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

Workshop on Open-Domain Retrieval Under Multi-Modal Settings, CVPR 2022, New Orleans

[1,2]Alberto Baldrati, [1]Marco Bertini, [1]Tiberio Uricchio, [1]Alberto Del Bimbo
[name.surname@unifi.it]

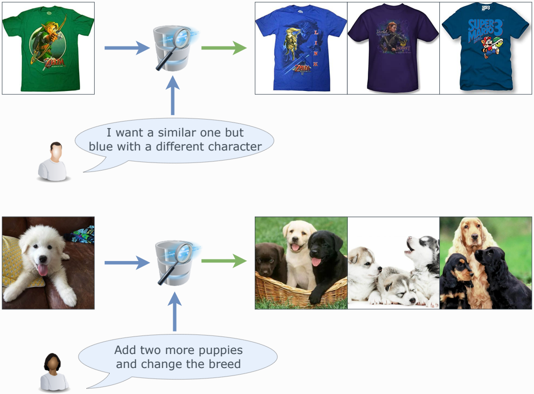
[1]Università degli Studi di Firenze - MICC
[2]Università di Pisa

Firenze, Italy - Pisa, Italy

Image retrieval with textual feedback

Conditioned and Composed image retrieval example

Conditioned and composed image retrieval extends traditional CBIR systems to improve their effectiveness by adding user feedback



Introduction
Overview

- To address the conditioned and composed image retrieval tasks we propose a two-stage approach based on CLIP [1] multimodal features:

Introduction
Overview

- To address the conditioned and composed image retrieval tasks we propose a two-stage approach based on CLIP [1] multimodal features:

  1. We fine-tune the CLIP text encoder using a simple combination of visual and textual features

Introduction
Overview

- To address the conditioned and composed image retrieval tasks we propose a two-stage approach based on CLIP [1] multimodal features:

  1. We fine-tune the CLIP text encoder using a simple combination of visual and textual features

  2. We train from scratch a Combiner network that learns to fuse the partially fine-tuned multimodal features

Introduction
Overview

- To address the conditioned and composed image retrieval tasks we propose a two-stage approach based on CLIP [1] multimodal features:

  1. We fine-tune the CLIP text encoder using a simple combination of visual and textual features

  2. We train from scratch a Combiner network that learns to fuse the partially fine-tuned multimodal features

- The proposed two-stage approach achieves state-of-the-art performance on FashionIQ [2] and CIRR [3] datasets

First Stage
Text encoder fine-tuning

In this stage we perform a fine-tuning of the CLIP text encoder to reduce the task mismatch between the large-scale image-text pre-training and the downstream task

First Stage
Text encoder fine-tuning

In this stage we perform a fine-tuning of the CLIP text encoder to reduce the task mismatch between the large-scale image-text pre-training and the downstream task

Reference Images

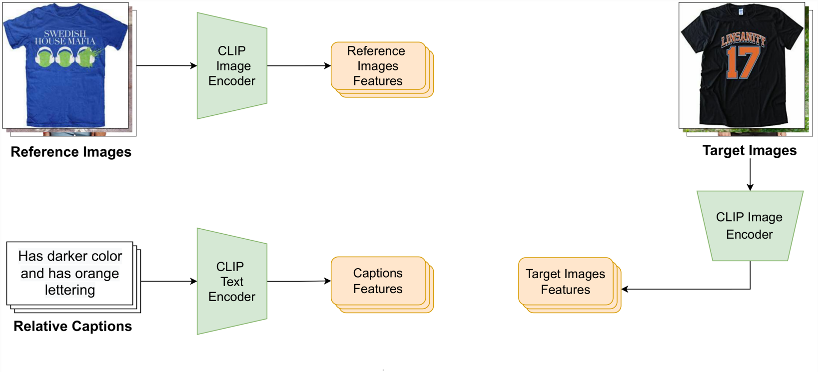
Target Images

Has darker color and has orange lettering

Relative Captions

First Stage
Text encoder fine-tuning

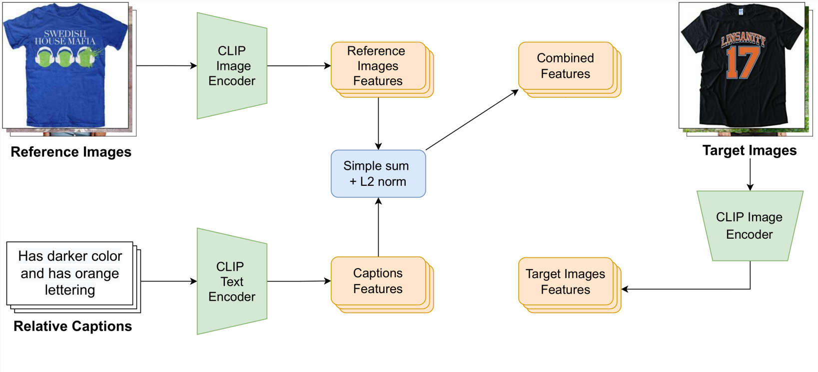
In this stage we perform a fine-tuning of the CLIP text encoder to reduce the task mismatch between the large scale image-text pre-training and the downstream task



First Stage

Text encoder fine-tuning

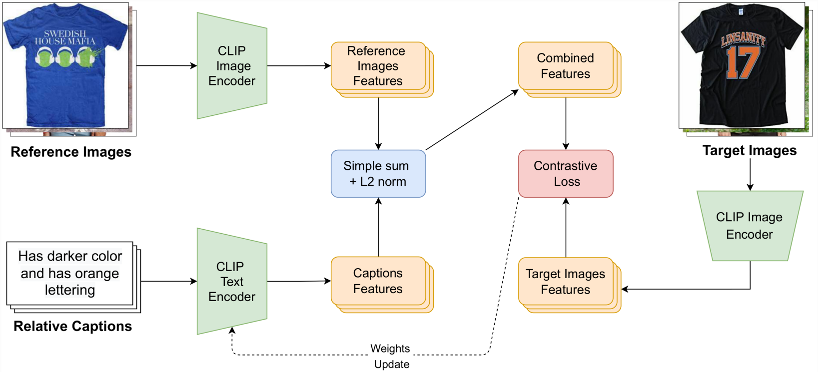
In this stage we perform a fine-tuning of the CLIP text encoder to reduce the task mismatch between the large scale image-text pre-training and the downstream task



First Stage

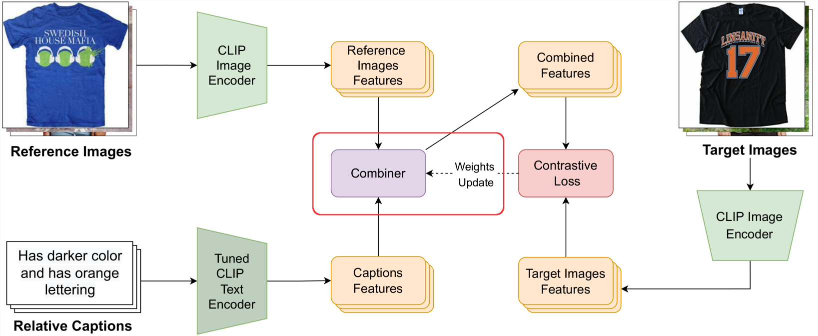
Text encoder fine-tuning

In this stage we perform a fine-tuning of the CLIP text encoder to reduce the task mismatch between the large scale image-text pre-training and the downstream task



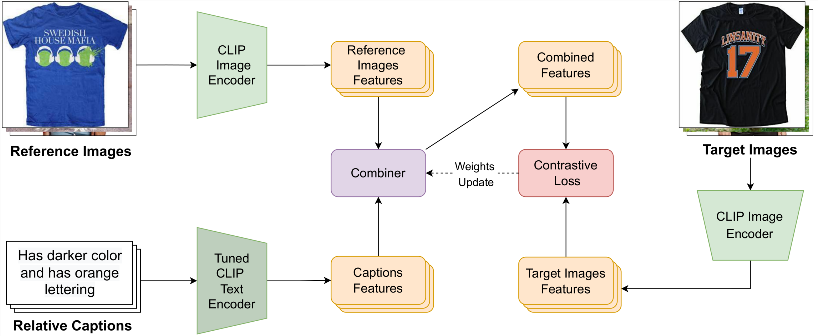
Second stage
Combiner training

In this second stage we train from scratch a Combiner network that learns to combine the multimodal query features



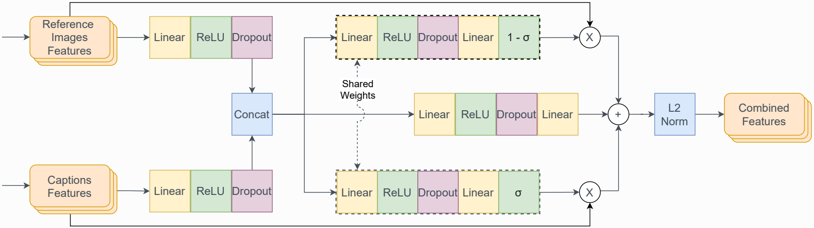
Second stage
Combiner training

In this second stage we train from scratch a Combiner network that learns to combine the multimodal query features



Combiner
Architecture

The Combiner network outputs a normalized sum of multiple components: a convex combination of text and image features and a learned text-image mixture



Comparison with sotA
FashionIQ dataset

Method	Shirt		Dress		Toptee		Average	
	R@10	*R@50*	*R@10*	*R@50*	*R@10*	*R@50*	*R@10*	*R@50*
ARTEMIS [4]	21.78	43.64	27.16	52.40	29.20	54.83	26.05	50.29
RTIC-GCN w/GloVe [5]	23.79	47.25	29.15	54.04	31.61	57.98	28.18	53.09
CoSMo [6]	24.90	49.18	25.64	50.30	29.21	57.46	26.58	52.31
AACL [7]	24.82	48.85	29.89	55.85	30.88	56.85	28.53	53.85
DCNet [8]	23.95	47.30	28.95	<u>56.07</u>	30.44	58.29	27.78	53.89
SAC w/BERT [9]	28.02	51.86	26.52	51.01	32.70	61.23	29.08	54.70
Baldrati et al (RN50x4) [10]	35.76	56.20	27.20	53.57	36.31	61.14	33.09	56.99
Proposed approach (RN50)	<u>35.77</u>	<u>57.02</u>	<u>31.73</u>	56.02	<u>36.46</u>	<u>62.77</u>	<u>34.65</u>	<u>58.60</u>
Proposed approach (RN50x4)	**39.99**	**60.45**	**33.81**	**59.40**	**41.41**	**65.37**	**38.32**	**61.74**

Table: Comparison between our method and current state-of-the-art models on the Fashion-IQ dataset. Best scores are highlighted in bold, second-best scores are underlined.

Comparison with sotA

CIRR dataset

Method	Recall@K				R_{subset}@K		
	$K = 1$	$K = 5$	$K = 10$	$K = 50$	$K = 1$	$K = 2$	$K = 3$
TIRG† [11]	14.61	48.37	64.08	90.03	22.67	44.97	65.14
MAAF† [12]	10.31	33.03	48.30	80.06	21.05	41.81	61.60
MAAF+BERT† [12]	10.12	33.10	48.01	80.57	22.04	42.41	62.14
ARTEMIS [4]	16.96	46.10	61.31	87.73	39.99	62.20	75.67
CIRPLANT † [3]	15.18	43.36	60.48	87.64	33.81	56.99	75.40
CIRPLANT w/OSCAR † [3]	19.55	52.55	68.39	92.38	39.20	63.03	79.49
Proposed approach (RN50)	<u>35.81</u>	<u>68.80</u>	<u>80.17</u>	<u>95.25</u>	<u>66.96</u>	<u>85.25</u>	<u>93.13</u>
Proposed approach (RN50x4)	**38.53**	**69.98**	**81.86**	**95.93**	**68.19**	**85.64**	**94.17**

Table: Comparison between our method and current state-of-the-art models on the CIRR test set. Best scores are highlighted in bold, second-best scores are underlined. † denotes results cited from [3]

Live demo

Scan the QR Code to try a LIVE DEMO

MICC

Media Integration and Communication Center

FLORENTINA STUDIORUM UNIVERSITAS

References I

[1] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *arXiv preprint arXiv:2103.00020* (2021). arXiv: 2103.00020 [cs.CV].

[2] Hui Wu et al. "Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback". In: *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. arXiv: 1905.12794 [cs.CV].

[3] Zheyuan Liu et al. "Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models". In: *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021. arXiv: 2108.04024 [cs.CV].

[4] Ginger Delmas et al. "ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity". In: *International Conference on Learning Representations*. 2021.

References II

[5] Minchul Shin et al. "RTIC: Residual Learning for Text and Image Composition using Graph Convolutional Network". In: *arXiv preprint arXiv:2104.03015* (2021).

[6] Seungmin Lee, Dongwan Kim, and Bohyung Han. "CoSMo: Content-Style Modulation for Image Retrieval With Text Feedback". In: *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 802–812.

[7] Yuxin Tian, Shawn Newsam, and Kofi Boakye. "Image Search with Text Feedback by Additive Attention Compositional Learning". In: *arXiv preprint arXiv:2203.03809* (2022).

[8] Jongseok Kim et al. "Dual Compositional Learning in Interactive Image Retrieval". In: *Proc. of AAAI Conference on Artificial Intelligence (AAAI)*. Vol. 35. 2. May 2021, pp. 1771–1779. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16271.

References III

[9] Surgan Jandial et al. "SAC: Semantic Attention Composition for Text-Conditioned Image Retrieval". In: *Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2022, pp. 4021–4030.

[10] Alberto Baldrati et al. "Conditioned Image Retrieval for Fashion using Contrastive Learning and CLIP-based Features". In: *Proc. of ACM Multimedia Asia (ACMMM Asia)*. 2021. DOI: 10.1145/3469877.3493593.

[11] Nam Vo et al. "Composing Text and Image for Image Retrieval - An Empirical Odyssey". In: *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018. arXiv: 1812.07119 [cs.CV].

[12] Eric Dodds et al. "Modality-Agnostic Attention Fusion for visual search with text feedback". In: *arXiv preprint arXiv:2007.00145* (2020). arXiv: 2007.00145 [cs.CV].