Report Ex4 Machine Learning

Lior Chemouhoum & Raphael Aben-Moha

Model Architecture:

After transforming the audio sounds into picture via GCommand_loader, we have builted a model composed of two main part:

1) **The CNN part :**
   the first CNN layer is composed of a filter with a kernel size 2x2, a padding of 2 at each side of the tensor, and a stride of 2.
   the activation map in this layer is 60 filters.
   Then we apply a Relu activation function on it.
   The Second CNN layer is composed of a layer is composed of a filter with a kernel size 6x6, a padding of 1, and a stride of 2x2.
   the activation map in this layer is 100 filters this time.
   After using a Relu as activation function, we applied a MaxPooling on this layer , with a kernel size of 2x5 and a stride of 2.
2) **The Fully Connected part:**
   After applying a standard Dropout over the last CNN layer, we did a 3 layer Fully Connected Neural network (i.e. 2 hidden layer) as follow:
   The first hidden layer get input from the previous dropout and has 2048 neurons, and relu activation function
   The second hidden layer has 512 neurons, and relu activation function as well.
   The last layer has 30 neurons which represent the 30 different words labels.

General Idea Behind the Model :

We noticed that the sounds where transformed into picture in a way it represent the evolution of the frequency over the time.
that is the reason why we decided to take advantage of this structured data , and therefore us a CNN layer as a first layer.
we naturally used a fully connected neural network to complete the learn from the CNN, and in order to balance with the 2 layer CNN , we decided to set 2 hidden layer in our fully connected.

The other hyper-parameters of our model such as the number of neurons and filters, learning rate, optimizer, etc.. where decided after number of experiment where we tried to improve and to reach the best accuracy for our model over the validation set.