

# Homework 6.2

Deep Learning 2024 Spring

Due on 2024/4/29

## 1 Q&A

**Problem 1.** (Transformer) Consider a vanilla transformer proposed in [1].

1. Which block is the most computationally expensive part of a vanilla transformer?
2. In practice, training a vanilla transformer requires a much larger GPU memory compared to RNN. We have learned in the lecture that we can reduce the computation cost of RNN via truncated BPTT. How can we utilize this practical trick for transformers? Try to design an algorithm and write down the pseudo-code.

**Hint:** You can read this paper.

**Problem 2.** (Pre-training and Fine-tuning) Consider two NLP tasks: *sentiment analysis* and *closed-book question answering*.

Sentiment analysis aims to systematically identify and study affective states and subjective information of natural languages. It usually focuses on the polarity of a text (positive, negative, neutral). For example, given a comment on a film, the NLP model should classify it as “positive”, “negative” or “neutral”.

Closed-book question answering is a task where the model can be asked arbitrary context-independent questions (e.g. well-known facts or historical details) and is not allowed to access any external knowledge whatsoever when answering questions. For example, one may ask “what is the highest mountain in the world”, and the model (actually Google) will answer “Mount Everest’s”.

Alice does not know much about deep learning or NLP, but she wants to utilize a pre-trained language model to perform the above two tasks. If the only two available pre-trained models are BERT and GPT-2, for each task, which model would you suggest Alice choose? Please design a fine-tuning procedure for Alice on these tasks respectively.

## References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.