Homework Assignment 3

- The one with the Transformer -

1. Description

The homework assignment focuses on getting some basic practical experience in working with transformer models (e.g. evaluation, pre-training, fine-tuning) using models, support code and tutorials from a popular NLP framework: Hugging Face.

In particular, the task will entail fine tuning a T5 model (link to paper) using two of the original training tasks of the model (span corruption and question answering) and observing *the degree to which* the model is able to *obtain knowledge* from an unsupervised task (span correction) and use it for another (question answering).

The objective is to evaluate the model on a small dataset of questions asking for the birthplace of known personalities.

2. Task List

The dataset we are using for this task consists of three files.

- `wiki.txt`: contains sentences extracted from Wikipedia, describing the birth date, birth place and known occupations of a person. It contains details about **one person per line.** The format of each line is:
  - **<person name> . <person details> \n**
  - **Example:** "`Khatchig Mouradian. Khatchig Mouradian is a journalist, writer and translator born in Lebanon .`"
- `birth_places_train.tsv, birth_places_test.tsv`: are tab-separated files containing a question about the birthplace of a person and the correct answer.

Task 1 [6pt]

Evaluate a pre-trained T5-small model on the question birth_places_test.tsv file. Compute the accuracy metric.

The question answering task is set up in a manner similar to the one from SQuAD, where the model is given a *question* and a *context* and is required to provide the answer from the context.

To build a (question, context, answer) dataset for evaluation you will have to **match** the *name* at the beginning of the `wiki.txt` dataset, with the question in which it appears from the `birth_places_test.tsv` file. This can be achieved using a simple substring matching call.

To set up the ***evaluation dataset*** structure in PyTorch, take inspiration from this repository, in the MyDataset.py modue.
To set up the evaluation procedure, take inspiration from the same repository by looking at the evaluation_t5.py module.

Task 2 [4 pt]

Fine tune the pre-trained T5-small model on the birth place type of questions and recompute the accuracy metric after this fine tuning procedure.
The fine tuning is with respect to a *question-answering* objective, as in the original T5 paper, where the question and context constitute the encoder input, and the answer is the decoder output.

Create a ***training dataset*** using the same approach as in Task 1, but this time using the birth_places_train.tsv file as source for *question* and *answer*, alongside `wiki.txt` for the *context*.

To set up the fine tuning procedure for this task take inspiration from the train_t5_selfrc.py module.

Notes on fine tuning procedure:

- Start with the default parameters used in train_t5_selfrc.py module.
- Take note of the *T5 fine tuning tips* from this Hugging Face discussion thread
- Use **at most 2h** of TPU training time on Google Colab for the fine-tuning of your model