Machine Learning 2018 Fall

Final Project Proposal

組別：NTU_b05901011_喔喔喔喔歐翰墨

組員：許秉倫、楊晟甫、歐瀚墨

題目選定：Video Caption

一、 題目敘述

給定一個影片，欲從四個選項中選取最符合影片內容的敘述。

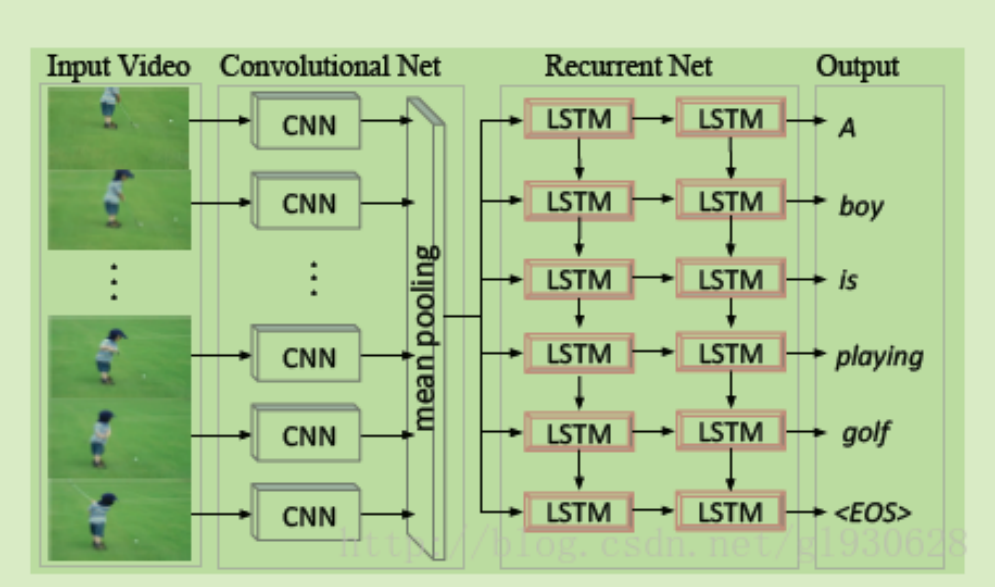
影片已經由 Feature Extractor 得到 feature，輸入的內容為 80 個 time slice，每個 slice 由 4096 維的向量表示。訓練資料有 1400 多部影片的 features，每一部影片都有數十句對應的敘述。另外，測試資料還包含四句敘述，程式必須能選出最接近影片內容的選項。

二、文獻探討

我們 survey 的論文是在 video to text 這個領域中相當經典的一篇，於 ICCV2015 所發表的”Sequence to Sequence – Video to Text”[2]。首先定義問題本身，我們可以視 Video to text 為一個每個 frame 的圖像序列映到句子序列的 seq2seq 的問題，在近幾年的文章中，大部分都使用了 RNN 中的 LSTM 模型來構造這個 encoder-decoder 的架構，本篇也不例外，惟其兩個是共用同一個 LSTM 的，詳細原因後面會做說明。

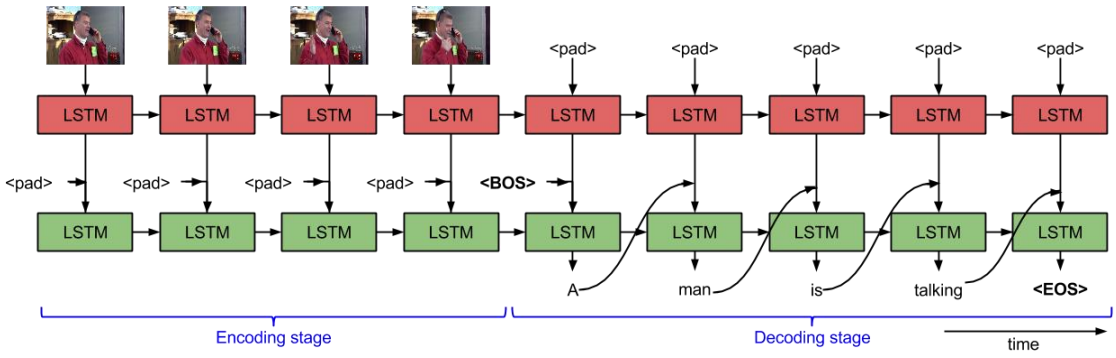
本篇論文首先闡釋了 video to text 這個問題的幾個挑戰：

1. 現實世界太過於複雜，一段影片中，不論是場景、物件、人物、行為、光線等都具有極高的變異性，如何確定主要內容是正確的並不容易。
2. 對於影片的描述是與每一個 frame 的時間先後有關的，像是在 Translating videos to natural language using deep recurrent neural networks 這篇論文中便沒有考慮到這個因素，其結構如下圖所示；此外，我們需要有可變的輸入以及輸出維度，因為輸入(frame 的數量)以及輸出(描述句子)的長度是不固定的。



圖一、模型結構，取自原文[2]

對此，作者提出了一個新的架構：S2VT，其架構如下圖所示：



圖二、S2VT 架構，取自原文[2]

其中此方法大致可分為三個區塊：

1. LSTM for sequence modeling

要使得模型能夠有可變的輸入以及輸出維度，本篇論文所採用的方法是在 encoding 階段一次輸入一個 frame，在 decoding 階段一次產生一個 word 的方式，而本文選用了 LSTM 作為 RNN 模型的架構。在 encoding 階段，給定一個輸入序列 X (x1, . . . , xn)，LSTM 會計算出一個隱藏狀態(h1, . . . , hn)；而在 decoding 階段，同樣給定一個輸入序列 X 可以得到一個輸出序列 Y (y1, . . . , ym)，其機率分布透過以下的式子說明：

$$p(y_1, \ldots\ldots, y_m) = \prod_{t=1}^{m} p(y_t | h_{n+t-1}, y_{t-1})$$

式一、機率分布

2. Sequence to sequence video to text

在其他論文的方法裡，會有兩個 LSTM，第一個 LSTM 負責將輸入序列編碼成一個固定長度的向量，而第二個 LSTM 負責將這些向量映射到輸出序列中。而本文只使用了一個 LSTM，好處是兩層可以做 weight sharing。其運算過程可以分為兩個部分：在影片的 frame 還在 input 時，第一層的 LSTM 進行 encoding，其計算出的隱藏狀態，再與空的 padded input words 連結，作為第二層 LSTM 的 input；而在影片輸入完畢之後，在第二層的 LSTM 嵌入一個<BOS>的標籤，開始做 decoding 的動作，以最大化預測的輸出結果以及前一個 word 的 log-likelihood 作為 loss function，用 SGD 去做 back propagation，最後通過一層 softmax 選擇機率最高的 output，由於輸出長度的不固定，所以我們需要一個<EOS>的標籤來結束每個句子。

3. Video and text representation

除了 RGB 的影像之外，為了加強模型的穩健度，作者也另外加入了一個光流的 feature：

對於 RGB，先輸入進 CNN 網絡，本文分別使用了 AlexNet 的變形 Caffe Reference Net 以及 VGG16 進行處理，值得一提的是，作者將最後一層的 fully connected layer 移除，將特徵 embedding 到一個較低的 500 維空間中，作為 LSTM 的輸入。

對於 optical flow，作者首先取出經典的光流特徵，接著創造 flow image，再計算光流強度，加入 flow image 成為除了 x,y 之外的第三個維度，接著放入在 UCF101 video dataset 上 pretrained 的 CNN 做 classification，同樣的將最後一層的 fully connected layer 移除，將特徵 embedding 到一個較低的 500 維空間中，作為 LSTM 的輸入。

最後，對於文字的處理，先以 one-hot vector 的方式表示，再 embedding 到一個較低的 500 維空間中，利用 back propagation 來調整他的參數，embedded 完的向量會跟第一層 LSTM 計算出的隱藏狀態結合在一起，作為第二層 LSTM 的輸入。除了這篇”Sequence to Sequence – Video to Text”之外，我們也有閱讀了其他相關的 paper，但是我們最後決定採用這篇作為我們主要的參考架構，因此做了更深入的介紹，下面是幾個我們在 survey 過程中看到的有趣架構：

1. Multi-Task Video Captioning with Video and Entailment Generation (ACL 2017)[3]

這篇論文的框架主要包含以下三個部分：

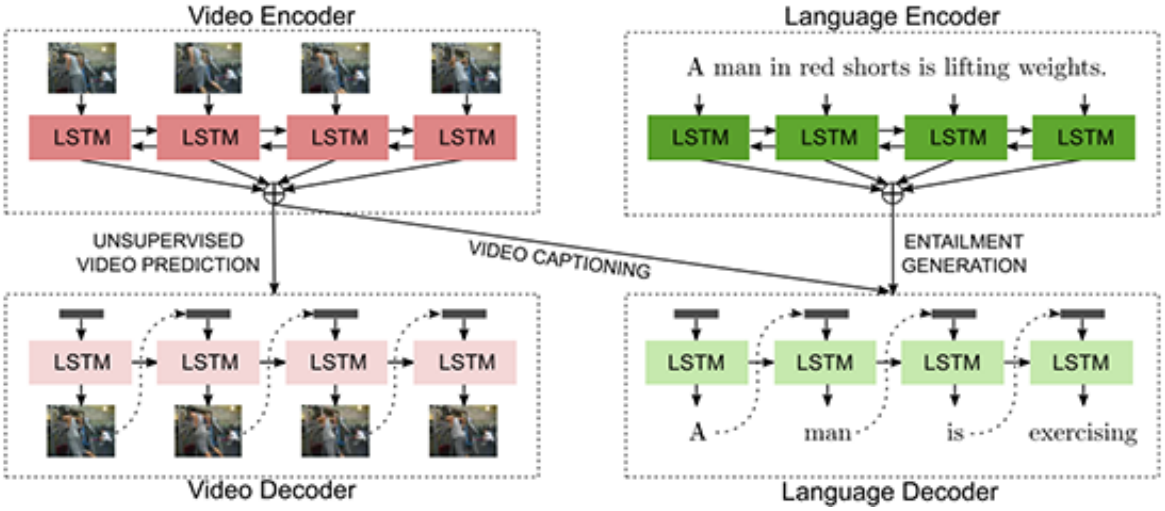
i. Unsupervised video prediction: 使用 video encoder 和 video decoder，來預測未來的 video frame

ii. Entailment Generation:使用 language encoder 和 language decoder，來生成句子含義相似的新句子

iii. Video Captioning: 使用了 video encoder 和 language decoder，來對影片進行描述。

在這幾個任務中，video encoder 和 language decoder 的參數是共用的。在訓練

中，該方法在 mini-batch 層面對三個部分進行迭代訓練（不同任務的訓練次數比例由參數確定）。其方法結構圖如下圖所示。



圖三、參數共用，取自原文[3]

2. Weakly Supervised Dense Video Captioning (CVPR 2017)[4]

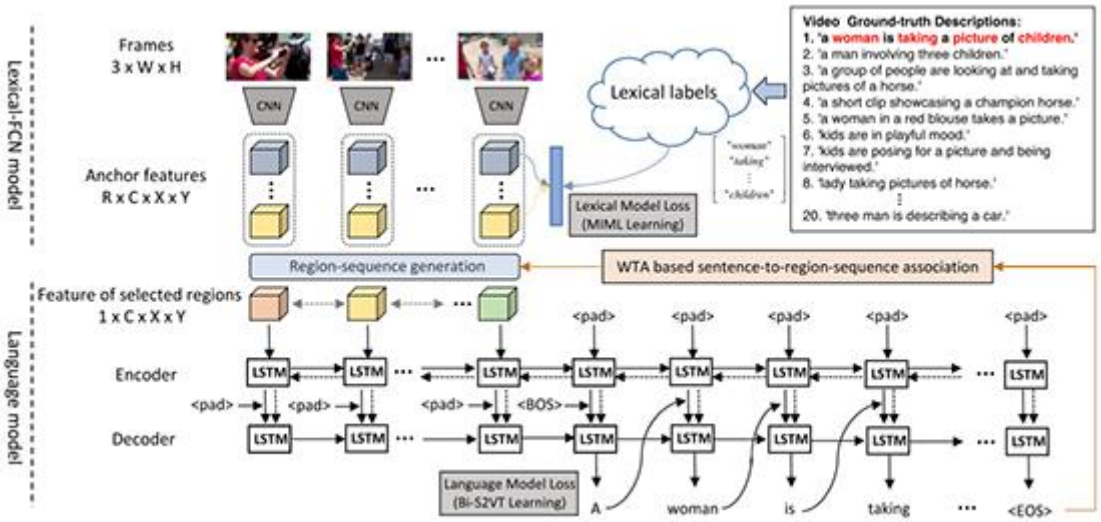
這篇文章主要研究的是 dense video captioning 問題，dense video captioning 的目的是要產生對一段視頻所有可能的描述，也就是對影片提供不同資訊以及面向的描述。本文提出了一種 weakly supervising 的方法，Multi-instance multi-label learning(MIMLL) 。MIMLL 直接從 video - level sentence annotations 中學習每個視頻圖象區域對應的描述詞彙向量。之後將這些詞彙描述向量結合起來作為 encoder-decoder 的輸入，實現 video captioning。作者宣稱其與現行的 video caption 有三大不同之處：

i. 使用 Lexical-FCN 以及 weakly supervising 的方法來 label 出不同影片區域對應到的 lexical labels

ii. 對於 1.的網絡輸出使用 submodular maximization scheme 來產生 region-sequence，這邊使用了 WTA(贏者全拿)的演算法來計算。

iii. 使用 s2s，也就是類似前面我們提及的 S2VT 的方法來產生想要的結果。

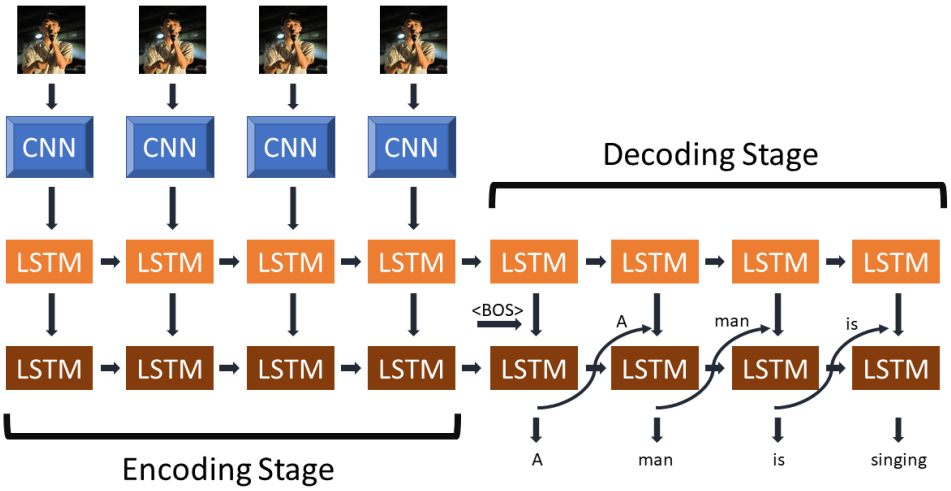
其方法結構如下圖所示。



圖四、s2s 架構，取自原文[4]

三、模型設計與實作

我們初始的計畫是採用 S2VT[1](Sequence to Sequence Video to Text)模型，由一個 LSTM encoder 和一個 LSTM decoder 組成，其架構參考自 Translating Videos to Natural Language Using Deep Recurrent Neural Networks[1]，其基本架構如下圖。



圖五、我們的原始架構

使用的套件包含 PyTorch、gensim 的 word2Vec 模組。

於撰稿時，這個基本架構已經完成，不過我們發現目前的模型還無法產生完整的句子，只會有很多”A man is playing a man”，但是我們有發現”kichen[*sic*]”是一個很容易被感測到的字，整體而言，模型的表現仍差強人意。

四、其他的方法與改良目標

1. 對於初始計畫的優化

目前的首要計畫便是對於我們已有雛型的基本模型作優化，打算嘗試的方法有

schedule sampling、attention 等等，對於文字轉向量的模型調整和預處理的調整也都是可行的方法。然而，因為我們發現直接 seq2seq 較為複雜，所以應該會先嘗試使用別的模型試試看：

2. Video to BOW 方法

這個方法的想法是只要模型能夠抓到關鍵的詞彙，就可以選出正確的答案，因此，我們就可以用 LSTM 的模型直接訓產生 bag of word 的向量。這個想法有很多地方可以調整。首先，每個單字的權重可以用語言模型去計算，把比較常出現的字權重條小一些。另外，如果一個字重複多次該怎麼處理也是一個可以嘗試的方向。目前的計畫是以 KL-divergence 作為 loss 去訓練。

3. Sentence embedding 方法

這個方法是由上面方法改良而成的，其基本思路是訓練一個 sentence embedding 模型，以 LSTM 把影片敘述轉成一個較短的向量，這方法比起上述 BOW 方法有一個好處，那就是文字的順序資訊可以被運用，而不只看文字的頻率。訓練時，我們有一個 LSTM 模型把影片轉成一維向量，另一個模型則把敘述轉成一維向量，盡量使得兩者接近。測試時則把影片與選項分別送進兩模型，就可以比較兩者相似度。

五、目前的工作分配

許秉倫：模型架設、W2V 訓練、S2VT 訓練
楊晟甫：文獻探討與探討部分撰寫、Attention 實作
歐瀚墨：測試與判斷部分實作、Proposal 底稿、其他方法實作

六、參考資料

1. Saenko, Kate(2017) Translating Videos to Natural Language Using Deep Recurrent Neural Networks Retrieved from : https://berkeley-deep-learning.github.io/cs294-131-s17/slides/saenko-talk.pdf?fbclid=IwAR1gDyidPZ-CPxipsG9sQw62qoWhfbVTP0yJP6YbRN_DiFiA-EQshASyz2c
2. Venugopalan and others Sequence to Sequence－Video to Text Retrieved from : http://www.cs.utexas.edu/users/ml/papers/venugopalan.iccv15.pdf?fbclid=IwAR0auU-S0yS5vrTIyf0Ynftr1AusoSc6r3-ULU8gqo7YQ-bfgaz4aPzIV8A
3. Ramakanth Pasunuru and Mohit Bansal(2017) Multi-Task Video Captioning with Video and Entailment Generation Retrieved from : https://arxiv.org/pdf/1704.07489.pdf
4. Shen and others(2017) Weakly Supervised Dense Video Captioning Retrieved from : https://arxiv.org/pdf/1704.01502.pdf