Assignment 1

Please submit the code and report (in pdf format) on Blackboard system before 23:59 Oct. 24. Report can be written either in English or Chinese. Name your report as `studentID_Name.pdf`.

Q1

Tokenization is an important step in modern NLP pipelines. In this task,

(1) [3 points] Explain the Byte Pair Encoding (BPE) algorithm. Then illustrate with code to show how BPE works.
(2) [2 points] Write code to train a BPE model (vocabulary size = 10000) and apply the tokenizer to the dataset. You can use the Subword-NMT toolkit. Please refer to link for the training data. You need to write python code to extract the data from the `conversations` data in the jsonl file and follow the ChatML format.

Note you can refer to the following code MinBPE and Subword-NMT, as well as the original paper Byte pair encoding. You can also ask for help from GPTs, however, you are responsible for checking the correctness of the responses.

Q2

[3 points] Complete the missing parts of `transformer_model.py`. Search for `TODO` in the code. There are 5 `TODOs` that correspond to key functions. Note that you can ask for help from GPTs, however, you are responsible for checking the correctness of the responses.

Q3

ImageBind-LLM: Multi-modality Instruction Tuning is a paper about large language model. Click 'Download PDF' button to read the pdf. The Figure-3 shows a bind network which consists of some feedforward network.

[2 points] Write the pytorch code for the bind network, the input is image_feature and output is transformered_image_feature. Define the `__init__()` and `forward()` function. Note that the bind network has three blocks ($\times 3$). You can refer to the source code here.