Report of assignment 3

Introduction

large language models (LLMs) have marked a significant milestone in natural language processing (NLP). In this assignment, we will use a large language model to solve the multiple-choice questions. By inputting the question and all the choices, we require the model to output the current answer and the reason why the answer is correct. The dataset contains 1119 training multiple-choice questions, 299 validation questions, and 1172 test questions.

Method

The backbone structure of our model is llama with 1 billion parameters. Due to the complexity of tuning all its parameters, we try to finetune the model by adding a Majority Voting technique to require the model to predict several times for one question and choose the most occurrence one to be the correct answer to the question. In the finetuning procedure, we mainly focus on the num_forward_passes (the number of retries in the Majority Voting process) and temperature (pretraining temperature for Tensor Parallelism).

Result

Model with **temperature 1** and **forward passes 8 (default) ------ Acc: 49.77%**

Model with **temperature 1** and **forward passes 4 ------ Acc: 48.21%**

Model with **temperature 1** and **forward passes 16 ------ Acc: 50.89%**

Model with **temperature 5** and **forward passes 16 (default) ------ Acc: 49.49%**

Model with **temperature 5** and **forward passes 8 ------ Acc: 48.81%**

Model with **temperature 5** and **forward passes 4 ------ Acc: 44.97%**

Model with **temperature 20** and **forward passes 8 ------ Acc: 48.81%**

Analysis

- Based on the results, we can find llama 1b can learn multiple questions and generate the correct answer to them to some degree (random guess accuracy rate is 25%, all the accuracy above is much higher this baseline).
- The number of forward passes in the majority voting process can be a great help in improving the accuracy of predictions, by adding the forward passes from 4 to 16, we can see the accuracy in predictions of the validation set increases from 48.21% to 50.89%. This result indicates that the majority voting technique is useful in improving the generation of large language models, especially in classification or question-answering tasks. The reason for this improvement is the retry procedure increases the robustness of the model, rather than predicting one answer, predicting multiple answers to one question can increase the choosing rate of the answer with the highest probability to the question. But I think there will be an upper boundary of this improvement since it can only approximate to predict the answer with the highest probability and it cannot improve the model's parameters.

- However, increasing the number of forward passes will tremendously increase the running time of the prediction process in linear. The experiment with 4 forward passes has taken 2 hours to finish the whole prediction process, meanwhile, the experiment with 16 forward passes has taken nearly 7 hours to finish. The reason is simple, the time will increase as the model increases the trials of prediction, and the major voting procedure increases the time to do the prediction of one question.
- Temperature will influence the prediction, and the influence is parallel to the impact of majority voting. From several experiments related to the temperature, we can find that when the temperature is 1, the prediction is the most accurate, however, the accuracy of prediction has insignificant change when increasing the temperature (acc 48.81% to both the 5 and 20 cases).

Final test

We finally adopt the model with **temperature 1** and **forward passes 16** to do the prediction on the test set, and the final accuracy is: **50.68%.** The predictions on the test set are stored in the file named model_prediction.jsonl