Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features

Alberto Baldrati, Marco Bertini, Tiberio Uricchio and Alberto Del Bimbo
Università degli Studi di Firenze – MICC, Florence, Italy

Introduction

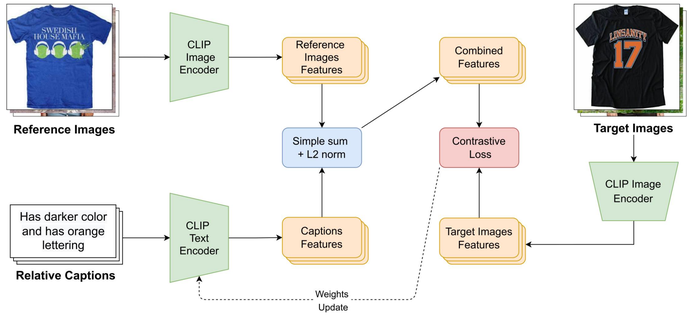
- In conditioned and composed image retrieval a query image is combined with a text that provides information regarding user intentions
- The proposed method is based on an initial stage where a simple combination of visual and textual features is used in order to fine-tune the CLIP text encoder
- Then in a second training stage, we learn a more complex combiner network that merges visual and textual features

Contributions

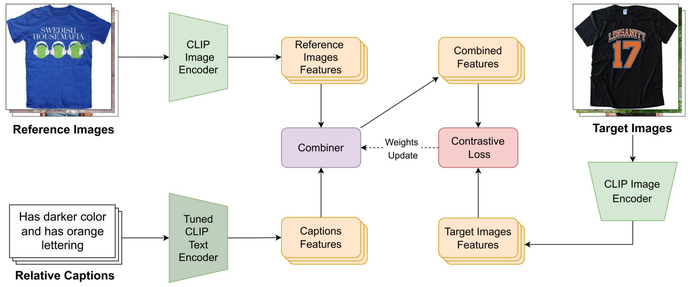
- We propose a novel CLIP fine-tuning scheme for conditioned and composed image retrieval which breaks up the symmetry between the two encoders
- We propose a novel two-stage approach that integrates the CLIP text-encoder fine-tuning with the training of a Combiner network
- We perform a study that tries to explain the effects of our approach on the distribution of the features in the embedding space

Proposed Method

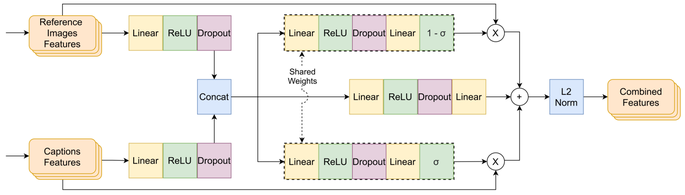
- The first stage of training: **CLIP text encoder fine-tuning**



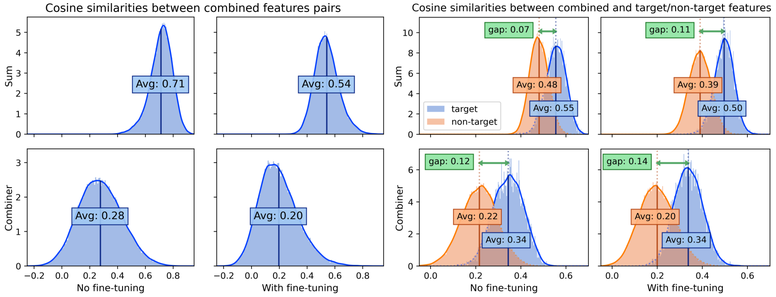
- The second stage of training**: Combiner training**



- The **Combiner architecture**



Feature Distribution Study



Comparison with SotA

- Our approach achieves state-of-the-art performances on both FashionIQ and CIRR datasets:
  - On **FashionIQ** the proposed method improves up to ~**9%** on average R@10 and ~**7%** on average R@50 over the best non-CLIP-based models
  - On **CIRR** the proposed method outperforms current methods by a significant margin, especially in **low-rank** recall metrics with an improvement of **~19%** on average R@1

Conclusions

- In this paper, we present a novel CLIP fine-tuning scheme tailored for conditioned and composed image retrieval
- The proposed novel two-stage approach manages to achieve state-of-the-art results on both FashionIQ and CIRR datasets
- **Scan the QR Code to try a LIVE DEMO**