R05922131 祝子軒
ADLxMLDS hw2 report

Video Captioning

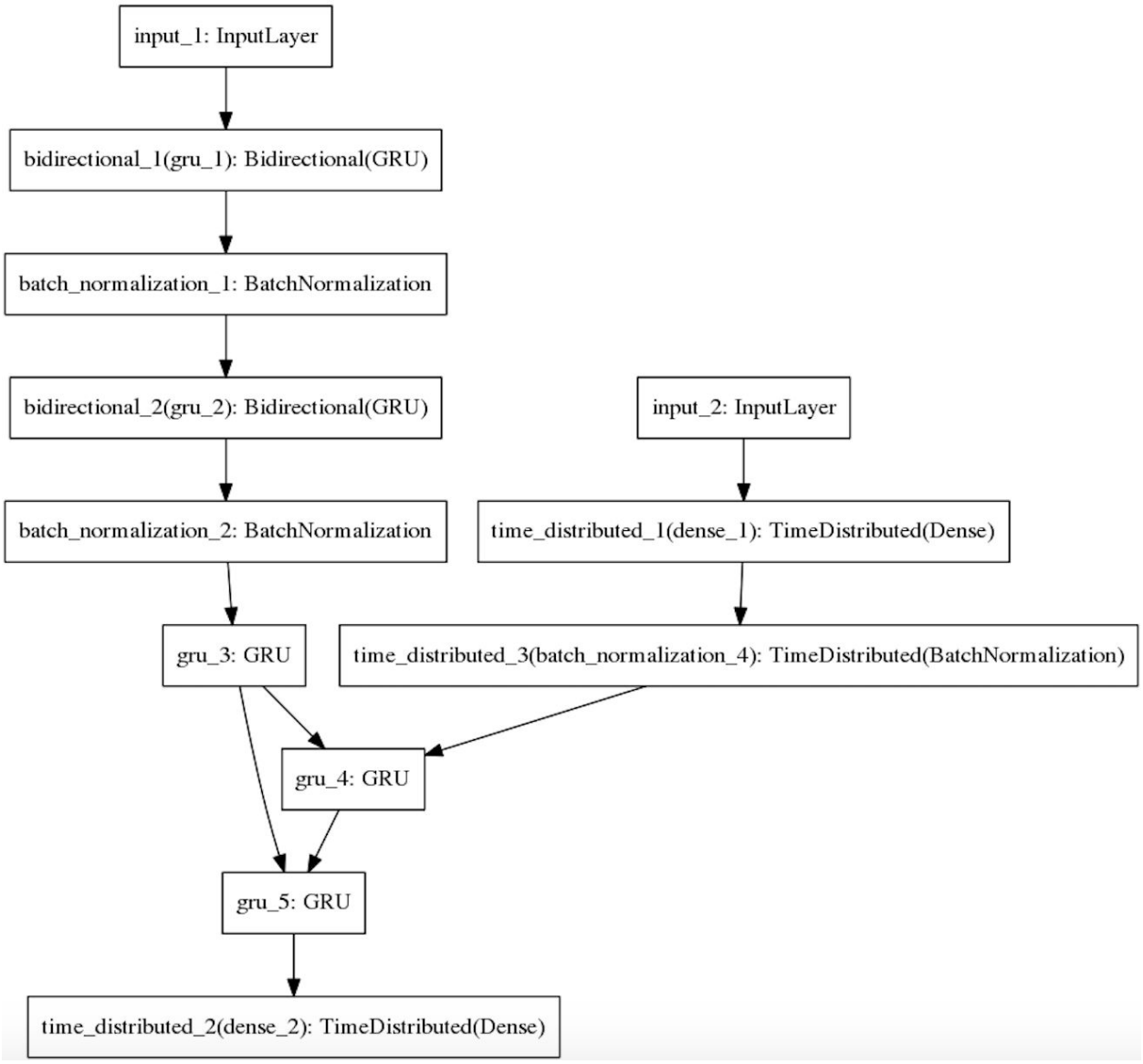
Model description

❖Seq2Seq Model （共四種變形）

- ➢ 使用encoder-decoder的架構
  - encoder
    - 將影片的feature的80個frame的features送入RNN, 因此RNN 的 input shape 為 (batch_size,80,4096), RNN 的output 為64-dim 的state
  - decoder
    - 將encoder輸出的state作為initial_state， 而將 <BOS>當作第一個輸入，再將predict出來的字當作第二個輸入，直到<EOS>出現或是到達預設的最大長度。
- ➢ 變形一
  - single ground truth model
    每段影片都有數個caption當作ground truth，在此只單純選其中一個caption拿來訓練。
- ➢ 變形二
  - attention model
    single ground truth model加上使用attention機制
- ➢ 變形三
  - multiple ground truth model
    每段影片選數個caption一起訓練。

➢ 變形四：

- multiple ground truth model with teacher force
  每段影片選數個caption一起訓練。並把caption直接當作decoder的input和output作訓練
  ex:
  input : <BOS>,w1,w2, …… ,wk,<padding>
  output : w1,w2, …… , wk, <EOS>, <padding>



Attention Mechanism

❖Q1 : How do you implement attention mechanism?

➢ 將decoder的input和encoder上面每個時間序上的output作elementwise相乘，在將所有的相乘結果作平均，將此結果當作新的decoder的input，也就是每一次的output理論上會把較相關的frame feature會得到較重的權重。

❖Q2 : Compare and analyze the results of models with and without attention mechanism.

➢ without attention model:

- BLEU_1 : 0.254　BLEU_2 : 0.622
- sample result :

```
ScdUht-pM6s_53_63.avi                A man is talking on a
wkgGxsuNVSg_34_41.avi                  A man is a a a head
BtQtRGI0F2Q_15_20.avi              A man is playing a bike
 k06Ge9ANKM8_5_16.avi                   A baby is a a ball
sZf3VDsdDPM_107_114.avi                   A man is on a a
```

➢ attention model :

- BLEU_1 : 0.265　BLEU_2 : 0.619
- sample result :

```
ScdUht-pM6s_53_63.avi                A man is cutting into and
wkgGxsuNVSg_34_41.avi                        A dog is in a pool
BtQtRGI0F2Q_15_20.avi              A man is playing cricket the
 k06Ge9ANKM8_5_16.avi                        A dog is a a couch
sZf3VDsdDPM_107_114.avi          A man is speaking on a microphone
```

➢ Discussion

- Attention可以決定哪個frame比較重要，但在這次作業的dataset上，因為影片並不長，所以每個frame的畫面差異不大，並且訓練的資料不多，因此多了attentin機制應該增加了不少overfitting的風險並且並沒有太多的優點，所以performance並沒有明顯變好。

How to improve your performance

- **Trade-off :**
  - caption的長度不一造成padding過長，因而造成記憶體空間及訓練難度的一些問題，因此選擇caption的長度在8以下的caption拿來做為training data，雖然少了一些可能的資訊，但卻能換來計算上以及語言架構上的一些好處。
  - 在BLEU score最高的model下(只train數個epoch），output的結果常常是"a man is a a a"，顯然不是合理的output，但因為是較為通用的output因此分數較高。但若train到fit training data的情況下，能夠學到語句的前後關係，因此能夠output出很通順的句子，但卻常常偏離影片的主題。因此最後選擇了在loss下降變慢前的model，可得到還算通順的句子，且能兼顧BLEU score。

Experimental results and settings

- **The better output model (multiple ground truth model with teacher force):**
  - epoch : 11
  - BLEU_1 : 0.279　BLEU_2 : 0.603
  - sample result :

```
   ScdUht-pM6s_53_63.avi              A woman is opening a carpet
   wkgGxsuNVSg_34_41.avi                  A man is jumping a dogs
   BtQtRGI0F2Q_15_20.avi                    A boy is is on a ball
    k06Ge9ANKM8_5_16.avi           A toddler ball playing a doll
sZf3VDsdDPM_107_114.avi              A woman is doing her doll
```

- **The highest BLUE score model (multiple ground truth model with teacher force)**
  - epoch : 3
  - BLEU_1 : 0.287　BLEU_2 : 0.732
  - sample result :

```
   ScdUht-pM6s_53_63.avi              A woman is a into a
   wkgGxsuNVSg_34_41.avi                 A man is a a the
   BtQtRGI0F2Q_15_20.avi                 A man is a a the
    k06Ge9ANKM8_5_16.avi          A cat is playing a the
sZf3VDsdDPM_107_114.avi                  A girl is a a
```