CS283: Assignment 2

February 3, 2022

Due: Sunday February 13th, 11:59pm
Submit by the **blakcboard system**

By turning in this assignment, I agree with the KAUST student protocol and declare that all of this is my own work.

Requirements:

- Put all files in a zip file and the file should be named LastnameFirstname_Assignment2.zip
- It is an individual assignment, independent write up, and submission in your own hand is required for credit.
- The work should be written in a **clear** way
- Late submission will receive credit penalty

1 Problem 1: Transformer Questions (60 points)

The Transformer model comes from the paper "Attention is all you need" [1] and it has achieved a lot of success in Natural Language Processing (NLP) domain. It shows several advantges over previous RNN model such as parallel processing and long-range context dependencies. BERT [2] and GPT-2 [3] are two famous extensions of the Transformer. You are going to answer several questions about your Transformer/BERT/GPT understanding.

The core idea of self-attention operation is the scaled dot-product attention. The features are first transformed into three different matrices Q, K and V, and the self-attention is calculated by the following:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Q1 (10 points): In the above self-attention operation, why we need to incorporate the scale factor $\sqrt{d_k}$ into calculation?

Q2 (10 points): When we train the Transformer on the word sequences, usually we need to add additional positional embedding for each word, why is this necessary?

Q3 (10 points): In the Transformer framework, there are two types of attention modules, which are self-attention and encoder-decoder attention. What is the difference between these two modules in terms of functionality and technical implementation?

Q4 (10 points): There are also other types of attention calculations such as the additive attention [4]. Additive attention computes the compatibility function using a feed-forward network with a single hidden layer. In the Transformer model, why the authors choose to use scaled-dot product attention instead of additive attention and what is the main advantages?

Q5 (10 points) : BERT and GPT model pretrain their model on a large-scale dataset in a self-supervising way. Please describe their pretraining tasks and discuss why it is useful.

Q6 (10 points) : In the BERT model design, there are two special tokens $[CLS]$ and $[SEP]$, what is the purpose of designing these two special tokens and how they are used during the training and evaluation.

2 Coding Implementation (40 points)

In this coding assignment, you need to implement a self-attention module which is the core of the Transformer. In details, you need to implement:

- scaled dot-product self-attention (**20 points**)
- residual connection (**5 points**)
- dropout (**5 points**)
- layer normalization (**5 points**)
- position-wise feed-forward network (**5 points**)

More detail instructions could be found in the ***readme.txt***

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.