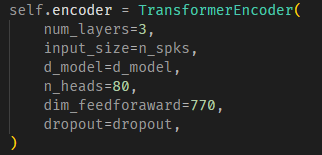
2023 Fall Deep Learning Lab3 Report

109651066林柏佑

**Screenshot of task1 (Transformer)**

* + Number of layers：3



* + Parameter size：497.43K



* + Accuracy：0.7781

**Screenshot of task2 (Conformer)**

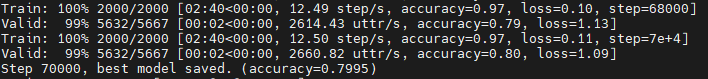
* + Number of layers：3



* + Parameter size：448.504K



* + Accuracy：0.7995



**In task2**

* + Which kind of transformer-like model do you choose？
  + The reason why you choose this model.
  + The advantage of chosen model.

我會選擇Conformer。因為Transformer是基於Self-attention設計，在針對大範圍前後有相關的特徵資訊，雖有較好的效果，但缺乏局部細微的特徵。而CNN提取局部細微特徵的效果非常好。而正好Conformer的架構是將Self-attention和Convolution layer這兩著結合，各自擷取其優點。

**Anything you do to improve the performance**

我認為在ADD和Norm那層，可以改成先做Norm再做Residual，效果上會有提升！而這個想法在2020年也有學者證實，並發表成論文。如下：

參考論文：[On Layer Normalization in the Transformer Architecture](https://arxiv.org/abs/2002.04745)

**Screenshot of your transformer code for both encoder layer and encoder**

* + **Plagiarism is forbidden !!!**
  + **There should be comment in your code**
  + **For both Task1 and Task2**

Task1：



Task2：





