MULTIMODAL REPRESENTATION ALIGNMENT

Giordano Cicchetti, Eleonora Grassucci, Danilo Comminiello, Aurelio Uncini

Dipartimento di Ingegneria dell'Informazione, Elettronica e Telecomunicazioni (DIET)
"Sapienza" Università di Roma
Via Eudossiana 18, 00184 Roma

Parole chiave: multimodal learning, modality alignment, representation learning, contrastive learning

Introduzione

L'apprendimento multimodale mira a integrare informazioni provenienti da diverse modalità di dati, come testo, audio e video [1], per ottenere rappresentazioni latenti condivise e semanticamente coerenti. Tuttavia, gli approcci tradizionali basati su allineamenti pairwise, come il contrastive learning con similarità coseno [2], presentano limitazioni quando si tratta di scalare a più di due modalità, poiché non garantiscono un allineamento simultaneo e coerente tra tutte le modalità coinvolte. In questo contesto, è stata proposto GRAM (Gramian Representation Alignment Measure), un nuovo framework che riformula l'allineamento multimodale tra due o più modalità direttamente nello spazio latente che gli embedding delle modalità definiscono [3]. GRAM misura l'allineamento tra le modalità calcolando il volume del parallelotopo k-dimensionale generato dai vettori di embedding delle diverse modalità. Questo volume è ottenuto come radice quadrata del determinante della matrice di Gram costruita a partire dagli embedding normalizzati. Un volume ridotto indica una forte coerenza semantica tra le modalità, mentre un volume elevato suggerisce disallineamento.

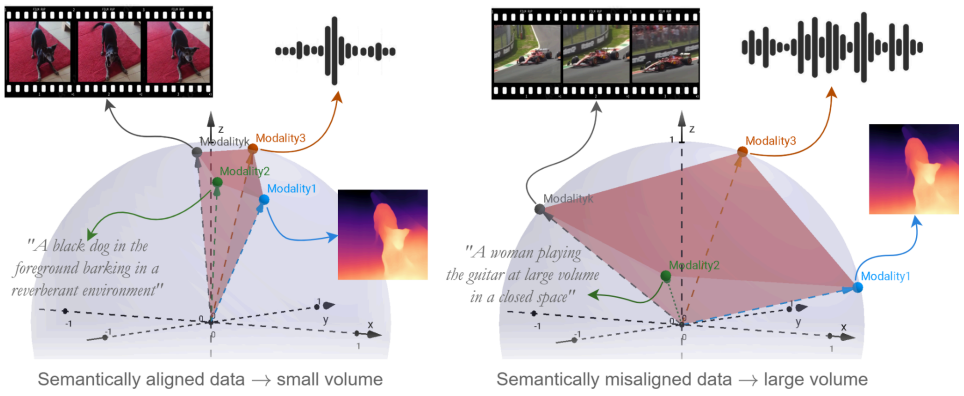

Figura 1: L''intuizione di GRAM, per dati semanticamente simili, il volume è ridotto, per dati semanticamente diversi, il volume è ampio.

GRAM introduce una nuova contrastive loss basata sul volume, che sostituisce la similarità coseno tradizionale. Questa loss minimizza il volume del parallelotopo per campioni positivi (modalità che rappresentano lo stesso contenuto) e lo massimizza per campioni negativi, favorendo così un allineamento simultaneo e coerente tra tutte le modalità.
I risultati sperimentali dimostrano che GRAM supera gli approcci state-of-the-art in compiti di retrieval e classificazione multimodale su benchmark consolidati. In particolare, l'uso della misura di volume come metrica di allineamento consente di ottenere prestazioni superiori senza modifiche architetturali significative, evidenziando la robustezza e la generalizzabilità del metodo proposto.

Bibliografia

[1] B . Zhu et al., “LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment”, ICLR 2024.
[2] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision”, ICML 2021.
[3] G. Cicchetti, E. Grassucci, L. Sigillo, D. Comminiello, “Gramian Multimodal Representation Learning and Alignment”, ICLR 2025.