學號：B03705049 系級： 資管四 姓名：廖寬璿

Model Description

- Encoder:
  - Input: 80x4096
  - LSTM(1024)
- Decoder:
  - Attention: Luong(1024) + Bahdanau(1024)
    - Input: Encoder output
  - LSTM(1024)
    - Initial state: Encoder final state
- Dense(6018) softmax output projection

Attention Mechanism

Implementation

使用 tensorflow.contrib.seq2seq 的 AttentionWrapper
簡化的計算如下：

For each timestep:

```
score = decoder_output * encoder_output
or
score = decoder_output + encoder_output

alignment = softmax( score * previous_alignment )
context = alignment * encoder_output
attention = Dense(decoder_output, context)

next_decoder_input = concat( attention, decoder_output )
```

Luong Attention

Score = decoder_output * encoder_output

Bahdanau Attention

Score = decoder_output + encoder_output

Comparison

Model:

- Encoder:
  - LSTM(1024)
- Decoder:
  - Attention (if exists)
  - LSTM(1024)
- Dense(6018)

Training:

- Schedule Sampling:
  - Inverse time decay
- 12 Epochs
  - 每個影片的每個caption都跑過一次 = 1 epoch

Results

Attention	**Bleu Score (new)**	**Bleu Score (old)**
None	0.6762	0.2792
Luong(1024)	0.7022	0.2815
Bahdanau(1024)	**0.7134**	0.2906
Luong(1024)+Bahdanau(1024)	0.7089	**0.2933**

有 attention 比沒 attention 好, 但多少 attention 跟評分方式有關

Performance Improvement

Technique

1. 先找到一個會跑 / 學得會東西的 model
2. 一次改一個 parameter (*2 *10 /2 /10…)
3. 比較改同一個 parameter 的 model 選出最好的
4. 選好的 parameter 全部用在一個新的model
5. 重複到不想做

Why

- 可以一次排很多個model下去跑再回來看結果
- 簡單
- 重複因為不同 parameter 的組合有不同的結果

Experiments

Scheduled Sampling

Training 時有 scheduled sampling 比沒有好
Inverse time decay 比其他 (polynomial, natural exponential, exponential) 好一點點

Greedy vs Sampling

Predict 時 greedy 比 sampling 好一點

Beam Search

Beam 10 ~ 100 結果跟 greedy (beam = 1) 一樣

S2VT

比較難寫
結果跟傳統 seq2seq + attention 插不多

GRU vs LSTM

LSTM 稍微好一些

Label embedding

6000 字作 embedding
50 > 500 > 5000

Layer Size

1024 > 512 > 128

Attention

Additive + Multiplicative (Bahdanau + Luong) > Luong > Bahdanau > No attention