Advancing Accessibility: ASL Visual Recognition Technology through EfficientNet

Priyanka Mazumdar[1] , **Anindya Biswas[1]** , **Anirban Naskar[2]**, **Aritra Mandal[1]**, **Soumya Sen[1]**

1. A.K. Choudhury School of Information Technology, University of Calcutta

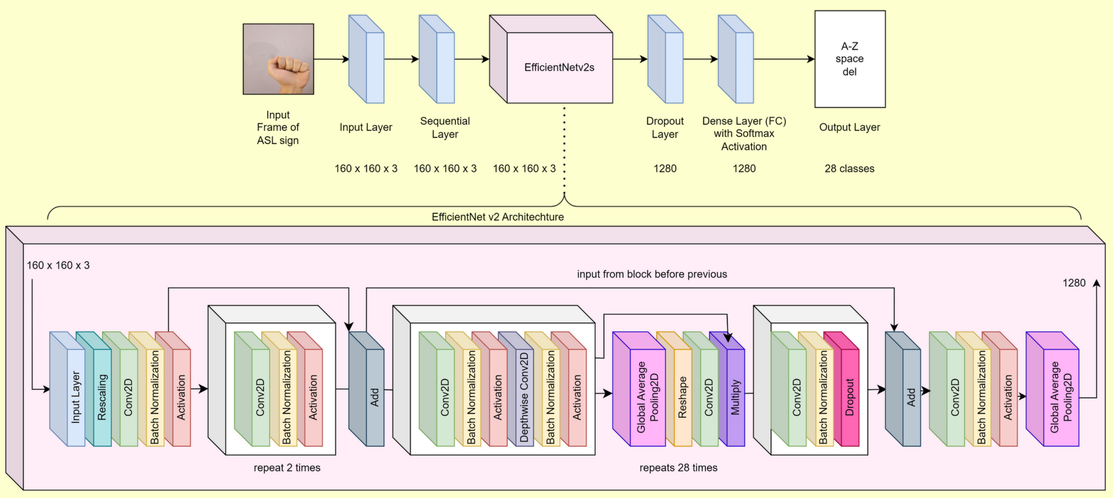
2. Department of Computer Science and Engineering, University of Calcutta

INTRODUCTION

American Sign Language (ASL) is a crucial form of communication for Deaf and Hard-of-Hearing individuals. However, the ability to accurately recognize ASL signs using machine learning remains an ongoing challenge. In this work, we explore the application of machine learning techniques to develop an ASL detection system capable of real-time sign recognition. We propose a deep learning model that utilizes a convolutional neural network (CNN) architecture, EfficientNet to identify hand gestures. Our model is trained on a comprehensive dataset of ASL signs, enabling it to identify a variety of hand gestures with high validation accuracy. The proposed system can bridge the communication gap between Deaf and hearing individuals, fostering seamless and inclusive interactions.

Data Augmentation methodology employed to increase diversity

METHODOLOGY



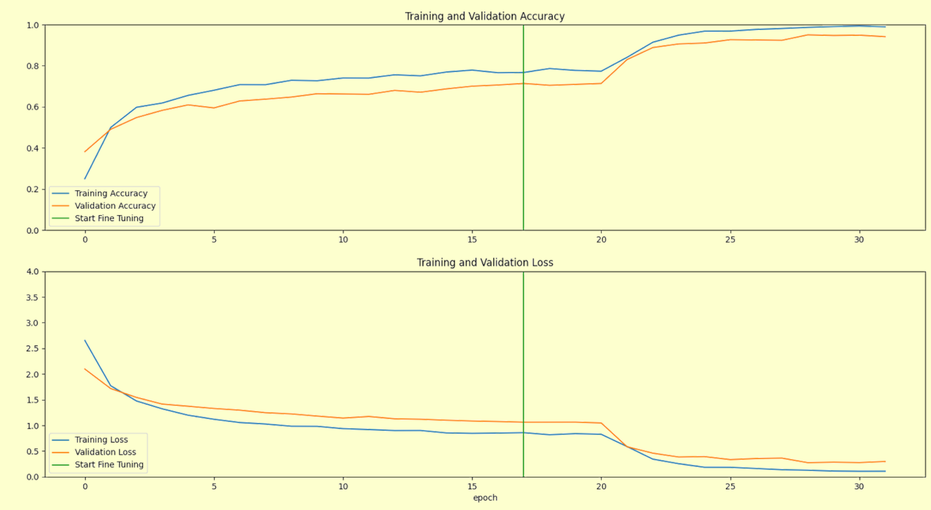
ASL Detection Model Architecture using EfficientNet

The previous models' shortcomings were due to limited dataset variation, hindering their ability to generalize across diverse instances. To bolster future model performance and robustness in problem-solving, we enhanced dataset diversity and employed the following methods:

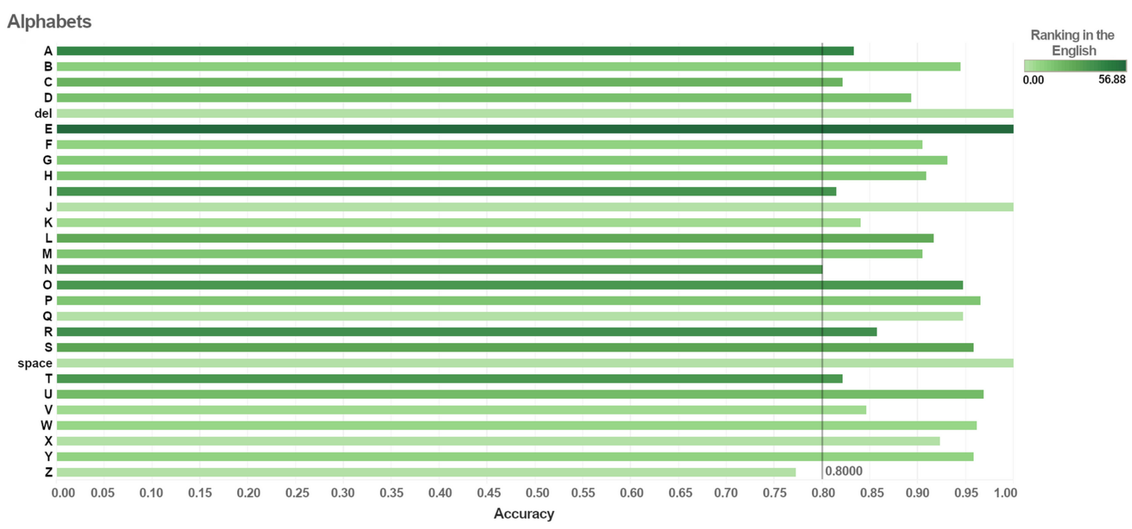
- **Dataset Curation:** Merged three diverse ASL hand gesture datasets, creating a final set of 28 classes (A-Z, *space*, *delete*) with variation in background, lighting and hand shape.
- **Data Augmentation:** Applied random rotation, translation, zoom, and flip to images for diversity and to prevent overfitting.
- **Base Model:** Utilized transfer learning on EfficientNetV2S, integrating a 20% dropout layer and a softmax dense layer. Trained for 64 epochs with early stopping at the 18th epoch, achieving 76.71% training and 71.37% validation accuracy.
- **Learning Setup:** Employed *Nadam* Optimizer with L2 Regularization and Exponential Decay Rate to expedite convergence and minimize weight oscillations.
- **Fine Tuning:** Retrained the top 120 layers of EfficientNetV2S, achieving 98.64% training and 95.01% validation accuracy.
- **Webcam App:** Developed real-time ASL symbol recognition via webcam, implementing a 90% confidence threshold for consecutive predictions. Integrated a text-to-speech API for inferred letter conversion to speech.

RESULTS

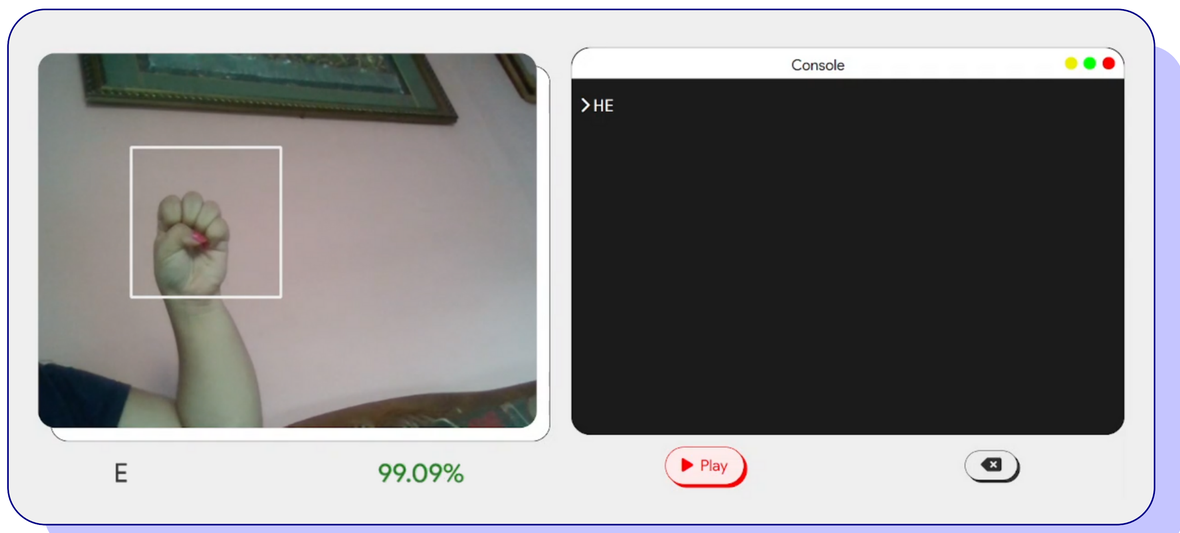
The final model achieved a 95.01% accuracy on the validation set (20% of the entire dataset with a loss of 0.2753) which is highlighted in the above chart, depicting Sparse Categorical Cross Entropy Loss across epochs. The steep decline post fine-tuning initiation, focusing on the top 120 layers, signifies EfficientNet's adaptability to our data domain during decoder phase. Early Stopping, selecting the lowest validation loss from 6 consecutive epochs, further enhances optimization.



This version of EfficientNetV2S fine-tuned on our curated dataset performs better than the state of the art not only on the validation set but also on images of our team members' hands, which were never presented to the model during its training.



Accuracy of each class



Web Application integrated with TTS

CONCLUSION

Our proposed solution improves over transfer-learning of pre-trained EfficientNetV2S by 33.12% through fine-tuning of the top 120 layers. The incorporated Text-to-Speech system makes communication between ASL and non-ASL speakers easier.

This work aims to bring inclusion to spaces which are ever-changing and developing to welcome diverse and differently-abled people. ASL recognition technology can enhance accessibility in digital platforms, such as video calls, social media, and online content. This inclusion in digital spaces promotes equal access to information and eases social interaction.