Artificial Intelligence Nanodegree

Voice User Interfaces

Project: Speech Recognition with Neural Networks

In this notebook, some template code has already been provided for you, and you will need to implement additional functionality to successfully complete this project. You will not need to modify the included code beyond what is requested. Sections that begin with **'(IMPLEMENTATION)'** in the header indicate that the following blocks of code will require additional functionality which you must provide. Please be sure to read the instructions carefully!

> **Note**: Once you have completed all of the code implementations, you need to finalize your work by exporting the Jupyter Notebook as an HTML document. Before exporting the notebook to html, all of the code cells need to have been run so that reviewers can see the final implementation and output. You can then export the notebook by using the menu above and navigating to \n", "**File -> Download as -> HTML (.html)**. Include the finished document along with this notebook as your submission.

In addition to implementing code, there will be questions that you must answer which relate to the project and your implementation. Each section where you will answer a question is preceded by a **'Question X'** header. Carefully read each question and provide thorough answers in the following text boxes that begin with **'Answer:'**. Your project submission will be evaluated based on your answers to each of the questions and the implementation you provide.

> **Note:** Code and Markdown cells can be executed using the **Shift + Enter** keyboard shortcut. Markdown cells can be edited by double-clicking the cell to enter edit mode.

The rubric contains *optional* "Stand Out Suggestions" for enhancing the project beyond the minimum requirements. If you decide to pursue the "Stand Out Suggestions", you should include the code in this Jupyter notebook.

Introduction

In this notebook, you will build a deep neural network that functions as part of an end-to-end automatic speech recognition (ASR) pipeline! Your completed pipeline will accept raw audio as input and return a predicted transcription of the spoken language. The full pipeline is summarized in the figure below.

- **STEP 1** is a pre-processing step that converts raw audio to one of two feature representations that are commonly used for ASR.

- **STEP 2** is an acoustic model which accepts audio features as input and returns a probability distribution over all potential transcriptions. After learning about the basic types of neural networks that are often used for acoustic modeling, you will engage in your own investigations, to design your own acoustic model!
- **STEP 3** in the pipeline takes the output from the acoustic model and returns a predicted transcription.

Feel free to use the links below to navigate the notebook:

The Data

We begin by investigating the dataset that will be used to train and evaluate your pipeline. LibriSpeech (http://www.danielpovey.com/files/2015_icassp_librispeech.pdf) is a large corpus of English-read speech, designed for training and evaluating models for ASR. The dataset contains 1000 hours of speech derived from audiobooks. We will work with a small subset in this project, since larger-scale data would take a long while to train. However, after completing this project, if you are interested in exploring further, you are encouraged to work with more of the data that is provided online (http://www.openslr.org/12/).

In the code cells below, you will use the `vis_train_features` module to visualize a training example. The supplied argument `index=0` tells the module to extract the first example in the training set. (You are welcome to change `index=0` to point to a different training example, if you like, but please **DO NOT** amend any other code in the cell.) The returned variables are:

- `vis_text` - transcribed text (label) for the training example.
- `vis_raw_audio` - raw audio waveform for the training example.
- `vis_mfcc_feature` - mel-frequency cepstral coefficients (MFCCs) for the training example.
- `vis_spectrogram_feature` - spectrogram for the training example.
- `vis_audio_path` - the file path to the training example.

In [1]:

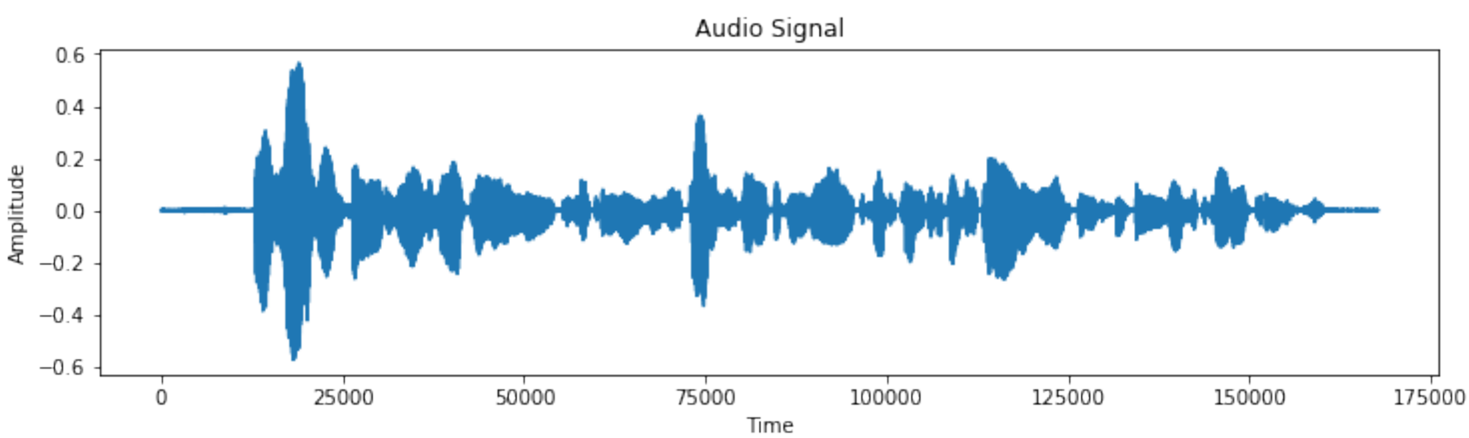
```
from data_generator import vis_train_features

# extract label and audio features for a single training example
vis_text, vis_raw_audio, vis_mfcc_feature, vis_spectrogram_feature, vis_audio_path = vis_train_features()
```

```
There are 2136 total training examples.
```

The following code cell visualizes the audio waveform for your chosen example, along with the corresponding transcript. You also have the option to play the audio in the notebook!

In [2]:

```
from IPython.display import Markdown, display
from data_generator import vis_train_features, plot_raw_audio
from IPython.display import Audio
%matplotlib inline

# plot audio signal
plot_raw_audio(vis_raw_audio)
# print length of audio signal
display(Markdown('**Shape of Audio Signal** : ' + str(vis_raw_audio.shape)))
# print transcript corresponding to audio clip
display(Markdown('**Transcript** : ' + str(vis_text)))
# play the audio file
Audio(vis_audio_path)
```



Shape of Audio Signal : (167691,)

Transcript : the houses seemed miserable in the extreme especially to an eye accustomed to the smiling neatness of english cottages

Out[2]:

STEP 1: Acoustic Features for Speech Recognition

For this project, you won't use the raw audio waveform as input to your model. Instead, we provide code that first performs a pre-processing step to convert the raw audio to a feature representation that has historically proven successful for ASR models. Your acoustic model will accept the feature representation as input.

In this project, you will explore two possible feature representations. *After completing the project*, if you'd like to read more about deep learning architectures that can accept raw audio input, you are encouraged to explore this research paper (https://pdfs.semanticscholar.org/a566/cd4a8623d661a4931814d9dffc72ecbf63c4.pdf).

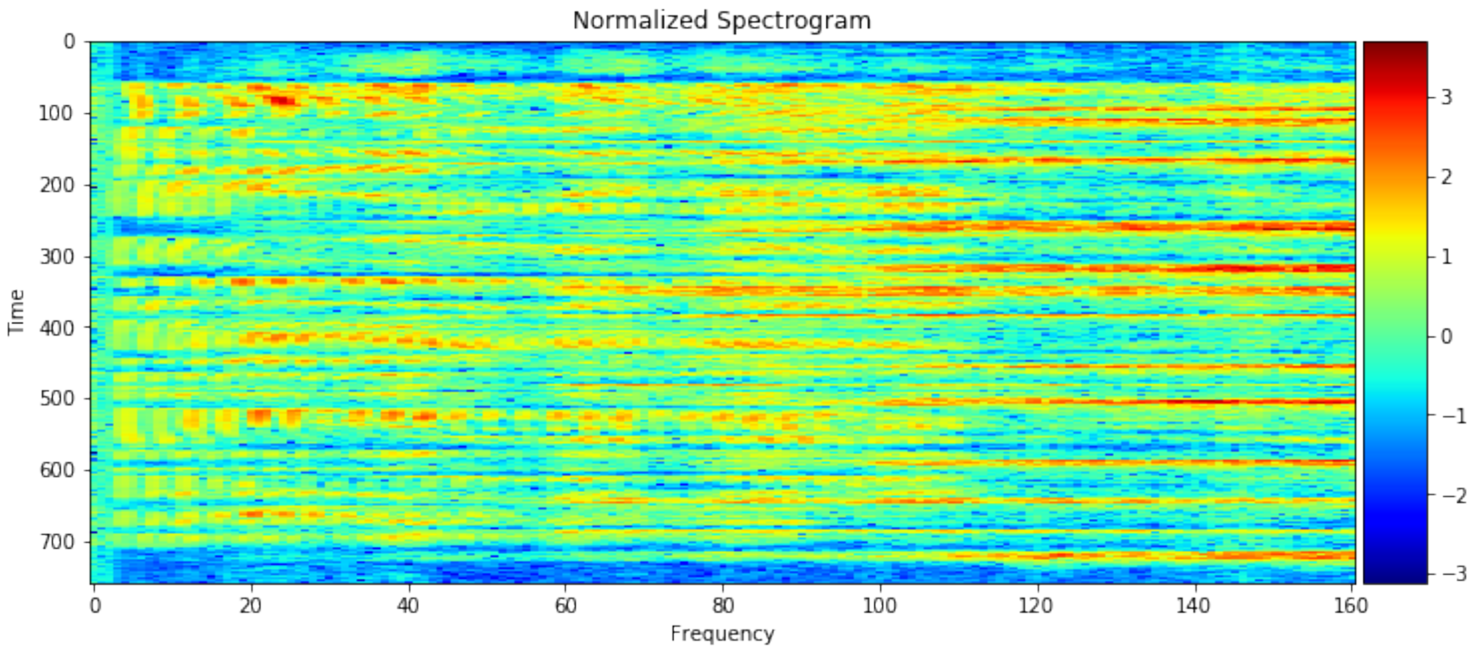
Spectrograms

The first option for an audio feature representation is the spectrogram (https://www.youtube.com/watch?v=_FatxGN3vAM). In order to complete this project, you will **not** need to dig deeply into the details of how a spectrogram is calculated; but, if you are curious, the code for calculating the spectrogram was borrowed from this repository (https://github.com/baidu-research/ba-dls-deepspeech). The implementation appears in the `utils.py` file in your repository.

The code that we give you returns the spectrogram as a 2D tensor, where the first (*vertical*) dimension indexes time, and the second (*horizontal*) dimension indexes frequency. To speed the convergence of your algorithm, we have also normalized the spectrogram. (You can see this quickly in the visualization below by noting that the mean value hovers around zero, and most entries in the tensor assume values close to zero.)

In [3]:

```
from data_generator import plot_spectrogram_feature

# plot normalized spectrogram
plot_spectrogram_feature(vis_spectrogram_feature)
# print shape of spectrogram
display(Markdown('**Shape of Spectrogram** : ' + str(vis_spectrogram_feature.shape)))
```



Shape of Spectrogram : (759, 161)

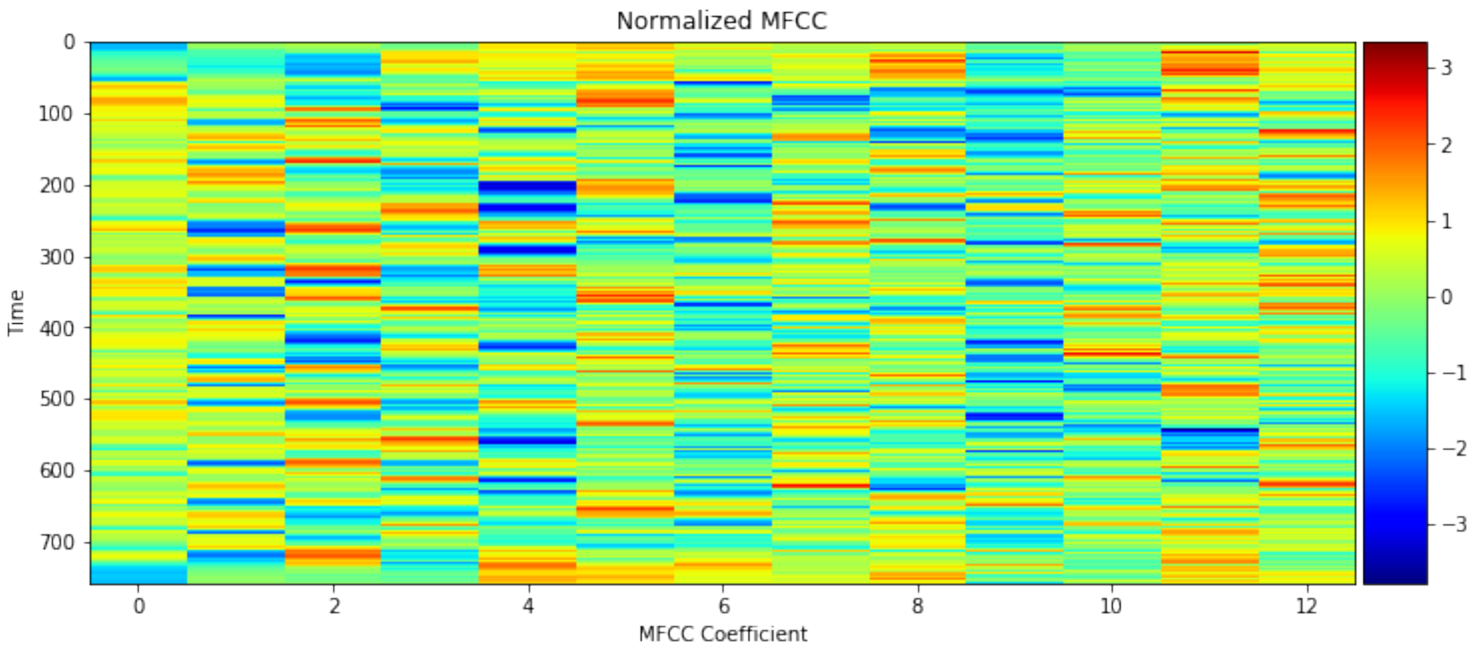
Mel-Frequency Cepstral Coefficients (MFCCs)

The second option for an audio feature representation is [MFCCs (https://en.wikipedia.org/wiki/Mel-frequency_cepstrum)](https://en.wikipedia.org/wiki/Mel-frequency_cepstrum). You do **not** need to dig deeply into the details of how MFCCs are calculated, but if you would like more information, you are welcome to peruse the [documentation (https://github.com/jameslyons/python_speech_features)](https://github.com/jameslyons/python_speech_features) of the `python_speech_features` Python package. Just as with the spectrogram features, the MFCCs are normalized in the supplied code.

The main idea behind MFCC features is the same as spectrogram features: at each time window, the MFCC feature yields a feature vector that characterizes the sound within the window. Note that the MFCC feature is much lower-dimensional than the spectrogram feature, which could help an acoustic model to avoid overfitting to the training dataset.

In [4]:

```
from data_generator import plot_mfcc_feature

# plot normalized MFCC
plot_mfcc_feature(vis_mfcc_feature)
# print shape of MFCC
display(Markdown('**Shape of MFCC** : ' + str(vis_mfcc_feature.shape)))
```



Shape of MFCC : (759, 13)

When you construct your pipeline, you will be able to choose to use either spectrogram or MFCC features. If you would like to see different implementations that make use of MFCCs and/or spectrograms, please check out the links below:

- This repository (https://github.com/baidu-research/ba-dls-deepspeech) uses spectrograms.
- This repository (https://github.com/mozilla/DeepSpeech) uses MFCCs.
- This repository (https://github.com/buriburisuri/speech-to-text-wavenet) also uses MFCCs.
- This repository (https://github.com/pannous/tensorflow-speech-recognition/blob/master/speech_data.py) experiments with raw audio, spectrograms, and MFCCs as features.

STEP 2: Deep Neural Networks for Acoustic Modeling

In this section, you will experiment with various neural network architectures for acoustic modeling.

You will begin by training five relatively simple architectures. **Model 0** is provided for you. You will write code to implement **Models 1**, **2**, **3**, and **4**. If you would like to experiment further, you are welcome to create and train more models under the **Models 5+** heading.

All models will be specified in the `sample_models.py` file. After importing the `sample_models` module, you will train your architectures in the notebook.

After experimenting with the five simple architectures, you will have the opportunity to compare their performance. Based on your findings, you will construct a deeper architecture that is designed to outperform all of the shallow models.

For your convenience, we have designed the notebook so that each model can be specified and trained on separate occasions. That is, say you decide to take a break from the notebook after training **Model 1**. Then, you need not re-execute all prior code cells in the notebook before training **Model 2**. You need only re-execute the code cell below, that is marked with **`RUN THIS CODE CELL IF YOU ARE RESUMING THE NOTEBOOK AFTER A BREAK`**, before transitioning to the code cells corresponding to **Model 2**.

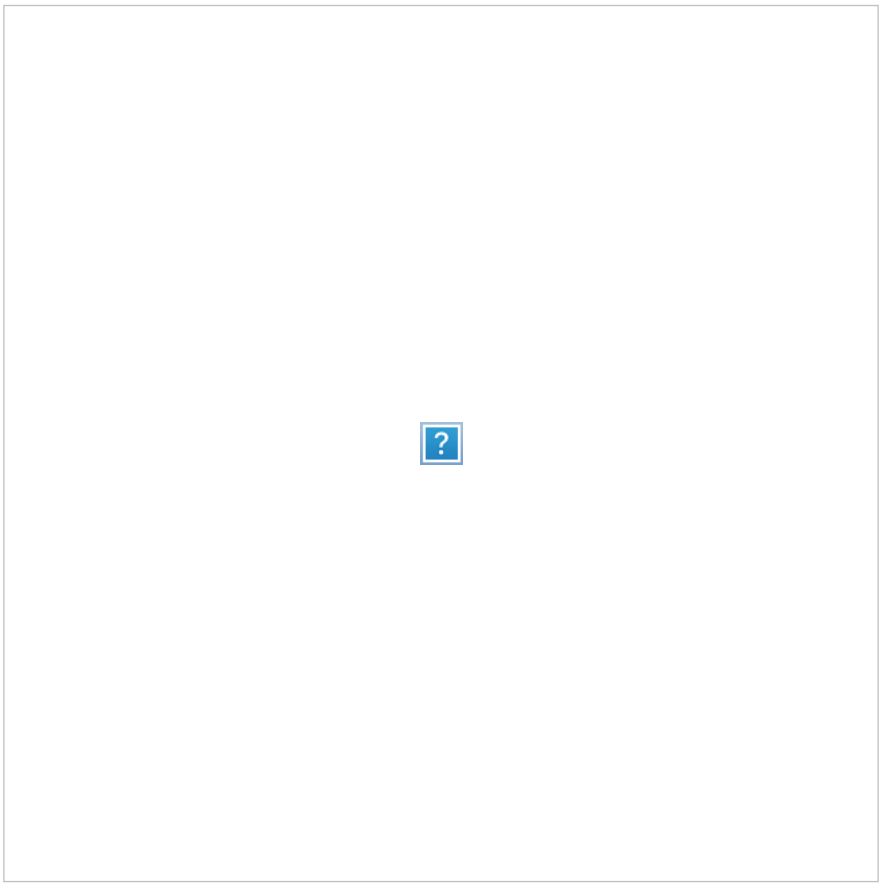
In [1]:

```
#####################################################################
# RUN THIS CODE CELL IF YOU ARE RESUMING THE NOTEBOOK AFTER A BREAK #
#####################################################################

# allocate 50% of GPU memory (if you like, feel free to change this)
from keras.backend.tensorflow_backend import set_session
import tensorflow as tf
config = tf.ConfigProto()
config.gpu_options.per_process_gpu_memory_fraction = 0.5
set_session(tf.Session(config=config))

# watch for any changes in the sample_models module, and reload it automatical
ly
%load_ext autoreload
%autoreload 2
# import NN architectures for speech recognition
from sample_models import *
# import function for training acoustic model
from train_utils import train_model
```

```
Using TensorFlow backend.
```

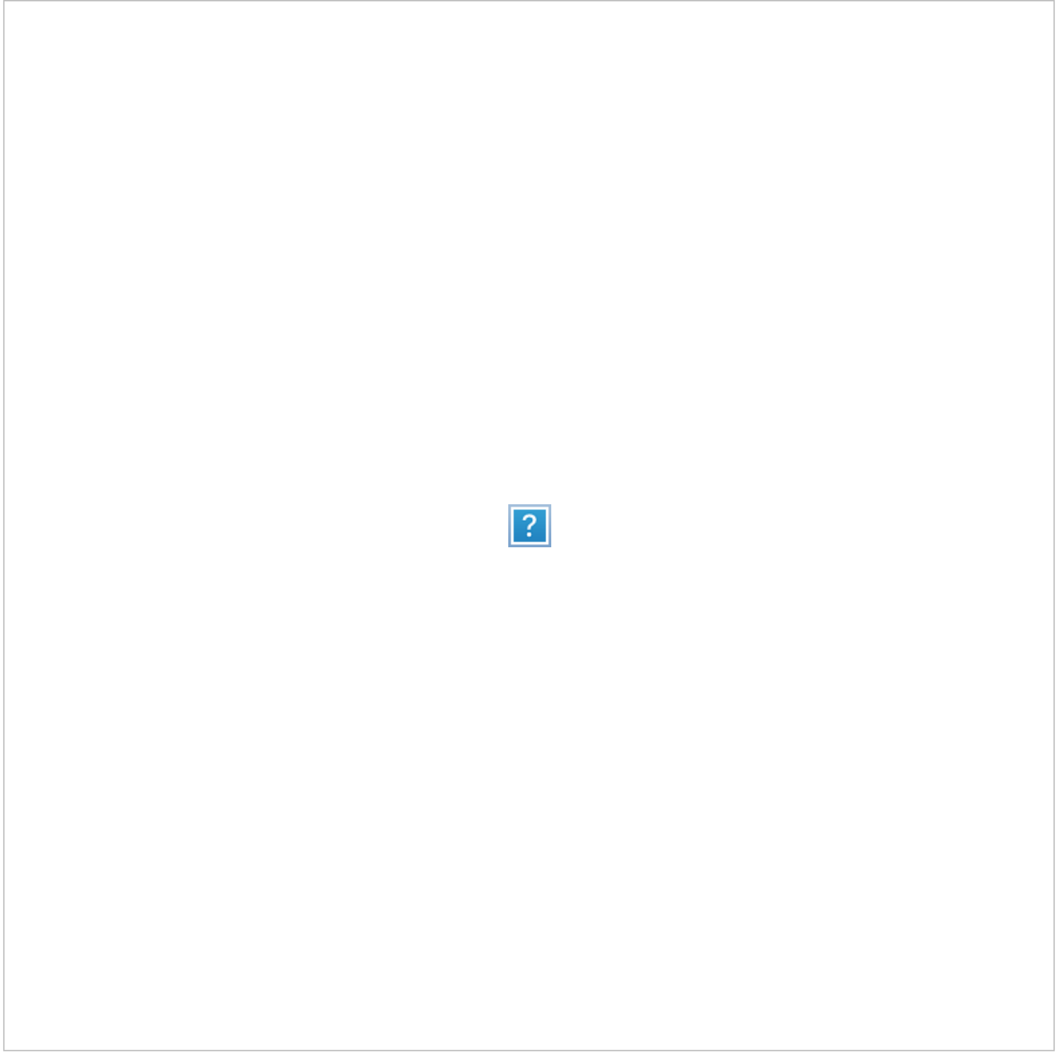
Model 0: RNN

Given their effectiveness in modeling sequential data, the first acoustic model you will use is an RNN. As shown in the figure below, the RNN we supply to you will take the time sequence of audio features as input.



At each time step, the speaker pronounces one of 28 possible characters, including each of the 26 letters in the English alphabet, along with a space character (" "), and an apostrophe (').

The output of the RNN at each time step is a vector of probabilities with 29 entries, where the i-th entry encodes the probability that the i-th character is spoken in the time sequence. (The extra 29th character is an empty "character" used to pad training examples within batches containing uneven lengths.) If you would like to peek under the hood at how characters are mapped to indices in the probability vector, look at the `char_map.py` file in the repository. The figure below shows an equivalent, rolled depiction of the RNN that shows the output layer in greater detail.



The model has already been specified for you in Keras. To import it, you need only run the code cell below.

In [14]:

```
model_0 = simple_rnn_model(input_dim=161) # change to 13 if you would like to use MFCC features
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
the_input (InputLayer)       (None, None, 161)         0         
_________________________________________________________________
rnn (GRU)                    (None, None, 29)          16617     
_________________________________________________________________
softmax (Activation)         (None, None, 29)          0         
=================================================================
Total params: 16,617
Trainable params: 16,617
Non-trainable params: 0
_________________________________________________________________
None
```

As explored in the lesson, you will train the acoustic model with the [CTC loss](http://www.cs.toronto.edu/~graves/icml_2006.pdf) (http://www.cs.toronto.edu/~graves/icml_2006.pdf) criterion. Custom loss functions take a bit of hacking in Keras, and so we have implemented the CTC loss function for you, so that you can focus on trying out as many deep learning architectures as possible :). If you'd like to peek at the implementation details, look at the `add_ctc_loss` function within the `train_utils.py` file in the repository.

To train your architecture, you will use the `train_model` function within the `train_utils` module; it has already been imported in one of the above code cells. The `train_model` function takes three **required** arguments:

- `input_to_softmax` - a Keras model instance.
- `pickle_path` - the name of the pickle file where the loss history will be saved.
- `save_model_path` - the name of the HDF5 file where the model will be saved.

If we have already supplied values for `input_to_softmax`, `pickle_path`, and `save_model_path`, please **DO NOT** modify these values.

There are several **optional** arguments that allow you to have more control over the training process. You are welcome to, but not required to, supply your own values for these arguments.

- `minibatch_size` - the size of the minibatches that are generated while training the model (default: `20`).
- `spectrogram` - Boolean value dictating whether spectrogram (`True`) or MFCC (`False`) features are used for training (default: `True`).
- `mfcc_dim` - the size of the feature dimension to use when generating MFCC features (default: `13`).
- `optimizer` - the Keras optimizer used to train the model (default: `SGD(lr=0.02, decay=1e-6, momentum=0.9, nesterov=True, clipnorm=5)`).
- `epochs` - the number of epochs to use to train the model (default: `20`). If you choose to modify this parameter, make sure that it is *at least* 20.
- `verbose` - controls the verbosity of the training output in the `model.fit_generator` method (default: `1`).
- `sort_by_duration` - Boolean value dictating whether the training and validation sets are sorted by (increasing) duration before the start of the first epoch (default: `False`).

The `train_model` function defaults to using spectrogram features; if you choose to use these features, note that the acoustic model in `simple_rnn_model` should have `input_dim=161`. Otherwise, if you choose to use MFCC features, the acoustic model should have `input_dim=13`.

We have chosen to use `GRU` units in the supplied RNN. If you would like to experiment with `LSTM` or `SimpleRNN` cells, feel free to do so here. If you change the `GRU` units to `SimpleRNN` cells in `simple_rnn_model`, you may notice that the loss quickly becomes undefined (`nan`) - you are strongly encouraged to check this for yourself! This is due to the exploding gradients problem (http://www.wildml.com/2015/10/recurrent-neural-networks-tutorial-part-3-backpropagation-through-time-and-vanishing-gradients/). We have already implemented gradient clipping (https://arxiv.org/pdf/1211.5063.pdf) in your optimizer to help you avoid this issue.

IMPORTANT NOTE: If you notice that your gradient has exploded in any of the models below, feel free to explore more with gradient clipping (the `clipnorm` argument in your optimizer) or swap out any `SimpleRNN` cells for `LSTM` or `GRU` cells. You can also try restarting the kernel to restart the training process.

In [7]:

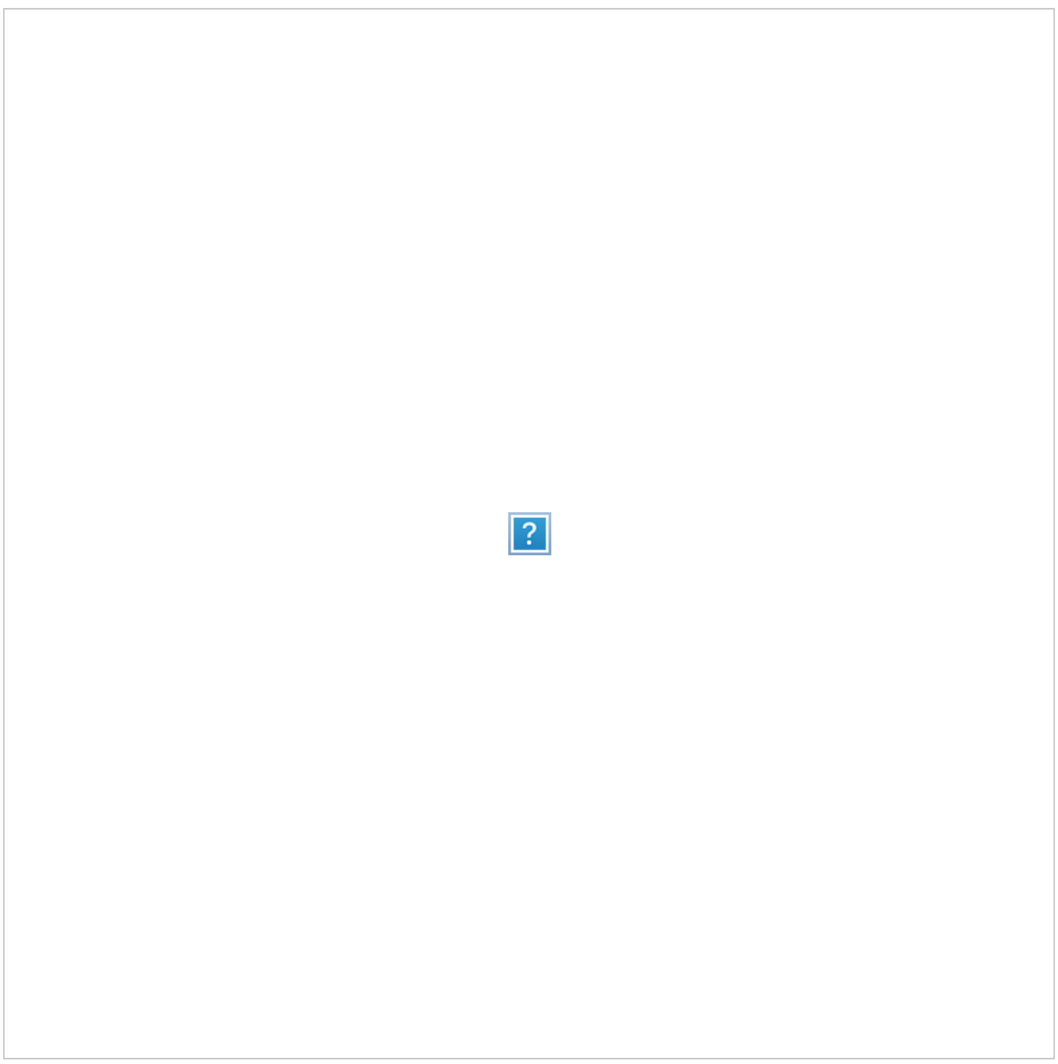
```python
train_model(input_to_softmax=model_0,
            pickle_path='model_0.pickle',
            save_model_path='model_0.h5',
            spectrogram=True) # change to False if you would like to use MFCC features
```

```
Epoch 1/20
106/106 [==============================] - 273s - loss: 826.9665 - val_loss: 727.5970
Epoch 2/20
106/106 [==============================] - 265s - loss: 752.5132 - val_loss: 729.8761
Epoch 3/20
106/106 [==============================] - 264s - loss: 751.5996 - val_loss: 719.1012
Epoch 4/20
106/106 [==============================] - 262s - loss: 751.6240 - val_loss: 728.2657
Epoch 5/20
106/106 [==============================] - 262s - loss: 751.4530 - val_loss: 722.8607
Epoch 6/20
106/106 [==============================] - 263s - loss: 751.5928 - val_loss: 725.5043
Epoch 7/20
106/106 [==============================] - 263s - loss: 751.5125 - val_loss: 736.7984
Epoch 8/20
106/106 [==============================] - 263s - loss: 751.7690 - val_loss: 717.5269
Epoch 9/20
106/106 [==============================] - 263s - loss: 750.9447 - val_loss: 723.7717
Epoch 10/20
106/106 [==============================] - 264s - loss: 752.1150 - val_loss: 730.9567
Epoch 11/20
106/106 [==============================] - 263s - loss: 752.1930 - val_loss: 725.8995
Epoch 12/20
106/106 [==============================] - 262s - loss: 751.3190 - val_loss: 732.9482
Epoch 13/20
106/106 [==============================] - 264s - loss: 751.6791 - val_loss: 721.6303
Epoch 14/20
106/106 [==============================] - 262s - loss: 751.6776 - val_loss: 728.1979
Epoch 15/20
106/106 [==============================] - 262s - loss: 751.2949 - val_loss: 728.0571
Epoch 16/20
106/106 [==============================] - 261s - loss: 749.6759 - val_loss: 718.1781
Epoch 17/20
```

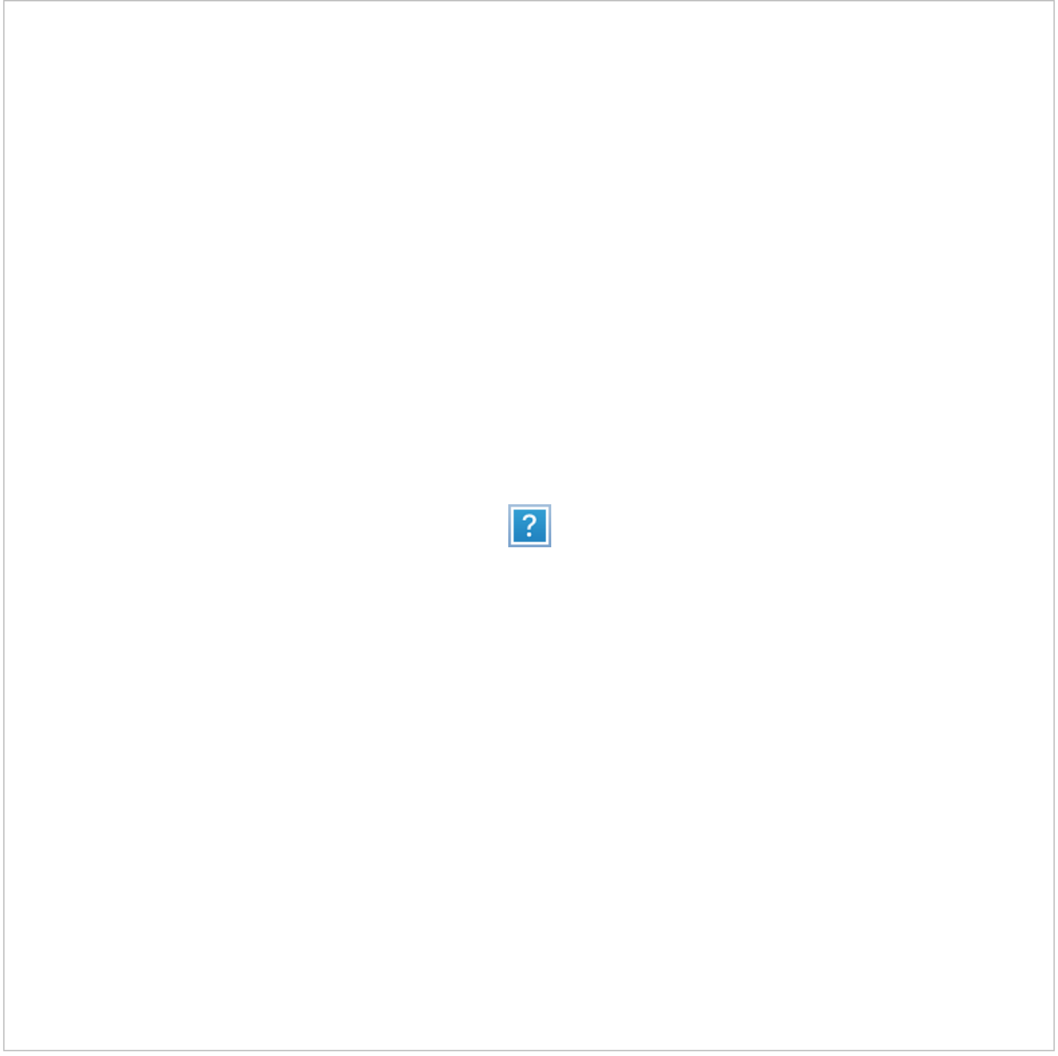
```
106/106 [==============================] - 263s - loss: 751.6380 -
val_loss: 726.3826
Epoch 18/20
106/106 [==============================] - 264s - loss: 751.4313 -
val_loss: 727.1105
Epoch 19/20
106/106 [==============================] - 262s - loss: 750.1568 -
val_loss: 722.1338
Epoch 20/20
106/106 [==============================] - 263s - loss: 751.3014 -
val_loss: 729.6760
```

(IMPLEMENTATION) Model 1: RNN + TimeDistributed Dense

Read about the TimeDistributed (https://keras.io/layers/wrappers/) wrapper and the BatchNormalization (https://keras.io/layers/normalization/) layer in the Keras documentation. For your next architecture, you will add batch normalization (https://arxiv.org/pdf/1510.01378.pdf) to the recurrent layer to reduce training times. The `TimeDistributed` layer will be used to find more complex patterns in the dataset. The unrolled snapshot of the architecture is depicted below.



The next figure shows an equivalent, rolled depiction of the RNN that shows the (`TimeDistrbuted`) dense and output layers in greater detail.



Use your research to complete the `rnn_model` function within the `sample_models.py` file. The function should specify an architecture that satisfies the following requirements:

- The first layer of the neural network should be an RNN (`SimpleRNN`, `LSTM`, or `GRU`) that takes the time sequence of audio features as input. We have added `GRU` units for you, but feel free to change `GRU` to `SimpleRNN` or `LSTM`, if you like!
- Whereas the architecture in `simple_rnn_model` treated the RNN output as the final layer of the model, you will use the output of your RNN as a hidden layer. Use `TimeDistributed` to apply a `Dense` layer to each of the time steps in the RNN output. Ensure that each `Dense` layer has `output_dim` units.

Use the code cell below to load your model into the `model_1` variable. Use a value for `input_dim` that matches your chosen audio features, and feel free to change the values for `units` and `activation` to tweak the behavior of your recurrent layer.

In [15]:

```
model_1 = rnn_model(input_dim=161, # change to 13 if you would like to use MFC
C features
                    units=200,
                    activation='relu')
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
the_input (InputLayer)       (None, None, 161)         0
_________________________________________________________________
rnn (GRU)                    (None, None, 200)         217200
_________________________________________________________________
bn_conv_1d (BatchNormalizati (None, None, 200)         800
_________________________________________________________________
time_distributed_6 (TimeDist (None, None, 29)          5829
_________________________________________________________________
softmax (Activation)         (None, None, 29)          0
=================================================================
Total params: 223,829
Trainable params: 223,429
Non-trainable params: 400
_________________________________________________________________
None
```

Please execute the code cell below to train the neural network you specified in `input_to_softmax`. After the model has finished training, the model is saved (https://keras.io/getting-started/faq/#how-can-i-save-a-keras-model) in the HDF5 file `model_1.h5`. The loss history is saved (https://wiki.python.org/moin/UsingPickle) in `model_1.pickle`. You are welcome to tweak any of the optional parameters while calling the `train_model` function, but this is not required.

In [3]:

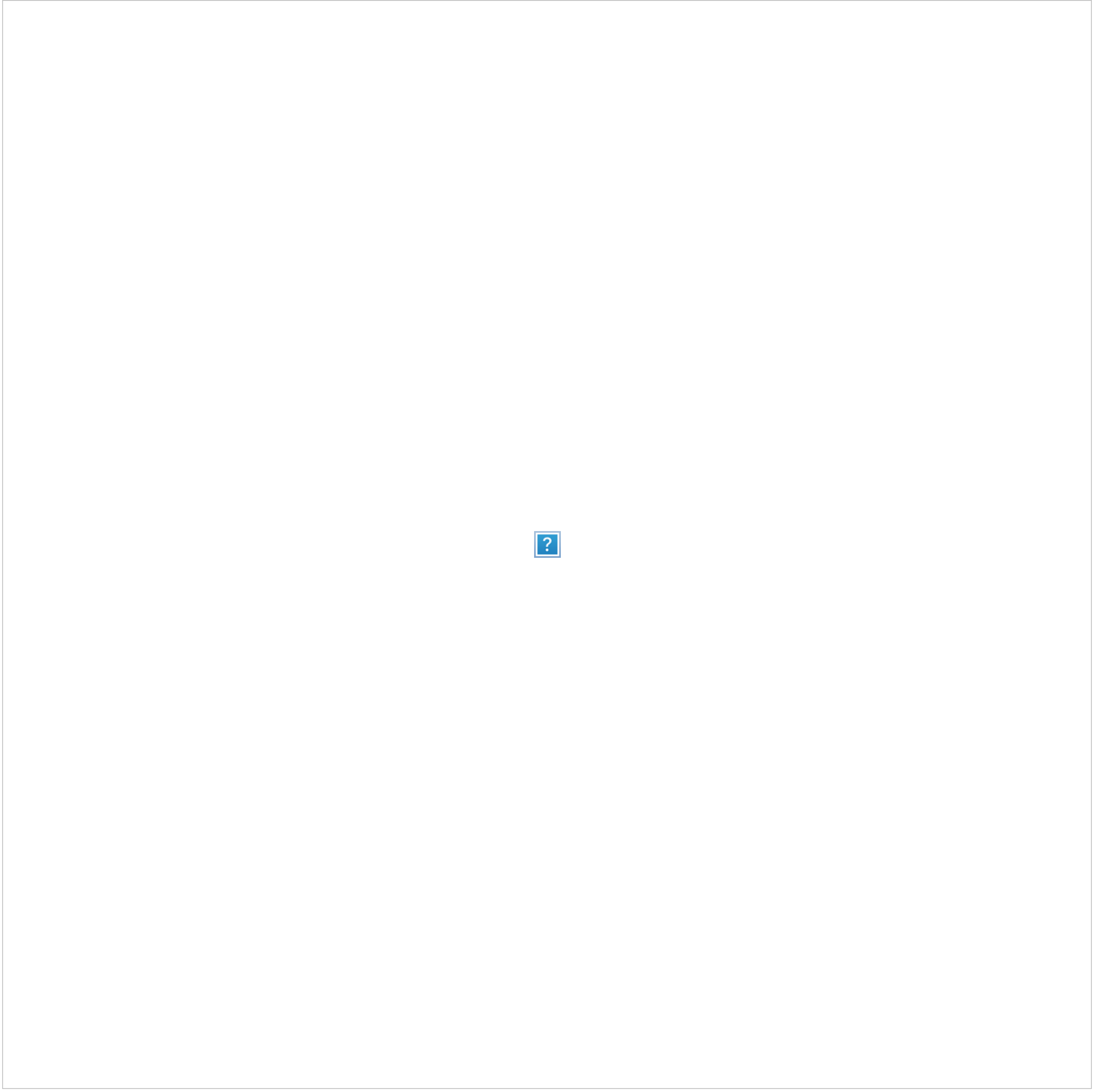
```
train_model(input_to_softmax=model_1,
            pickle_path='model_1.pickle',
            save_model_path='model_1.h5',
            spectrogram=True) # change to False if you would like to use MFCC
features
```

```
Epoch 1/20
106/106 [==============================] - 262s - loss: 305.0907 -
val_loss: 266.3746
Epoch 2/20
106/106 [==============================] - 263s - loss: 209.1511 -
val_loss: 193.7016
Epoch 3/20
106/106 [==============================] - 256s - loss: 184.8055 -
val_loss: 172.5797
Epoch 4/20
106/106 [==============================] - 256s - loss: 167.5047 -
val_loss: 164.1684
Epoch 5/20
106/106 [==============================] - 255s - loss: 157.6461 -
val_loss: 159.2960
Epoch 6/20
```

```
106/106 [==============================] - 255s - loss: 150.3956 -
val_loss: 150.9137
Epoch 7/20
106/106 [==============================] - 253s - loss: 145.5209 -
val_loss: 146.3591
Epoch 8/20
106/106 [==============================] - 255s - loss: 140.9787 -
val_loss: 144.4939
Epoch 9/20
106/106 [==============================] - 254s - loss: 138.2091 -
val_loss: 145.7322
Epoch 10/20
106/106 [==============================] - 255s - loss: 135.6048 -
val_loss: 141.8172
Epoch 11/20
106/106 [==============================] - 253s - loss: 132.7534 -
val_loss: 141.2671
Epoch 12/20
106/106 [==============================] - 255s - loss: 131.5465 -
val_loss: 145.4249
Epoch 13/20
106/106 [==============================] - 253s - loss: 130.7423 -
val_loss: 138.0322
Epoch 14/20
106/106 [==============================] - 255s - loss: 128.3839 -
val_loss: 138.0165
Epoch 15/20
106/106 [==============================] - 254s - loss: 126.3698 -
val_loss: 139.0408
Epoch 16/20
106/106 [==============================] - 255s - loss: 125.6943 -
val_loss: 135.4560
Epoch 17/20
106/106 [==============================] - 253s - loss: 125.1282 -
val_loss: 140.2005
Epoch 18/20
106/106 [==============================] - 253s - loss: 124.2635 -
val_loss: 137.4569
Epoch 19/20
106/106 [==============================] - 254s - loss: 126.7306 -
val_loss: 141.8771
Epoch 20/20
106/106 [==============================] - 255s - loss: 127.2249 -
val_loss: 140.6487
```

(IMPLEMENTATION) Model 2: CNN + RNN + TimeDistributed Dense

The architecture in `cnn_rnn_model` adds an additional level of complexity, by introducing a [1D convolution layer (https://keras.io/layers/convolutional/#conv1d)](https://keras.io/layers/convolutional/#conv1d).



This layer incorporates many arguments that can be (optionally) tuned when calling the `cnn_rnn_model` module. We provide sample starting parameters, which you might find useful if you choose to use spectrogram audio features.

If you instead want to use MFCC features, these arguments will have to be tuned. Note that the current architecture only supports values of `'same'` or `'valid'` for the `conv_border_mode` argument.

When tuning the parameters, be careful not to choose settings that make the convolutional layer overly small. If the temporal length of the CNN layer is shorter than the length of the transcribed text label, your code will throw an error.

Before running the code cell below, you must modify the `cnn_rnn_model` function in `sample_models.py`. Please add batch normalization to the recurrent layer, and provide the same `TimeDistributed` layer as before.

In [16]:

```
model_2 = cnn_rnn_model(input_dim=161, # change to 13 if you would like to use
MFCC features
                        filters=200,
                        kernel_size=11,
                        conv_stride=2,
                        conv_border_mode='valid',
                        units=200)
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
the_input (InputLayer)       (None, None, 161)         0         
_________________________________________________________________
conv1d (Conv1D)              (None, None, 200)         354400    
_________________________________________________________________
bn_conv_1d (BatchNormalizati (None, None, 200)         800       
_________________________________________________________________
rnn (SimpleRNN)              (None, None, 200)         80200     
_________________________________________________________________
bn_simple_rnn (BatchNormaliz (None, None, 200)         800       
_________________________________________________________________
time_distributed_7 (TimeDist (None, None, 29)          5829      
_________________________________________________________________
softmax (Activation)         (None, None, 29)          0         
=================================================================
Total params: 442,029
Trainable params: 441,229
Non-trainable params: 800
_________________________________________________________________
None
```

Please execute the code cell below to train the neural network you specified in `input_to_softmax`. After the model has finished training, the model is saved (https://keras.io/getting-started/faq/#how-can-i-save-a-keras-model) in the HDF5 file `model_2.h5`. The loss history is saved (https://wiki.python.org/moin/UsingPickle) in `model_2.pickle`. You are welcome to tweak any of the optional parameters while calling the `train_model` function, but this is not required.

In [7]:

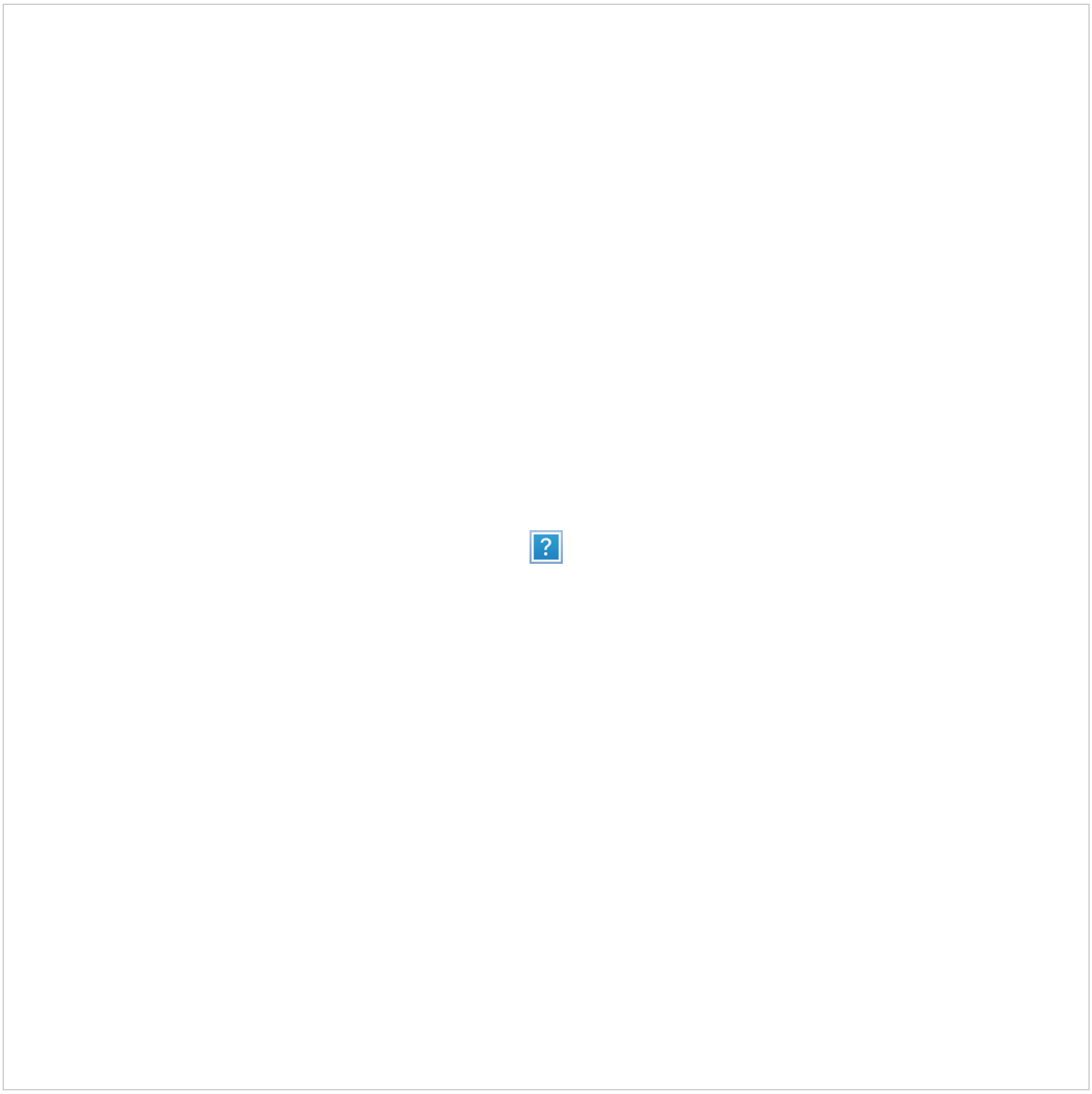
```
train_model(input_to_softmax=model_2,
            pickle_path='model_2.pickle',
            save_model_path='model_2.h5',
            spectrogram=True) # change to False if you would like to use MFCC
features
```

```
Epoch 1/20
106/106 [==============================] - 61s - loss: 244.8068 -
val_loss: 201.9662
Epoch 2/20
106/106 [==============================] - 56s - loss: 172.9962 -
val_loss: 160.0219
Epoch 3/20
106/106 [==============================] - 55s - loss: 148.6235 -
val_loss: 152.1978
```

```
Epoch 4/20
106/106 [==============================] - 55s - loss: 137.1272 -
val_loss: 145.7348
Epoch 5/20
106/106 [==============================] - 54s - loss: 128.6343 -
val_loss: 136.1398
Epoch 6/20
106/106 [==============================] - 55s - loss: 122.4205 -
val_loss: 134.3668
Epoch 7/20
106/106 [==============================] - 54s - loss: 117.7383 -
val_loss: 131.1280
Epoch 8/20
106/106 [==============================] - 54s - loss: 113.3958 -
val_loss: 132.7172
Epoch 9/20
106/106 [==============================] - 55s - loss: 109.1009 -
val_loss: 130.8541
Epoch 10/20
106/106 [==============================] - 54s - loss: 105.4067 -
val_loss: 130.3045
Epoch 11/20
106/106 [==============================] - 54s - loss: 102.6690 -
val_loss: 130.7059
Epoch 12/20
106/106 [==============================] - 55s - loss: 99.1126 - v
al_loss: 130.3630
Epoch 13/20
106/106 [==============================] - 55s - loss: 96.1172 - v
al_loss: 133.5040
Epoch 14/20
106/106 [==============================] - 54s - loss: 93.3058 - v
al_loss: 133.3372
Epoch 15/20
106/106 [==============================] - 55s - loss: 90.7179 - v
al_loss: 132.8380
Epoch 16/20
106/106 [==============================] - 55s - loss: 88.2505 - v
al_loss: 134.1363
Epoch 17/20
106/106 [==============================] - 55s - loss: 85.6168 - v
al_loss: 135.9311
Epoch 18/20
106/106 [==============================] - 54s - loss: 83.6011 - v
al_loss: 137.5858
Epoch 19/20
106/106 [==============================] - 55s - loss: 81.2098 - v
al_loss: 142.1707
Epoch 20/20
106/106 [==============================] - 55s - loss: 79.7823 - v
al_loss: 138.1723
```

(IMPLEMENTATION) Model 3: Deeper RNN + TimeDistributed Dense

Review the code in `rnn_model`, which makes use of a single recurrent layer. Now, specify an architecture in `deep_rnn_model` that utilizes a variable number `recur_layers` of recurrent layers. The figure below shows the architecture that should be returned if `recur_layers=2`. In the figure, the output sequence of the first recurrent layer is used as input for the next recurrent layer.



Feel free to change the supplied values of `units` to whatever you think performs best. You can change the value of `recur_layers`, as long as your final value is greater than 1. (As a quick check that you have implemented the additional functionality in `deep_rnn_model` correctly, make sure that the architecture that you specify here is identical to `rnn_model` if `recur_layers=1`.)

In [17]:

```
model_3 = deep_rnn_model(input_dim=161, # change to 13 if you would like to us
e MFCC features
                         units=200,
                         recur_layers=2)
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
the_input (InputLayer)       (None, None, 161)         0         
_________________________________________________________________
rnn0 (GRU)                   (None, None, 200)         217200    
_________________________________________________________________
bn_conv_1d0 (BatchNormalizat (None, None, 200)         800       
_________________________________________________________________
rnn1 (GRU)                   (None, None, 200)         240600    
_________________________________________________________________
bn_conv_1d1 (BatchNormalizat (None, None, 200)         800       
_________________________________________________________________
time_distributed_8 (TimeDist (None, None, 29)          5829      
_________________________________________________________________
softmax (Activation)         (None, None, 29)          0         
=================================================================
Total params: 465,229
Trainable params: 464,429
Non-trainable params: 800
_________________________________________________________________
None
```

Please execute the code cell below to train the neural network you specified in `input_to_softmax`. After the model has finished training, the model is saved (https://keras.io/getting-started/faq/#how-can-i-save-a-keras-model) in the HDF5 file `model_3.h5`. The loss history is saved (https://wiki.python.org/moin/UsingPickle) in `model_3.pickle`. You are welcome to tweak any of the optional parameters while calling the `train_model` function, but this is not required.

In [5]:

```
train_model(input_to_softmax=model_3,
            pickle_path='model_3.pickle',
            save_model_path='model_3.h5',
            spectrogram=True) # change to False if you would like to use MFCC
features
```

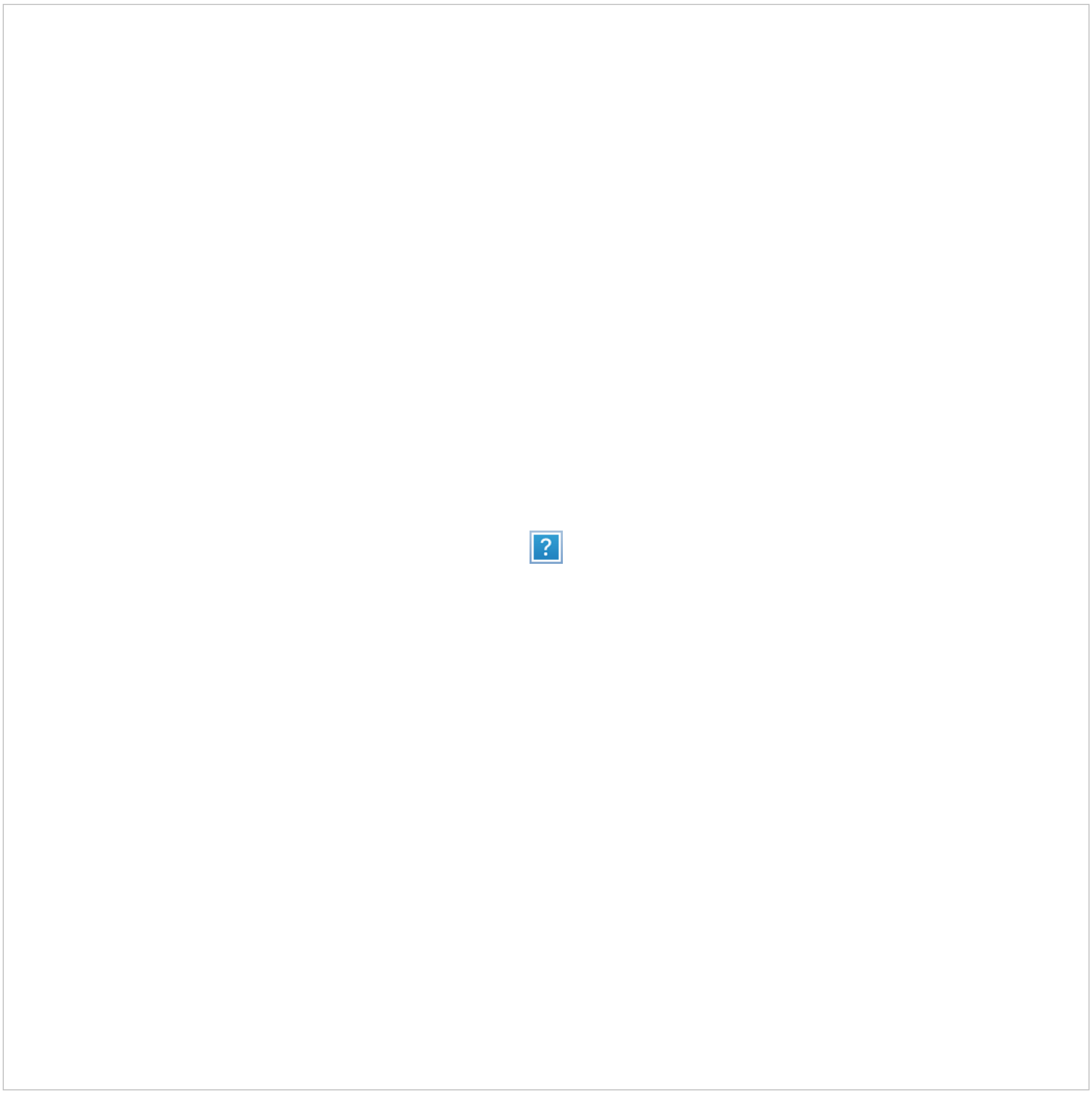
```
Epoch 1/20
106/106 [==============================] - 507s - loss: 312.5819 -
val_loss: 248.3049
Epoch 2/20
106/106 [==============================] - 493s - loss: 236.6660 -
val_loss: 207.2741
Epoch 3/20
106/106 [==============================] - 489s - loss: 209.0108 -
val_loss: 204.3543
Epoch 4/20
106/106 [==============================] - 492s - loss: 179.5015 -
val_loss: 168.4720
```

```
Epoch 5/20
106/106 [==============================] - 510s - loss: 162.8399 -
val_loss: 154.0702
Epoch 6/20
106/106 [==============================] - 508s - loss: 153.1116 -
val_loss: 148.0676
Epoch 7/20
106/106 [==============================] - 500s - loss: 145.6790 -
val_loss: 148.0781
Epoch 8/20
106/106 [==============================] - 495s - loss: 139.2540 -
val_loss: 143.3244
Epoch 9/20
106/106 [==============================] - 499s - loss: 133.7628 -
val_loss: 143.1459
Epoch 10/20
106/106 [==============================] - 507s - loss: 129.5097 -
val_loss: 136.5200
Epoch 11/20
106/106 [==============================] - 487s - loss: 124.8304 -
val_loss: 134.8822
Epoch 12/20
106/106 [==============================] - 491s - loss: 120.8744 -
val_loss: 134.9585
Epoch 13/20
106/106 [==============================] - 494s - loss: 117.4814 -
val_loss: 131.4986
Epoch 14/20
106/106 [==============================] - 492s - loss: 114.0314 -
val_loss: 130.4268
Epoch 15/20
106/106 [==============================] - 489s - loss: 111.0547 -
val_loss: 129.3562
Epoch 16/20
106/106 [==============================] - 494s - loss: 108.2449 -
val_loss: 130.7531
Epoch 17/20
106/106 [==============================] - 494s - loss: 105.4015 -
val_loss: 132.6674
Epoch 18/20
106/106 [==============================] - 487s - loss: 102.5714 -
val_loss: 129.1994
Epoch 19/20
106/106 [==============================] - 486s - loss: 99.5560 -
val_loss: 130.0924
Epoch 20/20
106/106 [==============================] - 486s - loss: 97.1715 -
val_loss: 130.3358
```

(IMPLEMENTATION) Model 4: Bidirectional RNN + TimeDistributed Dense

Read about the [Bidirectional (https://keras.io/layers/wrappers/)](https://keras.io/layers/wrappers/) wrapper in the Keras documentation. For your next architecture, you will specify an architecture that uses a single bidirectional RNN layer, before a (`TimeDistributed`) dense layer. The added value of a bidirectional RNN is described well in [this paper (http://www.cs.toronto.edu/~hinton/absps/DRNN_speech.pdf)](http://www.cs.toronto.edu/~hinton/absps/DRNN_speech.pdf).

> One shortcoming of conventional RNNs is that they are only able to make use of previous context. In speech recognition, where whole utterances are transcribed at once, there is no reason not to exploit future context as well. Bidirectional RNNs (BRNNs) do this by processing the data in both directions with two separate hidden layers which are then fed forwards to the same output layer.



Before running the code cell below, you must complete the `bidirectional_rnn_model` function in `sample_models.py`. Feel free to use `SimpleRNN`, `LSTM`, or `GRU` units. When specifying the `Bidirectional` wrapper, use `merge_mode='concat'`.

In [18]:

```
model_4 = bidirectional_rnn_model(input_dim=161, # change to 13 if you would l
ike to use MFCC features
                                  units=200)
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
the_input (InputLayer)       (None, None, 161)         0         
_________________________________________________________________
bidirectional_11 (Bidirectio (None, None, 400)         434400    
_________________________________________________________________
time_distributed_9 (TimeDist (None, None, 29)          11629     
_________________________________________________________________
softmax (Activation)         (None, None, 29)          0         
=================================================================
Total params: 446,029
Trainable params: 446,029
Non-trainable params: 0
_________________________________________________________________
None
```

Please execute the code cell below to train the neural network you specified in `input_to_softmax`. After the model has finished training, the model is saved (https://keras.io/getting-started/faq/#how-can-i-save-a-keras-model) in the HDF5 file `model_4.h5`. The loss history is saved (https://wiki.python.org/moin/UsingPickle) in `model_4.pickle`. You are welcome to tweak any of the optional parameters while calling the `train_model` function, but this is not required.

In [ ]:

```
train_model(input_to_softmax=model_4,
            pickle_path='model_4.pickle',
            save_model_path='model_4.h5',
            spectrogram=True) # change to False if you would like to use MFCC
features
```

```
Epoch 1/20
106/106 [==============================] - 463s - loss: 328.6116 -
val_loss: 282.8362
Epoch 2/20
106/106 [==============================] - 466s - loss: 250.4266 -
val_loss: 212.6035
Epoch 3/20
106/106 [==============================] - 463s - loss: 208.5978 -
val_loss: 192.9973
Epoch 4/20
106/106 [==============================] - 465s - loss: 194.6460 -
val_loss: 182.7656
Epoch 5/20
106/106 [==============================] - 466s - loss: 184.4290 -
val_loss: 178.0158
Epoch 6/20
106/106 [==============================] - 465s - loss: 175.5660 -
val_loss: 169.9380
Epoch 7/20
```

```
106/106 [==============================] - 467s - loss: 167.6150 -
val_loss: 162.1675
Epoch 8/20
106/106 [==============================] - 462s - loss: 160.4788 -
val_loss: 159.1697
Epoch 9/20
106/106 [==============================] - 463s - loss: 153.5957 -
val_loss: 156.2226
Epoch 10/20
106/106 [==============================] - 463s - loss: 147.2709 -
val_loss: 151.9135
Epoch 11/20
106/106 [==============================] - 469s - loss: 141.4530 -
val_loss: 147.6786
Epoch 12/20
106/106 [==============================] - 464s - loss: 136.3628 -
val_loss: 144.7224
Epoch 13/20
106/106 [==============================] - 468s - loss: 131.7110 -
val_loss: 142.0638
Epoch 14/20
106/106 [==============================] - 467s - loss: 127.2308 -
val_loss: 142.0087
Epoch 15/20
106/106 [==============================] - 469s - loss: 123.0587 -
val_loss: 141.5730
Epoch 16/20
106/106 [==============================] - 467s - loss: 119.2000 -
val_loss: 138.3363
Epoch 17/20
106/106 [==============================] - 465s - loss: 115.7211 -
val_loss: 135.5671
Epoch 18/20
106/106 [==============================] - 465s - loss: 112.1719 -
val_loss: 140.9935
Epoch 19/20
106/106 [==============================] - 463s - loss: 109.1201 -
val_loss: 138.8769
Epoch 20/20
106/106 [==============================] - 462s - loss: 105.8156 -
val_loss: 138.7642
```

(OPTIONAL IMPLEMENTATION) Models 5+

If you would like to try out more architectures than the ones above, please use the code cell below. Please continue to follow the same convention for saving the models; for the i-th sample model, please save the loss at **`model_i.pickle`** and saving the trained model at **`model_i.h5`**.

In [ ]:

```
## (Optional) TODO: Try out some more models!
### Feel free to use as many code cells as needed.
```

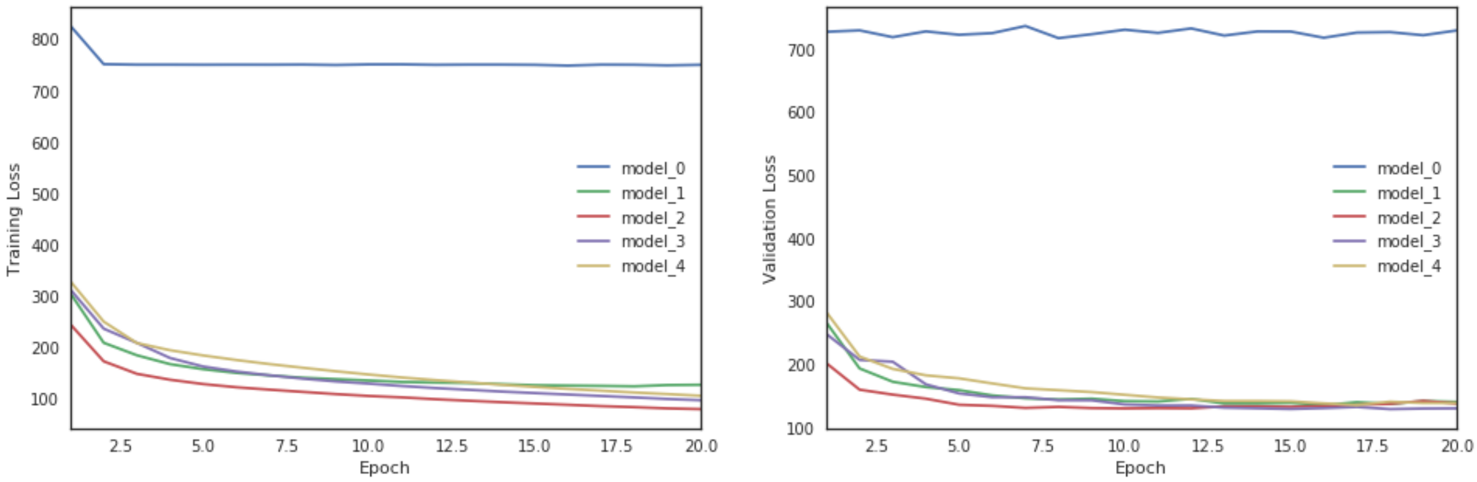
Compare the Models

Execute the code cell below to evaluate the performance of the drafted deep learning models. The training and validation loss are plotted for each model.

In [2]:

```
from glob import glob
import numpy as np
import _pickle as pickle
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set_style(style='white')

# obtain the paths for the saved model history
all_pickles = sorted(glob("results/*.pickle"))
# extract the name of each model
model_names = [item[8:-7] for item in all_pickles]
# extract the loss history for each model
valid_loss = [pickle.load( open( i, "rb" ) )['val_loss'] for i in all_pickles]
train_loss = [pickle.load( open( i, "rb" ) )['loss'] for i in all_pickles]
# save the number of epochs used to train each model
num_epochs = [len(valid_loss[i]) for i in range(len(valid_loss))]

fig = plt.figure(figsize=(16,5))

# plot the training loss vs. epoch for each model
ax1 = fig.add_subplot(121)
for i in range(len(all_pickles)):
    ax1.plot(np.linspace(1, num_epochs[i], num_epochs[i]),
            train_loss[i], label=model_names[i])
# clean up the plot
ax1.legend()
ax1.set_xlim([1, max(num_epochs)])
plt.xlabel('Epoch')
plt.ylabel('Training Loss')

# plot the validation loss vs. epoch for each model
ax2 = fig.add_subplot(122)
for i in range(len(all_pickles)):
    ax2.plot(np.linspace(1, num_epochs[i], num_epochs[i]),
            valid_loss[i], label=model_names[i])
# clean up the plot
ax2.legend()
ax2.set_xlim([1, max(num_epochs)])
plt.xlabel('Epoch')
plt.ylabel('Validation Loss')
plt.show()
```

```
/home/aind2/anaconda3/envs/aind-vui/lib/python3.5/site-packages/ma
tplotlib/font_manager.py:1297: UserWarning: findfont: Font family
['sans-serif'] not found. Falling back to DejaVu Sans
  (prop.get_family(), self.defaultFamily[fontext]))
```



Question 1: Use the plot above to analyze the performance of each of the attempted architectures. Which performs best? Provide an explanation regarding why you think some models perform better than others.

Answer: Model 3 seems to be the best performing architecture. It continues to decrease the validation loss the more it is trained (unlike for instance model 2 that decreases the training loss but increases validation loss in the last epochs, signaling some overfitting). Model 3 is the one with the lowest validation loss at the end.

Model's 2 improvement regarding model 1 is the CNN. Taking more hidden layers into account will allow the model to take into account more features and therefore potentially be more accurate. The amount of params from one model to another almost doubled. Some of the benefits can be found in this link: http://karol.piczak.com/papers/Piczak2015-ESC-ConvNet.pdf (http://karol.piczak.com/papers/Piczak2015-ESC-ConvNet.pdf). Model 3 sustitutes the CNN layer for an RNN layer (potentially for n RNN layers). RNN seem more appropiate for the given task as they perform better by taking into account the order series. A great summary of their difference from the following link https://datascience.stackexchange.com/questions/11619/rnn-vs-cnn-at-a-high-level (https://datascience.stackexchange.com/questions/11619/rnn-vs-cnn-at-a-high-level) is: 'A CNN will learn to recognize patterns across space. So, as you say, a CNN will learn to recognize components of an image (e.g., lines, curves, etc.) and then learn to combine these components to recognize larger structures (e.g., faces, objects, etc.).

You could say, in a very general way, that a RNN will similarly learn to recognize patterns across time. So a RNN that is trained to translate text might learn that "dog" should be translated differently if preceded by the word "hot".'

I expected intuitively for bidirectional rnn's (model 4) to perform better although the model used in this case seems to have overfitted the data in the last epochs (training loss decreased but validation loss increased). Adding dropout might help in this case.

It is worth mentioning that models 2,3 and 4 have approximately the same amount of parameters (and took about the same time to train). At the end of this notebook predictions of each model can be found. On a personal note, they seemed quite disappointing after all the training made but the final model seems to be the most accurate.

Model 0: RNN Total params: 16,617 Model 1: RNN + TimeDistributed Dense Total params: 223,829 Model 2: CNN + RNN + TimeDistributed Dense Total params: 442,029 Model 3: Deeper RNN + TimeDistributed Dense Total params: 465,229 Model 4: Bidirectional RNN + TimeDistributed Dense Total params: 446,029

(IMPLEMENTATION) Final Model

Now that you've tried out many sample models, use what you've learned to draft your own architecture! While your final acoustic model should not be identical to any of the architectures explored above, you are welcome to merely combine the explored layers above into a deeper architecture. It is **NOT** necessary to include new layer types that were not explored in the notebook.

However, if you would like some ideas for even more layer types, check out these ideas for some additional, optional extensions to your model:

- If you notice your model is overfitting to the training dataset, consider adding **dropout**! To add dropout to recurrent layers (https://faroit.github.io/keras-docs/1.0.2/layers/recurrent/), pay special attention to the `dropout_W` and `dropout_U` arguments. This paper (http://arxiv.org/abs/1512.05287) may also provide some interesting theoretical background.
- If you choose to include a convolutional layer in your model, you may get better results by working with **dilated convolutions**. If you choose to use dilated convolutions, make sure that you are able to accurately calculate the length of the acoustic model's output in the `model.output_length` lambda function. You can read more about dilated convolutions in Google's WaveNet paper (https://arxiv.org/abs/1609.03499). For an example of a speech-to-text system that makes use of dilated convolutions, check out this GitHub repository (https://github.com/buriburisuri/speech-to-text-wavenet). You can work with dilated convolutions in Keras (https://keras.io/layers/convolutional/) by paying special attention to the `padding` argument when you specify a convolutional layer.
- If your model makes use of convolutional layers, why not also experiment with adding **max pooling**? Check out this paper (https://arxiv.org/pdf/1701.02720.pdf) for example architecture that makes use of max pooling in an acoustic model.
- So far, you have experimented with a single bidirectional RNN layer. Consider stacking the bidirectional layers, to produce a deep bidirectional RNN (https://www.cs.toronto.edu/~graves/asru_2013.pdf)!

All models that you specify in this repository should have `output_length` defined as an attribute. This attribute is a lambda function that maps the (temporal) length of the input acoustic features to the (temporal) length of the output softmax layer. This function is used in the computation of CTC loss; to see this, look at the `add_ctc_loss` function in `train_utils.py`. To see where the `output_length` attribute is defined for the models in the code, take a look at the `sample_models.py` file. You will notice this line of code within most models:

```
model.output_length = lambda x: x
```

The acoustic model that incorporates a convolutional layer (`cnn_rnn_model`) has a line that is a bit different:

```
model.output_length = lambda x: cnn_output_length(
        x, kernel_size, conv_border_mode, conv_stride)
```

In the case of models that use purely recurrent layers, the lambda function is the identity function, as the recurrent layers do not modify the (temporal) length of their input tensors. However, convolutional layers are more complicated and require a specialized function (`cnn_output_length` in `sample_models.py`) to determine the temporal length of their output.

You will have to add the `output_length` attribute to your final model before running the code cell below. Feel free to use the `cnn_output_length` function, if it suits your model.

In [4]:

```
# specify the model
model_end = final_model(input_dim=161,
                        units=200,
                        recur_layers=2)
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
the_input (InputLayer)       (None, None, 161)         0         
_________________________________________________________________
bidirectional_1 (Bidirection (None, None, 400)         434400    
_________________________________________________________________
bidirectional_2 (Bidirection (None, None, 400)         721200    
_________________________________________________________________
time_distributed_1 (TimeDist (None, None, 29)          11629     
_________________________________________________________________
softmax (Activation)         (None, None, 29)          0         
=================================================================
Total params: 1,167,229
Trainable params: 1,167,229
Non-trainable params: 0
_________________________________________________________________
None
```

Please execute the code cell below to train the neural network you specified in `input_to_softmax`. After the model has finished training, the model is [saved (https://keras.io/getting-started/faq/#how-can-i-save-a-keras-model)](https://keras.io/getting-started/faq/#how-can-i-save-a-keras-model) in the HDF5 file `model_end.h5`. The loss history is [saved (https://wiki.python.org/moin/UsingPickle)](https://wiki.python.org/moin/UsingPickle) in `model_end.pickle`. You are welcome to tweak any of the optional parameters while calling the `train_model` function, but this is not required.

In [3]:

```
train_model(input_to_softmax=model_end,
            pickle_path='model_end.pickle',
            save_model_path='model_end.h5',
            spectrogram=True) # change to False if you would like to use MFCC 
features
```

```
Epoch 1/20
106/106 [==============================] - 1117s - loss: 300.5945 
- val_loss: 217.1167
Epoch 2/20
106/106 [==============================] - 1093s - loss: 214.2323 
- val_loss: 192.5075
Epoch 3/20
106/106 [==============================] - 1106s - loss: 192.0759 
- val_loss: 173.0631
Epoch 4/20
106/106 [==============================] - 1111s - loss: 171.6545 
- val_loss: 160.2459
Epoch 5/20
106/106 [==============================] - 1108s - loss: 154.2035 
- val_loss: 147.1542
Epoch 6/20
```

```
106/106 [==============================] - 1101s - loss: 141.9648
- val_loss: 139.8466
Epoch 7/20
106/106 [==============================] - 1111s - loss: 133.1456
- val_loss: 136.8613
Epoch 8/20
106/106 [==============================] - 1104s - loss: 125.5448
- val_loss: 129.4886
Epoch 9/20
106/106 [==============================] - 1105s - loss: 118.7642
- val_loss: 128.4629
Epoch 10/20
106/106 [==============================] - 1110s - loss: 113.3565
- val_loss: 122.9494
Epoch 11/20
106/106 [==============================] - 1109s - loss: 107.8659
- val_loss: 122.9961
Epoch 12/20
106/106 [==============================] - 1106s - loss: 103.1598
- val_loss: 122.4884
Epoch 13/20
106/106 [==============================] - 1110s - loss: 98.8958 -
val_loss: 121.4456
Epoch 14/20
106/106 [==============================] - 1111s - loss: 94.3830 -
val_loss: 121.1402
Epoch 15/20
106/106 [==============================] - 1100s - loss: 90.5160 -
val_loss: 118.9171
Epoch 16/20
106/106 [==============================] - 1104s - loss: 86.2843 -
val_loss: 122.2555
Epoch 17/20
106/106 [==============================] - 1099s - loss: 82.5558 -
val_loss: 118.5372
Epoch 18/20
106/106 [==============================] - 1108s - loss: 79.0190 -
val_loss: 121.0600
Epoch 19/20
106/106 [==============================] - 1105s - loss: 74.8587 -
val_loss: 122.5834
Epoch 20/20
106/106 [==============================] - 1115s - loss: 71.2669 -
val_loss: 123.9023
```

Question 2: Describe your final model architecture and your reasoning at each step.

Answer: Intuitively, bidirectional RNN seem the most powerful concept as they take more into account in the prediction the following words. This seems very useful in speach (if I had to guess the previous word from "... love you" I would guess "I" instead of "eye" for instance).

Taking the previous into account, I tried to measure the performance of adding more bidirectional layers. This resulted in a deep bidirectional rnn. Because of the very long training time I only tested with two recursive layers although this is parametrized and could be extended without any trouble. The recursivity was copied from the example of the deep rnn and the bidirectional layer from model 4.

Although it achieved a lower validation loss, it clearly overfitted as the testing loss decreased too much while validation loss reamined constant after epoch 12 approximately. To enhance the model I would try adding dropout.

STEP 3: Obtain Predictions

We have written a function for you to decode the predictions of your acoustic model. To use the function, please execute the code cell below.

In [5]:

```
import numpy as np
from data_generator import AudioGenerator
from keras import backend as K
from utils import int_sequence_to_text
from IPython.display import Audio

def get_predictions(index, partition, input_to_softmax, model_path):
    """ Print a model's decoded predictions
    Params:
        index (int): The example you would like to visualize
        partition (str): One of 'train' or 'validation'
        input_to_softmax (Model): The acoustic model
        model_path (str): Path to saved acoustic model's weights
    """
    # load the train and test data
    data_gen = AudioGenerator()
    data_gen.load_train_data()
    data_gen.load_validation_data()

    # obtain the true transcription and the audio features
    if partition == 'validation':
        transcr = data_gen.valid_texts[index]
        audio_path = data_gen.valid_audio_paths[index]
        data_point = data_gen.normalize(data_gen.featurize(audio_path))
    elif partition == 'train':
        transcr = data_gen.train_texts[index]
        audio_path = data_gen.train_audio_paths[index]
        data_point = data_gen.normalize(data_gen.featurize(audio_path))
    else:
        raise Exception('Invalid partition!  Must be "train" or "validation"')

    # obtain and decode the acoustic model's predictions
    input_to_softmax.load_weights(model_path)
    prediction = input_to_softmax.predict(np.expand_dims(data_point, axis=0))
    output_length = [input_to_softmax.output_length(data_point.shape[0])]
    pred_ints = (K.eval(K.ctc_decode(
                prediction, output_length)[0][0])+1).flatten().tolist()

    # play the audio file, and display the true and predicted transcriptions
    print('-'*80)
    Audio(audio_path)
    print('True transcription:\n' + '\n' + transcr)
    print('-'*80)
    print('Predicted transcription:\n' + '\n' + ''.join(int_sequence_to_text(p
red_ints)))
    print('-'*80)
```

Use the code cell below to obtain the transcription predicted by your final model for the first example in the training dataset.

In [11]:

```
get_predictions(index=0,
                partition='train',
                input_to_softmax=final_model(input_dim=161,##Adding params
                         units=200,
                         recur_layers=2),
                model_path='results/model_end.h5')
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
the_input (InputLayer)       (None, None, 161)         0
_________________________________________________________________
bidirectional_9 (Bidirection (None, None, 400)         434400
_________________________________________________________________
bidirectional_10 (Bidirectio (None, None, 400)         721200
_________________________________________________________________
time_distributed_5 (TimeDist (None, None, 29)          11629
_________________________________________________________________
softmax (Activation)         (None, None, 29)          0
=================================================================
Total params: 1,167,229
Trainable params: 1,167,229
Non-trainable params: 0
_________________________________________________________________
None
--------------------------------------------------------------------
--------------
True transcription:

the houses seemed miserable in the extreme especially to an eye ac
customed to the smiling neatness of english cottages
--------------------------------------------------------------------
--------------
Predicted transcription:

the huss teemed miserarbon the extriime i spelay to ani custim to
thi smaling nee bes f binglis cotijids
--------------------------------------------------------------------
--------------
```

Use the next code cell to visualize the model's prediction for the first example in the validation dataset.

In [10]:

```
get_predictions(index=0,
                partition='validation',
                input_to_softmax=final_model(input_dim=161,##Adding params
                        units=200,
                        recur_layers=2),
                model_path='results/model_end.h5')
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
the_input (InputLayer)       (None, None, 161)         0
_________________________________________________________________
bidirectional_7 (Bidirection (None, None, 400)         434400
_________________________________________________________________
bidirectional_8 (Bidirection (None, None, 400)         721200
_________________________________________________________________
time_distributed_4 (TimeDist (None, None, 29)          11629
_________________________________________________________________
softmax (Activation)         (None, None, 29)          0
=================================================================
Total params: 1,167,229
Trainable params: 1,167,229
Non-trainable params: 0
_________________________________________________________________
None
--------------------------------------------------------------------------------
True transcription:

the bogus legislature numbered thirty six members
--------------------------------------------------------------------------------
Predicted transcription:

the bo is it islag gu morberd pria sice evers
--------------------------------------------------------------------------------
```

In [20]:

```
get_predictions(index=0,
                partition='train',
                input_to_softmax=model_4,
                model_path='results/model_4.h5')
```

```
--------------------------------------------------------------------------------
True transcription:

the houses seemed miserable in the extreme especially to an eye accustomed to the smiling neatness of english cottages
--------------------------------------------------------------------------------
Predicted transcription:

te olzss se d masra en the xstrim s spelay to an  y costonto the smilingnet esof tinlih ce ts
--------------------------------------------------------------------------------
```

In [21]:

```
get_predictions(index=0,
                partition='train',
                input_to_softmax=model_3,
                model_path='results/model_3.h5')
```

```
--------------------------------------------------------------------------------
True transcription:

the houses seemed miserable in the extreme especially to an eye accustomed to the smiling neatness of english cottages
--------------------------------------------------------------------------------
Predicted transcription:

e olesendo misor e an theixtrerim mis pes lyto an i costanto the smilin ne nisof in lis coidtes
--------------------------------------------------------------------------------
```

In [22]:

```
get_predictions(index=0,
                partition='train',
                input_to_softmax=model_2,
                model_path='results/model_2.h5')
```

```
--------------------------------------------------------------------
--------------
True transcription:

the houses seemed miserable in the extreme especially to an eye ac
customed to the smiling neatness of english cottages
--------------------------------------------------------------------
--------------
Predicted transcription:

the heoes temed mes eroblein the extrme espesstlytoan  acustimto t
he smalling neat neso benglus cou gs
--------------------------------------------------------------------
--------------
```

In [23]:

```
get_predictions(index=0,
                partition='train',
                input_to_softmax=model_1,
                model_path='results/model_1.h5')
```

```
--------------------------------------------------------------------
--------------
True transcription:

the houses seemed miserable in the extreme especially to an eye ac
customed to the smiling neatness of english cottages
--------------------------------------------------------------------
--------------
Predicted transcription:

te hoos is se domas eran ea xtrm as pasly ton oacuster to the s ma
nne osof en lishcoe tons
--------------------------------------------------------------------
--------------
```

One standard way to improve the results of the decoder is to incorporate a language model. We won't pursue this in the notebook, but you are welcome to do so as an *optional extension*.

If you are interested in creating models that provide improved transcriptions, you are encouraged to download [more data (http://www.openslr.org/12/)](http://www.openslr.org/12/) and train bigger, deeper models. But beware - the model will likely take a long while to train. For instance, training this [state-of-the-art (https://arxiv.org/pdf/1512.02595v1.pdf)](https://arxiv.org/pdf/1512.02595v1.pdf) model would take 3-6 weeks on a single GPU!