Assignment2 Report

Name: 盧思樺

Institution (school): NTHU

Student ID: 110062272

Platform (Colab/Kaggle/Local): Colab

Python version: Colab

Operating system: Colab

CPU:

GPU: T4 GPU

Remark: For ease of grading, you are encouraged to present data in textual form rather than as images.

Present your hyper-parameters in training, including learning rate, batch size, hidden size, epochs(steps), etc. (5%)

Answer:

- Batch size: 64
- Epochs: 2
- Embedding dimension: 256
- Hidden dimension: 256
- Learning rate: 0.001
- Gradient clipping: 1

If you use RNN or GRU instead of LSTM, what will happen to the quality of your answer generation? Why? (10%)

Answer:

In this assignment, I experimented with both LSTM and RNN models. The RNN performed worse because it has weaker memory capability and is more prone to the vanishing gradient problem. In contrast, the LSTM can better retain long-term

dependencies, leading to a much higher accuracy.

If we construct an training set using three-digit numbers while the evaluation set is constructed from two-digit numbers, what will happen to the quality of your answer generation? (10%)

Answer:

If the training set contains only three-digit numbers while the evaluation set contains two-digit numbers, the model's performance will likely drop significantly. This is because larger numbers require the model to remember and reason over longer digit sequences. When the model has never seen two-digit arithmetic during training, it cannot generalize well to shorter inputs. As a result, the generated answers in evaluation will become less accurate and more inconsistent.

If we construct a training set that includes 20% incorrect answers, how will this affect the quality of the generated responses? Present some examples. (10%)

Answer:

If 20% of the training data contains wrong answers, the model receives conflicting supervision signals, making it harder to learn correct arithmetic rules. As a result, both training and evaluation accuracy decrease, and the generated outputs become less stable and less logically consistent.

For example:

3 + 5 = correct model outputs 8, noisy model may output 9.

(10 - 4) * 2 = correct model outputs 12, noisy model may output 8.

In summary, label noise reduces the model's ability to generalize and increases the probability of producing incorrect or inconsistent results.

Why do we need gradient clipping during training? (5%)

Answer:

We need gradient clipping to avoid situations where the gradients explode during training. It also helps the model converge more effectively.

… Anything that can strengthen your report. (5%)

Answer:

None.

The screenshot of your training logs and evaluation accuracy. (One Figure only) (10%)

```
Train epoch 1: 100%|██████████| 37020/37020 [06:36<00:00, 93.42it/s, loss=0.312]
0.592429249762583
Train epoch 2: 100%|██████████| 37020/37020 [06:34<00:00, 93.86it/s, loss=0.291]
0.6741690408357075
```