Deep Learning 2023040163 Zhao Han Hong

Homework 6

Homework 6.1

Problem 1 We can directly calculate

$$\begin{aligned}\lim_{k\to\infty} \nabla L_{\text{NCE}}^{(k)}(\theta;h) &= \lim_{k\to\infty}\Bigg[\sum_w \tilde{p}_{w|h}(w|h)\left(\frac{1}{u^\theta(w,h)} - \frac{1}{u^\theta(w,h)+kq_{\tilde{w}}(w)}\right)\cdot\nabla u^\theta(w,h)\\ &\qquad + \sum_{1\le i\le k,\bar{w}} q_{\tilde{w}}(\bar{w})\left(-\frac{1}{kq_{\tilde{w}}(\bar{w})}\right)\cdot\nabla u^\theta(\bar{w},h)\Bigg]\\ &= \sum_w \tilde{p}_{w|h}(w|h)\left(\frac{1}{u^\theta(w,h)}\right)\cdot\nabla u^\theta(w,h) - \sum_{\bar{w}}\nabla u^\theta(\bar{w},h)\\ &\approx \sum_w \left(\tilde{p}_{w|h}(w|h) - p^\theta_{w|h}(w|h)\right)\nabla\log u^\theta(w,h),\end{aligned}$$

where the final step uses the fact that the partion function Z is approximated to 1.

Homework 6.2

Problem 1

1. It is the self-attention and cross-attention parts, since the attention calculation has complexity $O(n^2)$, where n is the input sequence length.

2. The paper proposes to split the sequence into segments. Then, attention is only fully calculated within the segment; however, the hidden states in the previous segment is added in as a context without gradient. The training pseudo-code can be shown as below:

```
Split the sequence into segments
for each layer number $n$ do
    for each segment number $t$ do
        Concatenate the hidden states of the previous segment (but without gradient) NoGrad($h_{t-1}^{(n-1)}$) with the current segment's hidden state $h_t^{(n-1)}$
        Calculate the attention query based on solely $h_t^{n-1}$, but calculate the key and value based on the concatenated hidden states.
        Use attention mechanism to calculate the output $h_t^{(n)}$, which is the hidden state
        of the current segment at layer $n$.
    end for
end for
```

Problem 2

For the sentiment analysis task, we should use BERT, since it is a pretrained transformer encoder. BERT can extract features of the text bidirectionaly, so it can perform better on the task. For fine-tuning, we should add a MLP projection head on the top of the output hidden states of BERT and try to learn the sentiment from the output.

For the closed-book question answering task, we should use GPT-2. GPT-2 is a pretrained transformer decoder, which can generate text based on the context. For the fine-tuning, we can use the context as the input and the question as the output, and train the model on the corpus just as training a autoregressive language model.