A

L'**apprendimento contrastivo** è recentemente diventato un approccio predominante per apprendere **rappresentazioni** significative di concetti in ML. Il framework di apprendimento si basa sull'idea che concetti semanticamente correlati (ad esempio, due immagini dello stesso oggetto da punti di vista diversi) dovrebbero avere **rappresentazioni** simili, mentre quelli non correlati dovrebbero essere dissimili. Inizialmente concepito per **l'apprendimento della rappresentazione delle immagini auto-supervisionato**, l'**apprendimento contrastivo** è stato recentemente applicato anche al linguaggio. Lavori recenti hanno utilizzato l'addestramento contrastivo per collegare diverse modalità, ad esempio, visione e linguaggio, audio e linguaggio, o una combinazione delle tre. Questi modelli apprendono **rappresentazioni** di concetti da diverse modalità (ad esempio, un estratto testuale come "un cane che corre su un campo" e un'immagine che rappresenta la scena) e le ottimizzano per essere vicine in uno **spazio latente** condiviso. Fondamentalmente, la pipeline tipica è **auto-supervisionata**: poiché non è coinvolta alcuna **annotazione** manuale (ad esempio, nell'esempio precedente, si possono raccogliere coppie immagine-testo dal web), l'intervento umano è limitato a decidere quale compito di **pre-addestramento** debba essere utilizzato.

CLIP è una **rete neurale multi-modale visione-linguaggio** addestrata tramite CL per associare concetti visivi con il testo. Il modello comprende un **codificatore** di visione e uno di testo, ciascuno seguito da uno **strato lineare** per proiettare le **rappresentazioni** di immagini e testo nello stesso **spazio latente**. CLIP è addestrato a posizionare immagini e descrizioni corrispondenti (ad esempio, un'immagine di una maglietta rossa e la sua descrizione "una maglietta rossa") vicine nello **spazio vettoriale** (vedi Fig. [1] per un esempio). Quando addestrato su 400 milioni di <c**oppie**, **testo**> raccolte da internet, CLIP ha dimostrato un trasferimento competitivo zero-shot o few-shot a compiti **a valle** come OCR e classificazione fine-grained degli oggetti.

B

Il **Contrastive learning** è recentemente diventato un approccio predominante per apprendere **representation** significative di concetti in ML. Il framework di apprendimento si basa sull'idea che concetti semanticamente correlati (ad esempio, due immagini dello stesso oggetto da punti di vista diversi) dovrebbero avere **representation** simili, mentre quelli non correlati dovrebbero essere dissimili. Inizialmente concepito per il **self-supervised image representation learning**, il **contrastive learning** è stato recentemente applicato anche al linguaggio. Lavori recenti hanno utilizzato l'addestramento contrastivo per collegare diverse modalità, ad esempio, visione e linguaggio, audio e linguaggio, o una combinazione delle tre. Questi modelli apprendono **representation** di concetti da diverse modalità (ad esempio, un testo come "un cane che corre su un campo" e un'immagine che raffigura la scena) e li ottimizzano per essere vicini in uno stesso **latent space** condiviso. Fondamentalmente, il pipeline tipico è **self-supervised**: poiché non è coinvolta alcuna **annotation** manuale (ad esempio, nell'esempio precedente, si possono raccogliere coppie immagine-testo dal web), l'intervento umano è limitato a decidere quale compito di **pre-training** debba essere utilizzato.

CLIP è una **neural network vision-language multi-modal** addestrata tramite CL per associare concetti di visione con testo. Il modello comprende un **encoder** per la visione e uno per il testo, ciascuno seguito da un **linear layer** per proiettare le **representation** delle immagini e dei testi nello stesso **latent space**. CLIP è addestrato a posizionare immagini e descrizioni corrispondenti (ad esempio, un'immagine di una maglietta rossa e la sua descrizione "una maglietta rossa") vicini nel **vector space** (vedi Fig. [1] per un esempio). Quando addestrato su 400 milioni di coppie <**image**, **text**> raccolte da internet, CLIP ha dimostrato un trasferimento competitivo zero-shot o few-shot a compiti **downstream** come OCR e classificazione fine-grained degli oggetti.