CV for American Sign Language

Bhavya Samhitha Mallineni
mallinen@usc.edu

Dharitri Hemant Toshikhane
toshikha@usc.edu

***Abstract*— We propose to create a computer vision model to recognize the American Sign Language. Applying deep learning to a dataset of 29 classes including 26 English alphabets and 3 characters of SPACE, DELETE, and NOTHING, we intend to accurately interpret and translate gestures of ASL, using 3 different models of ResNet architecture.**

Keywords—American Sign Language (ASL), Computer Vision (CV), ResNet, Neural Network, CNN, ANN, MLP

I. INTRODUCTION

Roughly 10 million individuals are classified as hard of hearing, and approximately 1 million people are considered functionally deaf [1]. The Deaf community in the United States of America, primarily uses the American Sign Language. Various MNC's can improve their platform by integrating deep model techniques to enable convenient access to communication for people who are recognized as deaf. The model can also be applied in industries such as Health care and Education to improve opportunities and reduce the difficulties faced while communicating by hard of hearing people. The proposed model is based on computer vision that shall include convolution neural networks. CNNs excel in identifying patterns and are widely used in image processing related models. The CNN models accept the input images or feature maps and pass it through the convolution layer, where the data is filtered and ReLU is applied. The neurons have a weight and bias that can be adjusted according to the requirement. The final result is compared with the target values and Loss Function is generated. Optimization algorithms such as SGD are used to minimize the loss.

II. LITRETURE REVIEW

A. Sign Language Prediction using Machine Learning Techniques: A Review [2]:

Deepti et al, summarized different papers on using Machine Learning to comprehend ASL. They also compared the advantages and disadvantages of different models available. The paper concludes that the Video recognition for sign language translation lags behind photo recognition, with a model achieving 98% accuracy for static images but only 67% for videos. Harris corner Algorithm, was implemented in this experiment and it was fund that the accuracy was hight but was very sensitive to noise. To conclude, this paper helped in understanding the difference between Video and Photo recognition models.

B. Ml Based Real Time Sign Language [3]:

Aggarwal et al, used machine learning to overcome the obstacles faced by people learning ASL, by training a data set in the teachable machine, where each sign was trained to the machine with output being a text. They identified that all old studies used 26 letters, therefore they included the other dynamic ASL letters to their project. They used Open CV library, which helped in providing better tools for CV. This paper helped in understanding a potential to bridge the gap in the communication and the application of sign recognition system from communication aids to education.

C. DeepASLR: A CNN based human computer interface for American Sign Language recognition for hearing-impaired individuals [6]:

Ahmed et al, aimed to developed an American Sign Language recognition dataset and then use it in the deep learning model. The model depends on the neural networks to interpret hand signs. They used ResNet-34 CNN classifier. They achieved an accuracy of 99.38% and with a small loss of 0.025. The system is intended to be applied for the SLR systems.

III. MODEL ARCHITECTURE

A. ResNet :

We implemented ResNet model architecture. ResNet stands for Residual Network, which is a type of deep neural network architecture. The following is the description of the ResNet Architecture. The basic ResNet architecture is depicted in Figure.1.

Residual Block

Residual Block is a basic fundamental block of ResNet, which consists a shortcut connection that helps in skipping one or more layers and provides importance to the original input to the output layers.

Identity Mapping

This step helps in degradation issue that occurs due to accuracy saturation. Identity mapping is the ability of the model to learn the difference between the input and the required output.

Bottleneck Architecture

ResNet has a bottleneck structure, which is used to reduce computational complexity. In this structure 3 layers of convolution are stacked. This helps in reduction of parameters and helps improve training.

Deep Stacks

ResNet models are deeper and are prone to vanishing gradient problem.

Global Average Pooling

Global Average Pooling helps reduce the overfitting issues.

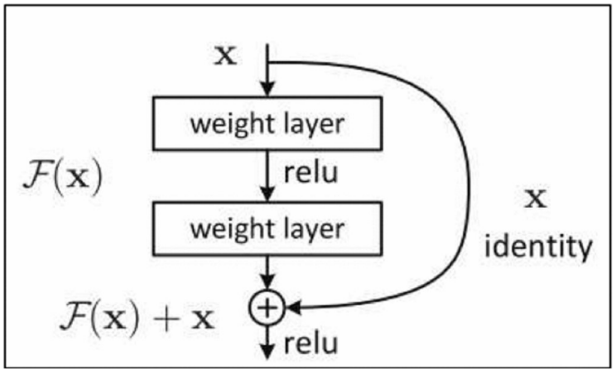

Figure.1. [4]

B. *Multi-Layer Perceptron:*

MLP is an artificial neural network which incorporates multi-layers of neurons that are 3 main layers: input, hidden and output. In this experiment it's a feedforward neural network, in which data flows from input to output layer. The following are the components of MLP architecture. Figure.2 is a depiction of Multi-Layer Perceptron.

Input Layer

This layer functions as input layer, in which the raw data is fed. It consists of multiple nodes. Each node represents a feature and the values of these nodes represent the input vector.

Hidden Layer

This is the middle layer which consists of neurons, which are responsible to add weights and biases, and further these are applied to the activation function. There can be more than one hidden layer in a model.

Output Layer

This layer is responsible to produce the final result of the model after the information is computed.

Activation and Loss Function

Activation function is mainly used to produce non-linear functions using the neural network which will help the model to learn complex patters.

Loss function is used to predict the difference between the produced and the expected outcome. The aim of training is to reduce the loss produced.

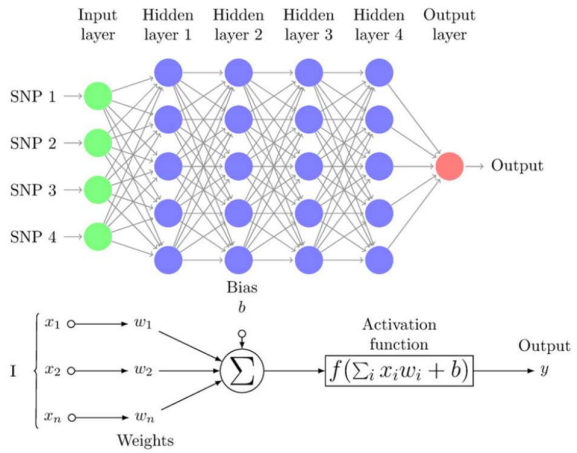

Figure.2.[5]

IV. METHODOLOGY

A. *Dataset Description:*

There are images from 29 classes which include 26 alphabets, SPACE, DELETE and NOTHING. Each of these images are 200 x 200 RGB. The training and testing sets have one folder for each class.

87,000 images were used to train the model, therefore each feature was trained using nearly 3000 images. This will help increase accuracy of identification of images during testing. 20 images are used for validation and the test set consists of 29 images.

B. *Training Parameters:*

- The input images were resized to 224 x 224 from the original image size of 64x64.
- We divided training dataset into 80% training and 20% validation dataset.
- The learning rate we used is 0.001
- Input parameters are image intensities.
- The experiment was conducted on 3 different models of ResNet.

C. *Activation Function: Softmax Function*

It is a widely used activation function for the output layer. It helps convert the input data into probability distribution. It is generally used to train the neurons with complex patters. It can be described using the following formula:

Softmax $(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$

Where:
zi is the raw input for class i
C is the total number of classes
E is the base of the natural log

D. *Loss Function: Multiclass Cross Entropy*

It is a commonly used loss function that is applicable to the multiclass problems. It calculates the difference between predicted and actual PDF. The following is the formula for the multiclass cross entropy

$$H(y, \hat{y}) = -\sum_{i=1}^{C} y_i \cdot \log(\hat{y}_i)$$

Where:
yi is the actual probability
yi cap is the predicted probability.

E. *Experiment*

We chose 3 different models of ResNet, namely ResNet50, ResNet101 and ResNet152.

ResNet50

This is a basic model of ResNet in comparison to that of ResNet101 and ResNet152. The model is relatively smaller and suitable for moderate computational constraints. It took less time to get trained and produced good results.

ResNet101

It is a more deeper architecture allowing it to capture more intrinsic details. ResNet101 is capable of performing high level tasks. This architecture model requires larger computational time but also produces satisfying results.

ResNet152

Among the three models ResNet152 has the deepest layers. Its performance is similar to ResNet101, which can be used for the computation of visual data. Due to its intensive depth, the training time is more.

V. ANALYSIS AND RESULT

A. Experiment 1

The first experiment was done by training the ResNet50 model with the training and validation dataset. The model took nearly 3610 seconds. The model presented a training accuracy of 99.11% and training loss of 0.0288. The validation accuracy and loss were 97.26% and 0.0726 respectively.

B. Experiment 2

The second experiment was conducted on ResNet101. The model took nearly 5400 seconds. The model presented a training accuracy of 99.04% and training loss of 0.0305. The validation accuracy and loss were 98.87% and 0.0396 respectively.

C. Experiment 3

ResNet152 was the model we trained using the training and validation dataset. The model took nearly 7200 seconds. The model presented a training accuracy of 99.05% and training loss of 0.0295. The validation accuracy and loss were 99.13% and 0.0247 respectively.

Figure.3 depicts the graphical representation of the accuracies of training and validation for all the three ResNet models.
The heat map along with confusion matrix for validation and test set for ResNet50 are presented in Figure.4 and Figure.5

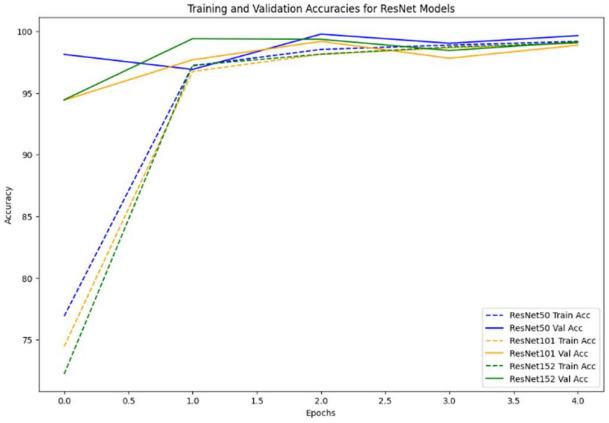

Figure.3

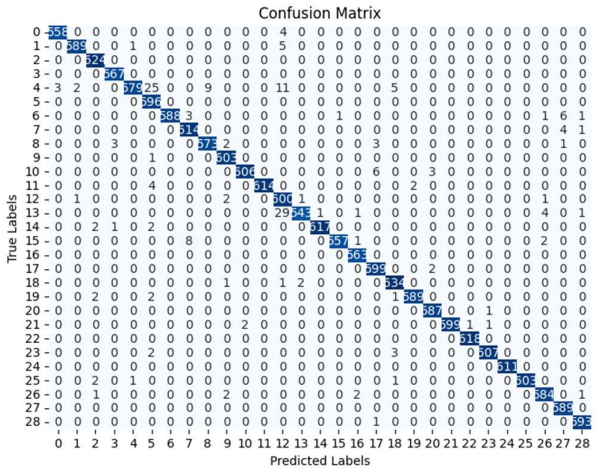

Figure.4

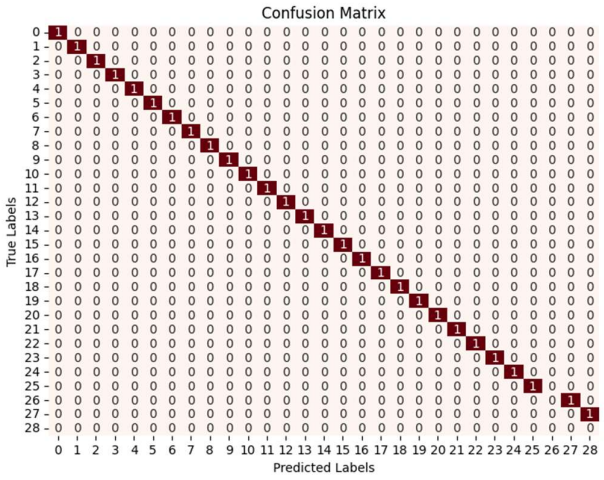

Figure.5

D. Additional Experiment

In the last experiment we added a special feature to the model, in which any hand gesture presented to the model can be presented with a speech output. We used the Google Text-to-Speech library. The predicted hand gesture is in the form of a text, from which it is converted to Speech. A MP3 file is generated at the end.
This feature can be used in improvising the similar existing models. Figure.6 depicts the detected image along with the MP3 audio file.

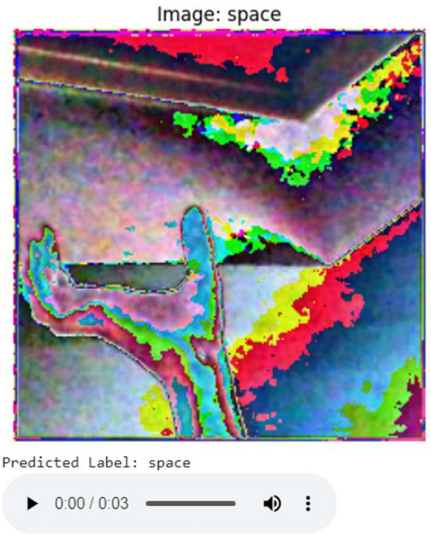

Figure.6

VI. Conclusion and Future Scope

A. Extension:

We believe that this is just the first step as the application of Computer Vision in recognition of American Sign Language has wide scope for development. A few things which we believe that would help extend this project are to create Real Time video recognition system, with a lesser processing time. Another application can be done by implementing

sensors and actuators in gloves of the users and then interpret not only the alphabets but also sentences.

B. Questions that were answered:

- We were excited to observe how the data was impacted by different types of ResNet models.
- It was also learning to know how this model can be implemented in various real life situations.

C. *Extended Curiosity*

In real-world circumstances, hand gestures frequently occur at a high speed, necessitating models that can accurately represent rapid motions. We find the task of developing such a model and investigating the potential for real-time training intriguing, since it will push the frontiers of gesture detection in hectic and dynamic settings. How can we create and refine a model that smoothly responds to real-time, fast gestures, helping to herald in a new era of flexible and responsive human-computer interaction?

D. Challenges Faced

One major issue was the slow calculation caused by the high computational needs of our models and the CPU's inability to handle the computation effectively. Next, we used GPU to tackle this problem instead of CPU.

VII. References

[1] Ross E. Mitchell, How Many Deaf People Are There in the United States? Estimates From the Survey of Income and Program Participation, *The Journal of Deaf Studies and Deaf Education*, Volume 11, Issue 1, Winter 2006, Pages 112–119, https://doi.org/10.1093/deafed/enj004

[2] D. Aggarwal, S. Ahirwar, S. Srivastava, S. Verma and Y. Goel, "Sign Language Prediction using Machine Learning Techniques: A Review," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2023, pp. 1296-1300, doi: 10.1109/ICEARS56392.2023.10084924.

[3] .N, Malligeswari & Sai, P.Vinay & S.Narendra, & Nagarathinam, K.P. (2023). Ml Based Real Time Sign Language. 11. 2814-2819. 10.15680/IJIRCCE.2023.1104208.

[4] "ResNet (34 50 101): Residual CNNs for Image Classification Tasks", *[Online]. Available:*, [online] Available: https://neurohive.io/en/popular-networks/resnet/.

[5] Pérez-Enciso, & Zingaretti, Laura. (2019). A Guide for Using Deep Learning for Complex Trait Genomic Prediction. Genes. 10. 553. 10.3390/genes10070553.

[6] Ahmed KASAPBAŞI, Ahmed Eltayeb AHMED ELBUSHRA, Omar AL-HARDANEE, Arif YILMAZ, DeepASLR: A CNN based human computer interface for American Sign Language recognition for hearing-impaired individuals,Computer Methods and Programs in Biomedicine Update,Volume 2, 2022, 100048, ISSN 2666-9900, https://doi.org/10.1016/j.cmpbup.2021.100048.

Git_hub link: https://github.com/BhavyaSamhithaMallineni/541_CV_Project/tree/main