Implementation CLIP

Corentin Clément, Victor Buthod, Guillaume Charvolin

19 fevrier 2024

Majeure SCIA-G

EPITA

Table des matières

1 Introduzione

Il modello CLIP (*Contrastive Language–Image Pretraining*), proposto da OpenAI, ha rivoluzionato il modo di apprendere rappresentazioni condivise per immagini e testi. L'idea principale è di allineare due spazi latenti—uno per le immagini e uno per i testi—per consentire compiti come il recupero di immagini basato su descrizioni testuali o la descrizione automatica di immagini.
In questo progetto, abbiamo riprodotto e adattato l'approccio descritto nell'articolo originale. Sono state esplorate tre varianti architetturali :

— **BERT + ResNet :** Combinazione di BERT per il testo e ResNet per le immagini.
— **BERT + ViT :** Utilizzo di BERT per il testo e di ViT (Vision Transformer) per le immagini.
— **GPT + ViT :** GPT-2 come encoder per il testo e ViT per le immagini.

Ogni modello è stato addestrato utilizzando una funzione di perdita contrastiva (*contrastive loss*) per massimizzare la similarità tra coppie immagine-testo corrispondenti e minimizzare quella tra coppie non corrispondenti.

2 Dataset e Preprocessing

2.1 Descrizione del Dataset

Il dataset utilizzato, **Fashion Product Images Small**, contiene circa 44,000 immagini di prodotti di moda accompagnate da descrizioni testuali nella colonna `productDisplayName`. Queste descrizioni sono generalmente semplici, come "Maglietta rossa da uomo". Sebbene il dataset sia utile per compiti di base, presenta alcune limitazioni :

— Le descrizioni testuali sono spesso ripetitive o vaghe, rendendo difficile distinguere tra immagini simili.
— La mancanza di informazioni contestuali o complesse può ostacolare modelli come GPT, che richiedono input più ricchi.

2.2 Preprocessing des Dati

Le seguenti fasi di preprocessing sono state applicate :

- Le immagini sono state ridimensionate a 224×224 per essere compatibili con le architetture ResNet e ViT.
- I valori dei pixel sono stati normalizzati nell'intervallo [-1, 1].
- Le descrizioni testuali sono state utilizzate direttamente senza arricchimenti.

A differenza dell'approccio originale di CLIP, non è stata creata una divisione tra set di addestramento e set di test a causa di vincoli di tempo e risorse.

3 Architettura dei Modelli

Ogni modello CLIP è composto da due encoder indipendenti : uno per le immagini e uno per i testi. Questi encoder proiettano gli input in uno spazio latente condiviso di dimensione 128, seguito da una normalizzazione L2. Un parametro di temperatura (*learnable parameter*) regola la distribuzione delle similarità durante l'addestramento.

3.1 Encoder per Immagini

Sono state testate due architetture per l'encoding delle immagini :

- **ResNet :** Un modello convoluzionale pre-addestrato su ImageNet.
- **ViT :** Un Vision Transformer che elabora le immagini come sequenze di patch.

3.2 Encoder per Testi

Per l'encoding dei testi, sono stati valutati due modelli :

- **BERT :** Un modello bidirezionale pre-addestrato su compiti di comprensione del linguaggio.
- **GPT-2 :** Un modello generativo in grado di produrre sequenze testuali.

4 Esperimenti e Risultati

4.1 Addestramento

Tutti e tre i modelli sono stati addestrati per 5 epoche sull'intero dataset con una dimensione di batch pari a 64. L'addestramento è stato eseguito su Google Colab, limitando il numero di epoche e la possibilità di esplorare ulteriori iperparametri.

4.2 Valutazione

Le performance sono state valutate utilizzando due metriche :

- **Accuracy :** Percentuale di coppie immagine-testo in cui il testo corretto aveva la similarità massima.
- **Top-100 Accuracy :** Percentuale di coppie in cui il testo corretto era tra i 100 testi più simili.

Modello	Accuracy (%)	Top-100 Accuracy (%)
BERT + ResNet	0.74	34.51
BERT + ViT	0.87	37.11
GPT + ViT	0.00	0.10

TABLE 1 – Confronto delle performance dei tre modelli testati.

4.3 Analisi dei Risultati

BERT + ResNet e **BERT + ViT** hanno mostrato prestazioni superiori, con un leggero vantaggio per BERT + ViT. Questo può essere attribuito a :

- **BERT :** Pre-addestrato su compiti di comprensione del linguaggio, è ben adatto per descrizioni testuali semplici.
- **ViT :** Supera ResNet catturando dettagli visivi più fini attraverso il suo approccio basato su patch.

Invece, **GPT + ViT** non ha raggiunto risultati significativi. Possibili spiegazioni includono :

- **Descrizioni testuali semplici :** GPT, progettato per sequenze complesse, ha difficoltà con input basilari.
- **Allineamento complesso :** La natura generativa di GPT complica il suo allineamento con rappresentazioni visive.
- **Sottoaddestramento :** Cinque epoche sono insufficienti per far convergere un'architettura così complessa.

5 Semplificazioni e Limitazioni

Per affrontare i vincoli, sono state introdotte diverse semplificazioni :

- **Input testuale semplice :** Uso diretto di `productDisplayName`.
- **Nessuna separazione train/test :** Il modello è stato valutato sugli stessi dati utilizzati per l'addestramento.
- **Epoche limitate :** Il numero ridotto di epoche ha influito sulla capacità del modello di convergere.

Queste semplificazioni spiegano in parte le prestazioni modeste, in particolare per GPT + ViT.

6 Conclusioni

Questo progetto ha permesso di riprodurre una versione semplificata di CLIP e di testare tre architetture. I risultati indicano che BERT + ViT è la combinazione migliore in questo contesto, bilanciando efficienza e accuratezza. Miglioramenti futuri potrebbero includere :

- Arricchire il dataset con annotazioni testuali più descrittive e diversificate.
- Estendere il tempo di addestramento ed esplorare la regolazione degli iperparametri.
- Creare una chiara separazione tra i set di addestramento e di test per una migliore valutazione della generalizzazione.

Questo lavoro evidenzia i punti di forza e le debolezze dei modelli CLIP in contesti limitati, aprendo la strada a miglioramenti futuri.