Report - Embedded System Project

Project Title - “Automatic recognition of gestures, sign languages, and interaction of humans from the camera input of mobile phones or microcontrollers. ”

Group Members -

- Krishi Patel (B20EE030)
- Preeti (B20EE044)

ABSTRACT

Automatic recognition of gestures, sign languages, and interaction of humans from the camera input of mobile phones or microcontrollers refers to the use of computer vision and machine learning techniques to analyze video streams from cameras on mobile phones or microcontrollers to recognize and interpret human gestures and movements.

The technology involves developing algorithms and models that can accurately recognize and classify gestures, such as hand movements, facial expressions, and body language, in real-time. These gestures can then be used to interact with the mobile device or microcontroller, enabling users to control various functions and applications without the need for traditional input methods such as keyboards or touchscreens.

PROCEDURE

We carried out the experiment by training 3 different models on the sign language MNIST dataset.

1. We trained CustomNet, which was a neural network made from scratch and checked how the accuracy for the model changes as we increase its size, change the weights and activation functions. Also we kept a note of how much time the model consumes.
2. We trained ResNet18 on the same dataset and checked the accuracy and time for different sizes of the same network (by incrementing the extra number of convolution layers).
3. Finally we trained MobileNet and observed the time and accuracy as the model size increases.

OBSERVATIONS AND RESULTS

1. Custom Net (Neural Network from scratch)

i) Convolution layer= 3 , Optimizer= Adam , Activation Function=Relu

train_acc=0.8016

test_acc=0.7912

ii)Convolution layer= 4 , Optimizer=Adam , Activation Function=Relu

train_acc=0.8916

test_acc=0.8900

iii) Convolution layer= 4 , Optimizer=Xavier , Activation Function=Relu train_acc=0.9016

test_acc=0.8924

iv) Convolution layer= 4 , Optimizer= Xavier , Activation Function=Sigmoid

train_acc=0.8524

test_acc=0.7024

2. ResNet 18

i) Convolution layer (extra) = 1

train_acc=0.9979

test_acc=0.9825

ii) Convolution layer (extra) = 2

train_acc=0.9976

test_acc=0.9900

iii) Convolution layer (extra) = 3

train_acc=0.9976

test_acc=0.9924

3. Mobile Net

i) Convolution layer (extra) = 1

train_acc=0.9874

test_acc=0.9777

ii) Convolution layer (extra) = 2

train_acc=0.9895

test_acc=0.9792

iii) Convolution layer (extra) = 3

train_acc=0.9910

test_acc=0.9821

From the above mentioned results (1 and 2) we can observe that **Custom Net** is neither robust nor accurate. Whereas **ResNet18** can be considered a better option when it comes to accuracy.

But we should also notice that **MobileNet** is a much better option than **ResNet18.** The accuracy for **MobileNet** is comparable to **ResNet18** but the difference comes in the context of resource consumption.

MobileNet is a lightweight neural network which is specially designed for systems like smartphones, microcontrollers, smartwatch etc. The fact that it is accurate and less resource consuming makes it the best choice for embedded systems.

THEORY

The technology we are dealing with in this project of gestures recognition, sign languages, and interaction of humans from the camera is used to enable natural and intuitive interaction between humans and machines, particularly in situations where the use of traditional input devices like keyboards and mouse are not practical. Examples of applications include sign language recognition, gesture-based gaming, and virtual reality interaction.

This technology has a wide range of applications, including accessibility for people with disabilities, gaming, virtual reality, and user interface design. It is an area of ongoing research and development, with many advancements being made in the field of computer vision and machine learning.

Model used in this projects are:-

1. **Custom Net (Neural Network from scratch)-** It was a simple CNN based neural network written from scratch. Initially it had 3 CNN layers and 3 fully connected layers which were further changed with the demand of the project.

2. **ResNet 18-** ResNet-18 is a convolutional neural network that is 18 layers deep. It is one of the most popular and widely used deep learning models in computer vision applications, particularly in image classification tasks.

The ResNet-18 architecture consists of 18 layers, which include convolutional layers, batch normalization layers, and fully connected layers. One of the key innovations in ResNet-18 is the introduction of residual blocks, which allow the network to better learn from its inputs by allowing information to flow directly from one layer to another. This helps to mitigate the problem of vanishing gradients, which can occur when training very deep neural networks.

ResNet-18 has achieved state-of-the-art results in a number of image classification tasks, including the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2015. It has since become a popular starting point for researchers and developers working on computer vision problems, and has been used as a baseline for many subsequent models.

3. **MobileNet-** MobileNet is a convolutional neural network architecture that was introduced in 2017 . It was designed specifically for use in mobile and embedded devices with limited computational resources, such as smartphones and IoT devices.

The MobileNet architecture consists of a series of depthwise separable convolutional layers, which reduce the number of parameters in the network while maintaining high accuracy. These layers are designed to factorize a standard convolutional layer into a depthwise convolutional layer and a pointwise convolutional layer, which significantly reduces the computational cost of the network.

MobileNet has achieved state-of-the-art results in a number of computer vision tasks, including image classification, object detection, and semantic segmentation. It has also been widely adopted in industry due to its low computational requirements, making it ideal for use in mobile and embedded devices.

CONCLUSION

1. We have checked the robustness of our model - Custom Net (Neural Network from scratch) by changing layers , weights, functions, etc. We have implemented our model for the 3 , 4 convolution layer and for Adam and Xavier optimizer and for activation function- Relu,sigmoid and then using the ResNet 18 model it was found that accuracy coming up for ResNet 18 is more than the Custom Net (Neural Network from scratch). Hence the ResNet 18 model is much better than Custom Net.

2. Further, we compared the two models ResNet 18 and MobileNet. We have compared these two by carrying out multiple experiments , and it was found that as we increase the number of extra convolution layers the size of the network also increases. As network size increases ResNet's accuracy tends to saturate but the computation time becomes large whereas for MobileNet accuracy increases continuously and also the time of computation is comparatively low. Hence we choose Mobile Net over ResNet 18.

3. As described in the results and observation section, it is clear that MobileNet is a better choice if we want to design a lightweight neural network. But apart from the results MobileNet offers multiple other advantages like low memory footprint, low computational requirements, fast and accurate, transfer learning, etc. Hence these are the reasons that we decided to choose MobileNet for gesture recognition tasks.