Sign Language Recognition and Translation System

# 1. Introduction

The Sign Language Recognition and Translation System aims to enhance accessibility for the deaf and hard-of-hearing community by providing real-time translation of sign language gestures into spoken and written language. This project integrates Computer Vision (CV) techniques for gesture recognition and Natural Language Processing (NLP) for translating recognized gestures into a target language.

# 2. Model Architecture and Working

## 2.1 Gesture Recognition Model (CNN)

The gesture recognition model is designed to classify hand gestures from the Sign Language MNIST dataset. The model architecture is as follows:  
1. \*\*Input Layer\*\*:  
 - \*\*Shape\*\*: (28, 28, 1) — The input consists of grayscale images of hand gestures, with a resolution of 28x28 pixels.  
2. \*\*Convolutional Layers\*\*:  
 - \*\*Conv2D\*\*: The model uses multiple 2D convolutional layers to extract spatial features from the images. Each convolutional layer applies a set of filters to the input, producing feature maps.  
 - \*\*Activation Function\*\*: ReLU (Rectified Linear Unit) is used to introduce non-linearity into the model.  
3. \*\*Pooling Layers\*\*:  
 - \*\*MaxPooling2D\*\*: Max pooling is applied to reduce the spatial dimensions of the feature maps, which helps in reducing computational complexity and the likelihood of overfitting.  
4. \*\*Flatten Layer\*\*:  
 - Converts the 2D feature maps into a 1D vector, preparing it for the fully connected layers.  
5. \*\*Fully Connected Layers (Dense)\*\*:  
 - \*\*Dense\*\*: The model uses fully connected layers to combine the features extracted by the convolutional layers. The final dense layer has 25 units corresponding to the 25 classes in the Sign Language MNIST dataset.  
 - \*\*Activation Function\*\*: The softmax activation function is used in the output layer to produce a probability distribution over the classes.  
6. \*\*Output Layer\*\*:  
 - \*\*Shape\*\*: (25,) — The output consists of a probability distribution over the 25 classes, representing the recognized gesture.

## 2.2 Translation Model (Transformer-based)

The translation model uses the mBART (Multilingual BART) architecture, a Transformer-based model for sequence-to-sequence tasks. The architecture consists of an encoder and a decoder, both of which include multiple layers of self-attention and feed-forward neural networks.  
1. \*\*Input Layer\*\*:  
 - The input consists of a sequence of tokens representing the recognized gesture in a source language (e.g., Romanian).  
2. \*\*Encoder\*\*:  
 - \*\*Self-Attention Mechanism\*\*: The encoder uses self-attention to capture dependencies between tokens in the input sequence. This mechanism allows the model to weigh the importance of different tokens when encoding the sequence.  
 - \*\*Feed-Forward Neural Network\*\*: A feed-forward network is applied to each position separately, followed by layer normalization.  
3. \*\*Decoder\*\*:  
 - \*\*Self-Attention Mechanism\*\*: Similar to the encoder, but it attends to the output sequence generated so far.  
 - \*\*Encoder-Decoder Attention\*\*: This layer attends to the output of the encoder, allowing the decoder to access the encoded information.  
 - \*\*Feed-Forward Neural Network\*\*: Applies to the output of the attention layers.  
4. \*\*Output Layer\*\*:  
 - \*\*Shape\*\*: Variable length, depending on the generated translation. The output consists of a sequence of tokens in the target language (e.g., English).

# 3. Evaluation and Results

## 3.1 Confusion Matrix

The confusion matrix provides a visualization of the performance of the gesture recognition model. It shows the counts of true positive, false positive, and false negative predictions for each class. From the confusion matrix, we can observe how well the model distinguishes between different gestures.  
- \*\*Diagonal Elements\*\*: Represent correct predictions.  
- \*\*Off-Diagonal Elements\*\*: Represent misclassifications.  
  
The test accuracy achieved by the model is 92.88%, indicating a high level of accuracy in recognizing gestures.

## 3.2 Classification Report

The classification report provides detailed metrics for each class, including precision, recall, and F1-score. Here are the key observations:  
- \*\*Precision\*\*: The ratio of correctly predicted positive observations to the total predicted positives.  
- \*\*Recall\*\*: The ratio of correctly predicted positive observations to all observations in the actual class.  
- \*\*F1-Score\*\*: The weighted average of precision and recall, providing a balance between the two.  
  
The model achieves high precision and recall across most classes, with an overall weighted average F1-score of 0.93.

# 4. Skills and Technologies Used

- \*\*Computer Vision\*\*: Skills in CNNs, image preprocessing, and feature extraction.  
- \*\*Natural Language Processing\*\*: Knowledge of Transformers, sequence-to-sequence models, and tokenization.  
- \*\*Deep Learning Frameworks\*\*: Proficiency in TensorFlow/Keras and Hugging Face Transformers.  
- \*\*Data Analysis\*\*: Experience with evaluating models using confusion matrices, classification reports, and other metrics.  
- \*\*Model Integration\*\*: Combining multiple models to create a cohesive system for end-to-end translation.

# 5. Conclusion and Future Work

The Sign Language Recognition and Translation System successfully recognizes and translates gestures into a target language. The integration of CNN for gesture recognition and Transformer for translation demonstrates the potential for creating accessible and inclusive technologies. Future work can focus on:  
- \*\*Real-Time Application\*\*: Developing a web or mobile application for real-time translation.  
- \*\*Multilingual Support\*\*: Expanding the system to support more languages.  
- \*\*Data Augmentation\*\*: Enhancing the dataset to include more gestures and variations to improve model robustness.  
  
This project not only highlights the technical capabilities of deep learning models but also emphasizes the importance of creating tools that foster inclusivity and accessibility. The skills and knowledge gained through this project are valuable for pursuing advanced roles in AI, machine learning, and data science.