北京大学
PEKING UNIVERSITY

人工智能的硬件基石：从物理器件到计算架构

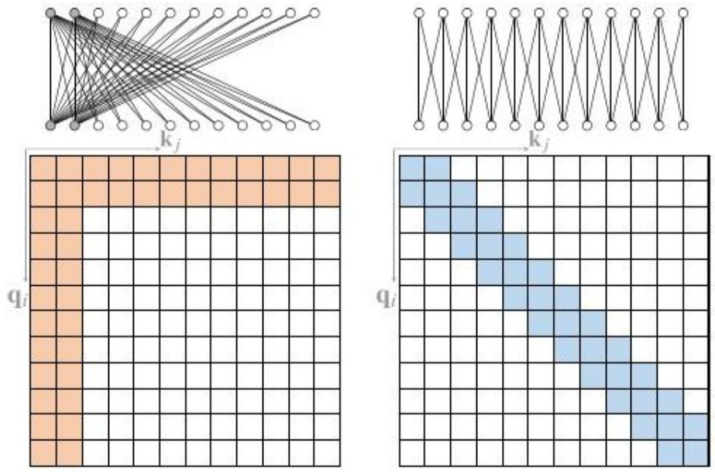
2025年春季 - 作业5

1、针对Transformer的稀疏性存储（60分）

（1）使用CSC按列压缩的方式，对下列稀疏矩阵进行压缩

1				
		3		
	6			
	5			8
	10			

（2）transformer中基于位置的稀疏注意力常用的有如下两种，global attention通过设置全局节点，让所有节点都可以通过全局节点进行信息集散，同时优化掉非全局节点之间的信息连接用于稀疏；band attention面向于特殊文本或图像序列，将注意力限制在局部区域内，更好的聚焦局部上下文与局部特征。假设下面两种原稀疏矩阵大小都是n*n的，给出用csc方式下的存储量。



(a) global　　(b) band

2、KV Cache（40分）

调研KV Cache相关的知识。假设在decode过程中，K和V向量大小为d，注意力头的个数为h，模型层数为l，连续token数为n，在没有KV Cache的情况下计算加载K和V需要的参数加载量；如果存在KV Cache且Cache容量充足，重新计算。若KV Cache容量不足以支撑全部token的K和V缓存存储，思考有什么方法对K和V的加载进行优化。