Topic 8: Attention Mechanism

1. Attention Is All You Need: https://arxiv.org/abs/1706.03762
2. The Illustrated Transformer: A beginner-friendly, visual introduction to the Transformer architecture and its self-attention mechanism: https://jalammar.github.io/illustrated-transformer/
3. Understanding the Self-Attention Mechanism: A short video that clearly explains how the self-attention mechanism works in Transformer models: https://www.youtube.com/watch?v=W28LfOld44Y
4. Multi-Query Attention: Overview of Multi-Query Attention, an attention variant where all heads share one set of key/value pairs to reduce memory usage and speed up decoder inference: https://pub.towardsai.net/multi-query-attention-explained-844dfc4935bf
5. MLA: Introduction to Multi-Head Latent Attention, a technique that compresses the key-value cache via low-rank projections: https://huggingface.co/blog/NormalUhr/mla-explanation
6. Explains what the key-value cache is and why it becomes a memory bottleneck for long sequences: https://huggingface.co/blog/nvidia/kvpress
7. FlashAttention: The original FlashAttention paper that reduces memory reads/writes: https://arxiv.org/abs/2205.14135
8. FlashAttention-2: Follow-up work to FlashAttention that improves parallelism and work partitioning to further speed up attention operations: https://arxiv.org/abs/2307.08691
9. A Gentle Introduction to Multi-Head Attention and Grouped-Query Attention: https://machinelearningmastery.com/a-gentle-introduction-to-multi-head-attention-and-grouped-query-attention/