Preliminary Research Idea (12): Multimodal Attention Consistency

Professor Xu Yida
Machine Learning Knowledge Disseminator

Introduction

In early computer vision (CV), this concept was typically used for Multi-view Geometry or object matching. If we see an object in view A, find it in view B, then find it in view C, and finally map it back to view A, it should still be the same object. Now we can migrate this concept to the Token level+ in multimodal Transformers+: If "dog's bark" (audio) maps to "dog's image" (video), and "dog's image" maps to the word "Dog" (text), then "dog's bark" must also directly map to the word "Dog".

Matrix Setup

Suppose we have M tokens and N modalities. As shown in the diagram below, the diagonal $M \times M$ blocks are all identity matrices, and the off-diagonal $M \times M$ blocks are permutation matrices. Together, they form a large matrix P.

Attention Matrix Structure

The attention matrix A has the following structure:

- **Diagonal Blocks (Self-Attention):** Identity matrices (I_M). These represent a modality attending to itself. Ideally, tokens attend perfectly to themselves (e.g., "dog" $\leftrightarrow$ "dog").

- **Off-Diagonal Blocks (Cross-Attention):** Permutation matrices (P_{ij}). These represent alignment between modalities. For example: Row 1 (Audio "Barking") attends to Column 4 (Video "Dog").

Core Mathematical Intuition

Let's examine the rank of the large matrix P:

1. **Ideal State (Rank = M):** If all modalities are perfectly aligned, the entire $NM \times NM$ super-matrix effectively contains only M independent "concepts" (e.g., dog, human, background). Regardless of how many modalities N you have, these rows and columns are linearly dependent. Therefore, $\text{Rank}(A) = M$.

2. **Inconsistent State (Rank > M):** If audio perceives "barking" as "dog" but video perceives "barking" as "human," this introduces a contradiction. This contradiction manifests as a new independent dimension in linear algebra, resulting in $\text{Rank}(A) > M$.

3. **Optimization Objective:** Since directly minimizing matrix rank (Rank Minimization) is an NP-hard problem, we typically use the Nuclear Norm+ ($\|P\|_*$) as a convex relaxation for the rank instead.

Example

Consider three modalities:

- **Modality 1 (Audio):** [Barking, Noise, Silence]
- **Modality 2 (Video):** [Dog, Human, Background]
- **Modality 3 (Text):** ["Dog", "Human", "None"]

In the ideal case, "Barking" (audio) should align with "Dog" (video) and "Dog" (text), maintaining consistency across all three modalities. The attention matrix should reflect this consistency through its rank structure.