Digital Speech Processing

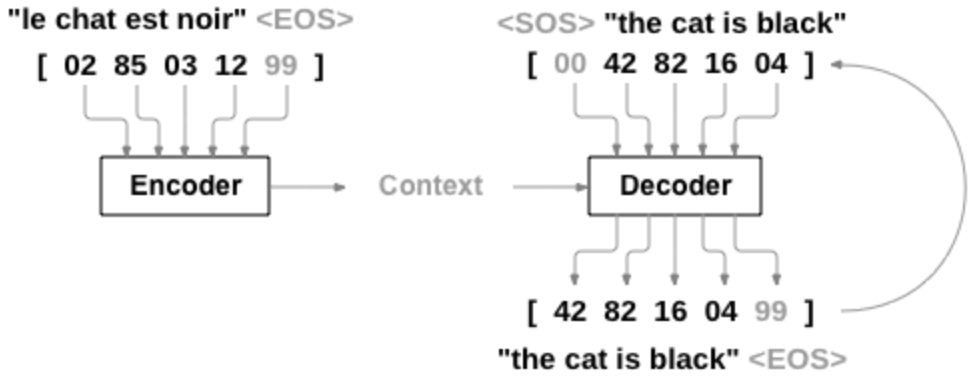
Final Project

Speech-to-Text by Seq2Seq

B04902004 王佑安

Seq2Seq model

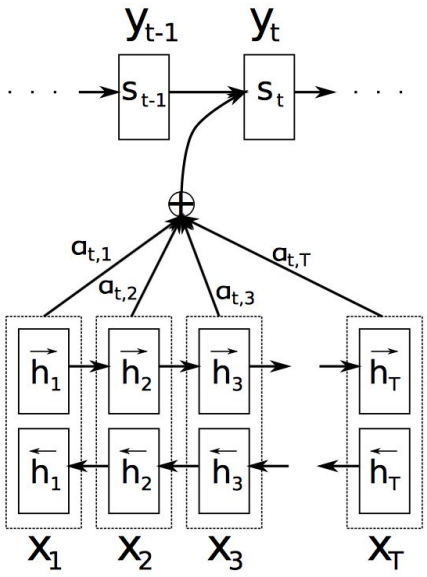
Seq2Seq可以將sequences從一個domain轉換到另外一個domain，常見的例子像是語言翻譯（英文->中文）、對話機器人（問句->回答）等等。也可以用在語音處理，像是Speech-to-Text跟Text-to-Speech都可以用Seq2Seq base的model來處理。



[0]

Seq2Seq的基本概念是利用一個encoder將sequence轉換成一段能代表這段sequence的context，再丟給decoder轉換成另一個domain的sequence。

Attention Technique



[1]

基礎的Seq2Seq只用context當作整個sequence的資訊，有時候sequence太長decoder很難只靠context decode出正確的sequence。

Attention透過將encoder在每個timestamp的output加權後傳給decoder解決這個問題，decoder在某個timestamp可以知道他的output應該對應到input sequence的哪些部分。在Speech-to-Text的task裡功能就類似HMM中的align，將output phone對應到某一段音訊。

Speech-to-Text by Seq2Seq Implementation

DSP hw2剛好有中文數字的speech dataset，我就試著用Seq2Seq+attention來解決這個task。

hw2的HTK tools有一個HList可以將wav直接轉成mfcc text data，再用python讀取。Encoder跟Decoder的部份都是用RNN來實作，由於output只有0-9的數字加上sil跟sp，再接一層DNN將decoder轉換成12維的output。

Result

	HMM test set	Seq2Seq train set	Seq2Seq test set
Sentence Accuracy	87.92%	96.87%	71.88%
Word Accuracy	96.38%	99.04%	90.10%

Seq2Seq在training set上已經達到將近100%的準確度，但在testing set上還是比HMM差了一節。hw2時是有調整HMM的參數才讓他準確度到95%以上，所以我也嘗試了調整Seq2Seq的參數，看能不能讓他表現提高 。

Improvement

Encoder type	RNN(128)	RNN(256)	CNN+RNN	BiRNN
Sentence Accuracy	71.88%	79.58%	77.08%	80.42%
Word Accuracy	90.10%	93.49%	92.40%	93.15%

hw2時調整了gaussian的數量，對應到Seq2Seq就是RNN的hidden size，越多的hidden state就能表示越複雜的資訊。將hidden state從128調整到256後準確度上升了，但還是比HMM差一點。

後來上網查了一些別人做過的Seq2Seq speech-to-text，發現有人使用CNN跟bidirectional RNN在encoder上，也試著加上去，不過效果並不好。因為我使用的feature在抽取mfcc時經過shift window的效果已經類似CNN了，因此推論CNN已經無法再讓表現上升。而bidirectional RNN看起來sentence accuracy有稍微提升一點點，推測是因為雙向的RNN可以讓sequence連貫性更高，所以sentence準確度可以上升，但目前的Seq2Seq跟HMM最大的差距可能是單個phone的準確度，因此整體表現還是不如HMM。

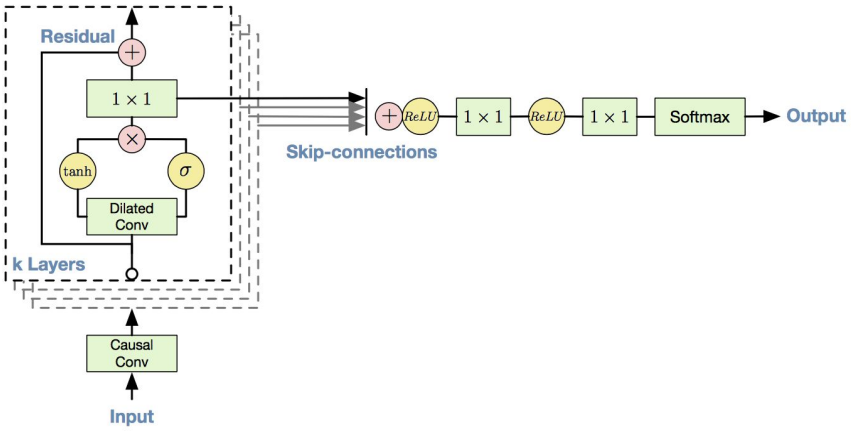
Future Work

End-to-end speech-to-text by Seq2Seq

本來是期望能用seq2seq在hw2的data上達到跟HMM相同甚至更好的表現，讓經過了很多嘗試與改良還是沒有成功。查了這方面相關的paper大多是end-to-end方面的研究，也就是用沒經過mfcc的raw data直接做seq2seq。大概是因為HMM在這方面的表現已經很好了沒什麼改良空間，因此研究大多是走HMM做不到的end-to-end，但end-to-end的model架構通常比較龐大，在training上需要耗費比較多時間，等有空再把model修改成end-to-end試試看。

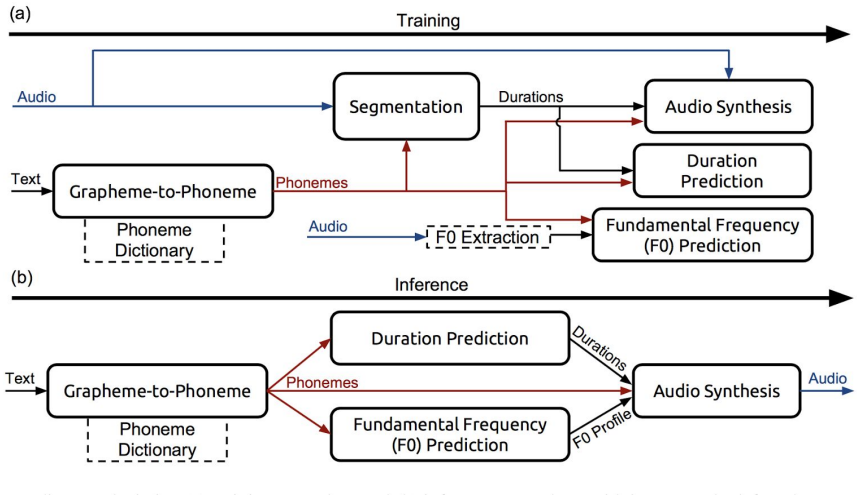
Work on text-to-speech task

一開始這個題目是因為讀了DeepMind發表的WaveNet[2]想做看看text-to-speech，後來看了更多paper後發現在語音生成的部分會有不少障礙，例如WaveNet在語音生成是用ResNet，ResNet是由很多層CNN構成的，因此需要龐大的運算資源才能在有限時間內做出成果。



WaveNet

另外一篇Deep Voice[3]則是將語音生成拆成frequency、phonemes、durations分別預測後在合成，實作起來也相當複雜。後來發現text-to-speech有很多seq2seq base的model [4][5]，才決定先試著做相對簡單的speech-to-text，之後再來實作text-to-speech。



Deep Voice

Reference

[0] Translation with a Sequence to Sequence Network and Attention

http://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html

[1] NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

https://arxiv.org/pdf/1409.0473.pdf

[2] WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

https://arxiv.org/pdf/1609.03499.pdf

[3] Deep Voice: Real-time Neural Text-to-Speech

https://arxiv.org/pdf/1702.07825.pdf

[4] CHAR2WAV: END-TO-END SPEECH SYNTHESIS

https://openreview.net/pdf?id=B1VWyySKx

[5] TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

https://arxiv.org/pdf/1703.10135.pdf