

# 2019 SDML HW2-1

Sequence-to-sequence model

# Goal of this task

- Learning how to implement/use a sequence-to-sequence model (HW2.0)
- Learning how to control the outputs of a sequence-to-sequence model (HW2.1)
- Understanding and explaining the capability of a GRU-based Seq2Seq model (HW2.2)

# HW2.0 Implementing a Seq2Seq Auto-encoder

- Due: 11/7
- Score: 10% of HW2
- Individual homework (everybody on your own)
- Data: Chinese lyric

# HW2.0 Implementing a Seq2Seq Auto-encoder

- Generate output sentence which is the same as input.
- Please use GRU as your basic cell (it will make your life easier for hw2.2)
- Dataset is a series of sentence, and sentence token is separated by space.
- Dataset example:
  - <SOS> Tom seemed annoyed <EOS>
  - <SOS> I was enthusiastic <EOS>
- Target accuracy: 80% (整句對才算對)

# HW2.1 Controlling the Outputs of a Seq2Seq

- Due: 11/7
- Score: 20% of HW2
- Individual homework (everybody on your own)
- Data: Chinese lyric

# HW2.1 Controlling the Outputs of a Seq2Seq

- Generate next sentence with previous sentence and some control signals
  - The control signal controls the position of a word to be generated
- source/target sentence token的間隔為一個空白
- Example
  - <SOS> 这样你的泪滴 <EOS> 1 能 □ <SOS> 能流得少一些 <EOS>
  - <SOS> 像空无一人一样华丽 <EOS> 2 渐 6 知 □ <SOS> 我渐渐失去知觉 <EOS>
- Target accuracy (指定正確數量 / 總指定數量)
  - HW2.1-1 Assign 1 word: 95%
  - HW2.1-2 Assign 1~2 words: 83%

# HW2.0 Evaluation

- 輸出成檔案 task1.txt
- <SOS>, <EOS> 需要被輸出。 一行一句, 每個token間用空白分隔。
- Example
  - <SOS> How was class <EOS>

# HW2.1 Evaluation

```
python3 hw2.1_evaluate.py --training_file=hw2.1-1_sample_testing_data.txt \  
--result_file=hw2.1-1.txt
```

result\_file格式:

一行一句，頭尾要有<SOS>和<EOS>，每個token間用空白分隔。

例: <SOS> 作業太簡單 <EOS>



# Data link

- Google drive link: <https://drive.google.com/drive/folders/1i00N26AB5a-BYU6nPIWN2MNysthoE74L?usp=sharing>
- HW 2.0 training dataset
  - You can split training and validation dataset by your own
- HW 2.1-1 and HW 2.1-2 sample testing dataset
- For HW2.1-1, we provide a function to test your accuracy
  - We will use the same function to evaluate your performance on testing dataset
- HW 2.0 & 2.1 true testing dataset
  - We will **release on 11/7 class**, and you need to submit output to ceiba before class ends
  - Your score is based on this dataset

# Warning & requirements (important)

- You need to use a Seq2Seq model (with GRU) for task2.0 and task2.1, you cannot simply write rules for this homework.
  - You will need to analyze and explain your Seq2Seq model for HW2.2
- You must **NOT use attention** or transformer for encoder/decoder
- You can use third party packages, but need to record them in the report
- Violation will lead to 0 score in this homework

# Scores

- Pass baseline on testing dataset
  - task 2.0: 10%
  - task 2.1-1: 10%
  - task 2.1-2: 10%
- Report 2.1: +5%
- Top 5 persons in 2-1 (presentation on 11/14): +5%
- May loss some points for any submission format problem

# Submission

- Submit to ceiba before **11/7** 17:00
- Don't upload any data to ceiba. The file should be less than 15MB.
- r08922xxx.zip (Your school ID)
  - report.pdf
  - src/
    - Your code
  - result/ (For testing dataset)
    - **task1.txt**
    - **task2-1.txt**
    - **task2-2.txt**

# Report

- Should be no more than 2 pages
- Accuracy of these two tasks on each dataset (training and testing)
- What package you use
- Your solution
- Experiments (optional)
  - Different layers or units
  - Parameter tuning

# TA

- [r07922035@ntu.edu.tw](mailto:r07922035@ntu.edu.tw) 李漪廷, [r07922119@ntu.edu.tw](mailto:r07922119@ntu.edu.tw) 吳達懿
- Please email to **both of us**
- Title: [SDML] hw2