CSE 599 Deep Learning: Project Milestone

Teammates: Arjun Singh (arjuns13), Vivek Kumar (vivekuma), Aditi Singhal (aditi91)

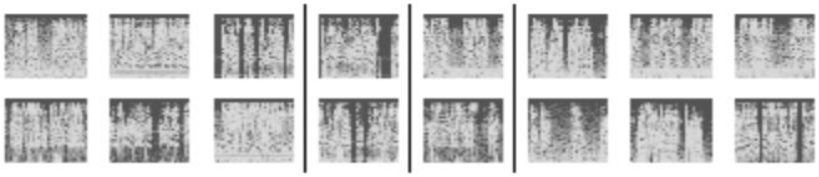
1. **Short description of the problem**

   Most biometric authentication systems rely on some kind of visual input. Two such systems, which are most popular for smartphones, are fingerprints and face ID. However, there are devices such as Alexa and Google Home which only receive audio as the input and hence need some kind of voice based authentication system.
   Our goal for the project is to design an authentication system using one shot learning based on both voice/audio from a speaker as well as the specto-grams generated from these audio samples. The idea is inspired by the DeepFace paper which does something similar for face recognition.

2. **What data are you using?**

   Out data source is the LibriSpeech dataset which has 460 hours of audio. This corresponds to about 1000 unique speakers. As of today, we have generated the specto-grams corresponding to 250 unique speakers, each with about 250 samples.

   Following is an example of some pairs of specto-grams and the labels corresponding to these pairs:



```
[[1.]
 [1.]
 [1.]
 [0.]
 [0.]
 [0.]
 [1.]
 [1.]]
```

3. **What simple baselines have you tried for this problem. Baselines need not be deep-learning methods?**

   Although we plan to begin the work on the audio samples, leveraging audio features such as MFCCs, in our immediate next steps, as a baseline we began with processing the data via images in the form of the specto-grams of these audios.

   Based on the things we've learned in class and the second homework, we were able to use this form of data and train Convolutional Neural Networks for the task.

4. **How are you tackling your problem? Describe the method or network design.**

   The problem is being tackled by using a Siamese network which takes a pair of images as its input with labels denoting whether the pairs come from the same source (specto-grams from the same speaker) or from two different sources (specto-grams from two different speakers).

   The loss used in our currently working implementation is the Contrastive Loss, where the goal is to predict the relative distances between inputs. In our case, this relative distance must be higher for specto-grams of audio originating from different speakers, and smaller for the ones originating from the same speaker. The function penalizes the model, if this isn't the case.

   The Network design used is provided as follows:

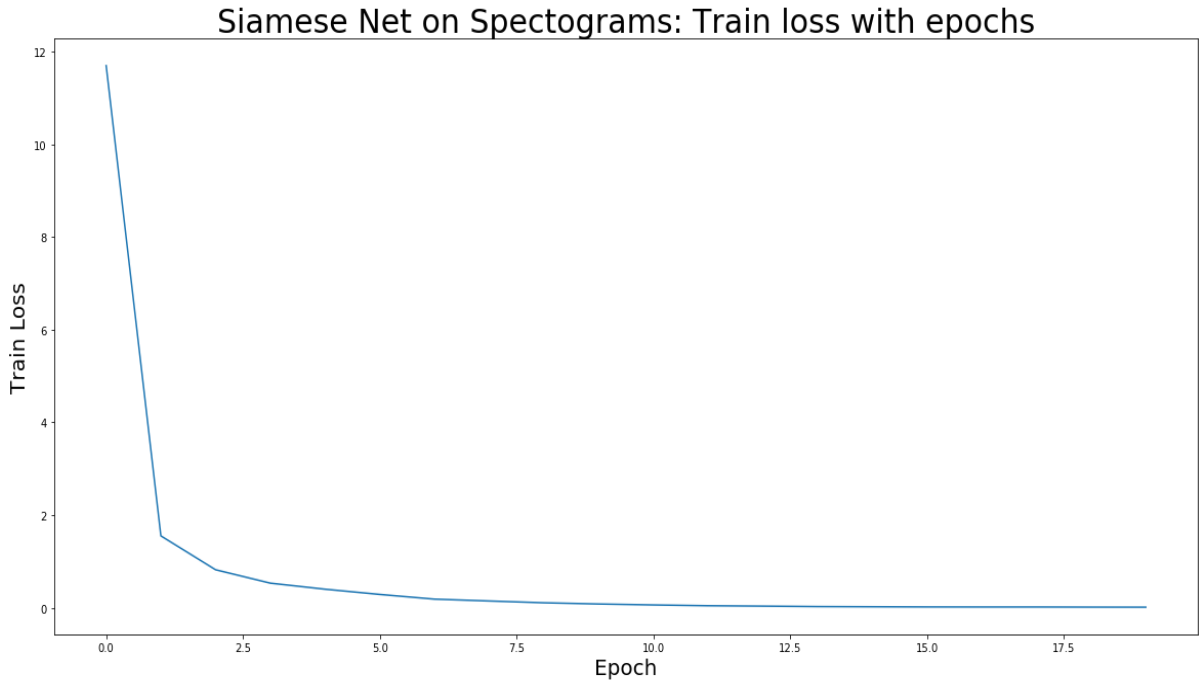
```
SiameseNetwork(
  (cnn1): Sequential(
    (0): ReflectionPad2d((1, 1, 1, 1))
    (1): Conv2d(1, 4, kernel_size=(3, 3), stride=(1, 1))
    (2): ReLU(inplace)
    (3): BatchNorm2d(4, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (4): ReflectionPad2d((1, 1, 1, 1))
    (5): Conv2d(4, 8, kernel_size=(3, 3), stride=(1, 1))
    (6): ReLU(inplace)
    (7): BatchNorm2d(8, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (8): ReflectionPad2d((1, 1, 1, 1))
    (9): Conv2d(8, 8, kernel_size=(3, 3), stride=(1, 1))
    (10): ReLU(inplace)
    (11): BatchNorm2d(8, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  )
  (fc1): Sequential(
    (0): Linear(in_features=412232, out_features=500, bias=True)
    (1): ReLU(inplace)
    (2): Linear(in_features=500, out_features=500, bias=True)
    (3): ReLU(inplace)
    (4): Linear(in_features=500, out_features=5, bias=True)
  )
)
```

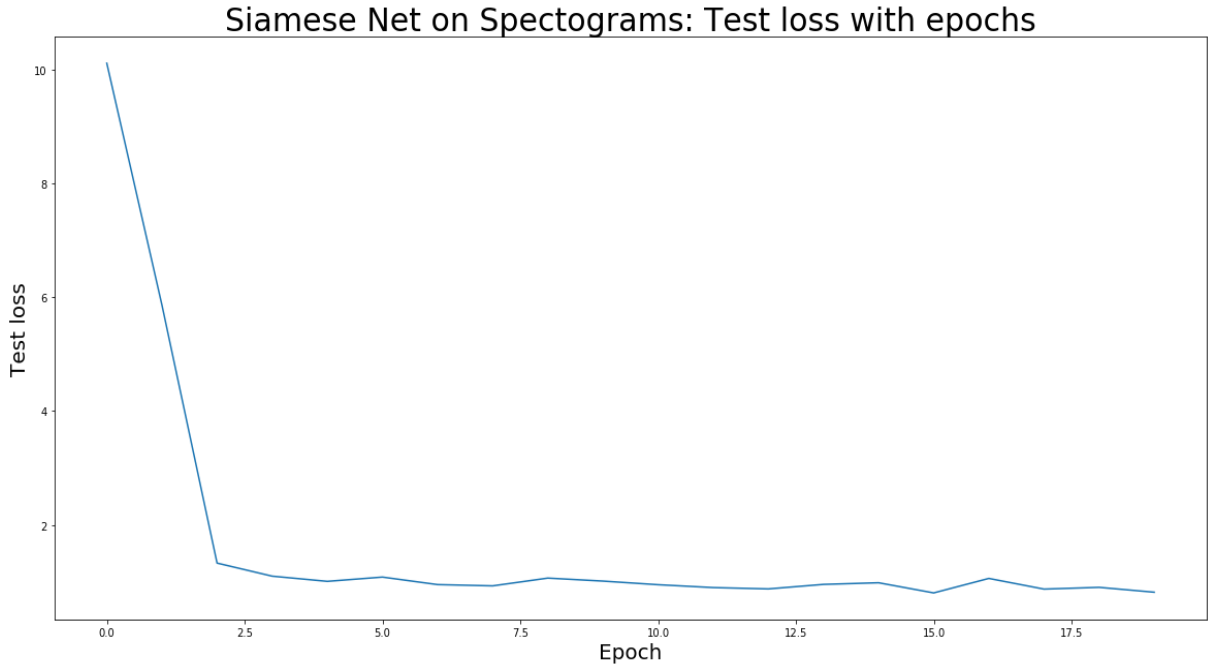
Details regarding the current training process:

- The training set includes pairs of samples from specto-grams belonging to 6 unique speakers
- The test set includes pairs of samples from specto-grams belonging to 3 unique speakers, different from any of the speakers in the train dataset
- All the images are used a grayscale for two reasons: (1) as a proof of concept we didn't want to use multiple input (RGB) channels and (2) since these are specto-grams, we don't think color should have a role to play in this, intuitively (let us know if we're missing something)
- Each speaker roughly has 250 specto-grams associated with them
- Train batch-size of 64 and test batch-size of 32 have been used
- Adam was used as the optimizer with learning rate = 0.0005
- Training has currently been done for 20 epochs

5. **How are your initial results?**

The following plots show the training and test loss over the 20 epochs for the aforementioned Siamese Network. Loss here refers to the Contrastive Loss.





6. **What issues have you faced? How have you solved them or how do you plan on solving them?**

   As we began the research and the implementation, there are a number of issues we've already faced, some of which we have a plan to mitigate, while

for others we'll continue to do our research, and consult the TAs, if need be. These are as follows:

- Since our loss is defined as contrastive loss, we're unable to figure out a way to use a metric like 'Accuracy'. This will be true, even if we were to append/replace this with a loss like Triplet Loss.
- Since this is, in a way, a classification problem, of whether the sample pair belongs to the same speaker or not, we may need to define a threshold for loss to determine positive versus negative predictions, in turn to figure out the accuracy. We don't have a great baseline for that threshold.
- We've also considered using the 'Cross Entropy' loss, just in order to be able to use accuracy as a metric, but will be happy to hear your thoughts/feedback for this.

Planned weekly timeline for the project

Week	**Target deliverable**
24th Nov	We aim to improve on the architecture of the current Siamese network, along with better tuning of its hyperparameters. We also aim to be using a much larger data set, rather than 6 and 3 speakers (which we're using right now). The MFCCs should also be generated for these audio samples.
28th Nov	The Network should be working for the MFCC data and we should be almost done with the optimal choice of the network for the specto-gram data. We also must have figured out exactly which loss and which metrics we'd be relying on.
2nd Dec	Final runs of the network with their corresponding documentation and working on getting the poster ready
4th Dec	Write-up for the final submission

Here is a link to our code on GitHub:
https://github.com/Speaker-Identification/IdentifySpeakersNet