Preliminary Research Idea (XIV): Breaking Boundaries: On the Synchronous Mixing of "Inter-Head" and "Intra-Head" Information in Multi-Head Attention Mechanisms

Professor Richard Yi Da Xu
Machine Learning Knowledge Propagator

Introduction

We know that the Multi-Head Attention (MHA) mechanism is a framework in which the internal mixing of matrices Q, K, and V first occurs within each respective head, and only after this internal mixing is complete does "inter-head" mixing take place.

Core Research Idea

This introduces a new idea: Can external mixing and internal mixing occur simultaneously? For example, the simplest approach could be to average the QK^T between two heads (or perhaps find other better methods?).

Analogy

To draw an analogy, traditional MHA is like first conducting "class internal discussions," and then conducting "inter-class discussions." However, if we allow "student exchanges" to occur occasionally, this could help achieve better fusion of information.