IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CONTENT BASED IMAGE RETRIEVAL COMBINING AND PARTIALLY FINE-TUNING CLIP-BASED FEATURES

[1]A. Jagadeesh Varma, [2]P. Vishnu, Mr. [3]A. Venkata Srinivasa Rao

Student, Student ,Associate Professor

Electronics and communication engineering,

SASI INSTITUTE OF TECHNOLOGY & ENGINEERING- TADEPALLIGUDEM

Abstract: In this work we present an approach for conditioned and composed image retrieval based on CLIP features. In this extension of content-based image retrieval (CBIR) an image is combined with a text that provides information regarding user intentions, and is relevant for application domains like e-commerce. The proposed method is based on an initial training stage where a simple combination of visual and textual features is used, 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. Contrastive learning is used in both stages. The proposed approach obtains state-of-the-art performance for conditioned CBIR on the Fashion-IQ dataset and for composed CBIR on the more recent CIRR dataset..

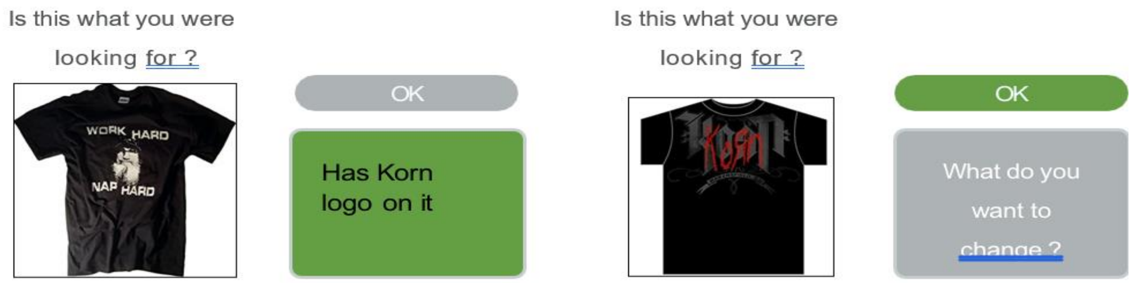
Keywords: Image Retrieval, Fine Tuning, Clip based Features, Visual and Image Pre training.

I. INTRODUCTION

Content-based image retrieval (CBIR) is a rapidly growing field driven by the explosion of visual content on the internet and the widespread use of high-resolution digital devices. Traditional image retrieval methods rely heavily on metadata and manual annotations which are error-prone, subjective, and infeasible for large datasets. CBIR aims to retrieve images based on their actual content such as color, texture, and shape.

This work presents an advanced CBIR approach that integrates image features with textual information to better capture user intent. This method is highly applicable to domains like fashion, advertising, and e-commerce where users often describe what they are looking for in natural language. By leveraging CLIP (Contrastive Language–Image Pre-training), which is a model trained to understand images and texts jointly, we enhance retrieval precision and relevance.

The contributions of this research include a two-stage training framework. In the first stage, we fine-tune the CLIP text encoder to adapt to our domain-specific retrieval task. In the second stage, a fusion model is trained to combine the image and text embeddings into a joint space optimized for similarity-based retrieval. The proposed system is evaluated on Fashion-IQ and CIRR datasets and achieves state-of-the-art performance.


Is this what you were looking for ?
OK
Has Korn logo on it
Is this what you were looking for ?
OK
What do you want to change ?

II. LITERATURE REVIEW

The CBIR field has evolved from early systems like QBIC and VisualSEEk that utilized simple color histograms, to more advanced systems that use complex feature representations. Researchers have introduced techniques such as texture and shape analysis, wavelet transforms, and semantic classification.CLIP by OpenAI has recently transformed the landscape by enabling zero-shot learning capabilities for vision-language tasks. Prior works have experimented with combining image and textual modalities for improved retrieval accuracy, often using late or early fusion methods. Our method builds on these advancements by incorporating domain-specific fine-tuning and learning a combiner network

Visual and language pretraining

The OpenAI CLIP network has very recently obtained remarkable results in multi-modal zero shot learning, and more in general it performs consistently well on different tasks despite not being directly optimized for a specific benchmark, thanks to its generalization capabilities of both images and text. CLIP learns associations between the im- ages and textual descriptions using 400 millions of image- text pairs scraped from the web for training. Effectiveness of CLIP is still subject of study , although it has already been successfully applied to different tasks like fine-grained art classification , image generation , zero shot video retrieval event and visual common- sense reasoning . Other approaches to learn image-text alignment have been proposed in. ALIGN uses a dual-encoder architecture and is trained on a huge dataset of 1 billion image-text pairs. Instead, the method proposed in is much more data efficient, exploiting contrastive dis- tillation, and requires a training dataset that is 133× smaller than that of CLIP.

Conditioned and combined image retrieval

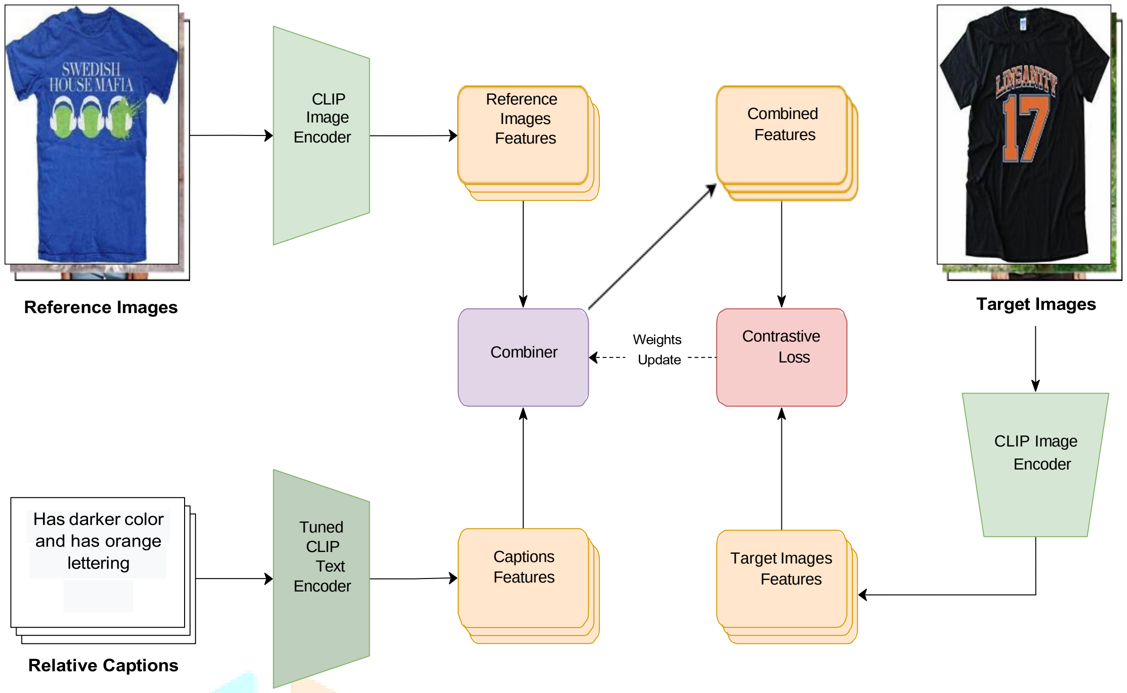
This work is related to the recent problem of conditioned fashion image retrieval , and with the very recent prob- lem of composed image retrieval of generic images .

The first task has been addressed in a large number of works. In, is presented a method based in a transformer that can be seamlessly plugged in a CNN to selectively pre- serve and transform the visual features conditioned on lan- guage semantics. In has been presented Text Image Residual Gating (TIRG), a method that combines image and text features using gating and residual features. In the authors combine graph neural networks and skip connec- tions. In , the authors use two different neural network modules, one for image style and one for image content. In a Correction Network is proposed to model explic- itly the difference between the reference and target image in the embedding space. In is proposed a model called Modality-Agnostic Attention Fusion (MAAF), designed for composed image retrieval, treating the convolutional spatial image features and learned text embeddings as modality- agnostic tokens, that are then passed to a Transformer

III. METHODOLOGY

The goal of text conditioned and composed image re- trieval is to retrieve the best matching image given a mul- timodal input consisting of an image-text pair. I.e., given an image (named *reference image*) and a text (named *rel- ative caption*) which expresses some modification with re- spect to the reference image, the aim of the retrieval is to find the best matching image which satisfy both the visual similarity constraints imposed by the reference image and integrates the changes expressed by the relative caption. In order to perform an effective retrieval, the system must be able to understand both the semantics of the image and the meaning of the text, to combine such multi domain infor- mation and finally to perform the retrieval using the fused representation.

Although having a common embedding space between text and images is a good starting point in the task we want to address, it is still not enough. Ideally, we would like to have a textual embedding space that contains displacement vectors in the image embedding space since the conditioned image retrieval task consists of moving between two points in image space using textual information



Experimental results

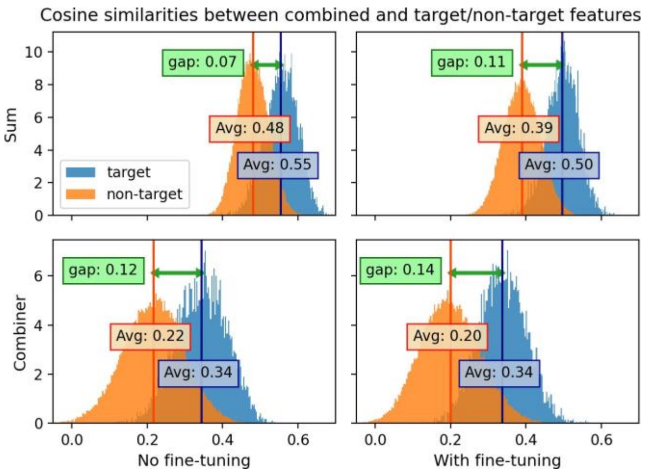
Datasets and metrics

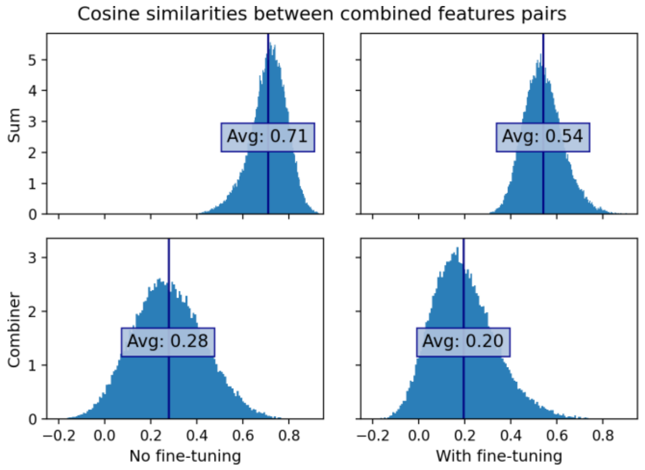
FashionIQ is a dataset for fashion conditioned image retrieval that contains 30,134 triplets from 77,684 images crawled from the web. The images are divided into three different categories: *Dress*, *Toptee* and *Shirt*.Following the standard experimental setting for this dataset, we report as evaluation metrics the average recall at rank K (Recall@K) at two different ranks: 10 and 50. All the results are on the validation set since, at the time of writing, test set ground-truth labels have not been publicly released.

CIRR(Compose Image Retrieval on Real-life im- ages) is a dataset containing 21,552 real-life images taken from the popular natural language reasoning $NLVR^2$ dataset [31]. It contains 36,554 triplets randomly assigned in 80% for training, 10% for validation and 10% for test.

Backbone	FT	CF	Recall@K				Rsubset @K		
			$K = 1$	$K = 5$	$K = 10$	$K = 50$	$K = 1$	$K = 2$	$K = 3$
RN50	✗	Sum	21.24	50.68	64.29	87.32	54.48	75.94	87.66
RN50	✓	Sum	31.57	65.10	77.47	94.47	65.31	84.64	93.06
RN50	✗	Combiner	31.28	64.84	77.88	94.90	62.04	81.58	91.60
RN50	✓	Combiner	37.00	70.94	82.28	96.13	67.47	85.39	93.66
RN50x4	✗	Sum	21.96	52.24	66.18	88.18	52.71	74.74	86.73
RN50x4	✓	Sum	32.62	67.02	79.74	95.31	65.41	84.67	92.54
RN50x4	✗	Combiner	33.63	67.16	80.22	95.58	63.62	82.85	92.15
RN50x4	✓	Combiner	39.75	73.71	83.90	96.87	70.92	87.42	94.19

Backbone	FT	CF	Shirt		Dress		Toptee		Average	
			$R@10$	$R@50$	$R@10$	$R@50$	$R@10$	$R@50$	$R@10$	$R@50$
RN50	✗	Sum	19.73	35.53	17.60	36.09	21.83	42.84	19.72	38.15
RN50	✓	Sum	30.71	52.21	27.52	51.66	33.61	58.95	30.61	54.27
RN50	✗	Combiner	31.80	53.38	26.82	51.31	33.40	57.01	30.67	53.90
RN50	✓	Combiner	35.77	57.02	31.73	56.02	36.46	62.77	34.65	58.60
RN50x4	✗	Sum	25.27	41.27	20.62	40.36	27.43	47.83	24.44	43.15
RN50x4	✓	Sum	35.77	57.41	31.14	55.18	38.09	61.04	35.00	57.88
RN50x4	✗	Combiner	36.36	58.00	31.63	56.67	38.19	62.42	35.39	59.03
RN50x4	✓	Combiner	39.99	60.45	33.81	59.40	41.41	65.37	38.32	61.74





V.Conclusions

In this work we present a fine-tuning scheme for conditioned image retrieval using CLIP-based features: we fine-tune the CLIP text encoder to adapt its embedding space to the task breaking up the symmetry between the en- coders. We then propose a novel two-stage approach which integrates the CLIP text encoder fine-tuning with the train- ing of a Combiner network which learns to combine the multimodal query features.

We conducted experiments on the fashion dataset Fash- ionIQ and on the open domain dataset CIRR. Experiments on both dataset show that our two-stage approach manages to reach state-of-the-art results by a consistent margin. The performances of the proposed method are particularly solid on low rank recall measures indicating the ability of captur- ing fine-grained modifications among similar images.

Finally we conducted a study which aims to explain the effects of our approach on feature distribution in the em- bedding space and how these effects are related to perfor- mance in the retrieval task. From the experiments we can notice that both the text encoder fine-tuning and the Com- biner network training led to a more efficient usage of the embedding space. Moreover it is shown that such increased sparsity in the embedding space helps to “move away” the combined features from the non-target ones improving the effectiveness of the retrieval.

Acknowledgement

We are the authors of this paper really convey our heartfull thanks to authos Alberto Baldrati,Marco Bertini,Tiberio Uricchio

For taking their paper as reference to do work on that.

References

[1] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Rad- ford, Jong Wook Kim, and Miles Brundage. Evaluating CLIP: Towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021. 2

[2] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kle- insteuber. Compositional learning of image-text query for image retrieval. In *Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1140– 1149, January 2021. 2

[3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Al- berto Del Bimbo. Conditioned image retrieval for fashion using contrastive learning and CLIP-based features. In *Proc. of ACM Multimedia Asia (ACMMM Asia)*, 2021. 2, 3, 4, 6, 7

[4] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided re- trieval. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 136–152, 11 2020. 7

[5] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learn- ing. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 7

[6] Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Va- jda, and Joseph E. Gonzalez. Data-efficient language- supervised zero-shot learning with self-distillation. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3119–3124, June 2021. 2

[7] Marcos V Conde and Kerem Turgutlu. CLIP-Art: Con- trastive pre-training for fine-grained art

classification. In *Proc. of Conference on Computer Vision and Pattern Recog- nition (CVPR)*, pages 3956–3960, 2021. 2, 3

[8] Ginger Delmas, Rafael S Rezende, Gabriela Csurka, and Di- ane Larlus. Artemis: Attention-based retrieval with text- explicit matching and implicit similarity. In *International Conference on Learning Representations*, 2021. 7

[9] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020. 2, 3, 7