ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447444

MULTI-LEVEL CONTRASTIVE LEARNING FOR HYBRID CROSS-MODAL RETRIEVAL

Yiming Zhao[⋆]*, Haoyu Lu*[⋆]*, Shiqi Zhao*[†]*, Haoran Wu*[†]*, Zhiwu Lu*[⋆*]

[⋆]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[†]China Unicom Research Institute, Beijing, China

ABSTRACT

Hybrid image retrieval is a significant task for a wide range of applications. In this scenario, the hybrid query for searching images consists of a reference image and a text modifier. The reference image provides a vital visual context and displays some semantic details, while the text modifier specifies the modifications to the reference image. To address such hybrid cross-modal retrieval, we propose a multi-level contrastive learning (MLCL) method for combining the hybrid query features into a fused feature by cross-modal contrastive learning with multi-level semantic alignment. Meanwhile, we additionally consider self-supervised contrastive learning to enhance the semantic correlation of the features at different levels of the combiner network. Extensive results on three public datasets (i.e., FashionIQ, Shoes, and CIRR) demonstrate that our proposed MLCL significantly outperforms the state-of-the-art methods under the hybrid cross-modal retrieval setting.

***Index Terms*—** Cross-modal retrieval, Multi-level semantic alignment, Feature fusion, Contrastive learning

1. INTRODUCTION

Multimodal retrieval is one of the most basic tasks in multimodal learning, aiming to retrieve data from one modality with data from another as a query. However, limiting search queries to a single modality is suboptimal in real-world applications. The textual description only provides an accurate but partial depiction of the desired result, as it is often difficult to ask the user to provide a complete text description. The visual queries are richer but more ambiguous because there is no clear definition of similarity between images. To alleviate this limitation, we can relax the restrictions on queries by allowing them to consist of data from more than one modality, i.e. retrieving images using a hybrid query consisting of a reference image and a text modifier, a task scenario we will call **Hybrid Cross-Modal Retrieval** (HCMR). The text modifier in the hybrid query explains how to modify the reference image to obtain the target image. The HCMR task has potential applications in e-commerce, where considering user intentions is crucial.

In this work, we propose a novel MLCL method to better fuse the hybrid query features for image retrieval. Two key components are designed to overcome the drawbacks of previous methods. The first is *cross-modal contrastive learning with multi-level semantic alignment*, which can more finely supervise the training process of the encoders and the combiner and extract more diverse features for retrieval. The second is *self-supervised semantic correlation learning* on the multi-level combined features, which helps to maintain the semantic representation capability of the encoders and the combiner. Extensive results show that our proposed MLCL significantly outperforms the state-of-the-art methods under the hybrid cross-modal retrieval setting. Our code is available at `https://github.com/JWargrave/MLCL`.

*Corresponding author

2. RELATED WORK AND CHALLENGES

The HCMR task has been addressed in a large number of works [1–16]. CoSMo [9] uses two different neural network modules, one for image style and one for image content. DCNet [10] proposes a Correction Network to model the difference between the reference and target image explicitly in the embedding space. CLIP4Cir [14] leverages CLIP [17] as the base network to integrate text and image features. The method employs a two-stage training approach. ARTEMIS [15] splits the hybrid cross-modal retrieval task into two separate retrieval tasks, namely text-to-image retrieval and image-to-image retrieval. However, this approach requires accessing all images in the database twice, leading to additional computational overhead during real deployment.

Despite significant progress in addressing the Hybrid Cross-Modal Retrieval (HCMR) task, current state-of-the-art methods still encounter three primary challenges. Firstly, the HCMR task involves data from multiple modalities, but many SOTA methods, such ARTEMIS [15], use a backbone pre-trained with unimodal data, which is not conducive to improving model performance. In our work, we employ the CLIP [17] model, a powerful model that achieves exceptional performance in multi-modal learning. Secondly, effective image retrieval requires understanding information at different levels of detail, from low-level visual features to high-level semantic concepts. However, many current models rely on feature fusion methods that only consider one level of detail, limiting their ability to extract information across different granularities. For example, recent models such as CLIP4Cir [14] use single-level matching methods, which do not facilitate the extraction of information at different granularities. Finally, with only simple supervision from cross-modal contrastive learning, the knowledge learned by the model from the original image or text is at risk of being lost. Our proposed MLCL method designs specialized mechanisms to address the above three major drawbacks.

3. METHODOLOGY

In this section, we give the details of our proposed **M**ulti-**L**evel **C**ontrastive **L**earning (MLCL) for hybrid cross-modal retrieval. In Section 3.1, we first introduce the hybrid cross-modal retrieval task setting. In Section 3.2, we further describe the overall architecture of our proposed MLCL. In Sections 3.3 and 3.4, cross-modal contrastive learning with multi-level semantic alignment and self-supervised semantic correlation learning are described in detail, respectively. In Section 3.5, we give the details of the training process and loss function.

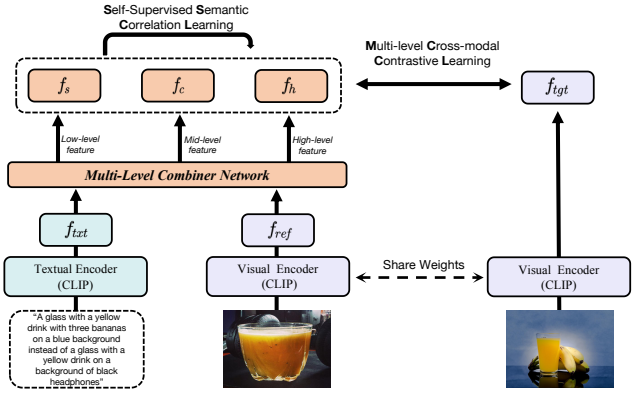

Fig. 1. The architecture of the proposed **MLCL** method, which consists of three trainable modules: the image encoder, the text encoder, and the combiner network. f_{ref}, f_{txt}, and f_{tgt} denote the L_2-normalized features of reference image, text modifier, and target image, respectively. The visual encoder used for feature extraction of the reference image and the target image is identical.

3.1. Hybrid Cross-Modal Retrieval

In this setting, a hybrid query is composed of a reference image $\mathbf{I_r}$ and a text modifier $\mathbf{T_m}$, which expresses some textual modification to the reference image. This task aims to retrieve the best matching image $\mathbf{I_t}$, which satisfies both the visual similarity constraints imposed by the reference image and the modification expressed by the text modifier.

3.2. Overall Architecture

As shown in Fig. 1, our MLCL consists of an image encoder $\Phi_I(\cdot)$, a text encoder $\Phi_T(\cdot)$, and a multi-level combiner network $\psi_{combiner}$ in general. Given the reference/target image and the text modifier, we first adopt the image encoder to extract the reference/target image features $f_{ref} = \Phi_I(\mathbf{I_r})/f_{tar} = \Phi_I(\mathbf{I_t})$ and a textual encoder to extract the text modifier feature $f_{txt} = \Phi_T(\mathbf{T_m})$. Then we fuse the reference image feature and the text modifier feature into three multi-level combined features $\{f_s, f_c, f_h\}$ by a *combiner* module: $\{f_s, f_c, f_h\} = \psi_{combiner}(f_{ref}, f_{txt})$.

Once all the features are extracted, we utilize two kinds of training objectives to optimize our MLCL: ***c**ross-modal **c**ontrastive **l**earning with **m**ulti-level semantic alignment* (**mCCL** for short), and ***s**elf-supervised **s**emantic **c**orrelation **l**earning* (**SSCL** for short). The two modules are described separately below.

3.3. Cross-Modal Contrastive Learning with Multi-Level Semantic Alignment

In order to perform effective image retrieval, both the encoders and the combiner module must well understand the semantics of the reference/target image and the text modifier for heterogeneous modality information fusion. To more finely supervise the training process of the encoders and the combiner, we propose three different fusion mechanisms to capture different degrees of integration features. In Fig. 2, we depict the combiner network, alongside three fusion mechanisms: **SimSum**, **CatCombined**, and **HybridFusion**. Each of these fusion mechanisms is described in detail below.

SimSum utilizes the concept proposed by CLIP to map various types of data into a standardized space and ensures the principle of additivity as described in [17]. Thus, in order to derive the first combined query feature f_s, we perform the summation of both visual and textual features:

$$f_s = f_{ref} + f_{txt} \tag{1}$$

where f_{ref}, f_{txt} denote the L_2-normalized features of reference image and text modifier extracted by the encoders, respectively.

CatCombined improves *SimSum* by introducing an additional learnable fusion module to better fuse the visual features and text features. This fusion module is based on the concatenation of the two query features. Specifically, we project the two query features via a linear layer followed by the ReLU function. Projected features are then concatenated and fed to three linear layers with the ReLU function to obtain the second combined features f_c:

$$\begin{aligned} f_m &= cat(MLP_{img}(f_{ref}), MLP_{txt}(f_{txt})) \\ f_c &= \text{Linear}(MLP_2(f_m)) \end{aligned} \tag{2}$$

where *MLP* denotes one to three linear layers with ReLU functions, *cat* denotes the concatenation operation, and Linear denotes one linear layer. While the Linear and MLP_2 modules can be merged into a single MLP, we have opted to present them as distinct entities for the sake of consistency with the notation employed later in this paper.

HybridFusion simultaneously combines *SimSum* and *CatCombined* for better information fusion. We replace the direct addition of two query features in *SimSum* with a convex combination of them. Specifically, the projected features mentioned above are concatenated and fed to two branches with a similar structure: two linear layers with a ReLU function. The first branch aims to compute the coefficients of a convex combination between the image and text features with a sigmoid function, while the second branch does not use the sigmoid function. Then f_h is obtained by adding the convex combination to the output of the second branch, which can be seen as a form of residual connection [18] and help to integrate different levels of features:

$$\begin{aligned} &\alpha = sigmoid(MLP_{con}(f_m)) \\ &Convex(f_{ref}, f_{txt}) = \alpha * f_{ref} + (1-\alpha) * f_{txt} \\ &f_h = MLP_2(f_m) + Convex(f_{ref}, f_{txt}) \end{aligned} \tag{3}$$

where *Convex* denotes a convex combination of two features, and its coefficient α is a learned scalar.

As shown in Fig. 2, we utilize the above three fusion mechanisms to fuse the reference image feature and the text modifier feature and obtain multi-level semantic integrated query features (f_h, f_c and f_s). During the training phase, we employ cross-modal contrastive learning to achieve multi-level semantic alignment. This process involves minimizing the distances between the three combined features and the target image feature, as described in Section 3.5. During the inference phase, we directly fuse the three combined features as the final query feature f_{com}:

$$f_{com} = f_h + f_c + f_s \tag{4}$$

3.4. Self-Supervised Semantic Correlation Learning

Simple supervision from cross-modal contrastive learning forces the data of different modalities to be aligned (i.e., the fused query features and the target image features), which decreases the semantic correlation of the fused features and increases their sensitivity to data

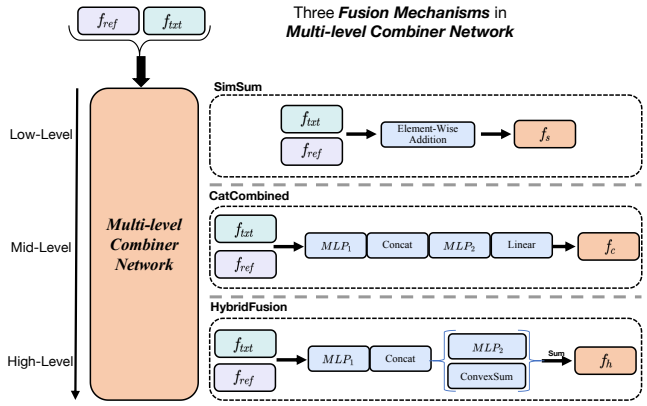

Fig. 2. The architecture of the **Multi-Level Combiner** network. f_{ref}, f_{txt}, and f_{tgt} denote the L_2-normalized features of reference image, text modifier, and target image, respectively. The MLP module has one to two linear layers with the ReLU function. MLPs with the same subscript j share the parameters ($j = 1, 2$). Moreover, MLP_1 is actually composed of two MLPs, namely MLP_{img} and MLP_{txt}, which operate on the reference feature and the text modifier feature, respectively. ConvexSum consists of two linear layers with the ReLU function followed by a sigmoid function and returns a convex combination of f_{txt} and f_{ref}, which is not depicted in the figure provided.

perturbations. Ensuring robustness to data perturbation is crucial. Therefore, we introduce self-supervised semantic correlation learning (SSCL) in our MLCL to enhance the semantic correlation of the fused features. Specifically, we conduct self-supervised contrastive learning on the three combined features (with augmented image-text pair as the target). To obtain the augmented fusion features, we apply data augmentation to the reference image, where a wide range of augmentation techniques (e.g., random resized cropping, rotation, horizontal and vertical flipping) are used.

3.5. Training Pipeline

The training of the whole model is performed with batches of triplets formed by: reference images, text modifiers, and target images. We employ the InfoNCE [19] loss to conduct contrastive learning:

$$L_{IN}(q,k) = \frac{1}{B}\sum_{i=1}^{B} -\log\left\{\frac{e^{\lambda * \psi_{sim}(q_i,k_i^+)}}{\sum_{j=1}^{B} e^{\lambda * \psi_{sim}(q_i,k_j)}}\right\} \quad (5)$$

where B denotes the batch size and the similarity function ψ_{sim} is implemented as cosine similarity. k_i^+ denotes the positive sample (paired with q_i) in the InfoNCE loss. λ is a temperature parameter that controls the range of the logits, and it is a learnable parameter.

According to Sections 3.3 and 3.4, the mCCL and SSCL losses can be formulated as:

$$\begin{aligned} L_{mCCL} &= L_{IN}(f_s, f_{tgt}) + L_{IN}(f_c, f_{tgt}) + L_{IN}(f_h, f_{tgt}) \\ L_{SSCL} &= L_{IN}(f_s, \widetilde{f_s}) + L_{IN}(f_c, \widetilde{f_c}) + L_{IN}(f_h, \widetilde{f_h}) \end{aligned} \quad (6)$$

where f_{tgt} denotes the target image features, $\{\widetilde{f_s}, \widetilde{f_c}, \widetilde{f_h}\}$ are the corresponding augmented fusion features.

The full loss of our MLCL is:

$$L_{MLCL} = L_{mCCL} + L_{SSCL} \quad (7)$$

Table 1. Comparative results on the **FashionIQ** validation set. The best score is bolded and the second best score is underlined.

Method	Shirt		Dress		Toptee		Average	
	$R@10$	$R@50$	$R@10$	$R@50$	$R@10$	$R@50$	$R@10$	$R@50$
VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$) [7]	21.03	42.75	21.47	43.83	26.71	51.81	23.07	46.13
VAL (GloVe) [7]	22.38	44.15	22.53	44.00	27.53	51.68	24.15	46.61
ARTEMIS (RN50) [15]	21.05	44.18	27.34	51.71	24.91	49.87	24.43	48.59
MAAF [11]	21.30	44.20	23.80	48.60	27.90	53.60	24.30	48.80
ARTEMIS (RN50) [15]	21.78	43.64	27.16	52.40	29.20	54.83	26.05	50.29
CurlingNet [13]	21.45	44.56	26.15	53.24	30.12	55.23	25.90	51.01
CoSmo [9]	24.90	49.18	25.64	50.30	29.21	57.46	26.58	52.31
AACL [23]	24.82	48.85	29.89	55.85	30.88	56.85	28.53	53.85
DCNet [10]	23.95	47.30	28.95	56.07	30.44	58.29	27.78	53.89
SAC w/BERT [24]	28.02	51.86	26.52	51.01	32.70	61.23	29.08	54.70
CLIP4Cir (RN50) [14]	35.77	57.02	31.73	56.02	36.46	62.77	34.65	58.60
CLIP4Cir (RN50x4) [14]	39.99	60.45	33.81	59.40	41.41	65.37	38.32	61.74
MLCL$_{(mCCL+SSCL)}$	**43.33**	**64.67**	**38.92**	**63.91**	**47.99**	**70.02**	**43.41**	**66.20**

4. EXPERIMENTS

4.1. Datasets and Metrics

We conduct comparative experiments on three public benchmark datasets, all of which make use of human-written textual modifiers in natural language. (1) The **Fashion IQ** [20] dataset contains 18k training triplets (46.6k images) and 12k test triplets (15.5k images). (2) The **Shoes** [21] dataset contains 9k training triplets (10k images) and 1.7k test triplets (4.7k images). (3) The **CIRR** [22] dataset is composed of 36k pairs of open-domain images, arranged in a 80%-10%-10% split between the train/validation/test.

In our performance evaluation on FashionIQ, we use the average recall at rank K (R@K) as the evaluation metric. Specifically, we consider two ranks: 10 and 50. For the Shoes dataset, we report the average recall at rank K for three ranks: 1, 10, and 50. As for CIRR, we report the average recall at rank K for four ranks: 1, 5, 10, and 50. Additionally, in line with prior work [22], we also report the subset recall at rank K ($Recall_{subset}@K$) for CIRR. This metric restricts the candidate target images to those that are semantically similar to the correct target image.

4.2. Implementation Details

Our MLCL model consists of three modules that need to be trained: the image encoder, the text encoder, and the combiner network. We initialize the encoders with the pre-trained CLIP [17] model. The computation of combined feature f_s only involves two encoders and is independent of the combiner network. Therefore, instead of training the three combined features simultaneously, we employed an asynchronous training strategy. Specifically, the training of the entire model is divided into two stages. In the first stage, we solely compute f_s and conduct a contrastive learning between it and the target image's features. During this stage, we only fine-tune the image and text encoders while keeping the parameters of the combiner network frozen. We refer to this stage as the fine-tuning stage. In the second stage, we exclusively train the combiner network while keeping the parameters of the encoders frozen. We compute f_c and f_h, which are extracted by the combiner network, and then conduct contrastive learning between these two combined features and the features of the target image. We refer to the second stage as the combiner training stage. In both stages, we use cosine similarity as a feature-to-feature similarity function.

Table 2. Comparative results on the **Shoes** validation set. The best score is bolded and the second best score is underlined.

Method	R@1	R@10	R@50	$(\Sigma R@K)/3$
FiLM [25]	10.19	38.89	68.30	39.13
MRN [26]	11.74	41.70	67.01	40.15
TIRG [1]	12.60	45.45	69.39	42.48
VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$) [7]	16.98	49.83	73.91	46.91
CoSmo [9]	16.72	48.36	75.64	46.91
DATIR [27]	17.20	51.10	75.60	47.97
VAL (GloVe) [7]	17.18	51.52	75.83	48.18
ARTEMIS (RN50-LSTM) [15]	17.60	51.05	76.85	48.50
ARTEMIS (RN50-BiGRU) [15]	18.72	53.11	79.31	50.38
MLCL$_{(mCCL+SSCL)}$	**22.71**	**57.41**	**81.43**	**53.85**

Table 3. Comparative results on the **CIRR** test set. The best score is bolded and the second best score is underlined. † denotes results cited from [22].

Method	$Recall@K$				$Recall_{subset}@K$		
	K=1	K=5	K=10	K=50	K=1	K=2	K=3
TIRG† [1]	14.61	48.37	64.08	90.03	22.67	44.97	65.14
TIRG+LastConv† [1]	11.04	35.68	51.27	83.29	23.82	45.65	64.55
MAAF† [11]	10.31	33.03	48.30	80.06	21.05	41.81	61.60
MAAF+BERT† [11]	10.12	33.10	48.01	80.57	22.04	42.41	62.14
MAAF-IT† [11]	9.90	32.86	48.83	80.27	21.17	42.04	60.91
MAAF-RP† [11]	10.22	33.32	48.68	81.84	21.41	42.17	61.60
ARTEMIS [15]	16.96	46.10	61.31	87.73	39.99	62.20	75.67
CIRPLANT† [22]	15.18	43.36	60.48	87.64	33.81	56.99	75.40
CIRPLANT w/O† [22]	19.55	52.55	68.39	92.38	39.20	63.03	79.49
CLIP4Cir (RN50) [14]	35.81	68.80	80.17	95.25	66.96	85.25	93.13
CLIP4Cir (RN50x4) [14]	38.53	69.98	81.86	95.93	68.19	85.64	94.17
MLCL$_{(mCCL+SSCL)}$	**43.18**	**76.77**	**87.16**	**97.88**	**70.84**	**87.40**	**95.18**

4.3. Comparison with State-of-the-Arts

Table 1, 2, and 3 show the comparative results between our MLCL method and the current state-of-the-art models on the FashionIQ, Shoes, and CIRR datasets, respectively. It can be clearly seen that: (1) Compared with the SOTA models, our method achieves 5.09%, 4.30% and 5.30% improvement in Recall@10 on three datasets, respectively. (2) The significant improvements over CLIP4Cir [14] and ARTEMIS [15] demonstrate that our devised SSCL and mCCL modules can indeed better learn the interaction and fusion between different modalities for hybrid cross-modal retrieval. (3) Our method is capable of achieving superior results, whether applied to a professional dataset within a designated field or a general dataset without any discernible preference.

4.4. Ablation Study

In this section, we conduct ablation experiments on the validation set of the open domain dataset CIRR and the fashion domain dataset Shoes to evaluate the influence of several design choices in our architecture. The obtained results are reported in Table 4 and 5. Concretely, to demonstrate the effectiveness of SSCL and mCCL, we start from several *single-level semantic alignment* baselines, where only one of the three combined features $\{f_h, f_c, f_s\}$ is employed during both training and evaluation phases. Note that employing different combined features activates the gradients of different parts of the model during the training process. Additionally, in the case of f_s, the effect of fine-tuning the image or text encoder is evaluated separately. After establishing baselines, we introduce one combined feature at a time until all three are incorporated, resulting in what we refer to as *multi-level semantic alignment*. Finally, we fuse SSCL into our full method.

Table 4. Ablation study results on the **CIRR** validation set. The best score is bolded, the second best score is underlined with a solid line, and the third best score is underlined with a dashed line. Note that "I" denotes fine-tuning the image encoder and "T" denotes fine-tuning the text encoder. $AR_{51} - (R@5 + R_s@1)/2$.

mCCL				SSCL	$Recall@K$				$Recall_{sub}@K$			AR_{51}	Avg
f_c	f_h	f_s I	f_s T		K=1	K=5	K=10	K=50	K=1	K=2	K=3		
✓					27.55	59.89	73.81	94.09	57.81	78.35	89.17	58.85	68.67
	✓				31.14	64.46	78.16	95.26	61.73	81.46	91.49	63.09	71.96
		✓			34.63	68.24	80.41	95.96	60.56	81.10	90.94	64.40	73.13
			✓		33.96	68.14	80.58	95.58	66.35	84.98	92.71	67.24	74.61
		✓	✓		41.33	74.46	84.79	96.80	68.12	86.41	94.26	71.29	78.02
✓	✓				32.17	65.73	79.72	95.43	61.83	81.87	91.10	63.78	72.55
✓	✓	✓	✓		44.42	76.87	87.28	**97.44**	70.89	**87.66**	**94.76**	73.88	79.90
✓	✓	✓	✓	✓	**45.75**	**78.14**	**87.75**	97.35	**71.68**	87.42	94.74	**74.91**	**80.40**

Table 5. Ablation study results on the **Shoes** validation set. The best score is bolded, the second best score is underlined with a solid line, and the third best score is underlined with a dashed line. Note that "I" denotes fine-tuning the image encoder and "T" denotes fine-tuning the text encoder.

mCCL				SSCL	R@1	R@10	R@50	$(\sum_K R@K)/3$
f_c	f_h	f_s I	f_s T					
✓					13.17	43.04	69.90	42.04
	✓				15.79	47.02	71.04	44.61
		✓			18.91	51.39	77.12	49.14
			✓		12.72	40.15	66.21	39.69
		✓	✓		20.56	54.51	79.67	51.58
✓	✓				16.75	47.36	72.12	45.41
✓	✓	✓	✓		22.49	56.27	81.03	53.27
✓	✓	✓	✓	✓	**22.71**	**57.41**	**81.43**	**53.85**

From Table 4 and 5, we find that when more levels of combined features are used, higher average recalls can be obtained. These results demonstrate that features at different levels can capture details at varying granularities, and their combination enables a more comprehensive understanding of multimodal data. Furthermore, in our experiments, we find that the proposed SSCL enhances the robustness of our model. Concretely, without using SSCL, the training process quickly converges to local minima, resulting in no further improvement in the recall rate.

5. CONCLUSION

In this paper, we propose a multi-level contrastive learning (MLCL) method for image retrieval with a hybrid query. Two key components are carefully designed to overcome the drawbacks of previous methods. The first is cross-modal contrastive learning with multi-level semantic alignment (mCCL), which can more finely supervise the training of the whole network and extract more diverse features for retrieval. The second is self-supervised semantic correlation learning (SSCL), which helps to enhance the semantic correlation of the multi-level combined features of the model. Experiments on Shoes, FashionIQ, and CIRR show that our MLCL method achieves new state-of-the-art performance.

6. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (62376274).

7. REFERENCES

[1] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays, "Composing text and image for image retrieval-an empirical odyssey," in *CVPR*, 2019, pp. 6439–6448.

[2] Mehrdad Hosseinzadeh and Yang Wang, "Composed query image retrieval using locally bounded features," in *CVPR*, 2020, pp. 3596–3605.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, vol. 28, 2015.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[5] Feifei Zhang, Mingliang Xu, Qirong Mao, and Changsheng Xu, "Joint attribute manipulation and modality alignment learning for composing text and image to image retrieval," in *ACM Multimedia*, 2020, pp. 3367–3376.

[6] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[7] Yanbei Chen, Shaogang Gong, and Loris Bazzani, "Image search with text feedback by visiolinguistic attention learning," in *CVPR*, 2020, pp. 3001–3011.

[8] Minchul Shin, Yoonjae Cho, Byungsoo Ko, and Geonmo Gu, "Rtic: Residual learning for text and image composition using graph convolutional network," *arXiv preprint arXiv:2104.03015*, 2021.

[9] Seungmin Lee, Dongwan Kim, and Bohyung Han, "Cosmo: Content-style modulation for image retrieval with text feedback," in *CVPR*, 2021, pp. 802–812.

[10] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim, "Dual compositional learning in interactive image retrieval," in *AAAI*, 2021, pp. 1771–1779.

[11] 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.

[12] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber, "Compositional learning of image-text query for image retrieval," in *WACV*, 2021, pp. 1140–1149.

[13] Youngjae Yu, Seunghwan Lee, Yuncheol Choi, and Gunhee Kim, "Curlingnet: Compositional learning between images and text for fashion iq data," *arXiv preprint arXiv:2003.12299*, 2020.

[14] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo, "Conditioned and composed image retrieval combining and partially fine-tuning clip-based features," in *CVPRW*, 2022, pp. 4955–4964.

[15] Ginger Delmas, Rafael S Rezende, Gabriela Csurka, and Diane Larlus, "Artemis: Attention-based retrieval with text-explicit matching and implicit similarity," in *ICLR*, 2022.

[16] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo, "Conditioned image retrieval for fashion using contrastive learning and clip-based features," in *ACM Multimedia Asia*, 2022.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[19] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[20] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris, "Fashion iq: A new dataset towards retrieving images by natural language feedback," in *CVPR*, 2021, pp. 11307–11317.

[21] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris, "Dialog-based interactive image retrieval," *NeurIPS*, vol. 31, 2018.

[22] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould, "Image retrieval on real-life images with pre-trained vision-and-language models," in *ICCV*, 2021, pp. 2125–2134.

[23] Yuxin Tian, Shawn Newsam, and Kofi Boakye, "Image search with text feedback by additive attention compositional learning," *arXiv preprint arXiv:2203.03809*, 2022.

[24] Surgan Jandial, Pinkesh Badjatiya, Pranit Chawla, Ayush Chopra, Mausoom Sarkar, and Balaji Krishnamurthy, "Sac: Semantic attention composition for text-conditioned image retrieval," in *WACV*, 2022, pp. 4021–4030.

[25] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018.

[26] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang, "Multimodal residual learning for visual qa," *NeurIPS*, vol. 29, 2016.

[27] Chunbin Gu, Jiajun Bu, Zhen Zhang, Zhi Yu, Dongfang Ma, and Wei Wang, "Image search with text feedback by deep hierarchical attention mutual information maximization," in *ACM Multimedia*, 2021, pp. 4600–4609.