CSCI 420 Assignment 2

Lynelle Chen
lschen01@wm.edu

Alex Zhou
ajzhou02@wm.edu

1. Dataset Preparation

With the provided datasets for training, validating, and testing, we fine-tuned the CodeT5 model using the first two. After that, we used the test dataset to evaluate the model. Processing for the dataset included removing whitespace and using a pre-trained tokenizer to remove the targeted if-statements and replace them with a "<mask>" token. The "<mask>" token would represent the entire if statement, meaning that the original starting "if" and ending ":" would be encompassed by it. Most of this process was completed in our `data_processing.py` file.

2. Fine-tuning Process

After masking and tokenizing the data, we fine-tuned two different models, one with ten epochs labeled model 1 and one with five epochs labeled model 2. We used the provided settings in the lab, making no significant changes. One minor change was simply to adjust logging. This process of fine-tuning the model(s) was done using the `model.py` file. We round both the loss and the evaluation loss to four decimal places.

Model 1 Epoch	Model 1 Loss	Model 1 Eval Loss	Model 2 Loss	Model 2 Eval Loss
1	0.0236	0.0134	0.0236	0.0132
2	0.0151	0.0130	0.0155	0.0121
3	0.0140	0.0127	0.0132	0.0116
4	0.0128	0.0125	0.0116	0.0112
5	0.0115	0.0123	0.0102	0.0109
6	0.0095	0.0118	N/A	N/A
7	0.0103	0.0120	N/A	N/A
8	0.0084	0.0115	N/A	N/A
9	0.0093	0.0114	N/A	N/A

10	0.0094	0.0113	N/A	N/A

As shown in the table above, although both were trained on the same settings, there is still some variance in the loss and validation loss. Over all epochs, although some were locally worse then the previous epoch, there is still a general trend downwards across all epochs.

3. Evaluation Results

We evaluated the 10 epoch model on the following metrics: BLEU-4, CodeBLEU, and exact match. This process, including the creation of the final dataset of test cases and answers, was done with the `evaluation.py` script.

BLEU-4	CodeBLEU	Exact match
0.353676	0.271895	0.0482

As expected exact match, a very stringent evaluation for this type of problem, has a very low score. BLEU-4 has a decent score of around 0.35, matching examples from lecture and also is as expected higher than the CodeBLEU metric being around 0.27.