Gradescope (4 pts)

t11902210 張一凡

1. Make a brief introduction about one of the variants of Transformer, and use an image of the structure of the model to help explain (2 pts)

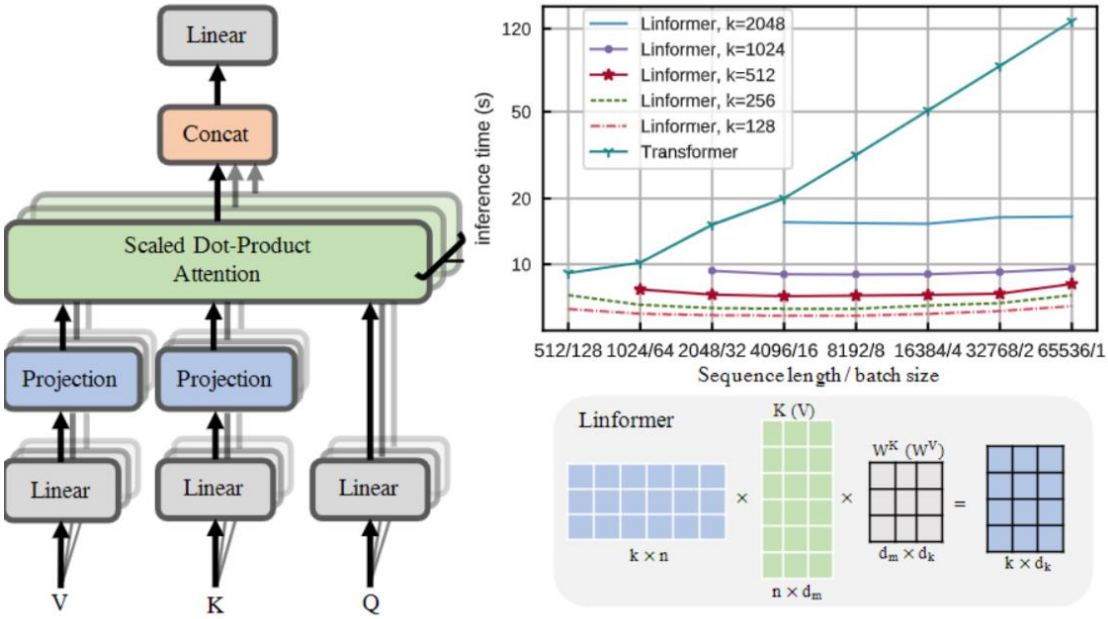
Answer:

Linformer: Self-Attention with Linear Complexity

Linformer is an efficient transformer based on the idea of low-rank self-attention. Its key idea is to map the sequence length dimension to a lower dimensional space of values and keys. The principle is that relevant findings have proven that the random matrix formed by self-attention can be approximated by a low rank matrix. Through this observation, (Wang et al., 2020b) introduced a new mechanism that can reduce self-attention to O (n) operations in both spatial and temporal complexity: by decomposing the original scaled dot product attention into multiple smaller attention operations through linear projection, the combination of these operations forms a low-rank decomposition of the original attention.

The specific method is to propose and implement a new self attention mechanism by using mixed precision training, knowledge extraction, sparse attention, local sensitive attention, and microbatch improving the efficiency of the optimizer. This mechanism allows us to calculate the context mapping matrix P * V * Wv within a linear time complexity, as well as the memory complexity regarding the length of the sequence, which is known as the Linformer model.

The main idea is to add a linear projection matrix to the calculation of Key and Value, as shown in the following figure:



Reference: Linformer: Self-Attention with Linear Complexity

Link: https://blog.csdn.net/ayayayayo/article/details/109693481

2. Briefly explain what're the advantages of this variant under certain situations. (2 pts)

Answer:

In short, Linformer runtime does not increase too quickly as the sequence grows longer. Large Transformer models have achieved very successful and recent results in many natural language processing applications. However, for long sequences, the cost of training and deploying these models can be prohibitive, as Transformer's standard self attention mechanism uses a square complexity of n in time and space relative to sequence length. Therefore, in this article, we attempt to answer the question: Can the Transformer model be optimized to avoid this secondary operation, or does this operation need to maintain strong performance? The writters further utilize this discovery to propose a new self attention mechanism that can reduce the complexity of self attention from the level of n to the level of n in both time and space. The resulting linear transformer, known as Linformer, has comparable performance to the standard transformer model, while providing greater memory and time efficiency.

The following table records the multiple improvements of Linformer in time and space under different sequence lengths and k values in this paper's experiment, with time on the left and space on the right.

length n	projected dimensions k				
	128	256	512	1024	2048
512	1.5x	1.3x	-	-	-
1024	1.7x	1.6x	1.3x	-	-
2048	2.6x	2.4x	2.1x	1.3x	-
4096	3.4x	3.2x	2.8x	2.2x	1.3x
8192	5.5x	5.0x	4.4x	3.5x	2.1x
16384	8.6x	7.8x	7.0x	5.6x	3.3x
32768	13x	12x	11x	8.8x	5.0x
65536	20x	18x	16x	14x	7.9x

length n	projected dimensions k				
	128	256	512	1024	2048
512	1.7x	1.5x	-	-	-
1024	3.0x	2.9x	1.8x	-	-
2048	6.1x	5.6x	3.6x	2.0x	-
4096	14x	13x	8.3x	4.3x	2.3x
8192	28x	26x	17x	8.5x	4.5x
16384	56x	48x	32x	16x	8x
32768	56x	48x	36x	18x	16x
65536	60x	52x	40x	20x	18x