Homework Assignment 7

Due Date: 6pm on Tuesday April 11th

1. Reading summary (6 points)

Please summarize the reading material in your own words. This exercise will help you comprehend the main objectives in the reading besides the technical details. Your summary should consist of three parts:

1. One-sentence summary
2. One-paragraph summary
3. Half-page summary

2. Questions (4 points)

Please select **one** question to answer in each of the following four question sets. (A total of four questions.)

2.1 Chapter 10.1.0

- What are the advantages of transformers compared with traditional RNNs such as LSTMs?
- Summarize the key properties of the self-attention mechanism.
- What are the three different roles that each input embedding plays during the course of the attention process?
- Why do transformers use scaled dot-product instead of dot product as the score function?
- How to solve the problem that the calculation of equation 10.10 results in a score for each query value to every key value, even including those that follow the query?
- Why do transformers computationally expensive?

2.2 Chapter 10.1.1

- What layers does a transformer block consist of?
- What are residual connections and why do transformer blocks need them?
- What are the roles of layer normalization in transformers?

2.3 Chapter 10.1.2 - 10.1.3

- What are multihead self-attention layers and what issues have they addressed?
- How to implement the notion of multihead self-attention layers?
- What are the roles of positional embeddings in transformers? The reading material introduces two different types of positional embeddings, what are they and what are their respective drawbacks?

2.4 Open Question (1 point)

- Apart from scaled dot-product, what are other alternatives of score functions and what are their pros and cons compared with scaled dot-product? Can you write down their definitions?
- Apart from the two positional embeddings discussed in the reading material, are there better approaches to represent positional information?