Intelligenza Artificiale per Ricerca in Big Multimedia Data

Fabio Carrara, Giuseppe Amato, Franca Debole, Marco Di Benedetto
Fabrizio Falchi, Claudio Gennaro, Nicola Messina

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI)
Consiglio Nazionale delle Ricerche (CNR)
nome.cognome@isti.cnr.it

Abstract

La diffusa produzione di immagini e media digitali ha reso necessario l'utilizzo di metodi automatici di analisi e indicizzazione su larga scala per la loro fruzione. Il gruppo AIMIR dell'ISTI-CNR si è specializzato da anni in questo ambito ed ha abbracciato tecniche di Deep Learning basate su reti neurali artificiali per molteplici aspetti di questa disciplina, come l'analisi, l'annotazione e la descrizione automatica di contenuti visuali e il loro recupero su larga scala.

1 Attività Scientifica

Il gruppo *Artificial Intelligence for Multimedia Information Retrieval* (AIMIR) dell'ISTI-CNR nasce storicamente in un contesto di gestione di dati multimediali ed ha quindi abbracciato le moderne tecniche di IA nella modellazione e rappresentazione di tali dati, sposandole con successo con molteplici aspetti di questa disciplina, in particolare con la gestione su larga scala di dati percettivi visuali, quali immagini e video. Tra le attività scientifiche sostenute e le competenze presenti nel gruppo, spiccano le seguenti:

Recupero di immagini su larga scala basati sul contenuto Data la mole di immagini prodotte quotidianamente dagli utenti del Web, lo sviluppo di tecniche automatiche e scalabili per la comprensione automatica ed il recupero di immagini risulta di vitale importanza. Sfruttando tecniche di modellazione profonda data-driven come il Deep Learning, il gruppo si è specializzato nello sviluppo e l'utilizzo di rappresentazioni vettoriali compatte ed efficaci per immagini estratte tramite reti neurali convoluzionali (Deep Features, R-MAC). L'adozione di questo tipo di rappresentazioni ci ha permesso di sviluppare tecniche di indicizzazione e ricerca per similarità visuale di immagini non etichettate con un alto grado di scalabilità (nell'ordine di centinaia di milioni di immagini[1]) mantenendo un alto livello di accuratezza dei risultati della ricerca [Amato *et al.*, 2016a].

In questo contesto, sono state svolte attività di ricerca sulla trasformazione di tali rappresentazioni tramite l'utilizzo di permutazioni [Amato *et al.*, 2014; Amato *et al.*, 2016b] e trasformazioni geometriche [Amato *et al.*, 2018a; Amato *et al.*, 2018b] per facilitarne l'indicizzazione. Le trasformazioni introdotte ci permettono utilizzare delle *rappresentazioni testuali surrogate* dei descrittori visuali e quindi di impiegare indici open source basati su liste invertite tradizionalmente usati per documenti testuali (e.g. Elasticsearch, Apache Lucene) per la gestione di database di immagini, favorendo il trasferimento tecnologico di tali tecniche [Amato *et al.*, 2017][2].

Inoltre, grazie alla flessibilità delle reti neurali profonde, sono state sviluppate tecniche di recupero di immagini che affrontano e risolvono problemi avanzati in questa disciplina, quali il *cross-media retrieval* [Carrara *et al.*, 2017], i.e. il recupero di immagini non etichettate partendo da una sua descrizione testuale , ed il *relational content-based image retrieval* [Messina *et al.*, 2018], dove si richiede di recuperare immagini raffiguranti oggetti con precise relazioni spaziali o semantiche tra loro[3].

Analisi visuale dell'emotività trasmessa Nel contesto dell'analisi dei dati provenienti dai social media, il gruppo ha sviluppato competenze e tecniche allo stato dell'arte di *visual sentiment analysis* [Vadicamo *et al.*, 2017], cioè nell'analisi del sentimento veicolato da media visuali, tramite l'utilizzo di reti neurali convoluzionali[4]. Sono state sviluppate techiche di allenamento cross-media che sfruttano la grande quantità di dati rumorosi provenienti dai social media (in particolare Twitter) per allenare modelli per la classificazione del sentimento visuale allo stato dell'arte senza indurre in costi di etichettatura o di creazione di dataset di training.

Sistemi di video-browsing Dall'unione delle competenze sopraelencate, il gruppo ha svolto attività di ricerca e sviluppo di tool per la ricerca interattiva di video su larga scala, partecipando alla competizione di Video Browsing Showdown (VBS 2019) con il sistema *VISIONE* [Amato *et al.*, 2019]. Il sistema integra moduli di analisi, annotazione e recupero del contenuto visuale basate su tecniche deep learning allo stato dell'arte e fornisce molteplici modalità di ricerca, come la ricerca per similarità visuale, per locazione spaziale di oggetti o per semplici keyword testuali. Tutte le informazioni risultanti dalle analisi sono codificate tramite rappresentazioni testuali surrogate ed indicizzate con motori di ricerca testuali performanti e scalabili.

[1] http://mifile.deepfeatures.org/

[2] http://melisandre.deepfeatures.org/LuceneGUI/

[3] http://rcbir.org/

[4] http://t4sa.it/