CS 584-A: Natural Language Processing

Homework 4

Due Date: Wednesday *11:59 AM (noon)* EST, Dec 6, 2023 **Total: 100 points**

Goals

The goal of HW4 is for you to get hands-on experience of utilizing pre-trained and large language model for the machine reading comprehension task. The questions are open questions and there is no fixed solution. Different data selection and processing, parameter initialization, data split strategies will lead to different predictions and results. Therefore, during the grading, the specific values in the results are not required. It is important that you focus on implementing and setting up the pipelines of applying these models to solve the tasks.

Task: Machine Reading Comprehension

1. **Data preparation (10 points)**
   1) **(5 points) Data download and preprocessing**: Download the Squad2 dataset, which is a benchmark data for question answering. Please select **15 question-answer pairs** for this homework. Each data sample will be a question-answer pair, which includes the following three primary components:
      a) **Context**: a paragraph about the context of the question.
      b) **Question**: a natural language question that needs to be answered based on the context.
      c) **Answer**: the answer to the question, which is a text span extracted from the context.

      You can directly download the data with the following codes from Hugging Face. You will need to install 🤗 Transformers and 🤗 Datasets first. A quick tour to set up and run Transformers on Hugging Face can be found here.

      ```
      from datasets import load_dataset
      dataset = load_dataset("squad_v2")
      ```

   2) **(5 points) Data statistics:** based on the data you selected, get basic data statistics about the following aspects.
      a) Statistics about the max/average/min length of the contexts
      b) Statistics about the max/average/min length of the questions
      c) Statistics about the max/average/min length of the answer

2. **Question answering (60 points, 20 points for each model)**
   1) Please use the **pre-trained RoBERTa model** to solve the question answering task on the data you prepared.

a) Implementation:
   - Link to the pretrained roberta-base-squad2 model along with instructions and examples: link
   - More details about how to use RoBERTa for question answering: link
   - Feel free to directly utilize the existing implementations on Hugging Face.
   - There is no specific requirement for the parameter settings. You are encouraged to try and test different settings and report the results. For other settings that are not specified here, you have the flexibility to select.

b) **Requirements about the output**. Please follow the examples provided in this link, and print out the following output.
   - The answer starts index and answer end index for each data sample.
   - Predicted answer tokens, which will be the token index in the vocabulary.
   - Decoded answers, which will be the predicted text spans from the text.

2) Please use design the prompts to interact with **ChatGPT** and solve the question answering task on the data you prepared.
   a) You can use the OpenAI API if you have access, or you can use the user interface of ChatGPT to work on this question.
   b) You can design your prompt accordingly. Feel free to explore different ways to design prompts. You can use the same prompts for all the 15 data samples, or you can explore different prompts for different data samples. Please explicitly mention your designed prompts in your submission.
   c) You will need to provide the instructions in the prompt and ask the model to output the answer to the question and explain the reason.
3) Like the requirements in 2), please design prompts to interact with **Claude 2** and solve the question answering task on the data you prepared.
   a) For each data sample, please use the same prompts as you designed in 2).

3. Results analysis and evaluation (30 points)

1) **(10 points)** Please manually check if the output answer for each data sample is correct or not. You can also compare the output answers with the ground truth answer to each question.
   a) The answers do not need to be exactly the same as the ground truth answer. It can be a correct answer if there is only a partial matching, but with the same meaning. You can decide based on your understanding.
   b) Based on your manual checking and decision, please calculate a general score, i.e., # of data samples provided correct answers/15, for each of the three models.
2) **(10 points)** Provide comparison discussions and analysis based on the output and evaluation results you obtained from each model. For example, which model performs best, which performs worst, and what is the possible reason for such results. Based on

this question, what is your experience about the advantages and disadvantages of the pre-trained language models and the large language models?

3) **(10 points)** Select two specific data samples and compare the returned answers by the three models and compare the answers with the ground truth answers. Provide some discussions based on your findings on these two specific examples.

Submission guidelines:

1. In your codes, please try to add ***clear comments and detailed steps*** to get all the partial points. Also, include the required discussions and results analysis in the text cell.
2. Please use Jupyter Notebook for your programming. Before submission, please make sure to clear all existing outputs and rerun all the codes. Once it is done to run all the codes, save it as a pdf file. You will need to ***submit both the .ipynb and .pdf files***. Both files should include all the codes, comments, discussions, and answers for the questions.
3. You can also organize all the results and analysis in a separate pdf file. This is optional.
4. Submissions should be made on ***Canvas***.
5. *Each student needs to submit their own answers and codes. Similar or the same answers/codes will not be graded. You will directly fail the course if cheating/plagiarism is found.*
6. ***Properly cite any resources*** you used for your submission, including but not limited to websites, GitHub repositories, publications, generative AI techniques, etc.