Module 7

Reading and Short Answer Assignment

Assigned: Wednesday, 05/29/2024
Due: Wednesday, 06/05/2024 by 11:59:59 PM

This week you will read a paper that was not discussed in class lecture: the Longformer paper. This paper is one of many that have attempted to make a more efficient attention mechanism.

Please read sections 1 through 4 of the paper:
https://arxiv.org/pdf/2004.05150.pdf

Questions

Please write a short response (4-5 sentences / short paragraph) to each of the following questions / the following question. Your responses will be graded for accuracy, critical thinking, and clarity. You may use any common word processing or text format. Please upload your answers by the due date.

Question 1: Why does the memory complexity of a transformer expand quadratically when the input sequence only expands linearly? How does this limit our ability to build larger and larger models?

Question 2: How does the Longformer try to improve on this complexity?