DL 2023040165 Zhicheng Jiang

Assignment 6

6.1

For simplicity, we denote $u^\theta(w,h) = u$, and $q(w), q(\bar{w}) = q$

Then,

$$\nabla L_{NCE} = \sum_w \nabla(\tilde{p}(w|h)\log\frac{u}{u+kq} + kq\log\frac{kq}{u+kq})$$

$$= \sum_w [\tilde{p}(w|h)\nabla\log\frac{u}{u+kq} + kq\nabla\log\frac{kq}{u+kq}]$$

$$= \sum_w [\tilde{p}(w|h)\cdot\frac{u+kq}{u}\cdot\frac{kq\nabla u}{(u+kq)^2} - kq\frac{u+kq}{kq}\frac{-kq\nabla u}{(u+kq)^2}]$$

$$= \sum_w [\tilde{p}(w|h)\cdot\frac{kq\nabla u}{u(u+kq)} + \frac{kq\nabla u}{u+kq}]$$

$$= \frac{kq}{u+kq}\sum_w((\tilde{p}(w|h)-u)\frac{\nabla u}{u})$$

$$= \frac{kq}{u+kq}\sum_w((\tilde{p}(w|h)-u)\nabla(\log u))$$

$$\approx \sum_w((\tilde{p}(w|h)-p^\theta(w|h))\nabla(\log u))$$

$$= \nabla L_{MLE}$$

The "$\approx$" sign achieves when$k \to \infty$, and $p^\theta(w|h) = u$

6.2

Problem 1

1. The most computationally expensive part of a vannila transformer is the self-attention mechanism, which has a time complexity of $O(n^2d)$, where n is the length of the sequence and d is the dimension of the input.

2. We can restrict the context window to a fixed size, and only attend to the tokens within the window. This can reduce the time complexity to $O(nwd)$, where w is the window size.

pseudo code:

```
for i in range(n):
```

```
for j in range(max(0, i-w), min(n, i+w)):
# calculate the attention score between i and j
```

Problem 2

For sentiment analysis, I would suggest Alice choose BERT. This is because BERT is a bidirectional transformer model, which can capture the context information of the input sequence, and in sentiment analysis, the context information is important.

Fine-tuning procedure:

1. Load the pre-trained BERT model.
2. Add a classification layer on top of the BERT model, based on the task description.
3. Re-train the BERT model on the sentiment analysis dataset using the classification layer.

For closed-book question answering, I would suggest Alice choose GPT-2. This is because GPT-2 is a generative transformer model, which can generate text based on the input sequence. In closed-book question answering, the model needs to generate the answer based on the input question without accessing any external knowledge.

Fine-tuning procedure:

1. Load the pre-trained GPT-2 model.
2. Re-train the GPT-2 model on the closed-book question answering dataset.