**Phase 2 Report**

In the secondphase of project, the following things are completed:

* Data extraction :The data was derived from three different sources
  + American Sign Language Lexical Dataset (ASSLVD) developed by Boston University and University of Arlington. This dataset consists of videos with different actions and CSV file corresponding to each video.CSV file consists of starting of the sign, ending of the sign and name of the sign [1].
  + National Center for Sign Language and Gesture Resources Corpus. This platform provided an interface to search different actions and download them. This dataset was generated under the supervision of Boston University [2].
  + Manually preparing the required gestures that were less in number by the team members.

* Data cleaning : Cleaning of the dataset can be divided into three major parts -
  + Dividing the videos of multiple actions into smaller videos of one action each with the help of CSV provided in the ASSLVD.
  + Dividing the videos of all datasets into frames and labelling them appropriately.
  + Cropping the extra regions of the frames that are not necessary for training the neural network. After cropping, changing the resolution of the frames to the desired dimensions 440x440 pixels.
* Changing the dataset into a format that is suitable for loading the dataset for training in the neural network.
  + Implementing ASL class
  + Creating Sign\_dataset instance for ASL class.
  + Implementing the DataLoader class of Pytorch in order to shuffle the dataset and define the batch size.
  + Creating the dataloader instance of the DataLoader class.

* Implementing the CNN network model class: Implementing the VGG style 3D CNN.
  + 11 convolution layers
  + Kernel size for all the layers is (3,3,3). The three dimensions are time, height and width except the first layer having the kernel size of (3,5,5).
  + The time dimension is preserved by placing one padding and one stride.
  + After each convolution layer there is a batch normalization layer followed by a ReLU activation function.
  + There are a total of 4 dropout layers used for computation purpose with probability of dropping a neuron in the layer equal to 0.2.
  + Averaging out the time dimension at the end of the computation.
  + The number of layers and the different types of layers such as pooling layer, convolution layer may be added or removed as per the results obtained.
* Input to the model:
  + This is composed of batch size of 2 and number of frames that are corresponding to the given video, 100 width and 100 height and 3 color channels (red, blue and green).

* Training the model:
  + Using a cross entropy loss function in order to calculate the loss and Stochastic gradient descent (SGD) as an optimizer function.
  + Learning rate = 0.001 and momentum = 0.9
  + No. of epochs = 20
  + The values of the above hyperparameters maybe changed in order to get better results.
* The training of the model is performed on Google Colab, Google machine learning platform which has 13 GB of RAM and Tesla K80 GPU.

**References**

[1] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan and A. Thangali, The ASL Lexicon Video Dataset, CVPR 2008 Workshop on Human Communicative Behaviour Analysis

[2] Carol Neidle and Christian Vogler [2012] "A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface," *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012, Istanbul, Turkey.*