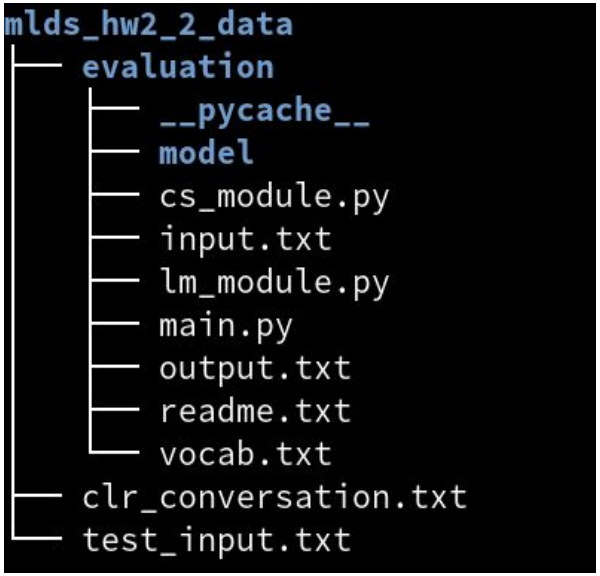
MLDS HW2-2

TAs
ntu.mldsta@gmail.com

HW2-2 UPDATE(4/27)

1. baseline code release (link)
2. Testing data release (link)
3. Perplexity baseline release **<= 100**
   Correlation baseline release **>= 0.45**
4. Whole dataset download: (link)



重要：本次 model evaluation 的結果都僅供參考而已，請同學不要在這上面做太多琢磨，只是給同學們寫報告時有個量化依據。

Outline

- Timeline
- Task Descriptions
- Q&A

Timeline

Two Parts in HW2

- (2-1) Video caption generation
  - Sequence-to-sequence model
  - Training Tips
- (2-2) Chatbot

Schedule

- 3/30：
  - Release HW2-1
- 4/13：
  - Release HW2-2
- 4/27：
  - Midterm
  - HW1 上台分享
- 5/4：
  - All HW2 due (including HW2-1, HW2-2)

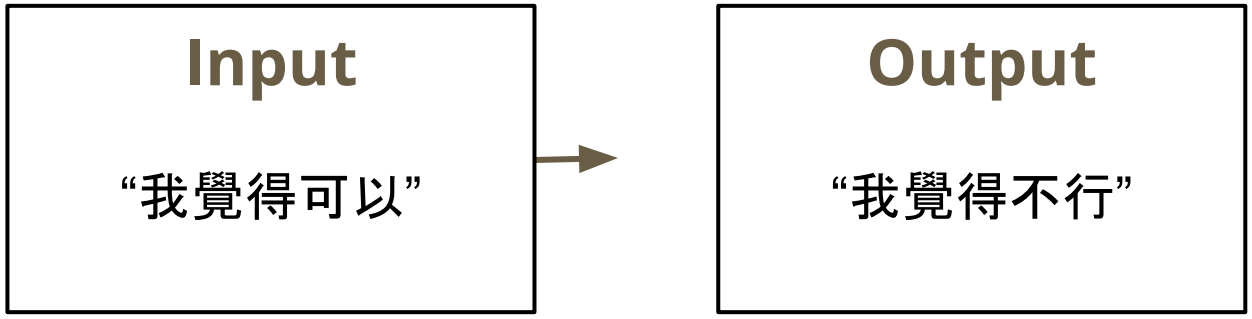
Task Descriptions

HW2-2: Chinese Chatbot

- Introduction
- Sequence-to-sequence model
- Training Tips
  - Attention
  - Schedule Sampling
  - Beamsearch
- How to reach the baseline ?

HW2-2 Introduction

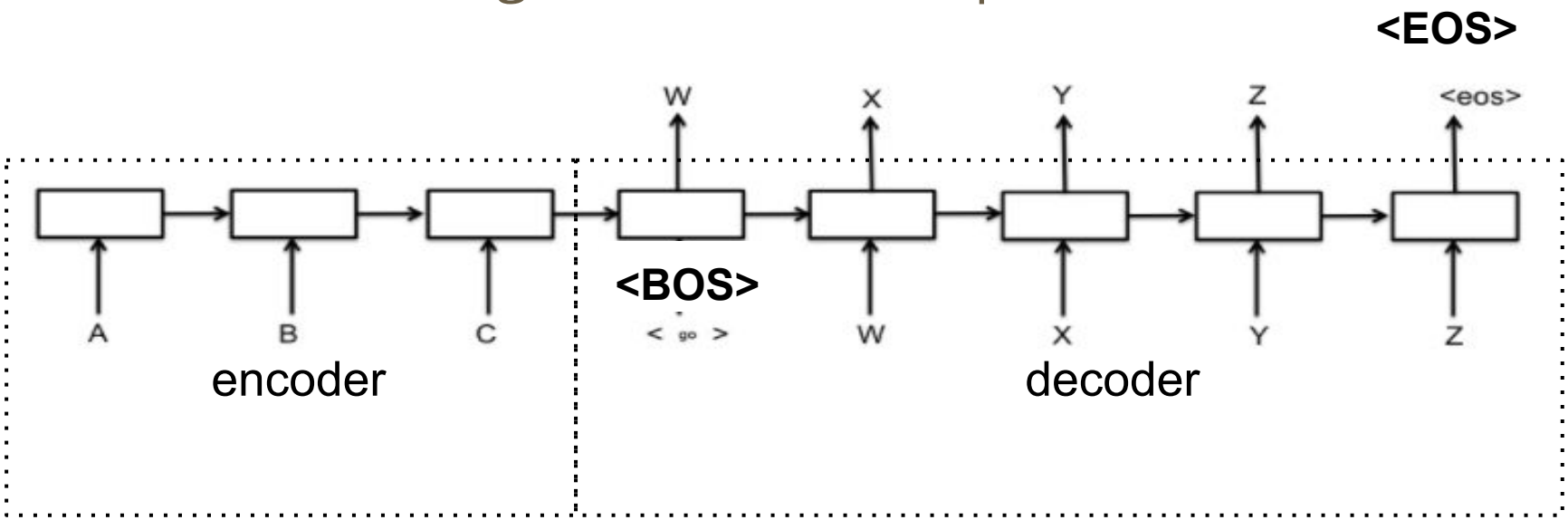
- Chatbot
  a. Input : A sentence
  b. Output: The corresponding reply.



- There are several difficulties including:
  a. Variable length of I/O

HW2-2 Sequence-to-sequence 1/5

- **Two recurrent neural networks (RNNs)**
  an encoder that processes the input
  a decoder that generates the output



https://www.tensorflow.org/tutorials/seq2seq

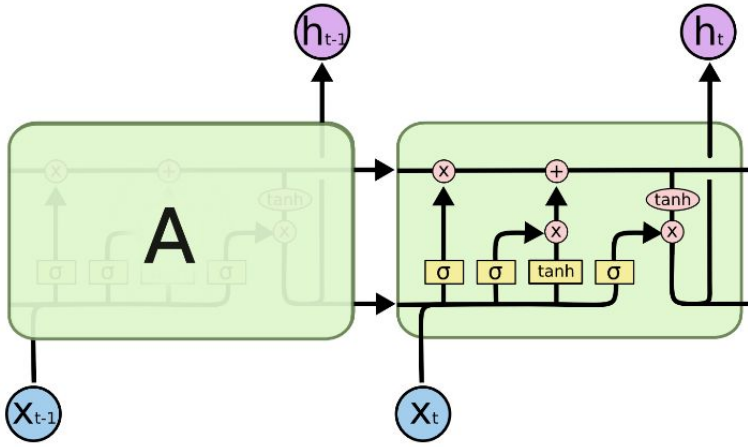
HW2-2 Sequence-to-sequence 2/5

- **Data preprocess:**
  - Dictionary - most frequently word or min count
  - other tokens: <PAD>, <BOS>, <EOS>, <UNK>
    - <PAD> : Pad the sentencen to the same length
    - <BOS> : Begin of sentence, a sign to generate the output sentence.
    - <EOS> : End of sentence, a sign of the end of the output sentence.
    - <UNK> : Use this token when the word isn't in the dictionary or just ignore the unknown word.

HW2-2 Sequence-to-sequence 3/5

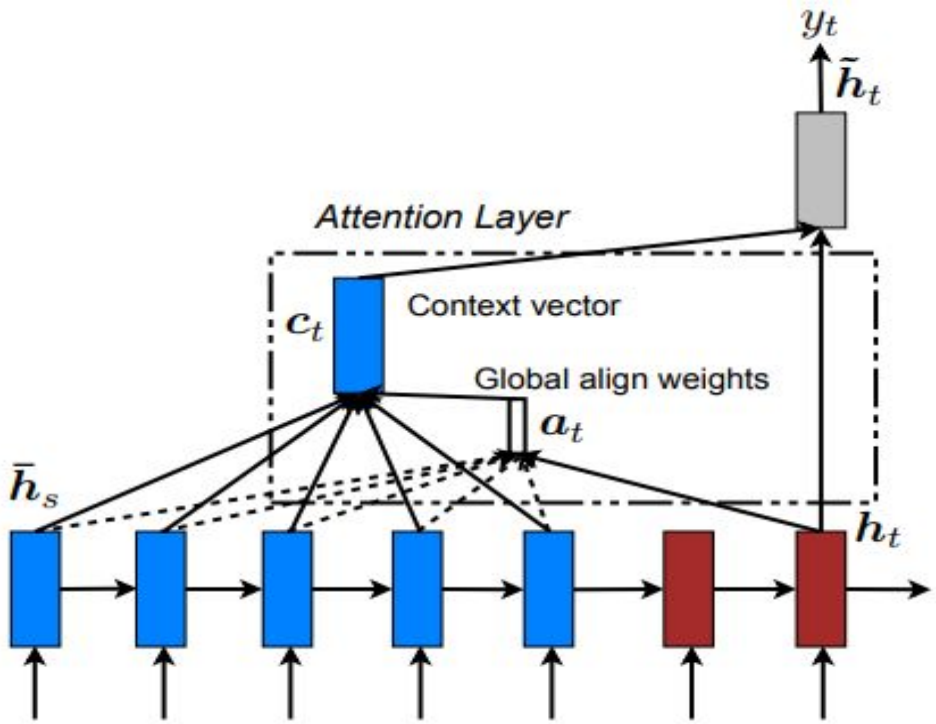
- **Text Input:** reference
  - One-hot Vector encoding
    (1-to-N coding, N is the size of the vocabulary in dictionary)
  - e.g.
    - neural = [0, 0, 0, ..., 1, 0, 0, ..., 0, 0, 0]
    - network = [0, 0, 0, ..., 0, 0, 1, ..., 0, 0, 0]

- **LSTM unit:**
  cell output than project to a vocabulary-size vector



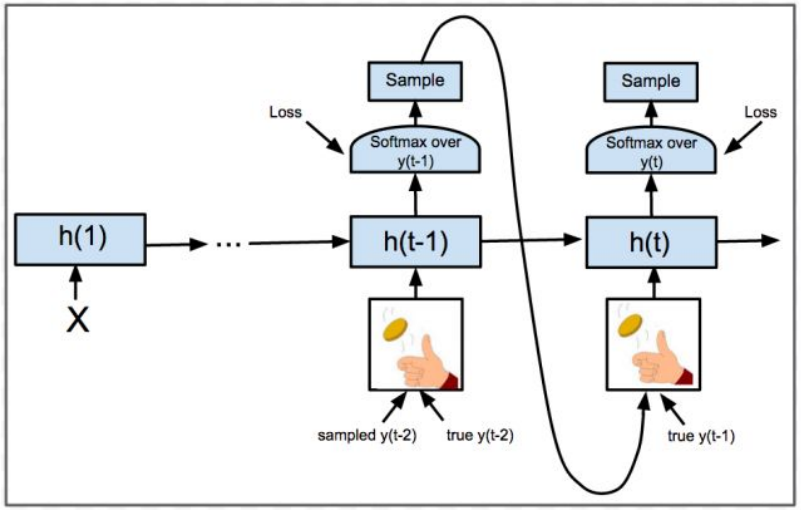
HW2-2 Training Tips - Attention 1/3

- Attention on encoder hidden states :
  - Allow model to peek at different sections of inputs at each decoding time step



HW2-2 Training Tips - Schedule Sampling 2/3

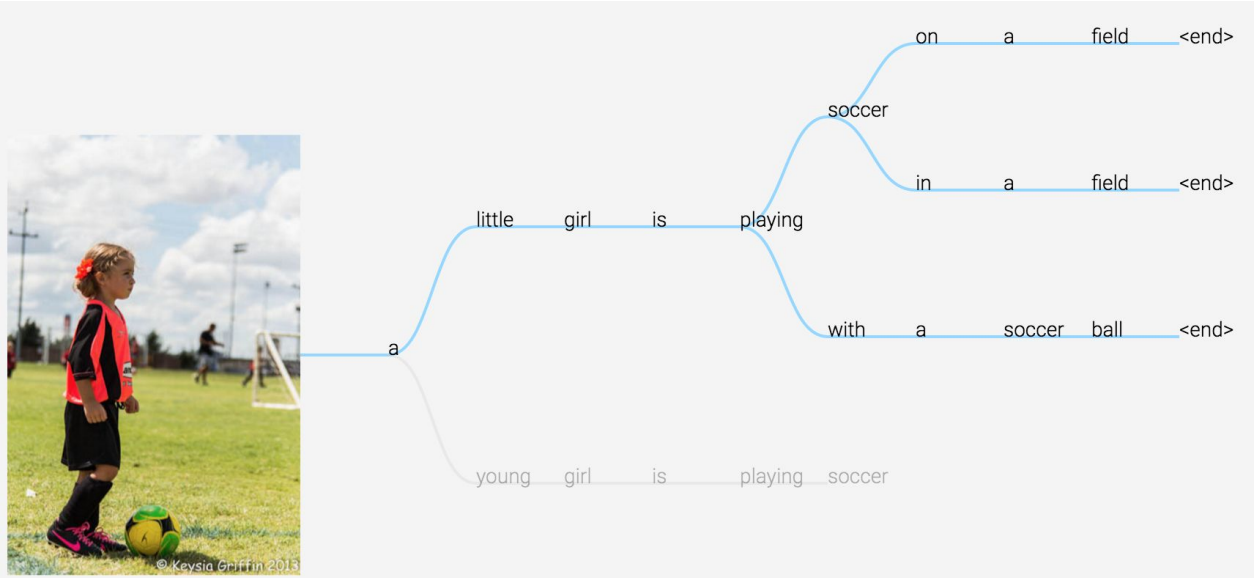
- Schedule Sampling :
  - To solve “exposure bias” problem,
    When training, we feed (groundtruth) or (last time step’s output) as input at odds



https://arxiv.org/abs/1506.03099

HW2-2 Training Tips - Beam search 3/3

- Beam search:
  - keep a fixed number of paths



Demo: http://dbs.cloudcv.org/captioning

HW2-2 How to reach the baseline ? 1/3

- **Baseline:**

  Perplexity < 100

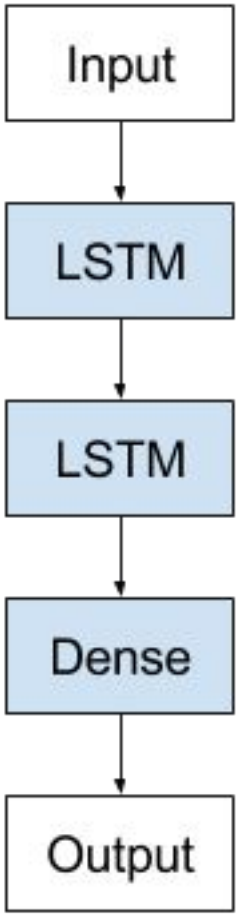
  Correlation Score >0.45

  Baseline model vocab

  Perplexity baseline code (link)

  Correlation baseline code(link)

  重要:本次 model evaluation 的結果都僅供參考而已，請同學不要在這上面做太多琢磨，只是給同學們寫報告時有個量化依據。



- Baseline model:

  Training iteration = 750000

  - Batchsize = 100
  - GRU dimension = 256 2 layers
  - Learning rate = 0.001
  - Sgd Optimizer
  - Training time = 8hrs on GTX1060

HW2-2 How to reach the baseline ? 2/3

- Evaluation : Perplexity

$$H(S) = -\sum_i p(x_i) \log[p(x_i)]$$

$$PP(S) = 2^{H(S)}$$

where H = entropy, PP= Perplexity

  - e.g. :

"I love NLP."

$$\prod_{i=1}^{n} p(w_i) = p(\text{'NLP'} \mid \text{'I', 'love'}) * p(\text{'love'} \mid \text{'I'}) * p(\text{'I'})$$

$$log_2 \prod_{i=1}^{n} p(w_i) = \sum_{i=1}^{n} log_2 p(s_i)$$

$$PP = 2^{\frac{-1}{N} \sum_{i=1}^{n} log_2 p(s_i)}$$

  - Language Model will be released soon.
  - 數位語音處理概論 lesson6

HW2-2 How to reach the baseline ? 3/3

- Evaluation : Correlation Score
  - Decided by Model.
  - The model is training by given dataset.
  - A kind of Discriminator.

- Model detail:
  - Correct scored 1, incorrect scored 0
  - Activation function sigmoid

Data & format

- Dataset:
  - 語音實驗室的電影字幕
    - 500萬句對話
- Format:
  - 一行一句話
  - 對話跟對話中間用+++$+++分隔
  - Download clr_conversation.txt
- Extra Data:
  - 以下為未整理data不符合上列格式
  - 連續劇data
  - 電影data(完整版)
  - 簡體corpus (baseline的language model不認得簡體 請自行轉換)

這 不是 一時 起意 的 行刺
而是 有 政治 動機
上校 , 這種 事
+++$+++
他 的 口袋 是 空 的
沒有 皮夾 , 也 沒有 身分證
手錶 停 在 4 點 15 分
大概 是 墜機 的 時刻
他 的 降落傘 被 樹枝 纏住 了

I/O Format

- Input:
  - 一行一句話

```
1  你好
2  今天天氣如何？
3  作業好多
```

- Output:
  - 一行一句話

```
1  你好
2  今天天氣很好
3  活該笑你
```

Submission & Rules

- Please implement one seq-to-seq model (or it's variant) to fulfill the task
- Extra dataset is allowed to use.
- Allow package:
  - python 3.6
  - **TensorFlow r1.6 ONLY** (CUDA 9.0)
  - PyTorch 0.3 / torchvision
  - Keras 2.0.7 (TensorFlow backend only)
  - MXNet 1.1.0, CNTK 2.4
  - matplotlib, Python Standard Library
  - If you want to use other packages, please ask TAs for permission first!
  - new allowed package:
    Gensim, pandas, tqdm

Submission & Rules

- Deadline : **2018/5/4 23:59 (GMT+8)**
- Upload **code** and **report** of HW2-1, HW2-2 to Github in **different** directory.
- For HW2-2 :
  - Your github must have directory hw2/hw2_2/, and there should be :

    (1) report.pdf (2) your_seq2seq_model **(3) hw2_seq2seq.sh**

    (4) model_seq2seq.py (*training code should include*)
  - If your model are too big for github, upload to a cloud space and **write it in your script to download the model.**
  - Please write shell script "**hw2_seq2seq.sh**" to run your code and follow the script usage below :

    - ./hw2_seq2seq.sh $1 $2
    - $1: input filename (format : .txt), $2: output filename (format : .txt)
    - Example ./hw2_seq2seq.sh input.txt output.txt

    Your script should be done within **10 mins** excluding model donwloading.
  - Please do not upload any dataset to Github (include external dataset).

Grading Policy

- HW2-1 : 15%
- HW2-2 : 10%
  - Baseline (2%):
    - Perplexity(1%)
    - Correlation Score(1%)
  - TAs review (2%):
    - Grammar score (1%)
    - Relative score (1%)
  - Report (6%)
- 分工表:0.5%
- 上台分享 : 1%
- 上台分享前三名:1%

Grading Policy - Report (6%)

- Do not exceed **4** pages and written in Chinese.
- Model description (2%)
  - Describe your seq2seq model
- How to improve your performance (3%)

  (Please do the method different with hw2-1)

  (e.g. Attention, Schedule Sampling, Beamsearch...)
  - Write down the method that makes you outstanding (1%)
  - Why do you use it (1%)
  - Analysis and compare your model without the method. (1%)
- Experimental results and settings (1%)
  - parameter tuning, schedual sampling ... etc
- README : please specify library and the corresponding version in README

Grading Policy - NOTICE

- Late submission (link):
  - Please fill the late submission form first only if you will submit HW late.
  - Please push your code before you fill the form
  - There will be 25% penalty per day for late submission, so you get 0% after four days
- Bug:
  - You will get 0% in Baseline and TAs review if the required script has bug.
  - If the error is due to the format issue, please come to fix the bug at the announced time, or you will get 10% penalty afterwards.

Q&A

ntu.mldsta@gmail.com