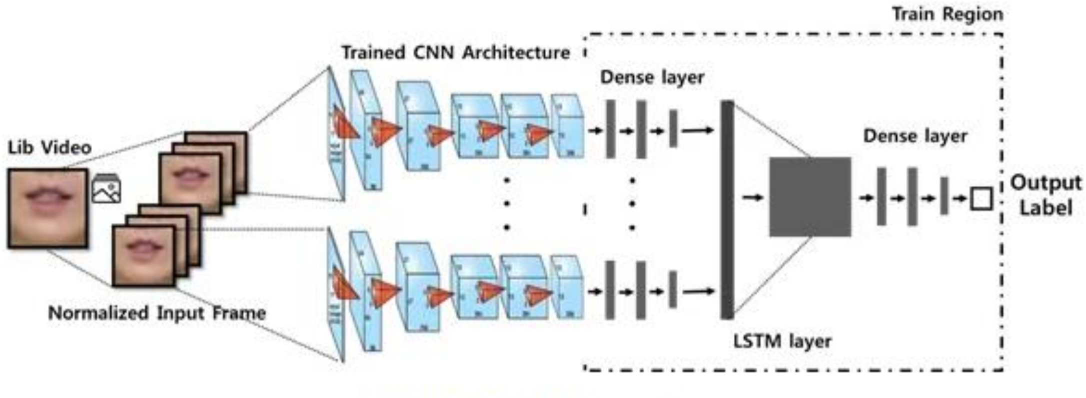
Project Design Phase-I
Solution Architecture

Date	21 November 2023
Team ID	591890
Project Name	Project – Lip Reading Using Deep Learning
Maximum Marks	4 Marks

Solution Architecture:

We use CNN and LSTM with several dense layers for lip reading. For lip reading, one must realize the change in lip shape. So, we choose CNN and LSTM. CNN will realize lip shape and then each of the outputs of CNN become transformed to the sequence. Finally, LSTM realizes the pattern of the sequence that contains the change of lip shape. Using this difference of sequences to classification.

Solution Architecture Diagram:



Simple Model Architecture

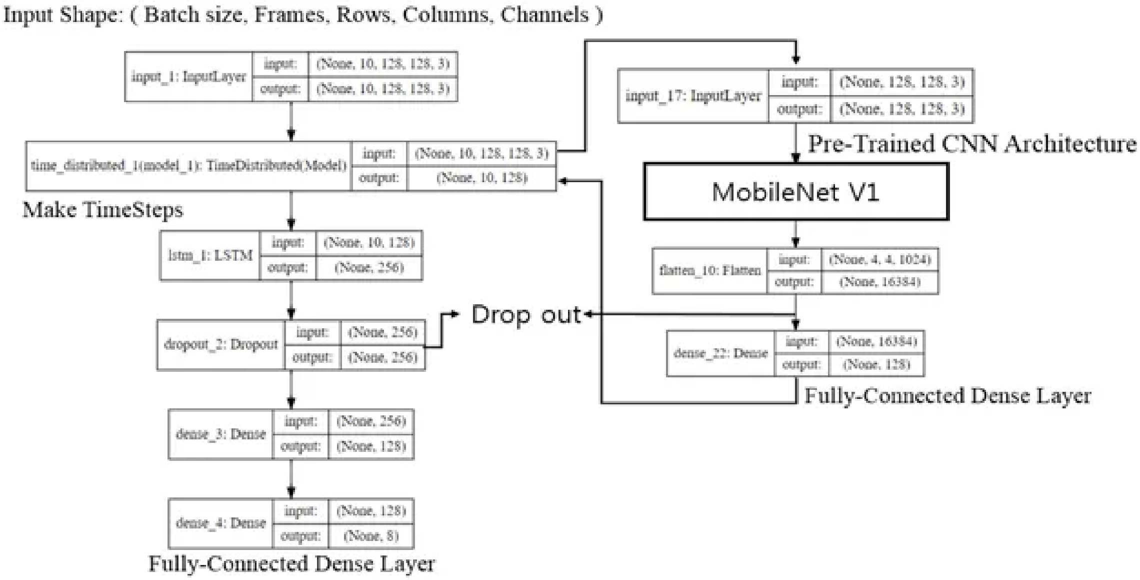
The above picture is the simple architecture of our model. First, we transform the lip video into several frame images. We used several frame images as input. Each fame image passes through the already trained CNN architecture. And then the output of CNN passes through the Dense layer for transforming to LSTM layer input. The output of the LSTM layer become the next dense layer input. Finally, receive the output label by a softmax activation function. Our train region is a dense layer of CNN architecture to the end part. This region is conceived in the above picture.

We judged that the extraction feature of lip shape is almost the same as the extraction feature of appearance. So, we decide to use the transfer learning from the ImageNet trained model. And then only train the dense layer in the CNN. We choose MobileNet for the transfer learning model. MobileNet is a compact model and users can resize within a regular range.

When we use the VGG model, take about 3 hours for one epoch on our dataset and an error occurred in the LSTM layer. But, it takes about 15 minutes for one epoch. As a result, we choose MobileNet.

For reading data, firstly count each label's number of datasets. And then, pile the lip shape image along the TimeSteps. For Keras input, transform the list to a Numpy array. Finally, shuffle data and make a batch for training. The final shape of dataset is (586,20,128,128,3) 586 is the number of datasets and 20 is TimeSteps(frame), 128X128 is size of image. 3 is the channel. (Color image)

The region for making TimeSteps is composed of MobileNet and dense layer. In the dense layer, applied to drop out. Drop out is one of the methods to avoid overfitting. The output of this region enters the LSTM layer as input. The output of LSTM passes through a dense layer with one more dropout. Finally, the classification label is output. The detailed model architecture is conceived in the below picture



CNN + LSTM Model Architecture