CV for American Sign Language

Bhavya Samhitha Mallineni
mallinen@usc.edu

Dharitri Hemant Toshikhane
toshikha@usc.edu

Introduction

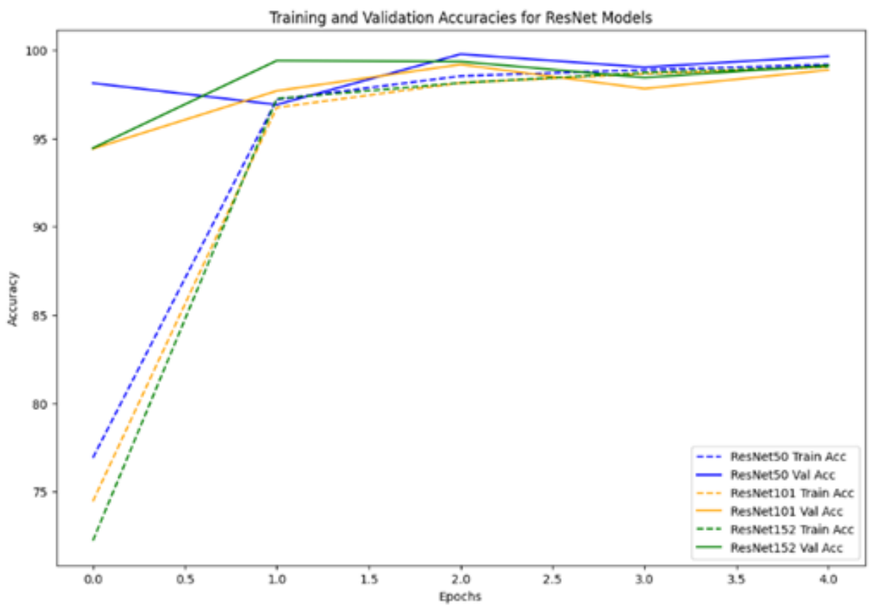
Approximately 10 million individuals are hard of hearing, with 1 million functionally deaf in the USA. To enhance accessibility, MNCs can integrate deep learning, specifically convolutional neural networks (CNNs), into their platforms. Focused on computer vision, CNNs excel in pattern recognition, making them ideal for applications in communication, healthcare, and education. The proposed model processes input images through convolution layers with adjustable weights and biases, applying ReLU activation. Loss function and optimization algorithms like SGD are employed to minimize discrepancies between the model's output and target values.

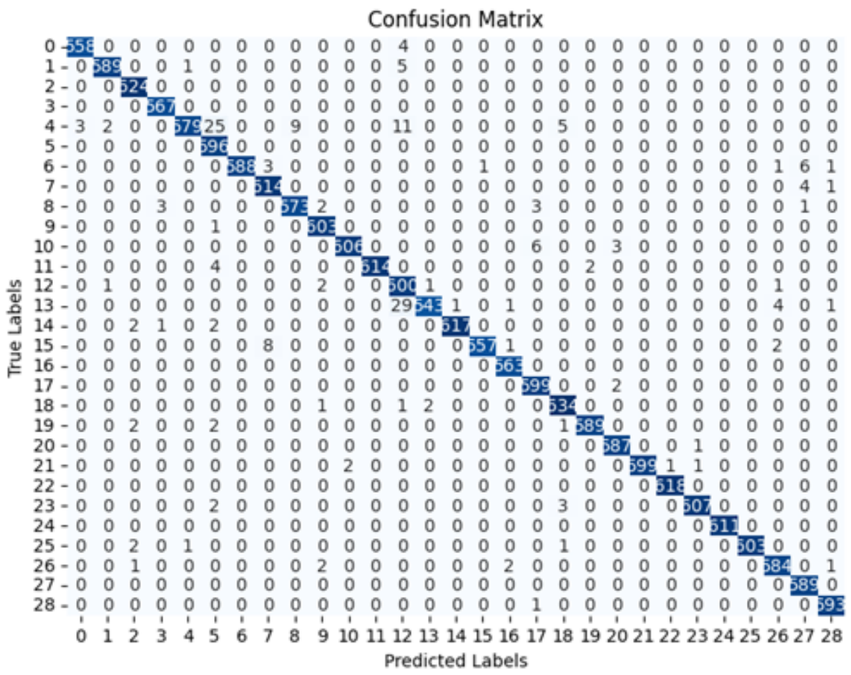
Methodology

A dataset of 87,000 RGB images (200x200) comprising 29 classes, including 26 alphabets, SPACE, DELETE, and NOTHING, was used for training ResNet models. The input images were resized to 224x224, and the dataset was split into 80% training and 20% validation sets. A learning rate of 0.001 was employed, and three ResNet models (ResNet50, ResNet101, ResNet152) were tested. The activation function used was Softmax, converting input data into a probability distribution. The Multiclass Cross Entropy loss function was applied to measure the difference between predicted and actual probabilities. ResNet50, a smaller and faster model, yielded good results, while ResNet101 and ResNet152, with deeper architectures, demonstrated enhanced performance at the cost of longer training times.

Analysis and Results

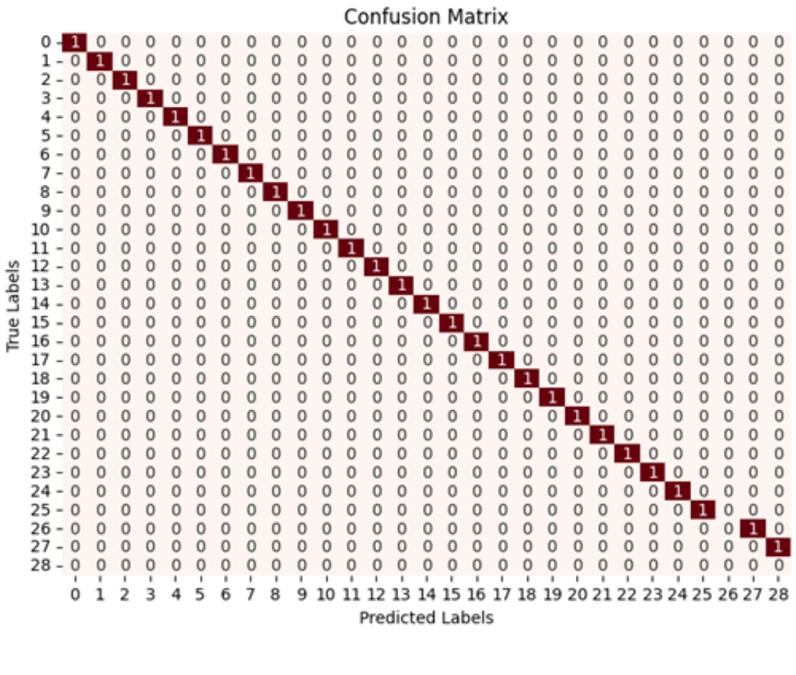
Experiment 1 involved training ResNet50, taking 3610 seconds, yielding a training accuracy of 99.11%, and validation accuracy of 97.26%. In Experiment 2, ResNet101 was trained for 5400 seconds, achieving a training accuracy of 99.04% and validation accuracy of 98.87%. Experiment 3 utilized ResNet152, taking 7200 seconds, with a training accuracy of 99.05% and validation accuracy of 99.13%. Figure.3 illustrates the graphical representation of accuracies for all three models. Heat maps and confusion matrices for ResNet50 validation and test sets are presented in Figure.4 and Figure.5. An additional experiment added a speech output feature using Google Text-to-Speech, converting detected hand gestures to MP3 audio files, enhancing model usability (Figure.6).





Conclusion

The project on American Sign Language recognition using Computer Vision marks an initial step, and we envision expanding its scope by developing a Real-Time video recognition system for faster processing. Additionally, implementing sensors and actuators in user gloves could extend the application to interpret not only individual alphabets but complete sentences. The project provided insights into the impact of ResNet models on data and showcased the versatility of implementing such models in real-life scenarios.



Architecture

Implemented in our project is the ResNet model architecture, short for Residual Network, a deep neural network design. The ResNet architecture includes a Residual Block, a fundamental component featuring a shortcut connection that skips one or more layers, emphasizing the original input in the output layers. Three ResNet models were utilized:

1. ResNet50: A foundational model, smaller and suitable for moderate computational constraints, exhibiting efficient training with favorable results.
2. ResNet101: A deeper architecture capturing intrinsic details, capable of high-level tasks but requiring larger computational time, offering satisfying results.
3. ResNet152: With the deepest layers among the models, its performance is comparable to ResNet101, suitable for visual data computation despite an intensive training time.