Titolo: Multimodal Large Language Models per il processamento di dati 3D

Oggetto attività di ricerca:

I Multimodal Large Language Models (MLLMs) hanno recentemente dimostrato capacità straordinarie nel comprendere e ragionare su dati eterogenei, integrando linguaggio naturale con modalità visive come immagini e video. Tuttavia, l'estensione di questi modelli al dominio 3D rappresenta ancora una frontiera di ricerca aperta e di grande rilevanza, sia per le applicazioni in robotica, realtà aumentata e veicoli autonomi, sia per la crescente disponibilità di rappresentazioni neurali implicite come i Neural Radiance Fields (NeRF) e i 3D Gaussian Splats.
Alcuni lavori recenti hanno aperto una direzione innovativa, dimostrando come sia possibile interfacciare direttamente MLLMs con rappresentazioni neurali 3D, processando i pesi delle reti NeRF come token e abilitando capacità di captioning, question answering e ragionamento spaziale su oggetti 3D. Tuttavia, gli approcci attuali presentano alcune limitazioni significative: si concentrano prevalentemente su una singola modalità di rappresentazione 3D per volta, sono progettati per scene contenenti singoli oggetti isolati, e si limitano a contesti indoor o a oggetti di piccola scala.
L'estensione di questi modelli a scenari più complessi e realistici è ancora in uno stadio iniziale, sia per la difficoltà di progettare encoder che gestiscano modalità 3D eterogenee in modo unificato, sia per la complessità intrinseca delle scene multi-oggetto e outdoor, dove le rappresentazioni neurali assumono dimensioni e caratteristiche radicalmente differenti. In questo contesto, l'attività di ricerca si propone di sviluppare modelli linguistici multimodali capaci di comprendere e ragionare su scene 3D complete ed eterogenee, contribuendo all'avanzamento verso assistenti AI in grado di interagire con il mondo fisico in modo sempre più ricco e generale.

Dettaglio attività da svolgere

L'attività di ricerca sarà organizzata secondo un percorso progressivo che, partendo da un consolidamento delle conoscenze nel campo, procederà verso lo sviluppo di soluzioni originali di crescente complessità.
In una prima fase, verrà condotta un'analisi approfondita dello stato dell'arte sui modelli linguistici multimodali applicati a dati 3D, con particolare attenzione alle rappresentazioni neurali implicite ed esplicite e alle strategie di tokenizzazione di dati tridimensionali. L'analisi sarà accompagnata da una valutazione sperimentale sistematica delle soluzioni più promettenti su benchmark pubblici, finalizzata a identificare i limiti delle metodologie esistenti e le opportunità di innovazione.
Successivamente, l'attività si concentrerà sull'integrazione di molteplici rappresentazioni neurali 3D all'interno di un unico modello multimodale, attraverso la progettazione di un'architettura unificata capace di processare congiuntamente formati eterogenei e

sfruttarne la complementarità. Saranno esplorate strategie di tokenizzazione cross-modale e meccanismi di proiezione in spazi di rappresentazione condivisi, con validazione su task di comprensione, descrizione e ragionamento.
La fase successiva affronterà il passaggio dalla comprensione di singoli oggetti alla comprensione di scene 3D complete, caratterizzate da molteplici entità in relazione spaziale e semantica tra loro. Sarà necessario ripensare i meccanismi di rappresentazione per catturare la struttura compositiva e relazionale dell'intera scena, sperimentando architetture adeguate alla complessità di ambienti realistici, a partire da dataset consolidati per il dominio indoor.
Come ulteriore estensione, si affronterà la generalizzazione a scene 3D outdoor, dove emergono sfide specifiche legate alla scala estesa, alla variabilità delle condizioni di acquisizione e alla presenza di elementi dinamici, costituendo un importante banco di prova per la generalità dell'approccio proposto.
Trasversalmente a tutte le fasi, i risultati saranno disseminati attraverso pubblicazioni in conferenze e riviste internazionali di riferimento e, ove possibile, mediante il rilascio open-source di codice e modelli prodotti