Encoder-Decoder Attention模型

現存的諸多模型皆存在著Encoder-Decoder的架構，例如：Seq2Seq、語音辨識、機器翻譯等領域，都以Encoder-Decoder模型，進行相應的深度學習領域研究[1][2][3]。Encoder可以視為利用時間數列資料的每一個時間點t-1進行預測，最後輸出一個t的結果，可表示成Encoder = （e0, e1, …, et-1），而其輸出結果為（ê1, ê2, …, êt），其中Encoder的輸入稱為Input，通常是欲預測的資料。而Decoder亦是利用時間數列資料的時間點T-1去預測T，可以表示成Decoder = （d0, d1, …, dT-1），其相對應輸出為（d̂1, d̂2, …, d̂T）。訓練完成後，Encoder的最終輸出êt即包含（e0, e1, …, et-1）序列的資訊，而其會作為輸入Decoder的一部份輸入，放入Decoder模型當中，以期得出Decoder的輸入在每一個時間點的輸出。本模型的Encoder輸入層為經過MFCC轉換過後的一秒英文數字檔案，為99個時間點的39維MFCC特徵，而Decoder的輸入為該音檔的音素（Phoneme）。以Decoder在每個時間點的輸出作為特徵，透過全連接層（Fully-connected）辨識該時間點的音素為何。Attention可視作嫁接Encoder與Decoder的橋樑，由於目前的RNN皆存在梯度消失以及梯度爆炸的問題，而Attention為一個十分簡單的全連接層架構，會對每一個時間點的RNN輸出進行特徵提取，並得出該輸出的注意力分配。使用RNN + Attention模型能夠過濾、提取出，對於輸出來說重要且具有時間相關性的特徵。

Encoder

Encoder分為三大部分，全連接層、ResNet層[4]，以及LSTM層。全連接層先對每個時點的39維特徵進行線性投射至高維度，突顯該時點的特徵。經由三個ResNet層將鄰近的時間也納入考量，此處的ResNet將模型輸入經由三層卷積神經網路層進行特徵擷取，且將輸入與第三層卷積神經網路的輸出結果相加，再透過ReLU輸出，可以有效避免模型的梯度消失，而為避免梯度爆炸問題，在所以卷積神經網路層的後面，都加上了一層Batch Normalization。將第一與第二層的ResNet每三個時間點的特徵進行連接，使時間長度t壓縮成t/3，也就是每個新的時間t/3，皆包含3個時間長度的資訊。時間的壓縮，有助於減少模型於LSTM層的運算時間，如此一來便可以加速模型訓練的進行。而最後一層的ResNet輸出，會作為Encoder的LSTM層的每一個時間點的輸入（e0, e1, …, et-1）得到（ê1, ê2, …, êt）。

Attention

Attention可以簡單分為兩個全連接層，與一個輸出，稱為Context Vector（C）。依據[5][6]所述，將Encoder的輸出（ê1, ê2, …, êt-1）稱作query（Q），而êt稱為value（V）。透過兩個不同的全連接層f與g，分別對query與value進行轉換過後，以element-wise的方式相加，最後透過tanh進行轉換，得出第一層的結果（output, O）。接著再次通過一層僅有一個結果的全連接層h，便可得出Encoder輸出所形成的score（S）。而對score取softmax轉換後的結果，為該Attention的weights（W）。將weights與value以element-wise的方式相乘並加總過後，稱為Context Vector。

Decoder

Decoder主要為一層LSTM，其輸入為音素序列。音素序列首先會先通過一個Embedding層，將每個音素投射至指定維度的空間，使得音素不會因為音素的編號數字影響輸出結果，音素序列（）前後會加入序列開始（Start of sequence）與序列結束（End of sequence）的信號，亦會同時透過Embedding層進行轉換，稱經過Embedding的音素序列為音素向量（）。接著將音素向量與Decoder的開始信號與Attention模型所輸出的C連接起來，作為LSTM層的輸入值，對開始信號之後的每一個時間步進行預測，直到預測的結果出現序列結束信號，又或者達到設定的預測個數上限時，才會停止。最後，將該預測的結果通過全連接層進行分類，進而得到每個時間點所預測的音素。

模型訓練

訓練Encoder-Decoder模型時，共讀取20000筆語音資料，每種類別皆讀取2000筆，並預留1000筆語音資料作為驗證集，進行訓練。每筆語音資料皆進行過MFCC轉換，並伴隨著相對應的標籤與音素。模型訓練是以四筆語音資料作為一個批次（Batch）進行，在一次Epoch中，總共有4750個批次，而模型每次的Epoch約需耗時15分鐘。經過多次反覆的測試過後，最終決定採用Adam優化器[7]，並將模型的學習率（Learning rate）設定為0.0001，Epoch最高上限訂為20次，終止條件為訓練集的損失小於0.02，進行模型訓練。在聲音模型（Acoustic Model）的訓練完成之後，以該結果進一步訓練字詞模型（Lexicon Model），並以字詞模型所輸出的結果與其他模型進行比較。

[1] Sequence to Sequence Learning with Neural Networks

[2] Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

[3] On the Properties of Neural Machine Translation: Encoder-Decoder Approaches

[4] Deep Residual Learning for Image Recognition

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

[6] Effective Approaches to Attention-based Neural Machine Translation

[7] ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION