Machine Learning HW5

Sequence to sequence

Outline

1. Introduction to sequence to sequence
2. Homework: machine translation
3. Workflow
4. Training tips
5. Requirements
6. Submission & Grading
7. JudgeBoi Guide
8. Links
9. Q&A

Introduction to sequence to sequence

Sequence to sequence

Generate a sequence from another sequence

Translation
text to text

ASR
speech to text

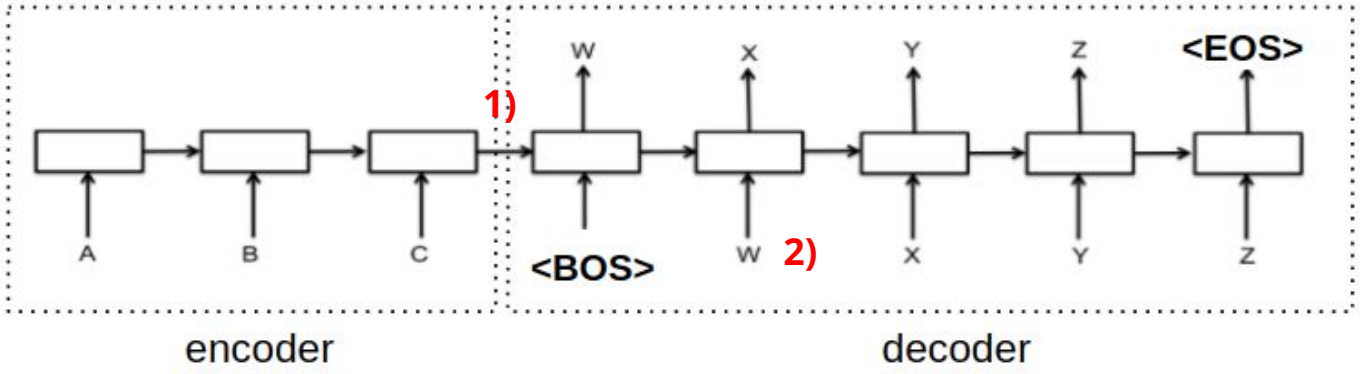
TTS
text to speech

and more...

Sequence to sequence

Often composed of encoder and decoder

- Encoder: encodes input sequence into a vector or sequence of vectors
- Decoder: decodes a sequence one token at a time, based on 1) encoder output and 2) previous decoded tokens



HW5: Machine Translation

Neural Machine Translation

We will translate from english to traditional chinese

- Cats are so cute. -> 貓咪真可愛。

A sentence is usually translated into another language with different length. Naturally, the seq2seq framework is applied on this task.

Training datasets

- Paired data
  - TED2020: TED talks with transcripts translated by a global community of volunteers to more than 100 language
  - We will use (en, zh-tw) aligned pairs
- Monolingual data
  - More TED talks in traditional Chinese

Evaluation

source: Cats are so cute.

target:貓咪真可愛。

output: 貓好可愛。

BLEU

- Modified[1] n-gram precision (n=1~4)
- Brevity penalty: penalizes short hypotheses

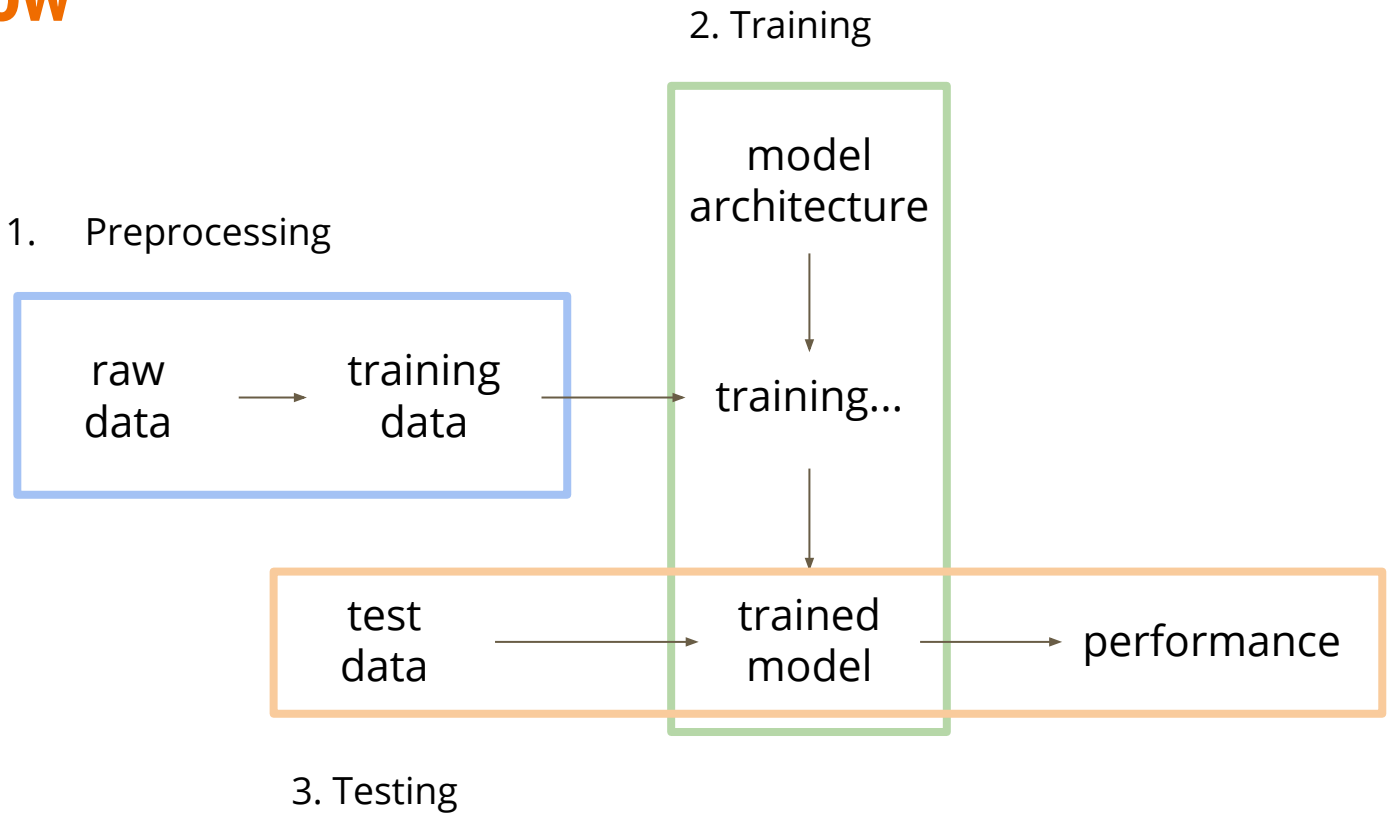
$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

  - c is the hypothesis length, r is the reference length
- The BLEU score is the geometric mean of n-gram precision, multiplied by brevity penalty

[1]the precision is clamped to # occurence in reference.

Workflow

Workflow


2. Training
model architecture
1. Preprocessing
raw data
training data
training...
test data
trained model
performance
3. Testing

Workflow

1. Preprocessing
    a. download raw data
    b. clean and normalize
    c. remove bad data (too long/short)
    d. tokenization
2. Training
    a. initialize a model
    b. train it with training data
3. Testing
    a. generate translation of test data
    b. evaluate the performance

1. 笑聲、標題等等把他clean掉
2. 把標點符號統一

或是長度差異太大

Training tips

Training tips

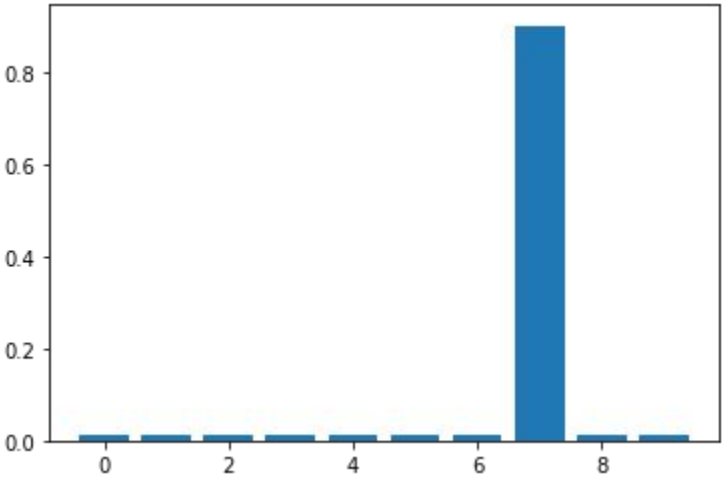
- Tokenize data with sub-word units
- Label smoothing regularization
- Learning rate scheduling
- Back-translation

Training tips

- Tokenize data with sub-word units
  - For one, we can reduce the vocabulary size (common prefix/suffix)
  - For another, alleviate the open vocabulary problem
  - example
    - `_new _ways _of _making _electric _trans port ation _.`
    - `new ways of making electric transportation.`

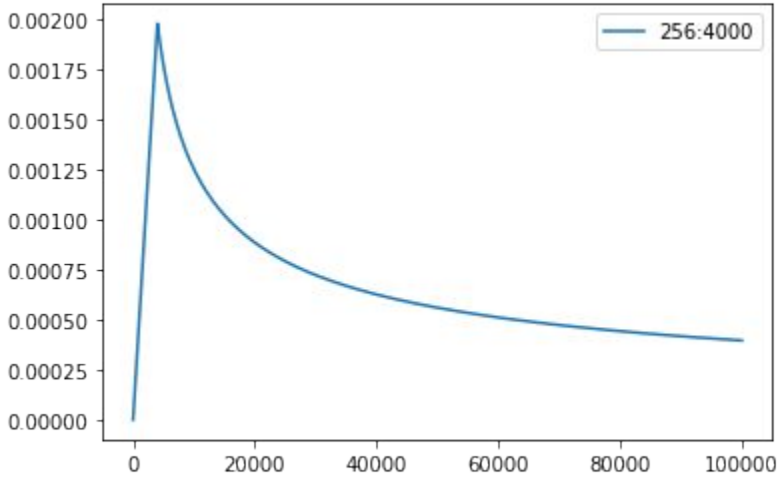
Training tips

- Label smoothing regularization
  - When calculating loss, reserve some probability for incorrect labels
  - Avoids overfitting



Training tips

- Learning rate scheduling
  - Linearly increase lr and then decay by inverse square root of steps
  - Stablilize training of transformers in early stages

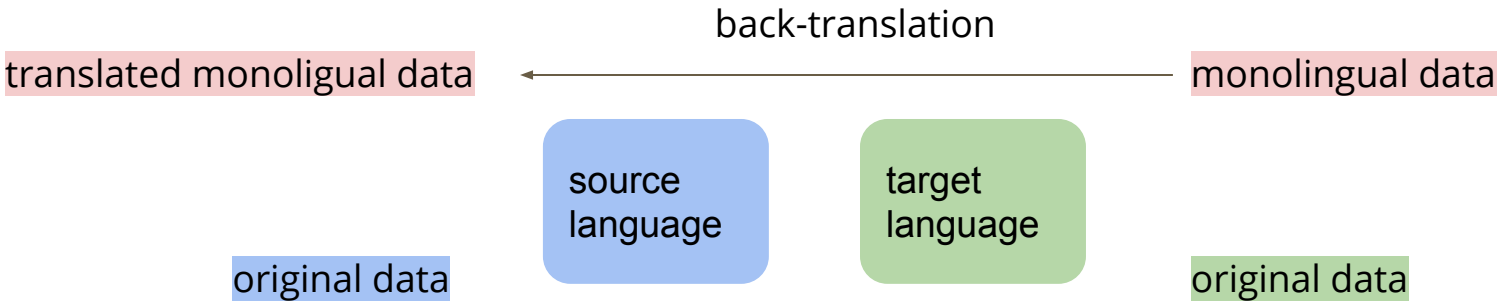


Back-translation (BT)

original goal: english -> chinese
bt: train a model for chinese -> english, and generation training data from monolingual data

Leverage monolingual data by creating synthetic translation data

1. Train a translation system in the **opposite direction**
2. Collect monolingual data in target side and apply machine translation
3. Use translated and original monolingual data as additional parallel data to train stronger translation systems



Back-translation

Some points to note about back-translation

資料來源需要很像（e.g. 都是新聞, 都是學術文章, ...）

1. Monolingual data should be in the same domain as the parallel corpus
2. The performance of the backward model is critical
3. You should increase model capacity (both forward and backward), since the data amount is increased.

Requirements

Requirements

You are encouraged to follow these tips to improve your performance in order to pass the 3 baselines.

1. Train a simple RNN seq2seq to acheive translation
2. Switch to transformer to boost performance
3. Apply back-translation to furthur boost performance

Baseline Guide

Train a simple RNN seq2seq to acheive translation

- Running the sample code should pass the baseline!

Baseline Guide

Switch to transformer to boost performance

1. Change the encoder/decoder architecture to transformer based, according to the hints in sample code
   - `RNNEncoder -> TransformerEncoder`
   - `RNNDecoder -> TransformerDecoder`
2. Change architecture configurations
   - `encoder_ffn_embed_dim -> 1024`
   - `encoder_layers/decoder_layers -> 4`
   - `#add_transformer_args(arch_args)  ->  add_transformer_args(arch_args)`

Baseline Guide

Apply back-translation to furthur boost performance

1. Train a **backward** model by switching languages
   - `source_lang = "zh"`
   - `target_lang = "en"`
2. Remember to change architecture to transformer-base
3. Translate monolingual data with backward model to obtain synthetic data
   - complete TODOs in the sample code.
   - all the TODOs can be completed by using commands from earlier cells.
4. Train a stronger forward model with the new data
   - if done correctly, ~30 epochs on new data should pass the baseline.

Submission & Grading

Prediction Submission

- Submit to JudgeBoi
- One example per line, in the original order
- Punctuation will be normalized by JudgeBoi with this script
- **Deadline: 4/30 (Fri.) 23:59**

Code Submission

- **NTU COOL** (4pts)
  - **Deadline: 5/2 (Sun.) 23:59**
  - Compress your code and report into

    <student ID>_hwX.zip

    *** e.g. b06901020_hw5.zip**

    *** X is the homework number**

  - We can only see your last submission.
  - **Do not submit your model or dataset.**
  - If your code is not reasonable, your semester grade x 0.9.

Code Submission

- Your .zip file should include only
  - **Code**: either .py or .ipynb
  - **Report**: .pdf (only for those who got 10 points)

- Example:

Regulation

- You should NOT plagiarize, if you use any other resource, you should cite it in the reference. (*)
- You should NOT modify your prediction files manually.
- Do NOT share codes or prediction files with any living creatures.
- Do NOT use any approaches to submit your results more than 5 times a day.
- **Do NOT search or use additional data or pre-trained models.**
- Your **final grade x 0.9** if you violate any of the above rules.
- Prof. Lee & TAs preserve the rights to change the rules & grades.

(*) Academic Ethics Guidelines for Researchers by the Ministry of Science and Technology

Grading

Baseline	BLEU	Points
Code submission		+4
Simple (public)	18.43	+1
Simple (private)	17.61	+1
Medium (public)	24.04	+1
Medium (private)	23.43	+1
Strong (public)	29.32	+1
Strong (private)	28.27	+1
Total		**10**

Grading -- Bonus

- **If you got 10 points**, we make your code **public** to the whole class.

- In this case, if you also submit **a PDF report briefly describing your methods** (<100 words in English), you get a bonus of **0.5 pt**. (your report will also be available to all students)

- Report template

JudgeBoi Guide

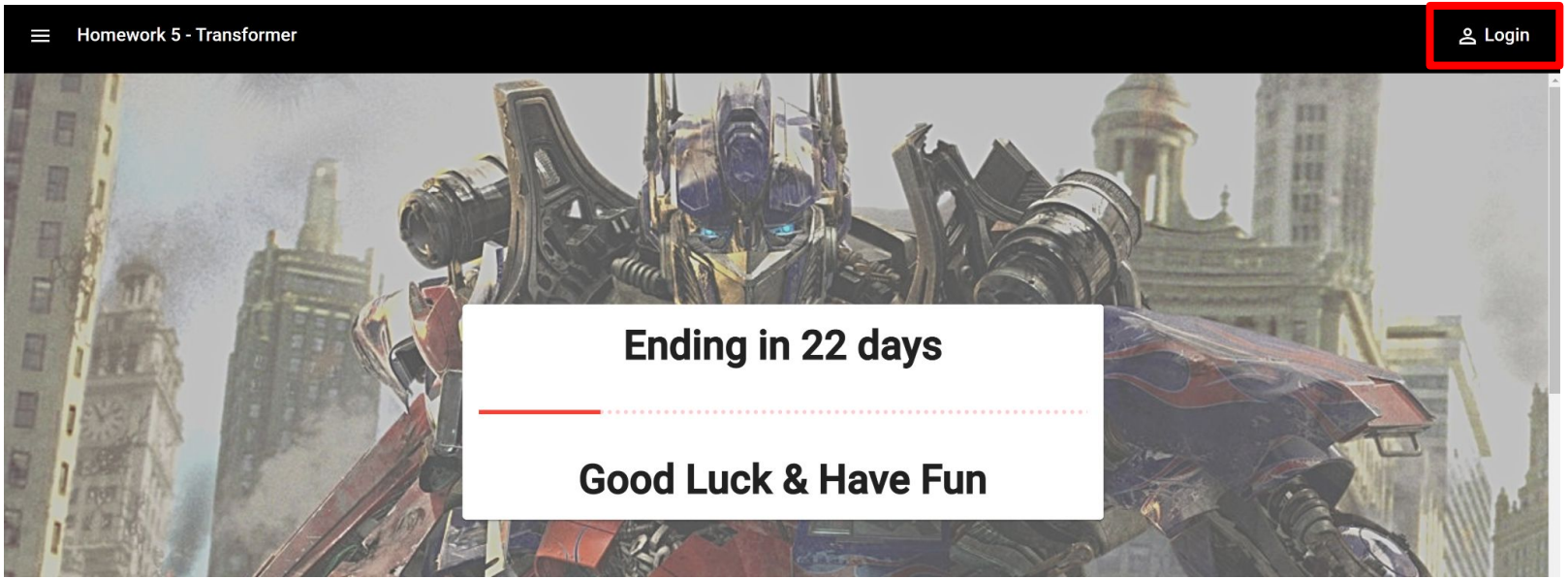
Previously... Github Account Survey

We have kindly requested everyone to report your github username and ID.

IMPORTANT: You must take this survey in order to submit to JudgeBoi server.

Step 1: Register for Submission

Go to JudgeBoi to login.

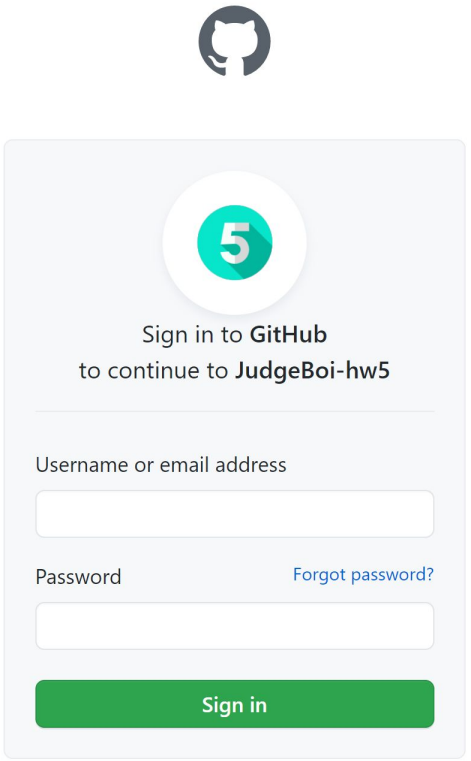


Step 2: Sign-in with Github

You need to sign in with the account you reported to us. Or you won't be able to upload your submissions.

fill in username >

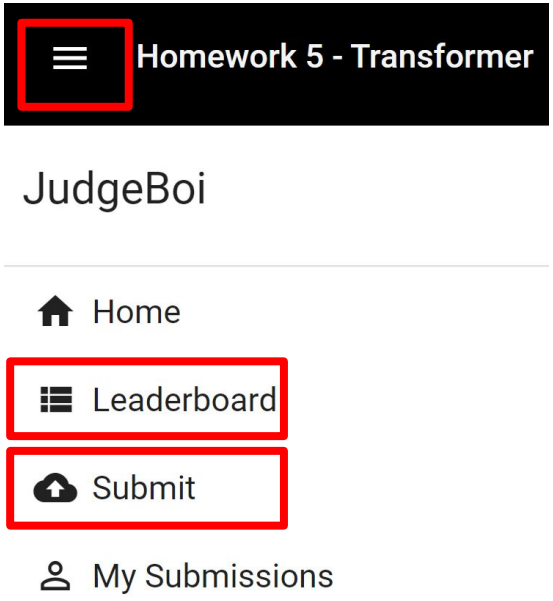
fill in password >



Step 3: Submit your Results

You can now submit results to the server and view the leaderboard.

1) click here

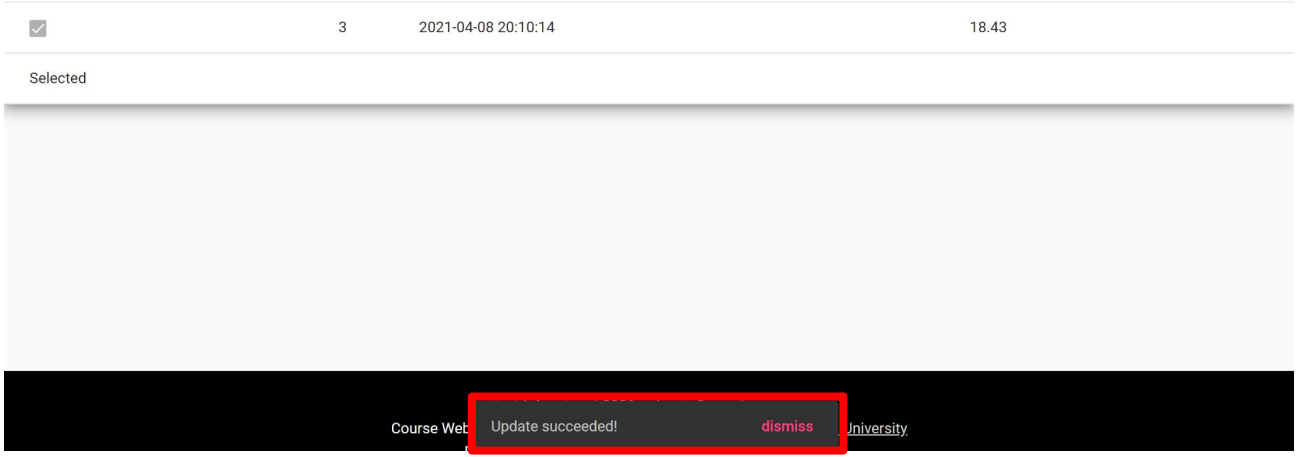


3) view leaderboard here

2) Submit result here

Step 4: Select your submissions

- You can select up to 2 submissions.
- If none of your submissions is chosen, we will use your first submission to calculate your private score.
- If your selection is successful, you will see a message box as follows:



More about JudgeBoi

- 5 submission quota per day, reset at **midnight**. Users not in whitelist will have no quota.
- Only ***.txt*** file is allowed, filesize should be smaller than **700kB**.
- The countdown timer on the homepage is for reference only.
- We do limit the number of connections and request rate for each ip. If you cannot access the website temporarily, please wait patiently.
- Please do not attempt to attack JudgeBoi, thank you.
- Every **Wednesday** and **Saturday** from **0:00 to 3:00** is our system maintenance time. If the website cannot be used during this time, please wait patiently for the completion of the maintenance.

Links

Sample code Colab Colab(chinese version)

Parallel data TED2020

Testing data Testdata

Monolingual TED_ZH

If any questions, you can ask us via...

- NTU COOL (recommended)
  - https://cool.ntu.edu.tw/courses/4793
- Email
  - ntu-ml-2021spring-ta@googlegroups.com
  - The title should begin with “[hw5]” or “[JudgeBoi]”
- TA hour
  - Each Friday during class

FAQ: BT

Q: My backward (zh-en) model is significantly weaker than forward (en-zh) model, what's going on?

A: BLEU scores aren't comparable across languages. However, your backward model should be as strong as possible for BT to work properly.

Q: Larger models or synthetic data requires long training time, but colab has limited usage?

A: The sample code saves model checkpoints each epoch, see next page.

Save checkpoints and data to drive

1. Mount your drive by clicking 
2. Save your preprocessed DATA to your drive

```
!mkdir -p /content/drive/MyDrive/ML2021-hw5/DATA
!cp -r ./DATA /content/drive/MyDrive/ML2021-hw5/DATA
```

3. Change checkpoint directory (under config) to your drive

```
savedir = "/content/drive/MyDrive/ML2021-hw5/checkpoints/transformer-back",
```

Next time, load preprocessed data quicky with

```
!cp -r /content/drive/MyDrive/ML-BTtest/DATA ./DATA
```

Change resume (under config) to following to resume from checkpoint.

```
resume='checkpoint_last.pt'
```