SignVision: Real-time American Sign Language Recognition using Deep Learning and Computer Vision

Abinav Anantharaman - 002774223
Satwik Shridhar Bhandiwad - 002920338

Introduction

1. "SignVision" is a proposed ASL recognition system that uses computer vision and deep learning techniques to recognize ASL gestures in real-time with high accuracy.
2. The system employs a CNN-based model with transfer learning on a pre-trained ResNet18 architecture, which outperforms other custom models due to its ability to learn more complex features and prevent overfitting.
3. The project aims to classify 29 classes of American Sign Language (ASL) gestures from a dataset of 87,000 images, including A to Z, SPACE, DELETE, and NOTHING.
4. The ultimate goal of this project is to develop an accurate and efficient ASL gesture recognition system that can facilitate communication between hearing-impaired and hearing individuals and enable the development of various ASL-based applications.

Related Works

[1] Rathi, Pulkit and Kuwar Gupta, Raj and Agarwal, Soumya and Shukla, Anupam, Sign Language Recognition Using ResNet50 Deep Neural Network Architecture (February 27, 2020). 5th International Conference on Next Generation Computing Technologies (NGCT-2019).

[2] M. Al-Qurishi, T. Khalid and R. Souissi, "Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues," in IEEE Access, vol. 9, pp. 126917-126951, 2021, doi: 10.1109/ACCESS.2021.3110912.

[3] Abu-Jamie, Tanseem N. & Abu-Naser, Samy S. (2022). Classification of Sign-Language Using Deep Learning by ResNet. International Journal of Academic Information Systems Research (IJAISR) 6 (8):25-34.

[4] K. Bantupalli and Y. Xie, "American Sign Language Recognition using Deep Learning and Computer Vision" 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4896-4899, doi: 10.1109/BigData.2018.8622141.

Data Preparation and Pre-Processing

The ASL dataset:

- Collection of 87,000 images depicting hand gestures of American Sign Language alphabet.
- Divided into 29 categories, including 26 letters of English alphabet and 3 additional gestures for space, delete, and nothing.
- 3 channel RGB images with a resolution of 200 x 200 pixels.

Pre-Processing

- Images resized to 64 x 64 and normalized pixel values to [0, 1].
- Augmented training set with random horizontal flips and rotations to increase diversity and prevent overfitting.
- Dataset divided into training, validation, and testing sets with an 80:10:10 ratio.
- 80% of the dataset used for training, 10% for validation, and 10% for testing.

Training

Custom Sequential Model 1:

- Convolutional neural network architecture consisting of multiple layers.
- 2D convolutional layers, max pooling layers, fully connected layers, ReLU activation functions, and dropout layers used to produce a probability distribution over 29 classes.

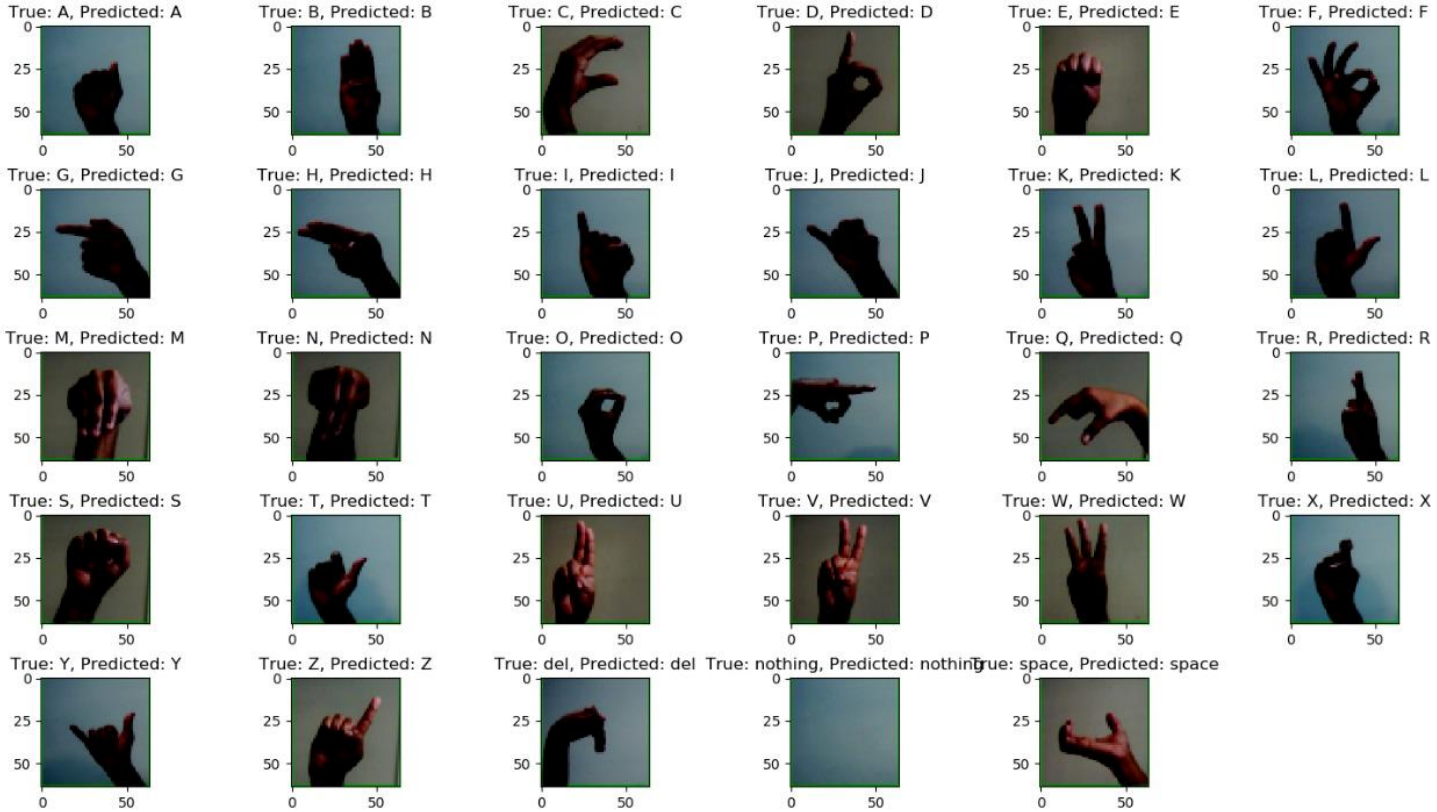
Custom Sequential Model 2:

- Deeper and more complex version of Model 1 with 5 convolutional layers, batch normalization layers, ReLU activation functions, and max pooling layers.
- Designed to be more robust and less prone to overfitting due to the use of batch normalization and dropout layers

ResNet18 with transfer learning:

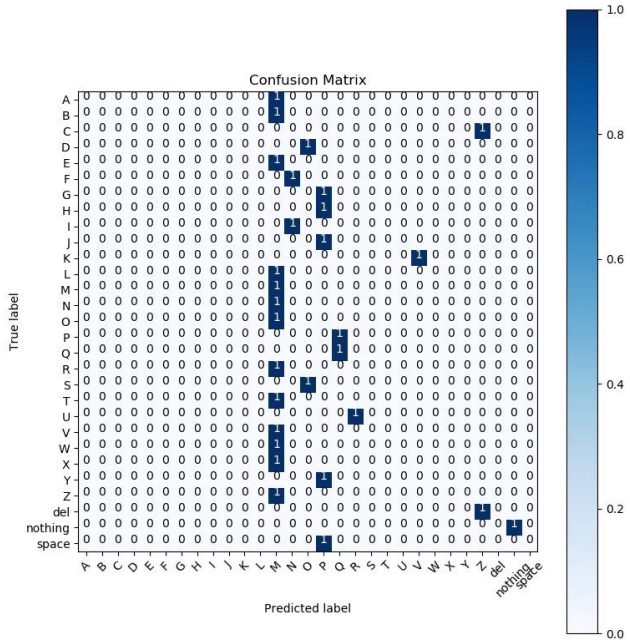
- Residual neural network architecture consisting of 18 layers with residual connections between the layers.
- Pre-trained ResNet18 model used as a feature extractor, and a new fully connected layer with 29 output classes added.
- During training, only the weights of the new fully connected layer are updated.

Results

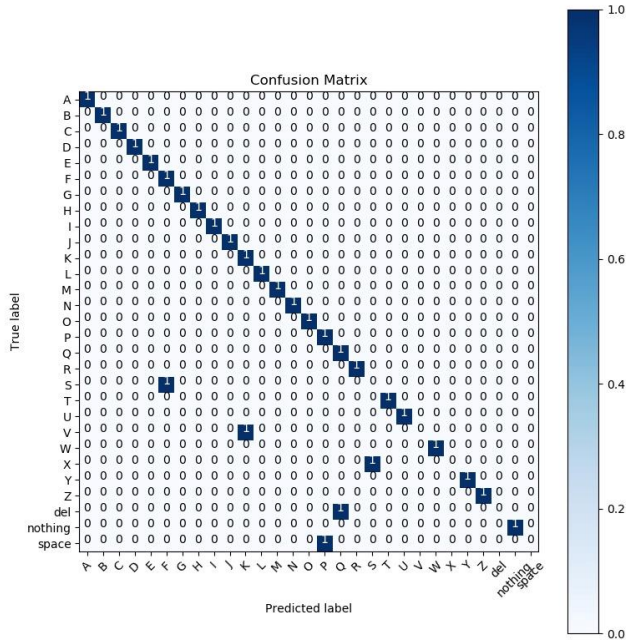

True: A, Predicted: A
True: B, Predicted: B
True: C, Predicted: C
True: D, Predicted: D
True: E, Predicted: E
True: F, Predicted: F
True: G, Predicted: G
True: H, Predicted: H
True: I, Predicted: I
True: J, Predicted: J
True: K, Predicted: K
True: L, Predicted: L
True: M, Predicted: M
True: N, Predicted: N
True: O, Predicted: O
True: P, Predicted: P
True: Q, Predicted: Q
True: R, Predicted: R
True: S, Predicted: S
True: T, Predicted: T
True: U, Predicted: U
True: V, Predicted: V
True: W, Predicted: W
True: X, Predicted: X
True: Y, Predicted: Y
True: Z, Predicted: Z
True: del, Predicted: del
True: nothing, Predicted: nothing
True: space, Predicted: space

Results Comparison

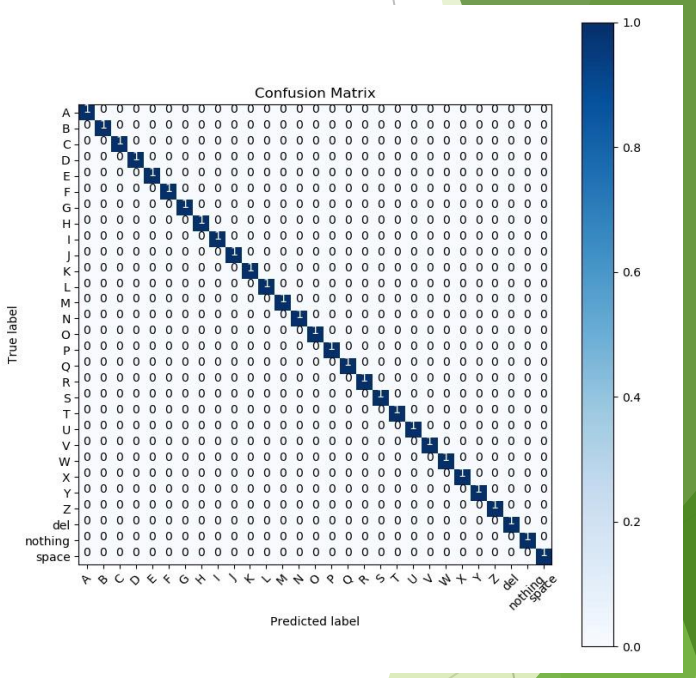
Confusion Matrices:



Model 1



Model 2



ResNet18

Final Demo