# MLDS 2017 Assignment 2

## Problem Description

Video Caption Generation

給予一個影片（此作業中使用助教對每個frame事先抽好的CNN features），我們要產生一個句子描述此影片內容，對產生的句子和答案算BLEU 分數來當作衡量的標準 。

Example:

Input

Output

A man is chopping butter into a container.

訓練資料的來源是MSVD dataset，共1450部影片，每部影片有多句對應的句子。公開測試資料來自同樣的資料集，共50部影片。

## Environment

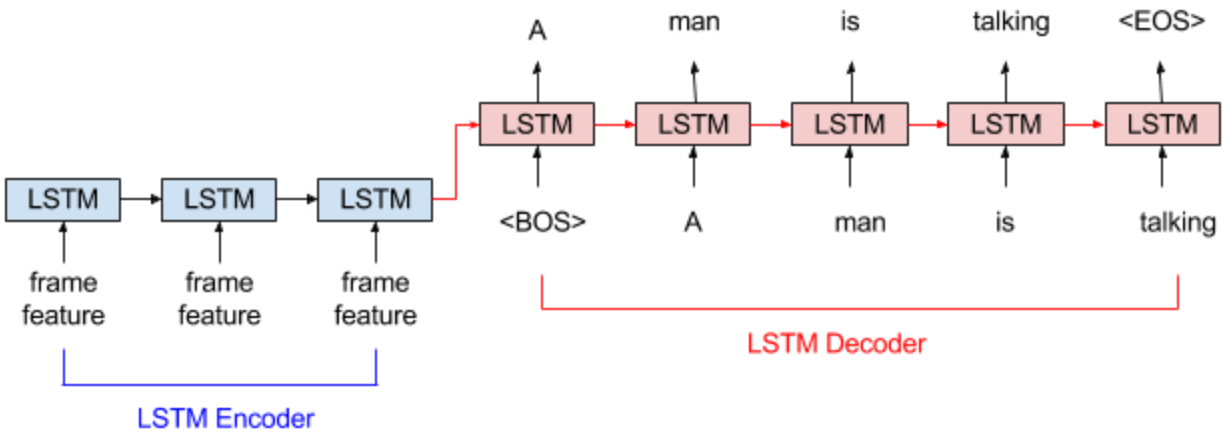
- OS: Linux
- CPU: Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz, Memory: 64GB
- GPU: GeForce GTX 1080, Memory: 8GB
- Libraries:
  - Python - 3.5
  - Tensorflow - 1.0
  - NumPy - 1.12.0
  - Progressbar2 - 3.18.1

## Model Description

**Data Preprocessing**

將訓練資料裡的每個句子開頭加上<BOS>，結尾加上<EOS>，不足15字的句子做padding，大於15字的句子捨去，接著建構所有句子所包含的字彙，由於訓練資料裡的句子已十分乾淨，我們決定不刪減字彙。最後將句子裡的每個字彙轉成其對應的ID，生成訓練資料。

**Model Structure**



我們採用一個單層LSTM當作encoder，encode影片裡80個frame的CNN features，另一個單層LSTM當作decoder，產生影片對應的句子。此外我們加上attention的機制，decoder每次在產生字時，除了考慮該時間點的輸入，也考慮encoder所有時間點的hidden states。損失函數 (loss function)使用cross entropy，back propagation用RMSPropOptimizer來更新模型的參數。

**Inference**

Encoder 輸入影片每個frame的CNN feature，接著decoder讀到<BOS>時，開始產生字，在t時間點輸出過softmax後機率最大的字，接著將此字當作t+1時間點的輸入，重複此步驟直到decoder輸出<EOS>結束，最後產生的句子當作此影片的caption。

**Model Details**

基本和best的版本使用相同的模型。
訓練資料：6450個字彙，22893個句子。
Pre-trained word embeddings: GloVe 300維 (glove.6B.300d.txt)
模型：單層LSTM encoder + 單層LSTM decoder + attention，LSTM dimension=128。Training epoch=10，batch size=100，initial learning rate=1e-3， RMSpropOptimizer。

## Improvement

**Hidden Layer Size**

我們同時調整encoder 和decoder LSTM cell的hidden layer size，觀察BLEU分數和產生的句子有無變好，詳細實驗數據列於experiments中。根據實驗結果，我們最佳模型選用hidden layer size=128，加大size並沒有讓結果變好，推測使用太大的hidde layer size會造成overfitting。

**Reinforcement Learning**

我們嘗試實作MIXER模型(Ranzato '15)，MIXER先使用一般的Seq2Seq 以及cross entropy作為initial，接著用border將decoder分為前後兩段，前段為原本的seq2seq模型（使用training data作為decoder的input），後段將decoder的input改為step的預測結果，這樣較符合predict時的情況。而後段訓練的方式使用reinforcement learning，將上一個step的hidden state、預測的字、得到的reward、以及得到的hidden state作為訓練資料，使用Deep Q-Network模型，要model的是輸入一個state，要預測接下來所有可能的action（也就是預測的字）的reward，這部分作為一個regression問題，用square loss作為loss function。這個border會隨著訓練時間逐漸往前移動，直到整個decoder都採用前一個step預測的字作為decoder的input。

我們嘗試的模型結果並沒有比較好，有可能是target score(BLEU-1)不能精確的衡量預測的好壞，或者一些實作上的細節(training tips)導致，故沒有將此模型放入實驗中。

## Experiments

**Hidden Layer Size**

這個實驗主要是將LSTM的hidden layer size做更動並且固定其他欄位的參數，來觀察hidden layer size對結果所產生的影響，測試在公開測試集上。

**Settings**

Epoch : 10
Decoder input length : 14
Vocabulary size : 6450
Initial learning rate : 1e-3
Model : Basic Attention Seq2Seq

| Hidden Size | BLEU Score With Max Out Inference | BLEU Score With Beam Search@3 |
|---|---|---|
| 128 | 0.35707 | 0.35072 |
| 256 | 0.33095 | 0.32547 |
| 512 | 0.35543 | 0.32825 |
| 1024 | 0.34654 | 0.33795 |

下面圖表會列出將不同hidden size分別透過不同inference方式對測試資料所產生出來的句子及其對照的影片截圖，以便透過直接觀察句子的方式來判斷結果的好壞，而從實驗的結果可以看出hidden size為128的時候所產生出來的句子，都有對影片較好的詮釋。

Example 1: **ScdUht-pM6s_53_63.avi**

| Hidden Size | Caption with Max Out Inference | Caption with Beam Search@3 |
|---|---|---|
| 128 | **a man is cutting a piece of meat** | a man is cutting a pineapple |
| 256 | a man is eating a banana | a man is eating a banana |
| 512 | a man is putting a vegetable into a bowl | a man is putting food into a pan |
| 1024 | a woman is kneading a pineapple | a woman is kneading a pineapple |

Example 2: **rl1rVk_xIOs_1_16.avi**

| Hidden Size | Caption with Max Out Inference | Caption with Beam Search@3 |
|---|---|---|
| 128 | **two women are dancing** | **two girls are dancing** |
| 256 | the man is doing a soccer ball | a woman is exercising |
| 512 | a woman is doing exercise | a woman is doing exercise |
| 1024 | a woman is doing some exercises | a woman is exercising |

## Team Division

| 組員 | 分工 |
|---|---|
| 江東峻 r05922027 | data preprocessing, reinforcement learning model, report |
| 陳翰浩 r05922021 | data preprocessing, Seq2Seq+attention model, report |
| 鄭嘉文 r05922036 | experiments, report |