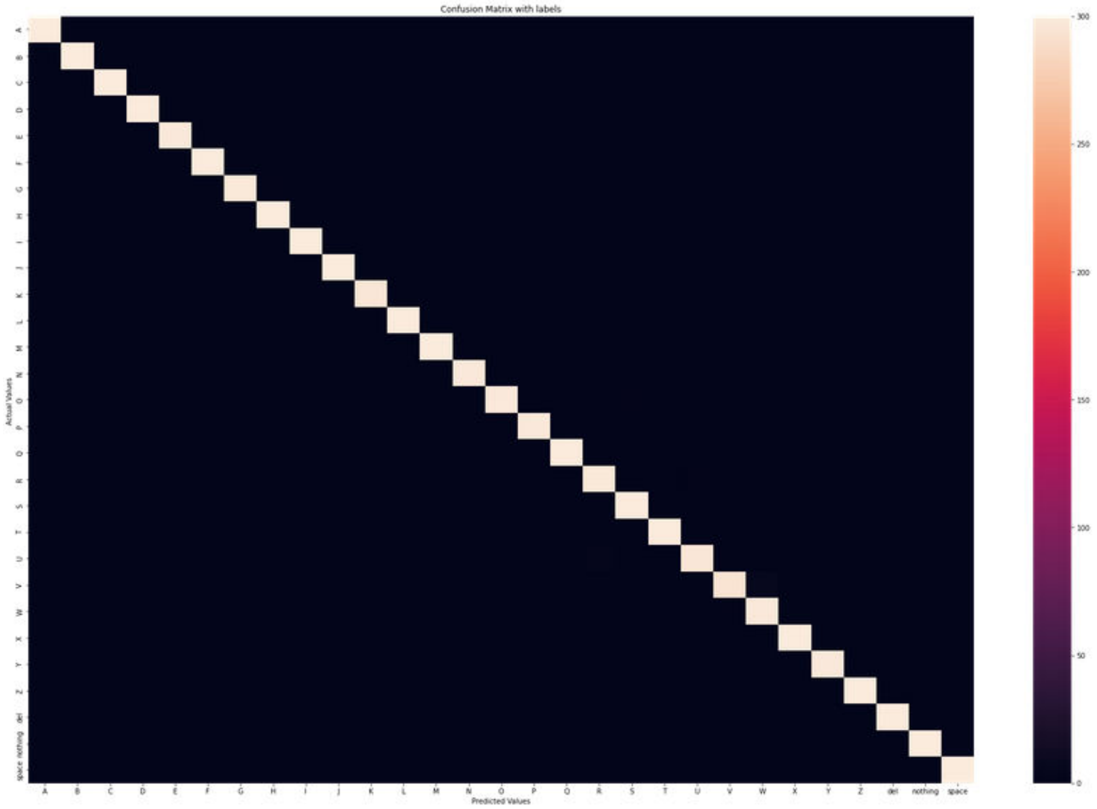
Fig. 7 Graph of (a) Validation Accuracy against Training Accuracy (b) Validation Loss against Training Loss for Simple CNN

Confusion matrices were also plotted by comparing the values obtained from the models with the ground truth as shown in Fig. 8.

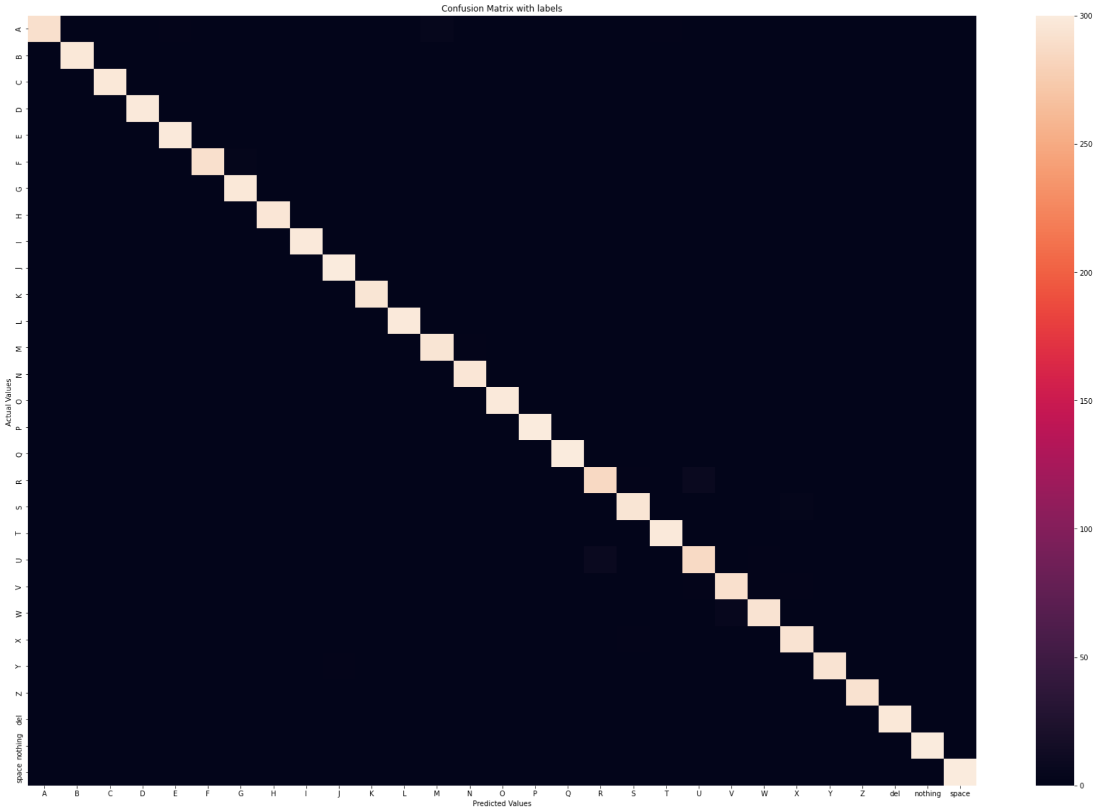


(a)

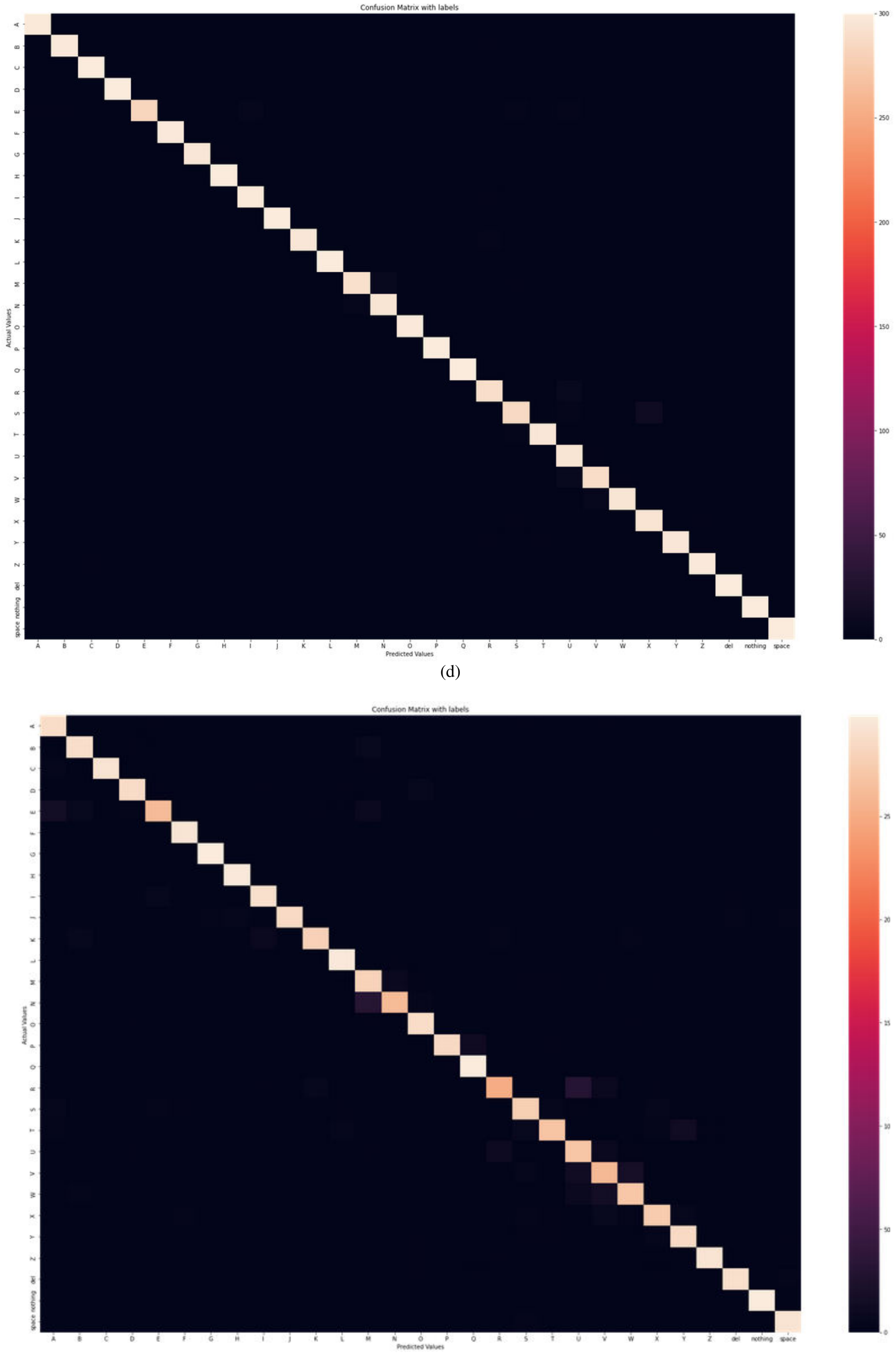
| DOI: 10.15680/IJIRCCE.2022.1005086|



(b)



(c)



(d)

(e)

Fig.8 Confusion Matrices of (a) ResNet50 (b) ResNet50V2 (c) InceptionV3 (d) VGG16 (e) Simple CNN

The results obtained were compared as shown in Table 1.

	Training Accuracy	Validation Accuracy	Test Accuracy
Resnet50	59.35%	64.95%	65.0%
Resnet50V2	99.37%	99.54%	100%
InceptionV3	98.21%	98.61%	99.0%
VGG16	96.91%	98.67%	99.0%
CNN	93.71%	94.39%	94%

Table 1. Results Summary of all models

V. CONCLUSION

The deep learning models proposed in this study for the translation of sign language have significant results. ResNet50V2, InceptionV3 and VGG16 performed the best, the Simple CNN model performed decently and ResNet50 performed poorly. The deep learning models were able to efficiently extract features from the images of various sign language gestures and thus can be utilizedto develop a translator that can output a fully understandable translation on being fed sign language gestures. As a result, it would help in the development of a sign language translator that would be able to assist Non Sign Language Speakers to be able to understand what is being said by Sign Language Speakers. In-turn helping the population of Sign Language speakers to communicate and be understood easily.

REFERENCES

1. Rajarshi Bhadra and Subhajit Kar in 2021, “Sign Language Detection from Hand Gesture Images using Deep Multi-layered Convolution Neural Network” in IEEE Second International Conference on Control, Measurement and Instrumentation (CMI).
2. Md. Jahangir Hossein and Md. Sabbir Ejaz - “Recognition of Bengali Sign Language using Novel Deep CNN.” in 2020 2nd International Conference on Sustainable Technologies for Industry 4.0.
3. Manuel Eugenio Morocho Cayamcela and Wansu Lim in 2019, “Fine-tuning a pre- trained Convolutional Neural Network Model to translate American Sign Language in Real-time” in International Conference on Computing, Networking and Communications (ICNC).
4. Aditya Das, Shantanu Gawde, Khyati Suratwala1 and Dr. Dhananjay Kalbande in 2018, “Sign Language Recognition Using Deep Learning on Custom Processed Static Gesture Images” in International Conference on Smart City and Emerging Technology (ICSCET).
5. Kshitij Bantupalli and Ying Xie in 2018, “American Sign Language Recognition using Deep Learning and Computer Vision” in IEEE International Conference on Big Data (Big Data).
6. Murat Taskiran, Mehmet Killioglu and Nihan Kahraman in 2018, “A Real-Time System For Recognition Of American Sign Language By Using Deep Learning” in 41st International Conference on Telecommunications and Signal Processing (TSP).
7. Saleh Ahmad Khan, S. M. Asaduzzaman, Amit Debnath Joy, and Morsalin Hossain in 2019, “An Efficient Sign Language Translator Device Using Convolutional Neural Network and Customized ROI Segmentation” in 2nd International Conference on Communication Engineering and Technology (ICCET).
8. Mark Borg and Kenneth P. Camilleri in 2019, “Sign Language Detection “In The Wild” with Recurrent Neural Networks” in ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
9. Jinalee Jayeshkumar Raval and Ruchi Gajjar in 2021, “Real-time Sign Language Recognition using Computer Vision” in 3rd International Conference on Signal Processing and Communication (ICPSC).
10. Siming He in 2019, “Research of a Sign Language Translation System Based on Deep Learning” in International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM).
11. Gaurav Labhane, Rutuja Pansare, Saumil Maheshwari, Ritu Tiwari and Anupam Shukla in 2020, “Detection of Pediatric Pneumonia from Chest X-Ray Images using CNN and Transfer Learning” in 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE-2020)

12. Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun in 2016,"Deep Residual Learning for Image Recognition" in 2016 IEEE Conference on Computer Vision and Pattern Recognition
13. Shruti Mohanty, Supriya Prasad, Tanvi Sinha and B. Niranjana Krupa in 2020, "German Sign Language Translation using 3D Hand Pose Estimation and Deep Learning" in 2020 IEEE REGION 10 CONFERENCE (TENCON)
14. Gautham Jayadeep, N.V. Vishnupriya, Vyshnavi Venugopal, S. Vishnu, M. Geetha in 2020, "Mudra: Convolutional Neural Network based Indian Sign Language Translator for Banks" in 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)
15. Xiaoling Xia, Cui Xu & Bing Nan. In 2017 "Inception-v3 for flower classification" in 2017 2nd International Conference on Image, Vision and Computing (ICIVC).

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING