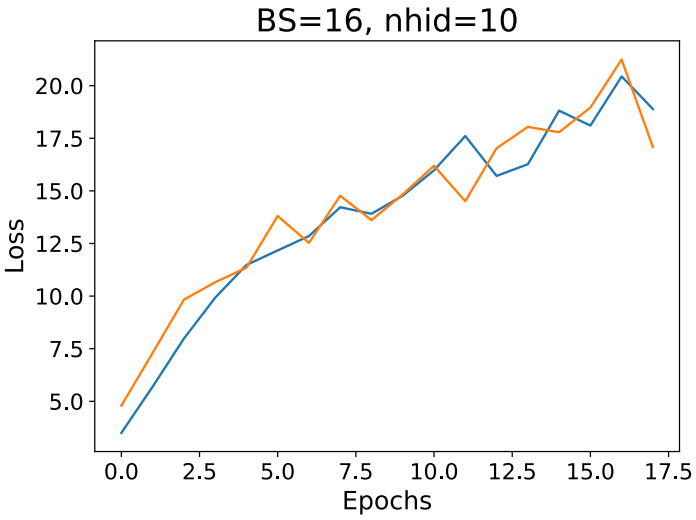
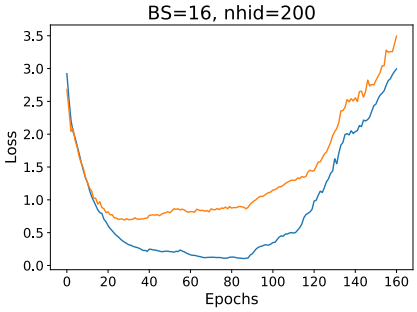
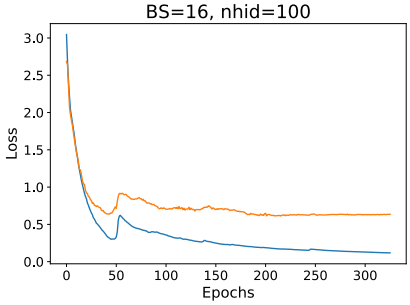
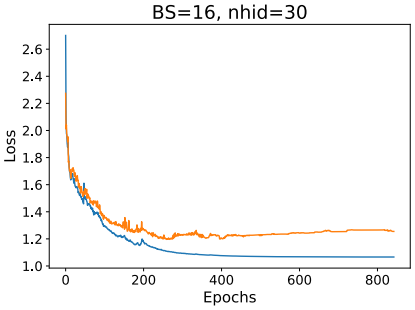
CSE 291-G Assignment report/ FatemehSadat Mireshghallah
Experiments:
When I used learning rate of 0.01 with 10 hidden units, it diverged, this is the plot:



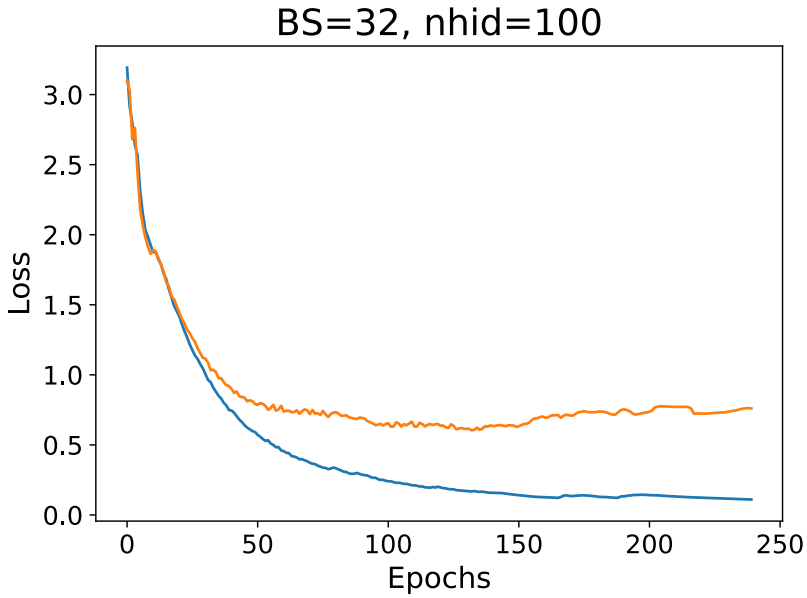
Based on the experiments, the 0.0001 learning rate yields better results. These are plots of some experiments using lr = 0.0001:







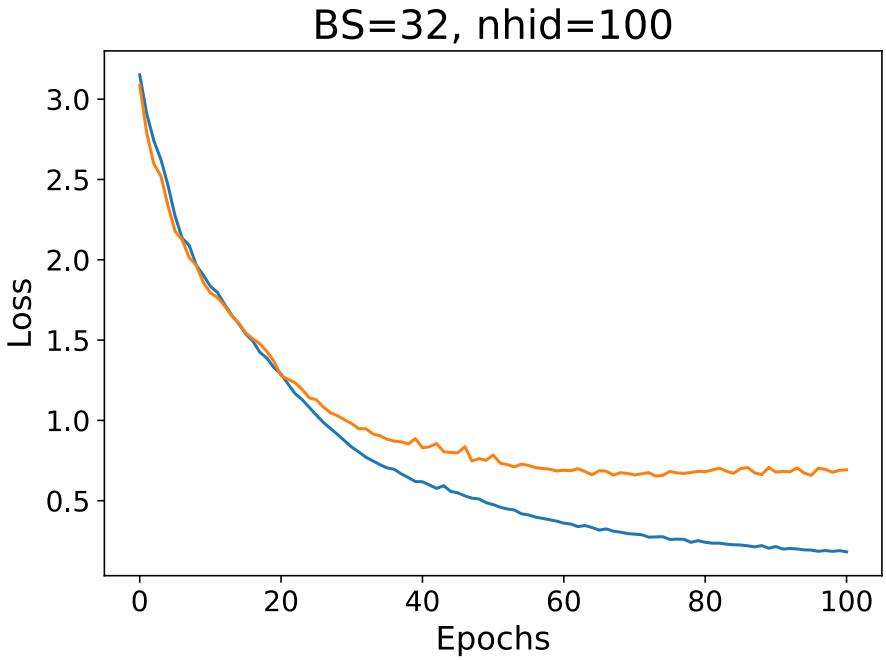
As it can be seen, for the number of hidden units, 100 is better, since 30 is too low (slow convergence to loss of about 1.0) and 200 diverges with lr of 0.0001. For batch size, 32 yielded relatively faster convergence:



The parameters here are BS=32, hidden= 100, lr=0.0001 wich yield training loss of 0.11 after 240 epochs of training which is the best I got. I also tried with BS=64 but didn't really make that much difference.

Something that really didn't work for me was a sigmoid activation I had added for the output layer of the decoder. It really slowed down the convergence. Also, another issue I encountered was with getting the loss of validation, which I realized has to be measured in a force teaching manner as well. I was sampling which caused the accuracy to keep going up.

I think one way that would help improve the model is to add layers to the gru cell (stack layers up). Also, I tried the bi-directional GRU, this is the result (for lr=0.0001):



One would expect it to be much better, since it also captures the context from future. But it seems similar. I think it might be because there is not much context, I mean the context doesn't have that much effect on the recognition (or maybe I have implemented it wrong).

In the codes file submitted, there are two notebooks, one is the original assignment, and another one which has bidirectional in its name, which is uses the bidirectional GRU.