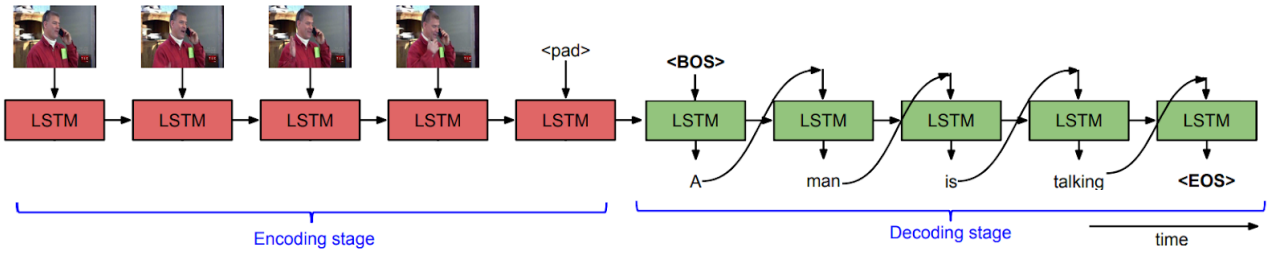
HW2-1

ID & Name : R06946009 林庭宇、R06946006 李筑真、R06946015 黃永翰

- **Model description (3%)**
  - **Describe your seq2seq model**

    我們使用兩個 GRU，一個作為 encoder，另一個作為 decoder．首先將已抽取好的 feature 依序輸入 encoder 後，取 encoder 最後一個 time step 的 output 向量來代表該 Video，並將此向量餵入 decoder 中，讓 decoder 在每個 time step 輸出一個 word，直到輸出結束字元表示完成這個句子．



model 訓練完成後，在 testing 的過程中，decoder 每個 time step 的 input 為前一個 time step 的 output；而 training 時則是取 video 所對應的 ground truth 作為 input，並且以最大化每個 time step 預測出正確字的機率作為目標式．

此外，我們亦加上 attention 的機制，讓 decoder 學習在每個 time step 注意相對重要的 feature．

```
Model(
  (encoder): Encoder(
    (vid2hid): Linear(in_features=4096, out_features=512, bias=True)
    (input_dropout): Dropout(p=0)
    (rnn): GRU(512, 512, batch_first=True)
  )
  (decoder): Decoder(
    (rnn): GRU(1024, 512, batch_first=True, dropout=0.2)
    (input_dropout): Dropout(p=0)
    (embedding): Embedding(1961, 512)
    (attention): Attention(
      (linear1): Linear(in_features=1024, out_features=512, bias=True)
      (linear2): Linear(in_features=512, out_features=1, bias=False)
    )
    (out): Linear(in_features=512, out_features=1961, bias=True)
  )
)
```

- **How to improve your performance (3%)**
  (e.g. Attention, Schedule Sampling, Beamsearch...)
  - **Write down the method that makes you outstanding (1%)**

    Schedule Sampling: decoder 在我們餵入一個字(input)時會預測下一個字，在未使用 schedule sampleing 的方法中，decoder 每個 time step 都直接使用 ground truth 作為輸入的字(input)，而 schedule sampleing 則是給定一個機率 p 以 ground truth 作為 input，相對的有(1-p)的機率我們會使用上一個 time step 預測出的字作為當前這個 time step 的 input 來預測下一個字．

  - **Why do you use it (1%)**

    在原始的方法中，decoder 在 training 與 testing 的運作過程並不一致，因為 testing 時除了起始字元外，每個 time step 的 input 都是來自上一個 time step 所 predict 的字，然而在 training 時卻是直接使用 ground truth 的字，為了讓 model 能以更接近 testing 的情境進行訓練又不至於壞掉，所以我們使用 schedule sampling 的方式來進行訓練．

  - **Analysis and compare your model without the method. (1%)**

是否使用 Schedule Sampling	Bleu
否	0.6223
是(p = 0.7)	0.6842

在其他條件不變下，訓練 150 epochs，由表中可以見到 Schedule Sampling 可以有效的提升 Bleu 分數，然而由實際產生的句子看並無法看出明顯差異．

- **Experimental results and settings (1%)**
  (parameter tuning, schedual sampling … etc)

  Setting:

  optimizer 使用 adam、batch size 設為 32、GRU 的 hidden layer 與 word embedding 之 dimension 均設定為 512、字典大小為 1961(取自 training set label 中出現次數超過 5 次者，在額外加上起始字元跟結束字元)

Experimental results:

比較 Schedule Sampling 在不同 teacher forcing ratio(使用 ground truth 作為 input 的比例，以下使用 p 代表)下的 bleu 分數

以下皆使用相同 model 與參數設定，並且分別在訓練 50、100 與 150 epochs 後計算 testing set 的 bleu 分數：

P	0.5	0.7	0.9	1
50epochs	0.6966	0.6966	0.6396	0.6575
100epochs	0.6841	0.6995	0.6251	0.6256
150epochs	0.6763	0.6569	0.6491	0.6205

由實驗結果可以發現使用 schedule sampling 可以提高 bleu 分數，而在 p=0.7 時，訓練 100 epoch 有最好的結果．

此外從上表中，我們可以發現似乎越多 epoch 的 training 反而使 bleu 分數下降(然而此時 training loss 其實仍在下降中)，因此我們亦比較 training 3000 epochs 與 training 50 epochs 的成果(不使用 schedule sampling，其他設定與上同)：

由 testing data 的 bleu 分數來看 50 epochs 為 0.65 而 3000 epochs 只有 0.62，然而若直接看兩者產生的內容則是 3000 epochs 的結果明顯優於 50 epochs，由此可見或許 Bleu1 並不能完美的對 video caption 成果進行評估．

分工表

學號姓名	負責工作	比例(%)
R06946009 林庭宇	HW2-2	25
R06946006 李筑真	HW2-2	25
R06946015 黃永翰	HW2-1	50