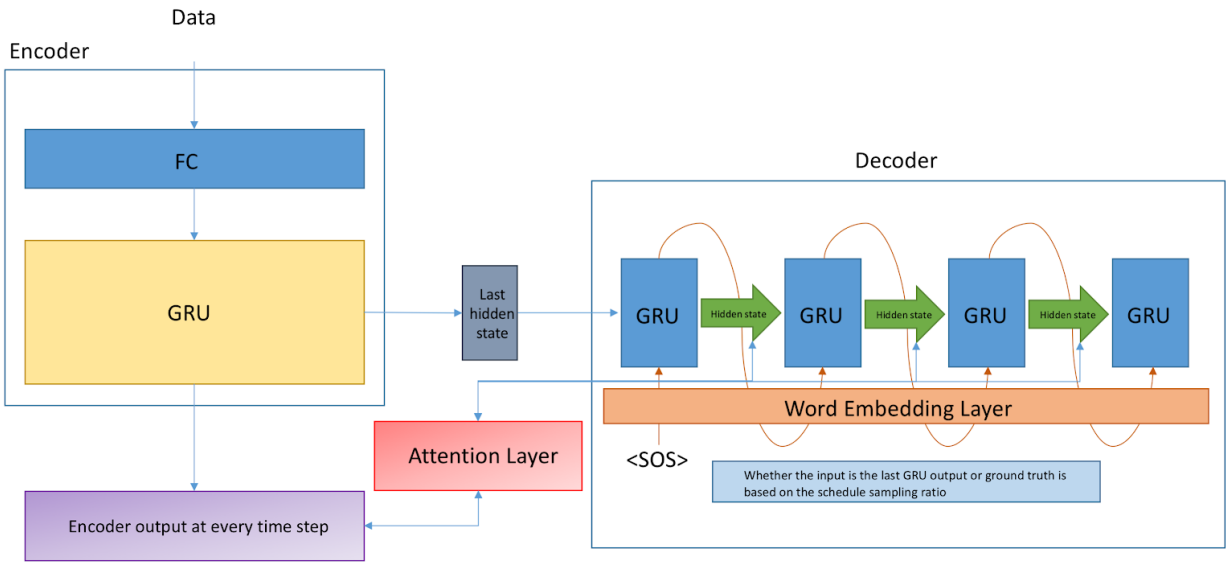
MLDS HW2 Report

2-1 Video Caption Generation

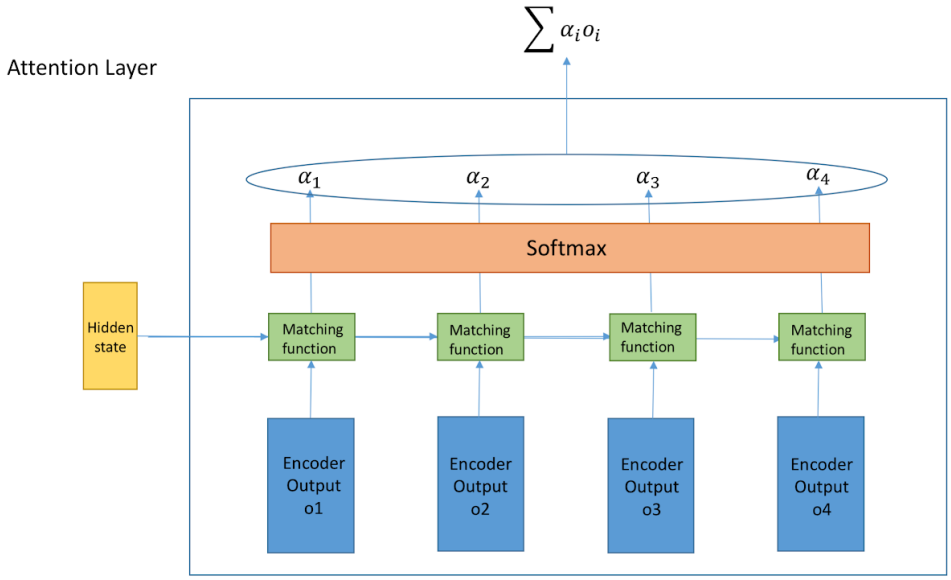
1. Model Description(3%)

我們的model有三個主要的部分，分別為encoder, decoder跟中間的attention layer。 以下分別對這三個部分做詳細的介紹：

- Encoder: 輸入為一個(80, 4096)的影片feature向量，經過一層dense將維度壓縮成(80, 1000)之後送進gru。因為我們的model設計所以不用取得中間的hidden state，因此能夠一次將整個矩陣送進gru中得到 last hidden state跟encoder 每一個time step的輸出。
- Decoder: 將encoder last hidden state和<SOS>當作decoder的initial state。這邊因為我們要取得decoder每一個time step 的hidden state來做attention，因此沒辦法使用matrix operation而必須使用for loop來寫。在training的過程中我們會使用schedule sampling因此model必須決定這個time step 要使用自己上一個time step解出來的字還是ground truth當作這一個time step的輸入。decoder每一個time step解出來的向量經過一層dense還原成真實的vocabulary size大小的機率向量經過argmax取得最後的輸出文字。



- Attention Layer: attention的實作是最基本的方法。將decoder每一個time step的hidden state 跟encoder的所有輸出做一個matching function得到一個scalar，再將這一把scalar經過softmax得到每一個time step output的的百分比，再經過weighted sum後得到最後的new hidden state再送進decoder 的下一個time step。在我們的model中matching function為一層小nn。



2. How to improve your performance? (3%)

- Write down the method that makes your model outstanding? (1%)
  - Attention layer
  - Schedule sampling

- Why do you use it? (1%)
  - Attention: 傳統rnn是將整個sequence的資訊壓縮在encoder hidden state當中。這個作法的問題是當sequence長度太長model會開始忘記前面看過的資訊。這個時候加入一層attention能夠在decoder開始解字的時候選取相對重要的input 資訊，進而增加decode出來字句的完整度。
  - Schedule sampling: 在rnn decoder當中每個time step會輸入上一個time step的hidden state跟output。這個作法的缺點是只要有一個time step輸入的output是錯的接下來的所有輸出都會是錯的，有一點一步錯步步錯的感覺。為了解決這個問題我們引入schedule sampling的做法，就是每一個time step讓model去選要使用上一個time step自己解出來的字還是要用ground truth。這

樣就算有一些輸入的output是錯的model還是有機會透過ground truth把結果修正回來。

- Analysis and compare your model without the method (1%)
    - 以下我們對照有加attention跟沒有attention的結果：

ID:UXs3eq68ZjE_250_255.avi	Result
with attention	Someone is pouring a pitcher of milk into a pan containing a rice
without attention	A is being added to a pot of rice

ID:Je3V7U5Ctj4_569_576.avi	Result
with attention	A man is adding shredded cheese to a tortilla
without attention	A man is a mixture of a bowl

ID:mtrCf667KDk_134_176.avi	Result
with attention	A woman is cutting a carrot into pieces
without attention	A man is <UNK> a piece of raw

可以看見有沒有加attention其實對model最後的輸出影響很大。加了attention的model解出來的句子跟沒有attention的句子比起來長了許多，細節描述的表現也比沒有加attention的來的好。

3. Experimental results and settings (1%)

- 以下說明不同實驗參數的結果：(ssr為schedule sampling ratio)

ID:Je3V7U5Ctj4_569_576.avi	Ground truth: A man is adding shredded cheese to a tortilla
w/o attention + 0.5 ssr	A man is a mixture of a bowl
w/o attention + 0.7 ssr	A man is adding shredded to a tortilla
Attention + 0.7 ssr	A man is adding shredded cheese to a tortilla
w/o attention + 1 ssr	A man is putting a something into a bowl of sauce
w/o attention + 0 ssr	A man is a something a a

Attention + 1 ssr	A man is putting shredded cheese in a bowl

ID:mtrCf667KDk_134_176.avi	Ground truth: A woman is peeling a papaya
w/o attention + 0.5 ssr	A man is something a piece of raw
w/o attention + 0.7 ssr	A man is cutting a
Attention + 0.7 ssr	A woman is cutting a carrot into pieces
w/o attention + 1 ssr	A woman is cutting a carrot
w/o attention + 0 ssr	A woman is the a
Attention + 1 ssr	A man is cutting a piece of peeled carrots

可以看見有加入attention的model會描述比較多影片中的細節，句子跟文法也比較完整。比較值得注意的是attention加上training時所有的input都為ground truth的狀況下結果不會比schedule sampling還要差，但是這個model在測試bleu score時分數卻是比較低的，我們推測應該是因為model在training時沒有吃過自己的輸出所以在inference的時候因為沒有ground truth的輸入造成輸出的字句跟影片沒有關聯，進而導致bleu score降低。

分工表：

組員	分工
B04901040 電機三 林哲賢 (33%)	寫HW2-2report，hw2-2的testing，beamsearch，train model，調整參數，整理實驗結果
B04901117 電機三 毛弘仁 (33%)	改寫 seq2seq model；增加功能（可調整 vocab size、幾層 RNN、training data 量）；訓練 model；協助整理 Github repo 及實驗結果
B04901118 電機三 王克安 (33%)	HW2-1 model construction, experimenting, model training, report writing