Preliminary Research Idea (4): Optimizing Attention Heads Using Determinantal Point Process

Professor Richard Yi Da Xu
HKBU and TadReamk Limited

Using DeepSeek's Lightning Selector, we can identify sparse relationships between the current token and all preceding tokens. This can be viewed as filtering relevant information along the token sequence dimension. However, can we also track the similarity of self-attention scores between different attention heads? For example, if the evolution trajectories of Q, K, V across several heads exhibit very similar behavior, we could simply select one of them as a representative. This approach can reduce redundancy and save computational resources from the attention head dimension. Such optimization can be achieved by introducing Determinantal Point Process (DPP) into the mechanism.