

# Neural Machine Translation

英文翻譯中文

# 目標

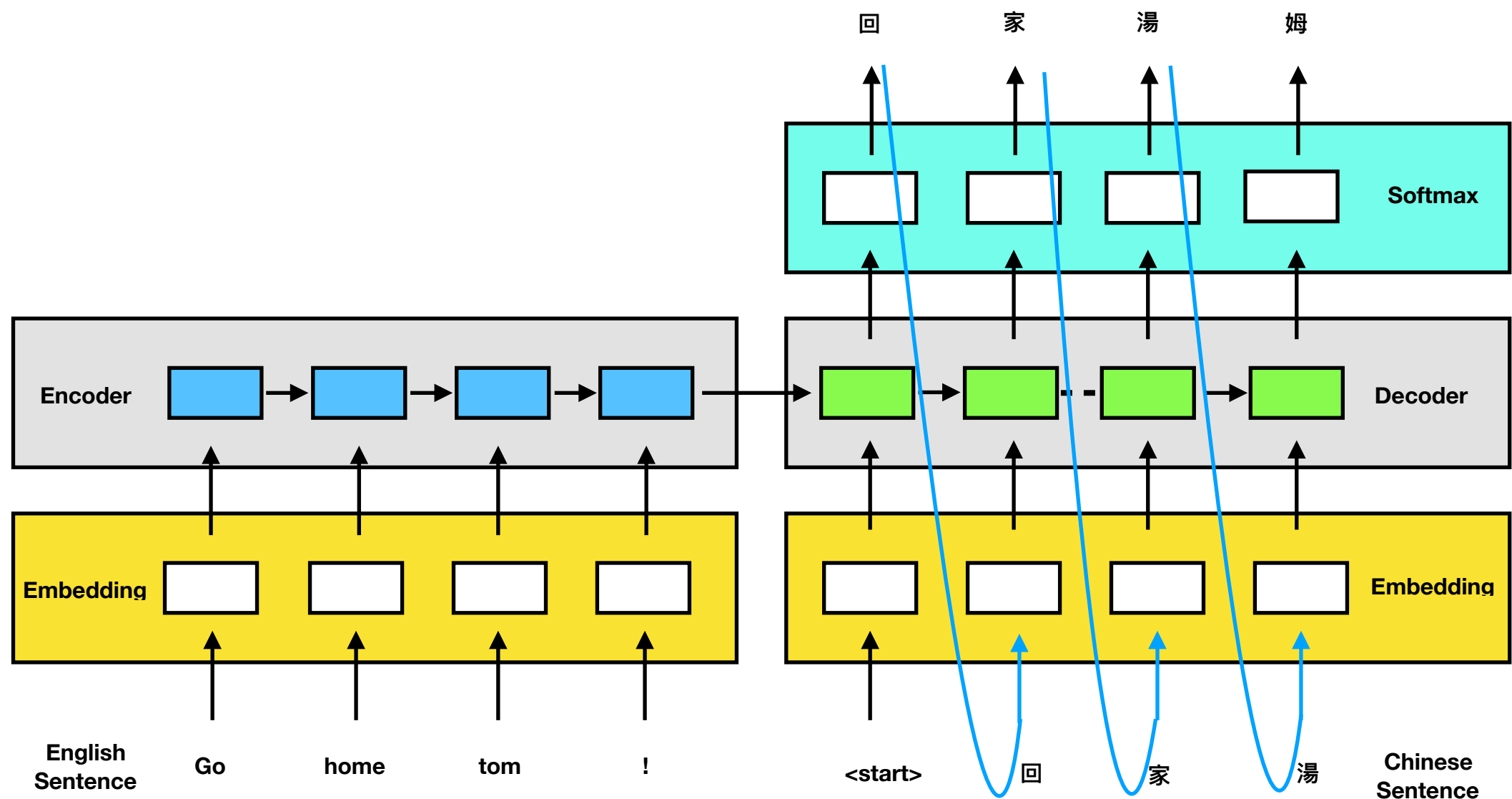
- 將character-based neural network 改成word-based neural network。
- 英文的 word embedding 使用pre-trained 的word embedding。



# Starter Code

- Character-based NMT
  - [https://github.com/keras-team/keras/blob/master/examples/lstm\\_seq2seq.py](https://github.com/keras-team/keras/blob/master/examples/lstm_seq2seq.py)
- Keras pre-trained word embedding
  - <https://blog.keras.io/using-pre-trained-word-embeddings-in-a-keras-model.html>

# Neural Network結構



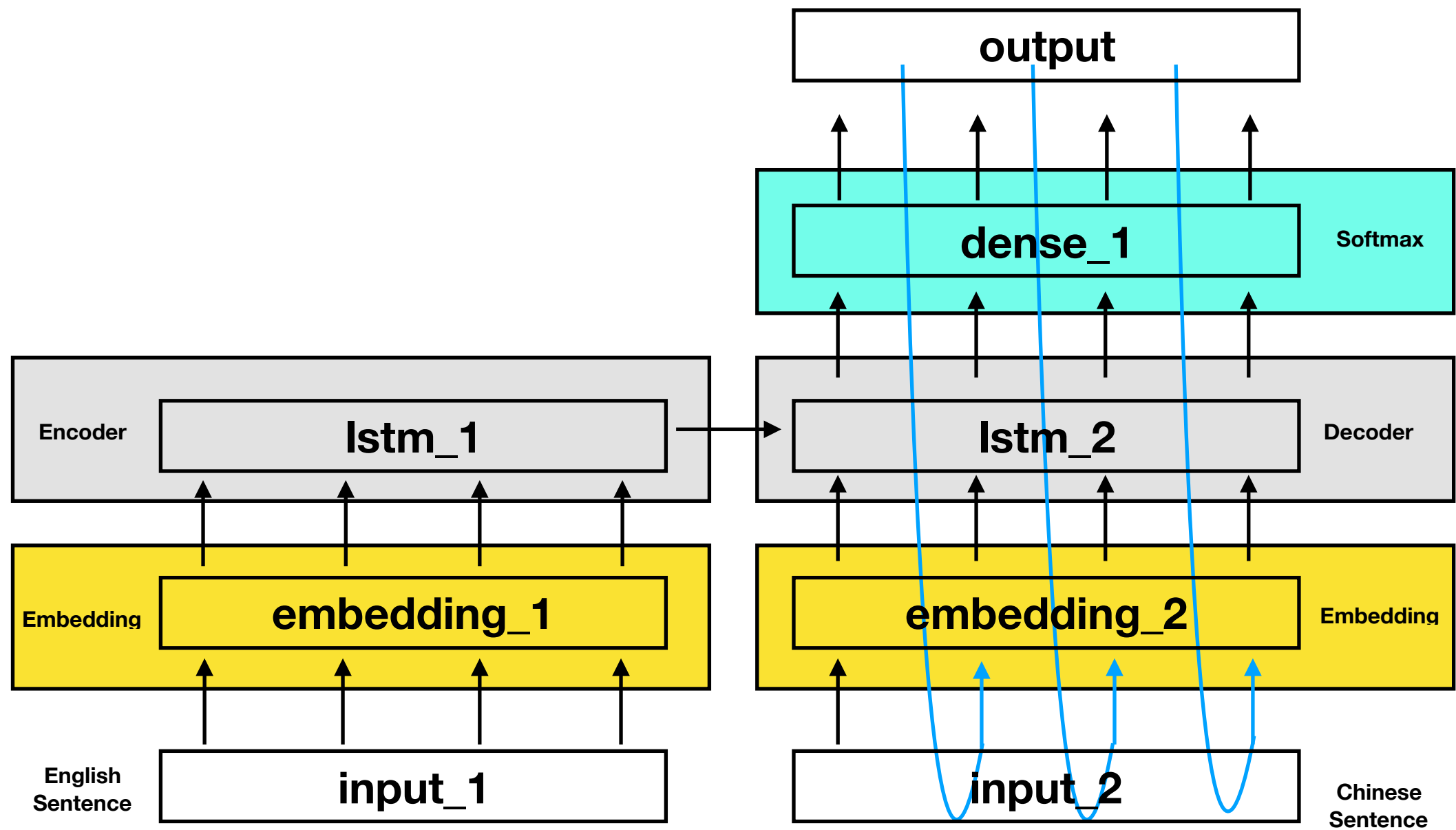
# Keras Layers

- Input
- Embedding
- LSTM
- Dense

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, None)	0	
input_2 (InputLayer)	(None, None)	0	
embedding_1 (Embedding)	(None, None, 100)	1390200	input_1[0][0]
embedding_2 (Embedding)	(None, None, 100)	2731200	input_2[0][0]
lstm_1 (LSTM)	[(None, 100), (None, 80400		embedding_1[0][0]
lstm_2 (LSTM)	[(None, None, 100), 80400		embedding_2[0][0] lstm_1[0][1] lstm_1[0][2]
dense_1 (Dense)	(None, None, 27312)	2758512	lstm_2[0][0]

**NMT model summary**

# Keras Model



# Dataset

- <http://www.manythings.org/anki/cmn-eng.zip>

英文 <tab> 中文

Hi.	嗨。
Hi.	你好。
Run.	你用跑的。
Wait!	等等！
Hello!	你好。
I try.	让我来。
I won!	我赢了。
Oh no!	不会吧。
Cheers!	乾杯！
He ran.	他跑了。

# Tokenize

- jieba(chinese): <https://github.com/fxsjy/jieba>
- nltk(english)

["他害怕蛇。"] → ["他", "害怕", "蛇", "。"] → [10, 17, 23, 4]

word	Index
。	4
他	10
害怕	17
蛇	23

word2index dictionary



# Tokens

- <start> : 句子的開始
- <end> : 句子的結束
- <pad> : 維持句子長度一致
- <unk> : unknown word

word	Index
<pad>	0
<start>	1
<end>	2
<unk>	3

Encoder Sentence(input) [10, 17, 23, 4]  $\longrightarrow$  [10, 17, 23, 4, 0, 0, 0, 0, 0]

加<end> token在最後

Decoder Sentence(input) [5, 18, 38, 40, 44, 4]  $\longrightarrow$  [1, 5, 18, 38, 40, 44, 4, 0, 0]

加<start> token在開始

Decoder Sentence(output) [5, 18, 38, 40, 44, 4]  $\longrightarrow$  [5, 18, 38, 40, 44, 4, 2, 0, 0]

訓練資料最大長度為 9的情況

# Embedding

- `Embedding(input_dim, output_dim, weights=[embedding_matrix], trainable=[True | False])`
  - `input_dim` : vocabulary size
  - `output_dim` : output size, 文字向量的維度。
  - `weights` : pre-trained weights
  - `trainable` : 如果是False, freeze the layer.
- 如果字為<unk>或者是不在pre-trained word embedding中，word vector為0。

# LSTM

- LSTM(units, return\_sequences=[True | False], return\_state=[True | False])
  - units : hidden dimension (output size)
  - return\_sequences : 如果是True，回傳全部字的output，反之，回傳最後字的output。
  - return\_state : 是否回傳最後一個cell state。

# Dense

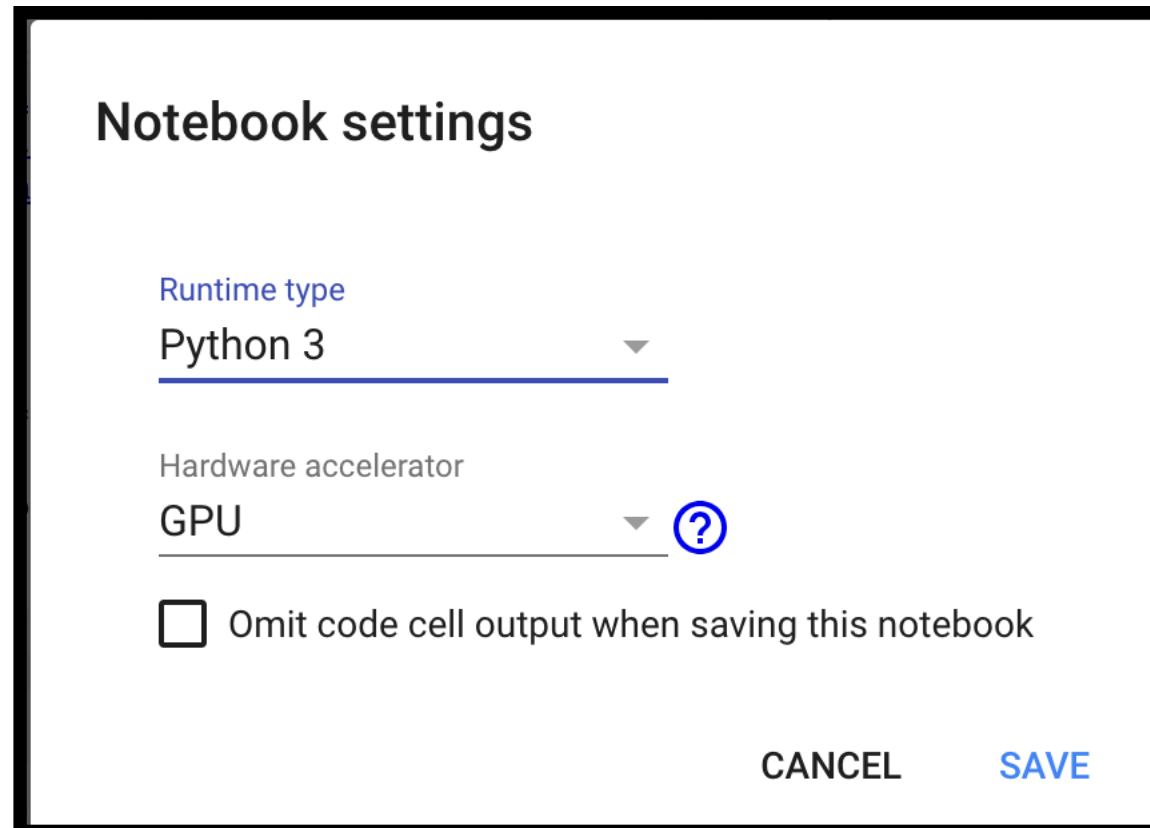
- Dense(units, activation=<activations>)
  - units : output size (在這次作業中，等於中文vocabulary size)
  - activation : 請用softmax

# Colab

- 沒有gpu的同學可以使用colab
- <https://colab.research.google.com/>

# Runtime

- 選擇使用CPU,GPU, TPU去跑你的程式。
- 務必用GPU，使用方法Runtime >> Change runtime type，下面的小視窗Hardware accelerator選擇用GPU。



Notebook settings

Runtime type  
Python 3

Hardware accelerator  
GPU

☐ Omit code cell output when saving this notebook

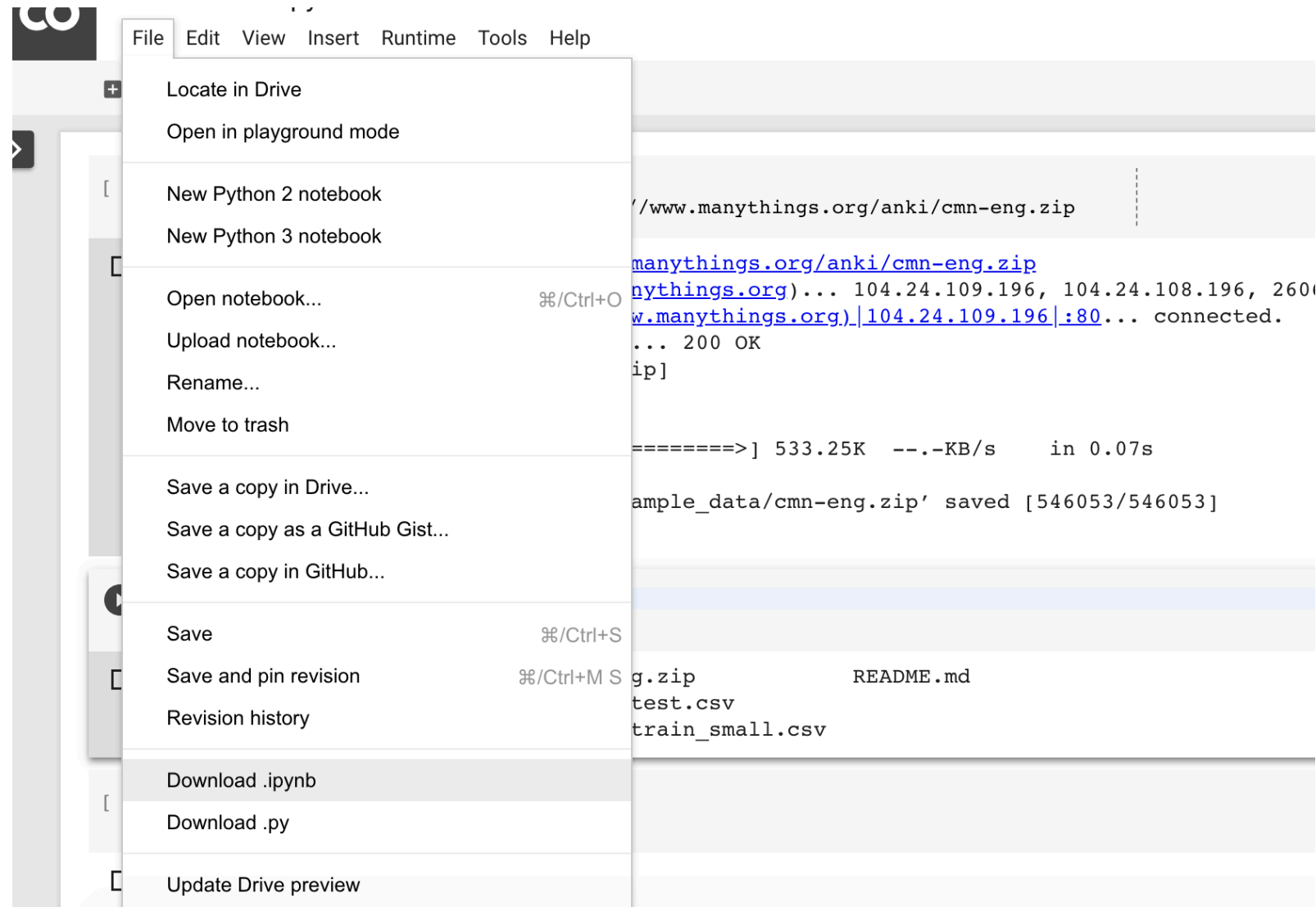
CANCEL SAVE

# Unix in Colab

- 在Colab的cell最前面加上！符號，可以使用Unix指令。
- 好用的Unix 指令
  - head, tail ,less, cat, grep: 查看文本
  - ls, find, mv, cp, rm, touch: 資料夾狀態和移動刪除檔案
  - nvidia-smi, nvcc: gpu 狀態, gpu toolkit

# Download ipynb

- File >> Download .ipynb





# 作業評分

- Preprocessing(padding & tokenize) (10%)
- Word-based encoder & Word-based decoder (20%)
- Embedding layer in encoder & Embedding layer in decoder (20%)
- Pre-trained word embedding (25%)
- Softmax layer (10%)
- Comment & Code readability (5%)
- Output.txt (10%)

# 作業繳交

- lab3\_<學號>.ipynb, output.txt
- 使用全部資料，建議train 50個epoch。
- 依順序輸出第4077, 2122, 3335, 1464, 8956, 7168, 3490, 4495, 5100, 119行到output.txt，第4077例子如下。

**Input sentence<tab>He is afraid of snakes.**

**Decoded sentence<tab>他害怕蛇。<end>**

- 翻譯的結果必須有一定成效，如果結果太差或自行修改output.txt，以零分算。

# TA's output

---

-  
Input sentence: he is afraid of snakes .  
Decoded sentence: 他 害怕 蛇 。 <end>  
-  
Input sentence: i miss you so much .  
Decoded sentence: 我 如此 想念 你 。 <end>  
-  
Input sentence: we 're going by train .  
Decoded sentence: 我们 要 乘火车 去 。 <end>  
-  
Input sentence: the sky is clear .  
Decoded sentence: 天空 很 晴朗 。 <end>  
-  
Input sentence: wearing a suit , he stood out .  
Decoded sentence: 他 穿著 西 裝 站 了 出 來 。 <end>  
-  
Input sentence: she made a serious mistake .  
Decoded sentence: 她 犯 了 一 個 嚴 重 的 錯 <end>  
-  
Input sentence: have you eaten dinner ?  
Decoded sentence: 你 吃 晚 飯 了 嗎 ? <end>  
-  
Input sentence: what do you want to be ?  
Decoded sentence: 你 想 成 为 什 么 ? <end>  
-  
Input sentence: tom is going to help us .  
Decoded sentence: 汤姆 要 帮助 我们 。 <end>  
-  
Input sentence: he 's lazy .  
Decoded sentence: 他 很 懒 。 <end>

# 作業繳交

- 繳交截止12/18
- 遲交一週扣10分，最多扣40分。