EasyChair Preprint

№ 9930

ITCONTRAST: Contrastive Learning with Hard Negative Synthesis for Image-Text Matching

Fangyu Wu, Qiufeng Wang, Qi Chen, Yushi Li, Bailing Zhang and Eng Gee Lim

April 6, 2023

ITCONTRAST: Contrastive Learning with Hard Negative Synthesis for Image-Text Matching

Abstract

Image-text matching aims to bridge vision and language so as to match the instance of one modality with the instance of another modality. Recent years have seen considerable progress in the research area by exploring local alignment between image regions and sentence words. However, how to learn modality-invariant feature embedding and make use of the hard negatives in the training set to infer more accurate matching scores are still open questions. In this paper, we attempt to solve these problems by introducing a new Image-Text Modality Contrastive Learning (abbreviated as ITContrast) approach for image-text matching. Specifically, a pre-trained vision-language model OSCAR is firstly fine-tuned to obtain the visual and textual features, and a hard negative synthesis module is then introduced to leverage the hardness of negative samples, which features of profiling negative samples in a mini-match and generating their representatives to reflect the hardness relations to the anchor. A novel cost function is designed to comprehensively combine the knowledge of positives, negatives and synthesized hard negatives. Extensive experiments on the MS-COCO and Flickr30K datasets demonstrate that our approach is effective for image-text matching.

Keywords: Multimodal Deep Learning, Hard Negative Synthesis, Contrastive Learning, Image-text Matching

1 Introduction

Modality often refers to a specific way in which people receive information. The technique of learning across multiple modalities such as vision and language simultaneously is important for many cross-modal tasks. One of the fundamental multi-model learning techniques is image-text matching, which is to measure the similarity between an image and a text. This is related to many important cross-modal tasks, such as semantic image retrieval, image description, visual QA and so on.

The main challenges in image-text matching include the heterogeneity gap and semantic gap. Heterogeneity gap means inconsistent feature representation of the data from different modalities of image and text while the semantic gap refers to the misalignment in capturing the cross-modal correspondence between image and text. Much effort has been endeavoured to find the solutions. Many works [1, 2] have been published with the shared aim to learn a joint embedding space where related image and sentence instances are located close to each other and to measure the image-text relevance by computing the distance between global representations. Currently, approaches [3, 4, 5] based on attention mechanism have shown advantages in aligning image and text by discovering more fine-grained cross-modal correspondence. While these approaches have made noticeable improvements, there are several limitations that we'd like to address and improve.

Firstly, most existing works bridge the heterogeneity gap by exploiting some pre-trained modules such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to extract the image and text features, and the joint embedding learning is performed on these extracted features. These feature extractors are not trained or fine-tuned for discriminative image-text embedding, and the optimality of the learned representation is questionable.

Secondly, many recent image-text matching approaches [3, 6] exploit a triplet loss to encourage the model to predict higher similarity scores between positive image-text pairs than negative ones. The cost function design does not sufficiently take the hardness of negative samples into consideration, which is one of the main causes of weak generalization. The work in [7] also shows that increasing the batch size to obtain more negatives will lead to a sharp increase in computational complexity and diminishing returns in terms of performance.

To address the aforementioned problems, we introduce ITContrast: an Image-Text Modality Contrastive Learning method that bridges the heterogeneity gap and learns a discriminative and modality-invariant embedding space. Specifically, we start with learning discriminative cross-modal embedding while proceeding with the semantic alignments between images regions and sentence fragments. The text and image inputs are jointly processed by multiple Transformer layers in the OSCAR model [8], which is pre-trained with 6.5 million text-image pairs. By leveraging the fine-tuning capabilities of OSCAR, we capture the intricate associations between text and image and learn more discriminative image-text embedding. Then we propose a SynHNC module, standing for "(Syn)thesizing the (H)ard (N)egative from the (C)lusters", to directly synthesize hard negatives in the embedding space, which is adaptive to each anchor. Specifically, given each anchor in the training set, we cluster its negative samples into the representative embedding groups of semantically similar vectors. After that, SynHNC compares the anchor with prototypes in each negative group to obtain the nonlinear relationship, and then synthesizes the hard negatives by aggregating the attentive contributions from all prototypes. Therefore, the negative prototypes in a cluster near the anchor acquire more weights in the synthesis process. This strategy flexibly increases the hardness of negative samples and enhances the discriminative ability of the image-text matching model. The training of the ITContrast is based on contrastive learning by proposing a novel InfoCMR loss with cross-modal data being taken into account.

Our main contributions are summarized as follows:

- We propose a novel contrastive learning method called ITContrast for training an image-text matching model, which consists of a hard negative synthesis (SynHNC) module for each positive query and integrating with feature learning in a plug-and-play manner.
- A new InfoCMR loss is introduced to enhance the embedding space where positive image-text pairs are close while dissimilar pairs are farther apart. We demonstrate that the pre-trained OSCAR model can be successfully fine-tuned using InfoCMR loss to boost the image-text matching capabilities.
- The state-of-the-art performance on two publicly datasets shown that contrastive learning is well-suited for image-text matching and that it results in modality-invariant embedding. We further demonstrate the generality of our method based on the popular models: SCAN [3] and SGRAF [5].

2 Related work

Image-Text Matching. Early image-text matching methods [1, 2] map the image and text into a shared global embedding space with deep neural networks, in which the image-text relevance can be directly measured using the cosine similarity or inner product. Recent methods [3, 6] further expanded the exploration of relationships between words and image objects. Lee et al. [3] proposed to discover the local alignments, which produce impressive results and inspire a surge of works [4, 5, 9, 10] to explore accurate fine-grained correspondence. However, these works pay less attention to the hardness of negative sample, leading to weak generalization. Chen et al. [11] sampled offline hard negatives from the training set, while the improvement of the model performance is limited as the hard negatives in the training set are still not sufficient. In this paper, we develop a hard negative synthesis method for image-text matching, which can effectively construct hard negative pairs in the embedding space.

Contrastive Learning. Contrastive learning is a self-supervised learning methodology that for-

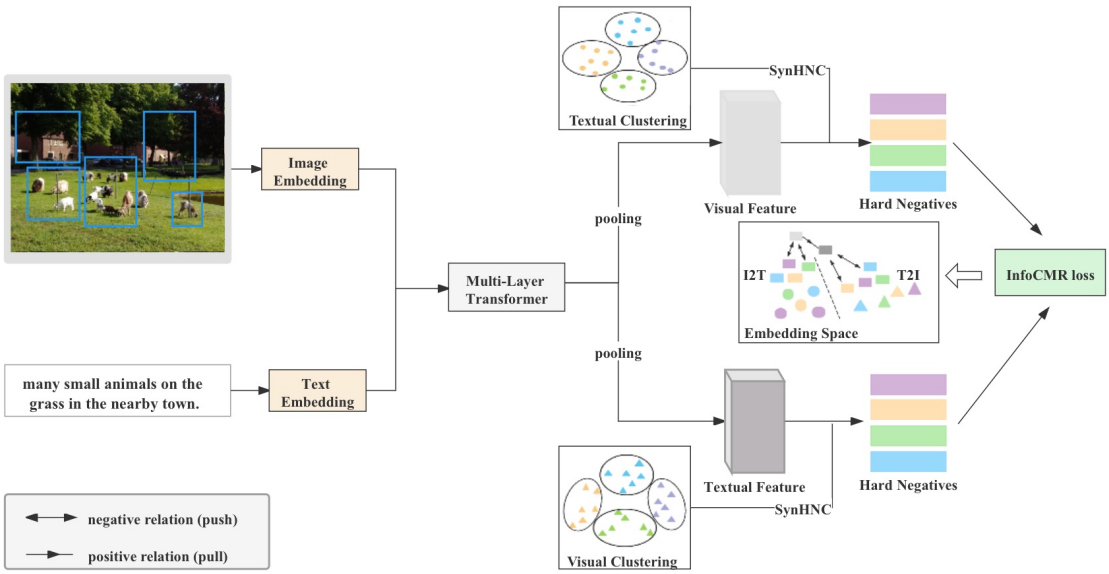

Figure 1: Illustration of the proposed method. ITContrast firstly encodes the image-text pair into the visual feature and textual feature, followed by the SynHNC module to synthesize challenging negatives based on clustering. The InfoCMR loss performs image-text contrastive matching from the embedding space and narrows the heterogeneous gap.

mulates the learning task as finding and encoding similar and dissimilar objects. The core idea is to map a data item (anchor) and its various augmented versions (positive samples) to an embedding space where they are close together and are separated from other different items (negative samples) [12]. Chen et al. introduced SimCLR [7] to achieve promising results compared with a supervised ResNet50 model. MoCo [12] utilized a memory bank to improve the efficiency of contrastive learning with small batch size. Khosla et al. [13] extended contrastive learning to supervised learning, allowing the model to leverage label information. Inspired by these works, we propose a novel contrastive learning method with effective implementations for image-text matching.

3 Proposed Method

In this section, we present our Image-Text Modality Contrastive Learning method (ITContrast). Fig. 1 illustrates an overview of our approach.

3.1 Joint Feature Learning

Text Embedding. Given a sentence c, we split it into s words with tokenization technique, and use the Oscar-base token embedding E_{tok} to represent each token. Let a sentence be $c = \{o_1, ..., o_z\}$ after tokenization, we have:

$$\hat{o}_i = E_{tok}(o_r) \tag{1}$$

where o_i is the i^{th} token of the sentence. Therefore, a sentence is represented as $\hat{c} = \{\hat{o}_1, ..., \hat{o}_z\}$, $\hat{o}_i \in \mathbb{R}^{d_H}$.

Image Embedding. For each input image i with n regions of objects, we pass it through Faster R-CNN [14] which is pre-trained on Visual Genome dataset to extract the regional visual features $v_{re} \in \mathbb{R}^{n \times d_{rcnn}}$ and region position $v_{pos} \in \mathbb{R}^{n \times d_{pos}}$. After that, the v_{re} and v_{pos} are concatenated as $\hat{v}$ using a linear projection to ensure that image has the same dimension of text embedding. Given a pair of image and text local embeddings, we use a single transformer in OSCAR to get the joint feature representation of the image-text pair. Afterwards, the global visual representation $\boldsymbol{v}$ and global textual representation $\boldsymbol{c}$ are computed by a mean-pooling operation over all the local features.

3.2 Hard Negative Synthesis Module

Fig.2 shows a t-SNE [15] plot after running SCAN [3] using online triplet loss [2] on feature embeddings. It can be clearly observed that anchor has more easy negatives and fewer hard ones, i.e. many negatives are too far to contribute to the online triplet loss. Therefore, it is significant to get more meaningful negative samples and increase their "hardness" degree.

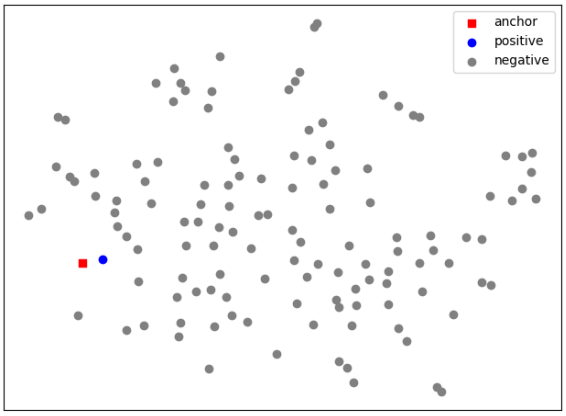

Figure 2: The t-SNE plot within a mini-batch from MSCOCO.

Hard Negatives Learning. To increase the discriminative ability of ITContrast, we propose the SynHNC module to synthesize hard negatives based on the clustering results. Without loss generality, the samples in each image-text pair are defined as anchor $\boldsymbol{q}$, which represent either image $\boldsymbol{v}$ or text $\boldsymbol{c}$. SynHNC firstly performs global semantic clustering for each sample in the training set, i.e., we perform k-means on all its non-corresponding negatives in a mini-batch to obtain clusters $G = \{g_1, g_2, ..., g_m\}$. After that, hard negatives are learnt based on a Kernel Associative Memory (KAM) [16] algorithm which involves two phases, an encoding phase and a reconstruction phase. During the encoding phase, kernel operations encode anchor $\boldsymbol{q}$ to obtain its nonlinear relationship with all of the prototypes in each negative cluster. In the reconstruction phase, the anchor are associated with the prototypes as expected hard negatives.

More specifically, with N negative prototypes $\{x_{j1}, x_{j2}, ..., x_{jN}\}$ from cluster g_j, a similarity measure can be defined between $\boldsymbol{q}$ and each prototype by employing a kernel function as,

$$k(\boldsymbol{q}, x_{jn}) =< \Phi(\boldsymbol{q}) \cdot \Phi(x_{jn}) > \tag{2}$$

where x_{jn} is the n^{th} sample in cluster j. A kernel corresponding to Φ, implicitly obtains some nonlinear relationships between negative pair $(\boldsymbol{q}, x_{jn})$. A popular option of k is Gaussian radial basis functions:

$$k(\boldsymbol{q}, x_{jn}) = exp(-\frac{||\boldsymbol{q} - x_{jn}||^2}{2\sigma^2}) \tag{3}$$

Then, a hard negative sample $\overline{\boldsymbol{q}}_h^j$ can be obtained for $\boldsymbol{q}$ based on the reconstruction process from the normalized kernels, e.g.,

$$\overline{\boldsymbol{q}}_h^j = \frac{\sum_{n=1}^{N} w_n k(\boldsymbol{q}, x_{jn})}{\sum_n k(\boldsymbol{q}, x_{jn})} \tag{4}$$

where w_n can be regarded as the reconstruction weight of each prototype in synthesizing hard negatives, which is determined by the following least square objective:

$$J(W) = min||\mathbf{X} - \mathbf{Wk}|| \tag{5}$$

where $\mathbf{X}$ is a matrix in which the n^{th} column is x_n, and $\mathbf{k}$ is a vector in which the n^{th} element is equal to $k(\boldsymbol{q}, x_n)$. Generally, the optimum values for the weights can be obtained by using LS approximation from Eq. (5). After that, the synthetic points $\overline{\boldsymbol{v}}_h = \{\overline{\boldsymbol{v}}_h^i\}_{i=1}^m$ and $\overline{\boldsymbol{c}}_h = \{\overline{\boldsymbol{c}}_h^i\}_{i=1}^m$ are used as hard negatives of $\boldsymbol{c}$ and $\boldsymbol{v}$ to perform contrastive learning.

3.3 InfoCMR Loss

Following the generic idea of cross-model contrastive matching, we propose an InfoCMR loss to contrast different image-text pairs, including $(\boldsymbol{v}, \boldsymbol{c})$ and negative pairs$\{(\boldsymbol{v}, \overline{\boldsymbol{c}}), (\boldsymbol{t}, \overline{\boldsymbol{v}}), (\boldsymbol{v}, \overline{\boldsymbol{c}}_h), (\boldsymbol{c}, \overline{\boldsymbol{v}}_h)\}$. Specifically, for the view of retrieving text with image, we get positive similarity s^{vc+} by calculating cosine similarity between $\boldsymbol{v}$ and $\boldsymbol{c}$. Then, we obtain negative similarity s^{vc-} by calculating cosine similarity among $\boldsymbol{v}$ and negative text representations in $\overline{C} = \overline{\boldsymbol{c}} \cup \overline{\boldsymbol{c}}_h$. Thus, we achieve $S^{vc} = s^{vc+} \cup s^{vc-}$. Similarly, from the view of retrieving image with text, we get $S^{cv} = s^{cv+} \cup s^{cv-}$. The loss function for an image-text pair can be written as:

Methods	MSCOCO						Flickr30K					
	Sentence Retrieval			Image Retrieval			Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
MMCA [17]	74.8	95.6	97.7	61.6	89.8	95.2	74.2	92.8	96.4	54.8	81.4	87.8
CAAN [18]	75.5	95.4	98.5	61.3	89.7	95.2	70.1	91.6	97.2	52.8	79.0	87.9
CASC [19]	72.3	96.0	99.0	58.9	89.8	96.0	68.5	90.7	95.9	50.2	78.3	86.3
IMRAM [20]	76.7	95.6	98.5	61.7	89.1	95.0	74.1	93.0	96.6	53.9	79.4	87.2
DP-RNN [4]	75.3	95.8	98.6	62.5	89.7	95.1	70.2	91.6	95.8	55.5	81.3	88.2
UWML [21]	76.8	96.2	98.5	60.9	89.0	95.5	73.1	92.7	96.8	54.2	79.9	87.3
GSMN [22]	78.4	96.4	98.6	63.3	90.1	95.7	76.4	94.3	97.3	57.4	82.3	89.0
CVSE [23]	74.8	95.1	98.3	59.9	89.4	95.2	73.5	92.1	95.8	52.9	80.4	87.8
SGRAF [5]	79.6	96.2	98.5	63.2	90.7	96.1	77.8	94.1	97.4	58.5	83.0	88.8
Unicoder-VL[24]	84.3	97.3	99.3	69.7	93.5	97.2	86.2	96.3	99.0	71.5	90.9	94.9
UNITER [25]	-	-	-	-	-	-	85.9	97.1	98.8	72.5	92.4	96.1
VSE∞ [26]	85.6	98.0	99.4	73.1	94.3	97.7	88.7	98.9	99.8	76.1	94.5	97.1
Fast and Slow [27]	-	-	-	-	-	-	-	-	-	72.1	91.5	95.2
OSCAR [8]	88.4	99.1	99.8	75.7	95.2	98.3	88.5	98.5	99.2	79.8	93.3	96.6
Ours (SGRAF)	90.1	99.3	99.3	76.9	95.9	98.5	89.3	98.3	99.1	81.8	94.8	97.0
Ours (ITContrast)	**92.5**	**99.6**	**99.8**	**79.6**	**97.4**	**99.4**	**92.8**	**98.9**	**99.3**	**84.1**	**96.2**	**98.1**

Table 1: Comparison of image-text matching results on MSCOCO 1K test set and Flickr30K test set.

$$\mathcal{L}(\boldsymbol{v},\boldsymbol{c}) = -log\frac{e^{(s^{vc+}/\tau)}}{\sum_{i=1}^{1+|\overline{C}|} e^{(S_i^{vc}/\tau)}} - log\frac{e^{(s^{cv+}/\tau)}}{\sum_{i=1}^{1+|\overline{V}|} e^{(S_i^{cv}/\tau)}} \tag{6}$$

where τ is a temperature parameter, $|\cdot|$ is the size of set.

We further introduce an additional penalty term to avoid overfitting. More specifically, we randomly sample Z Gaussian noise vectors from a Gaussian distribution with the same dimensions as the anchor vector. These vectors constitute high confident negative pairs with each sample in the batch to smooth the representation space. Note that these Gaussian noise vectors will not participate in the positive pair constitution. Accordingly, we defined the InfoCMR as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{v},\boldsymbol{c}) = &- log\frac{e^{(s^{vc+}/\tau)}}{\sum_{i=1}^{1+|\overline{C}|} e^{(S_i^{vc}/\tau)} + \sum_{j=1}^{Z} e^{(S_j^{vg}/\tau)}} \\ &- log\frac{e^{(s^{cv+}/\tau)}}{\sum_{i=1}^{1+|\overline{V}|} e^{(S_i^{cv}/\tau)} + \sum_{j=1}^{Z} e^{(S_j^{cg}/\tau)}}\end{aligned} \tag{7}$$

where S_j^{vg} represents the cosine similarity between v and j^{th} random Gaussian noise vector g. By minimizing Eq. (7), the ITContrast network is enforced to mitigating the heterogeneous gap between image and text modalities while excavating the apparent discrimination. The new InfoCMR provides several benefits (1) the gradient of loss function encourages learning from hard negatives; and (2) the denominator of InfoCMR introduces an additional penalty term to avoid overfitting.

4 Experiments

4.1 Datasets and Settings

Datasets. MS-COCO [28] and Flickr30K [29] contains 123,287 and 31,783 image samples, each labelled with 5 captions. Following [5], we used 11,328 images for training, 5,000 for validation, and 5,000 for testing in MS-COCO. For Flickr30K, we used 29,783/1,000/1,000 images for training, validation and testing.

Evaluation Metric. We take Recall@K that describes the proportion of ground truth instance being retrieved at the top K results as the evaluation metric. The results on image retrieval and sentence retrieval are reported.

Implementation Details. We fine-tune the OSCAR-base model under the ITContrast framework for 20 epochs, with a learning rate 0.00002 and batch size 128. The value of hyper-parameter σ, τ and Z in Eq. (3) and Eq. (7) are 0.1, 0.05 and 128. Our method is implemented on the PyTorch framework.

Methods	Sen. Ret.		Ima. Ret.	
	R@1	R@10	R@1	R@10
CAAN [18]	52.5	90.9	41.2	82.9
IMRAM [20]	53.7	91.0	39.7	79.8
MMCA [17]	54.0	90.7	38.7	80.8
SGRAF [5]	57.8	91.6	41.9	81.3
Unicoder-VL [24]	62.3	92.8	46.7	85.3
UNITER [25]	63.3	93.1	48.4	76.7
VSE∞ [26]	68.1	90.2	52.7	80.2
OSCAR [8]	70.0	95.5	54.0	88.5
Ours (SGRAF)	71.9	96.0	56.4	89.4
Ours (ITContrast)	**74.5**	**97.2**	**59.1**	**92.3**

Table 2: Comparison of image-text matching results on MSCOCO 5K.

4.2 Quatitative Results and Analysis

Table 1 and Table 2 shows the experimental results on MSCOCO dataset. We can observe that the proposed ITContrast outperforms the previous methods on both MSCOCO 1K and 5K test set. Specifically, ITContrast achieved the best R@1=92.5% for sentence retrieval and R@1=79.6% for image retrieval with 1K test set. As for 5K test images, the proposed approach also outperforms the latest algorithms. On the more challenging dataset, Flickr30K, our approach obtained an improvement of more than 7% in rank-1 over the latest pre-trained models [25, 24]. For fair comparison, we also implement the state-of-the-art method SGARF [5] by applying the OSCAR model to learn the feature embedding instead of bi-directional GRU and CNN. It can be observed that the ITContrast still outperforms the new version of SGRAF, which originates from the better feature embedding space provided by the hard negative synthesis as well as the proposed contrastive learning framework.

4.3 Ablation Studies

Impact of Each Component. We analyze the effectiveness of each component in ITContrast on Flickr30K dataset. 1) OSCAR model. We employ the SCAN t-i LSE method [3] as the baseline(#1). Comparing #1 with #5 based on R@1, better feature embeddings can be learned that achieves 27.2% improvement for sentence retrieval and 32.4% for image retrieval by introducing OSCAR. 2) SynHNC. Comparing #1, #5 with #2, #6, we discover that sampling hard negatives is beneficial for training a more effective model which improves 5.7% and 2.6% for top-1 sentence retrieval. 3) InfoCMR: Comparing #2, #4 and #6, #7, additional improvements are obtained with the proposed contrastive loss, regardless of whether OSCAR is applied or not.

Extension to Other Architectures. We further evaluate the generality of the proposed method by it them into the popular models: SCAN [3] and SGRAF [5]. As shown in Table 4, we can see that the performance of SCAN and SGRAF on Flickr30K are significantly improved with SynHNC and InfoCMR loss. Overall, the most significant improvements are achieved on SCAN. In particular, it outperforms the original SCAN by 4.3% and 6.0% in top-1 sentence retrieval and image retrieval. Experimental results further confirm the effectiveness and generality of our method which greatly improves the performance of existing state-of-the-art method SGRAF.

4.4 Qualitative Results and Analysis

Fig.3 exhibits the qualitative comparison between the models trained by different approaches on the Flickr30K dataset, including OSCAR and ITContrast. For sentence retrieval, our ITContrast guides the model to better distinguish the highly relevant descriptions of concepts in negative sample, such as the OSCAR mismatch of Query2, which contains highly relevant descriptions of actions (e.g., " jumping") and scene (e.g.,"stream") in the image. For image retrieval, our network can distinguish hard samples well, even if negative samples consist of the same semantic concepts, and attribute. Overall, these qualitative results further verify the robustness of our ITContrast on the small dataset that contains a limited number of hard negative pairs in the training set.

5 Conclusion

This paper proposes a new contrastive learning objective, ITContrast, for the representation of image-text matching. The key idea is to compare image-text pairs with reinforced negative samples generated at the feature level. Based on the synthesized hard negatives, the effective In-

Model	Base.	OSCAR	Syn.	Info.	Sen. Ret.		Ima. Ret.	
					R@1	R@10	R@1	R@10
1	✓				61.1	91.5	43.3	80.9
2	✓		✓		66.8	95.1	51.7	85.4
3	✓			✓	70.8	94.8	50.4	84.9
4	✓		✓	✓	73.3	97.2	54.2	87.0
5		✓			88.3	99.0	75.7	96.1
6		✓	✓		90.9	99.4	80.7	97.4
7		✓	✓	✓	**92.8**	**99.3**	**84.1**	**98.1**

Table 3: Ablation study on each component of ITContrast.

Figure 3: Qualitative retrieval results on the Flickr30K test set. The correct matches are colored in green for each query.

foCMR loss is proposed to learn the embedding space to better distinguish positive and negative pairs. The deep vision-language model boosts the image-text matching capabilities with better visual and textual embeddings. We thoroughly evaluated the proposed method on the image-text matching task and further demonstrated that (i) contrastive learning is well suited for cross-model matching, (ii) hard negative samples are crucial for learning discriminative representations of image and text modality. Experimental results showed that ITContrast obtains state-of-the-art results on two benchmark datasets.

References

[1] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, pages 686–701, 2018.

[2] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[3] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked

Model	Sen. Ret.		Ima. Ret.	
	R@1	R@10	R@1	R@10
SCAN [3]	67.4	95.8	48.6	85.2
SCAN + Syn.	68.5	96.7	51.4	86.4
SCAN + Info.	71.4	97.0	52.9	86.1
SCAN + Syn. + Info.	71.7	97.4	54.6	87.4
SGRAF [5]	77.8	97.4	58.5	88.8
SGRAF + Syn.	78.7	97.8	60.2	89.2
SGRAF + Info.	79.8	97.4	61.3	90.5
SGRAF+ Syn. +Info.	80.7	98.3	62.2	91.1

Table 4: Ablation study on the generality of ITContrast.

cross attention for image-text matching. In *ECCV*, pages 201–216, 2018.

[4] Tianlang Chen and Jiebo Luo. Expressing objects just like words: Recurrent visual embedding for image-text matching. In *AAAI*, pages 10583–10590, 2020.

[5] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, pages 1218–1226, 2021.

[6] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, pages 4654–4662, 2019.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.

[8] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137, 2020.

[9] Jônatas Wehrmann, Camila Kolling, and Rodrigo C Barros. Adaptive cross-modal embeddings for image-text alignment. In *AAAI*, pages 12313–12320, 2020.

[10] Chen Chen, Dan Wang, Bin Song, and Hao Tan. Inter-intra modal representation augmentation with dct-transformer adversarial network for image-text matching. *IEEE Transactions on Multimedia*, 2023.

[11] Tianlang Chen, Jiajun Deng, and Jiebo Luo. Adaptive offline quintuplet loss for image-text matching. In *ECCV*, pages 549–565, 2020.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 33, 2020.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28:91–99, 2015.

[15] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11):2579–2605, 2008.

[16] Bai-Ling Zhang, Haihong Zhang, and Shuzhi Sam Ge. Face recognition by applying wavelet subband representation and kernel associative memory. *TNNLS*, 15(1):166–177, 2004.

[17] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *CVPR*, pages 10941–10950, 2020.

[18] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *CVPR*, pages 3536–3545, 2020.

[19] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. Cross-modal attention with semantic consistence for image–text matching. *TNNLS*, 31(12):5412–5425, 2020.

[20] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *CVPR*, pages 12655–12663, 2020.

[21] Jiwei Wei, Yang Yang, Xing Xu, Xiaofeng Zhu, and Heng Tao Shen. Universal weighting metric learning for cross-modal retrieval. *TPAMI*, 2021.

[22] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *CVPR*, pages 10921–10930, 2020.

[23] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV*, pages 18–34, 2020.

[24] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020.

[25] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Zhe Ahmed, et al. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120, 2020.

[26] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, pages 15789–15798, 2021.

[27] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, pages 9826–9836, 2021.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Perona, et al. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[29] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.