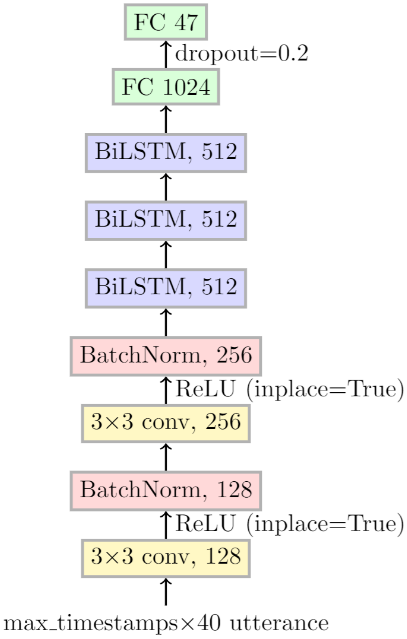
Homework 3 Part 2
Utterance To Phoneme Mapping

Yanjia Duan

1 Model Architecture

The model uses two convolution layers for utterance feature extraction. Then, three layers of bidirectional LSTMs with hidden size 512 are used for recognizing utterance. Then, two linear layers are used for classification. The architecture is shown below:



1.1 Loss Functions

CTCLoss with `blank=46`.

1.2 Hyper Parameters

Model Architectures:

- 1^{st} conv: `in_channel=40, out_channel=128, kernel_size=3, stride=1, padding=1, bias=False`
- 2^{nd} conv: `in_channel=128, out_channel=256, kernel_size=3, stride=1, padding=1, bias=False`
- LSTMs: `hidden_size=512, num_layers=3, bidirectional=True`
- 1^{st} linear: `in_features=1024, out_features=1024`
- 2^{nd} linear: `in_features=1024, out_features=47`
- Dropout: `p=0.2`

Loss: CTCLoss: `blank=46` Optimizer: Adam: `lr=1e-3, weight_decay=5e-5` Training: `batch_size=64`
Scheduler: ReduceLROnPlateau: `patience=3, threshold=1e-2, factor=0.5`

2 Other Interesting Facts

This task has 46 phoneme labels, and the number of training data is just about 25,000. The validation loss of a small model can quickly go down to ≈10 in five epochs. Therefore, using a large complex model is likely to overfit the training data. So I chose to use only 2 conv layers with small numbers of output channels, and only 1 linear layer (besides the classification layer) with a small output feature size.

The resulting loss function is very complex and has many local minimum within a very small range. So, I first train the model with a relatively large learning rate (`1e-3`) to explore the surface. For the next 20 epochs I gradually reduce the learning rate to find a good local minimum. Then, I use a fairly small learning rate (for example, `1e-6`) to fine tune the model and gradually reduce the learning rate for another 20 epochs. The validation loss could decrease to almost 0.32 for my best model.