REPORT EX 5

Aim: Train a model to classify a speech command using speech data. We want to classify our data which contains 30 different categories of commands.
To do so, we transformed the audio sounds into pictures using Gcommand_loader that was given to us. We tried different architectures to find the one with the best accuracy (more than 90%). We will show you in that report, an example which didn't give us the best one, and our final architecture that worked well.

Here is an example of architecture that we tried but that didn't give us the best accuracy :

CNN :

The first architecture was the CNN model with multiple Convolutional layers

First layer:

1) Convolution layer with input=1, outputs=60, kernel_size=3, stride=2, padding= 2
2) Activation function ReLU.

Second layer:

1) Convolution layer with input=60, outputs=100, kernel_size=6, stride=2, padding=1
2) Activation function ReLU.
3) Max Pool layer kernel_size=2x5, stride=2

Fully connected classifier with dropout and two hidden layers:

1) Fully connected layer 2048 neurons.
2) Activation function ReLU
3) Fully connected layer 512 neurons
4) Activation function ReLU.
5) Fully connected layer 30 neurons. They represent the words as labels.

We got an accuracy of 84%.

We tried different architectures. With one layer of fully connected, we found out that the accuracy wasn't good enough. Then, we added layers to separate the examples.

We also tried different values of learning rate until we found the one that gave us a fast increase without jumps betweens the epochs.

With a lot of manipulations we found that by adding max pool layer after each convolution has positive impact on accuracy we also found that optimal kernel size for each conv layer is 5 when the steps equals 1. Also by adding two additional blocks of convolution gave us better performance on training phase, with an accuracy of 91.2% :

CNN

1st layer:

1) Convolution layer with input=1, output=32, kernel_size=5x5, stride=1, padding=2
2) Activation function ReLU
3) Max Pool layer kernel_size=2, stride=2, padding=0, dilation=1

2nd layer:

1) Convolution layer with input=32, output=64, kernel_size=5x5, stride=1, padding=2
2) Activation function ReLU
3) Max Pool layer kernel_size=2, stride=2, padding=0, dilation=1

3rd layer:

1) Convolution layer with input=64, output=64, kernel_size=5x5, stride=1, padding=2
2) Activation function ReLU
3) Max Pool layer kernel_size=2, stride=2, padding=0, dilation=1

4th layer:

1) Convolution layer with input=64, output=32, kernel_size=5x5, stride=1, padding=2
2) Activation function ReLU
3) Max Pool layer kernel_size=2, stride=2, padding=0, dilation=1

Classifier:

1) Linear input = 1920, output = 500
2) Dropout with p = 0.5
3) Activation function ReLU
4) Linear input = 500, output = 100
5) Dropout with p = 0.5
6) Activation function ReLU
7) Linear input = 100, output = 30

Result of our accuracy and loss on train and validation sets

Epoch: 1

Train set: Accuracy: 24492/30000(**81.64%**), Average loss: **0.62309718**

Validation set: Accuracy: 5282/6798**(77.70%**), Average loss: **0.74002993**

Epoch: 2

Train set: Accuracy: 27406/30000(**91.35%**), Average loss: **0.31223658**

Validation set: Accuracy: 5912/6798(**86.97%**), Average loss: **0.44074333**

Epoch: 3

Train set: Accuracy: 28052/30000(**93.51%**), Average loss: **0.23457111**

Validation set: Accuracy: 6054/6798(**89.06%**), Average loss: **0.37903434**

Epoch: 4

Train set: Accuracy: 28695/30000(**95.65%**), Average loss: **0.15612289**

Validation set: Accuracy: 6119/6798(**90.01%**), Average loss: **0.33460078**

Epoch: 5

Train set: Accuracy: 28805/30000(**96.02%**), Average loss: **0.13044284**

Validation set: Accuracy: 6165/6798(**90.69%**), Average loss: **0.31851918**

Epoch: 6

Train set: Accuracy: 29121/30000(**97.07%**), Average loss: **0.10565013**

Validation set: Accuracy: 6190/6798(**91.06%**), Average loss: **0.31457543**

Epoch: 7

Train set: Accuracy: 29247/30000(**97.49%**), Average loss: **0.08476888**

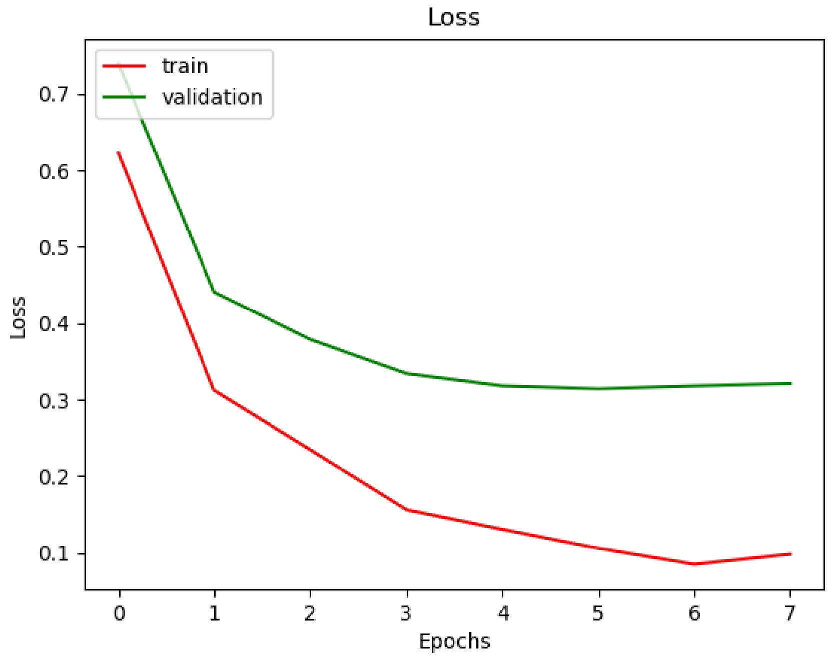
Validation set: Accuracy: 6232/6798(**91.67%**), Average loss: **0.31843066**

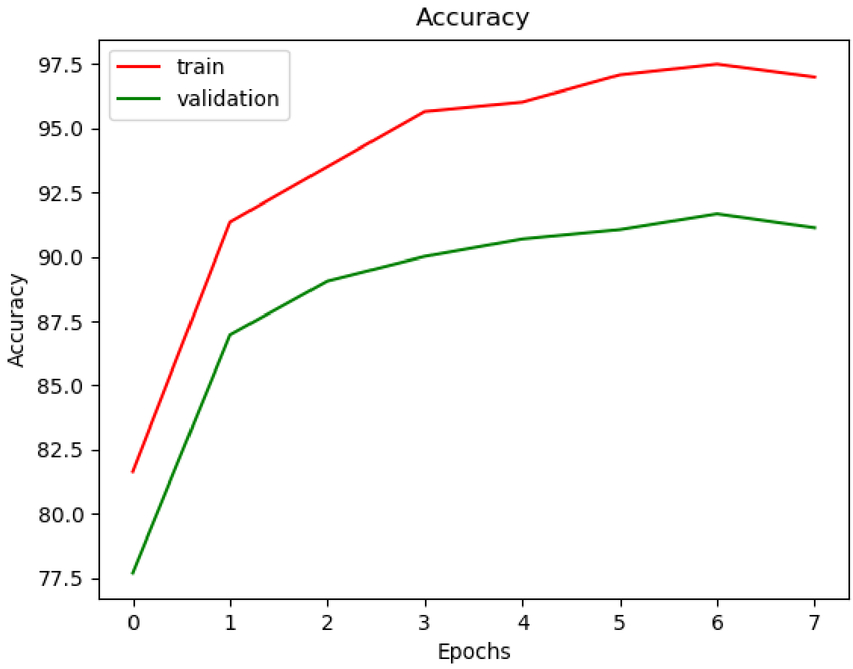
Epoch: 8

Train set: Accuracy: 29095/30000(**96.98%**), Average loss: **0.09786282**

Validation set: Accuracy: 6195/6798(**91.13%**), Average loss: **0.32149196**

Graphs of accuracy and loss


Loss
train
validation
Loss
Epochs


Accuracy
train
validation
Accuracy
Epochs