

CODER: Coupled Diversity-Sensitive Momentum Contrastive Learning for Image-Text Retrieval

Haoran Wang¹, Dongliang He^{1*}, Wenhao Wu^{1,5}, Boyang Xia², Min Yang¹, Fu Li¹, Yunlong Yu³, Zhong Ji⁴, Errui Ding¹, and Jingdong Wang¹

¹ Department of Computer Vision Technology (VIS), Baidu Inc., Beijing, China

² Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China

³ College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou, China

⁴ School of Electrical & Information Engineering, Tianjin University, Tianjin, China

⁵ The University of Sydney, Sydney, Australia

{wanghaoran09, hedongliang01, yangmin09, lifu}@baidu.com,
xiaboyang20@mails.ucas.ac.cn, yuyunlong@zju.edu.cn, jizhong@tju.edu.cn,
{dingerrui, wangjingdong}@baidu.com

Abstract. Image-Text Retrieval (ITR) is challenging in bridging visual and lingual modalities. Contrastive learning has been adopted by most prior arts. Except for limited amount of negative image-text pairs, the capability of contrastive learning is restricted by manually weighting negative pairs as well as unawareness of external knowledge. In this paper, we propose our novel Coupled Diversity-Sensitive Momentum Contrastive Learning (CODER) for improving cross-modal representation. Firstly, a novel diversity-sensitive contrastive learning (DCL) architecture is invented. We introduce dynamic dictionaries for both modalities to enlarge the scale of image-text pairs, and diversity-sensitiveness is achieved by adaptive negative pair weighting. Furthermore, two branches are designed in CODER. One learns instance-level embeddings from image/text, and it also generates pseudo online clustering labels for its input image/text based on their embeddings. Meanwhile, the other branch learns to query from commonsense knowledge graph to form concept-level descriptors for both modalities. Afterwards, both branches leverage DCL to align the cross-modal embedding spaces while an extra pseudo clustering label prediction loss is utilized to promote concept-level representation learning for the second branch. Extensive experiments conducted on two popular benchmarks, *i.e.* MSCOCO and Flickr30K, validate CODER remarkably outperforms the state-of-the-art approaches.

1 Introduction

Image-text retrieval (ITR) refers to searching for the semantically similar instance from visual (textual) modality with the query instance from textual (vi-

*indicates corresponding author.

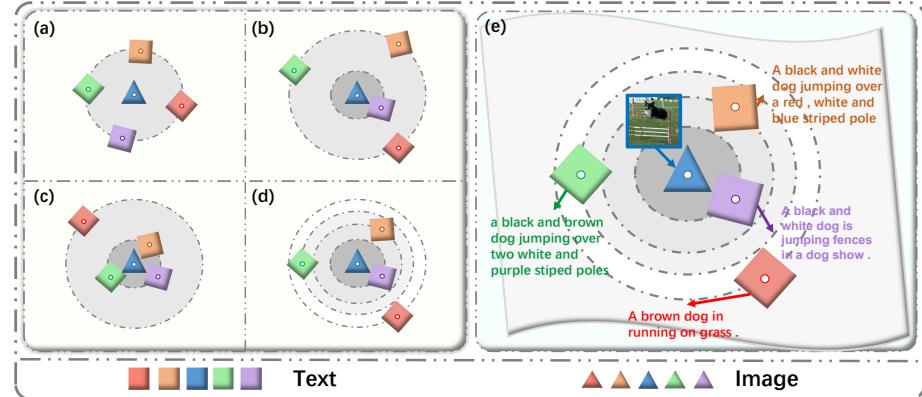


Fig. 1. Conceptual illustration of our proposed Diversity-sensitive Contrastive Learning (DCL) loss. Sub-figure (a), (b) and (c) depict three exemplary distributions of negative samples which are undesired because they do not show much similarity variations, respectively. Sub-figure (d) shows desired negative sample distribution given an anchor, where different negative samples are not equally pushed away. It demonstrate the joint space can well distinguish fine-grained semantic difference among negative samples. Sub-figure (e) illustrates the ideal joint embedding space affected by DCL.

sual) modality. Nowadays, it has become a compelling topic from both industrial and research community and is of potential value to benefit extensive relevant applications [2, 3, 19, 24, 25, 40, 51, 53, 62–65, 67]. In the past decade, tremendous progresses have been made with the prevalence of deep learning [30]. Early works typically associate image with text via learning global [11, 29, 55] or local cross-modal alignment [4, 31]. Follow-up studies attempt to introduce external knowledge information, including commonsense knowledge [49, 54] or scene graph [56] information, into visual-semantic embedding models. It remains challenging due to heterogeneous multi-modal data distributions, which requires pretty precise cross-modal alignment.

Loss functions play the central role in aligning multi-modal data. The prevailing bi-directional triplet ranking (BTR) loss used in [11, 13] can be regarded as one special case of contrastive loss [17], where only one negative sample is considered. Then, bidirectional Info-NCE loss [43] (BIN), as a typical contrastive loss, has been widely adopted in many tasks [7, 38, 47]. It exploits the whole paired relationships among a mini-batch of image-text samples when applied to the ITR task. Meanwhile, contrastive learning is well-known in limited negative sample scale [18], which acts as the bottleneck of its capability.

Another notable issue is both aforementioned contrastive losses manually design the weighting strategy for negative image-text pairs. They both enforce the negatives and anchor samples to be separated far away enough, whilst ignoring the relative differences between them. Consequently, the fine-grained discrepancies among negative pairs are hard to be fully captured.

In fact, the importance of each image/text instance is unequal [6] in contrastive learning. A critical factor determining the importance of instance is its semantic ambiguity [50]. In particular, the samples with high semantic ambiguity refers to those with multiple meanings/concepts. Oppositely, the samples with simple and clear meanings usually have low semantic ambiguity. To explicitly model the semantic ambiguity of sample, we present a term called “**Diversity**”. Concretely, the diversity of one sample is defined based on the distribution of cross-modal negatives around it. For example, as depicted in three typical cases in Figure 1(a-c), if a sample has multiple negatives with similar distances to it, we call this sample as low-diversity one. Obviously, the existence of low-diversity samples are undesirable, which will weaken the discrimination ability of the learned joint space. Conversely, if the data distribution around an anchor instance is well-spaced (see 1(d), this sample has high diversity), it could better measure the difference among different negative samples, which is more ideal.

To address the aforementioned limitations and questions, first of all, inspired by Momentum Contrastive Learning (MCL) paradigm [18], dynamic dictionaries of memory banks are introduced in coupled form for both visual and textual modality to enlarge interactions among image-text pairs. Furthermore, in this paper, we propose to extend contrastive learning to a novel **Diversity-sensitive Contrastive Learning (DCL)** paradigm. To achieve it, a novel diversity-sensitive contrastive loss is presented, which incorporates our defined diversity into contrastive loss. Specifically, in contrastive loss, a simple yet effective estimation function is designed to quantify the diversity of each anchor sample in a mini-batch of data, the diversity term is then used to dynamically weight negative samples of each anchor, enabling the training procedure to balance between diversity and total contrastive loss. With our DCL, on one hand, the image-text pairs built based on low diversity anchor sample can be allocated with larger weight and *vice versa*; on the other hand, given a negative sample, when it is paired with different anchors, it can be unequally weighted according to the anchor’s diversity. Doing so enables the original contrastive loss to be aware of semantic diversities of samples, and suppress the adverse impact brought by low-diversity ones. Accordingly, *instance-level* visual or textual representations can be learned with our DCL. As consequence, we can obtain a more structured and hierarchical joint embedding space. Taking Figure 1(e) as example, the subtle difference between the caption (marked in orange) and another one (marked in green) can be appropriately distinguished in their semantic distances.

Furthermore, how to leverage external knowledge into contrastive learning framework is worth exploring. To be complementary to the *instance-level* alignment, we achieve *concept-level* cross-modal feature alignment via exploiting commonsense knowledge. Different from the former, *concept-level* alignment is built by firstly learning to extract homogeneous concept-level visual and textual embeddings from commonsense graph, followed by aligning the cross-modal embeddings via adopting DCL along with a **Prototype-Guided Classification loss (PGC)**. In order to enable PGC, an online clustering procedure is performed on *instance-level* representations and each cluster id is treated as a prototype, then

a prediction head based on the *concept-level* image/text embedding is employed for classifying the cluster id of the input image/text. The final image-text matching score is a combination of similarities obtained from both instance-level and concept-level alignment. Extensive experiments conducted on MSCOCO [34] and Flickr30K [45] verify the superiority of our framework and show that our Coupled Diversity-Sensitive Contrastive Learning (CODER) method significantly outperforms recent state-of-the-art solutions.

To sum up, the main contributions are listed as follows:

- We incorporate coupled Momentum Contrastive learning (MCL) into image-text representation learning and further extend contrastive learning to a novel Diversity-Sensitive Contrastive Learning (DCL) paradigm, which can adaptively weight negative image-text pairs to further boost the performance.
- A Coupled Diversity-Sensitive Contrastive Learning (CODER) framework is proposed to exploit not only instance-level image-text representations but also concept-level embeddings with the aid of external knowledge as well as on-line clustering based prototype-guided classification loss.
- Extensive experimental results on two benchmarks demonstrate our approach considerably outperforms state-of-the-art methods by a large margin.

2 Related Work

2.1 Contrastive Learning

Recently, Contrastive Learning [7, 16, 18, 43, 47] has made remarkable progress in unsupervised representation learning. Chen *et al.* [7] shows that contrastive learning in unsupervised visual representation learning benefits from large batch size negatives and stronger data augmentation. He *et al.* [18] proposed Momentum Contrastive Learning (MCL) paradigm that obtains the new key representation on-the-fly by a momentum-updated key encoder, and maintains a dictionary as a queue to allow the training process to reuse the encoded key representations from the immediate preceding mini-batches. Recently, more Contrastive Learning based vision-language understanding studies [21, 33, 47, 66] are emerging. For video-text retrieval, Liu *et al.* [37] first introduces the vanilla info-NCE loss based MCL mechanism to enhance the cross-modal discrimination. Distinct from them, we integrate coupled MCL into our proposed Diversity-sensitive contrastive learning (DCL) paradigm for tackling ITR.

2.2 Image-Text Retrieval

Along with the renaissance of deep learning, a surge of works have been proposed for ITR. Early attempts [13, 39, 42, 55] typically employ global features to represent both image and text in a common semantic space. For instance, Kiros *et al.* [29] encoded image and text by CNN and RNN respectively, utilizing BTR loss to train the model. Afterwards, another line of research [4, 10, 31, 58, 60]

employed multi-modal attention mechanism [4, 23, 31, 61] or knowledge aided representation learning [14, 20, 35, 49, 54] to achieve cross-modal alignment by exploiting more fine-grained associations. For instance, Lee *et al.* [31] developed Stacked Cross Attention Network that aligns image regions and textual words.

Except for focusing on representation architecture designing, some studies [6, 11, 36, 59] endeavored to improve the learning objectives. As a seminal work, Faghri [11] *et al.* proposed to introduce one on-line hard negative mining (OHNM) strategy into BTR loss, which is very prevailing for ITR. Liu *et al.* [36] proposed to tackle hubness problem by imposing heavy punishment on the hard negatives in triplets. Afterwards, Chen *et al.* [6] further improved the BTR loss by searching for more hard negatives in off-line way to constitute the quintuplet. Overall, the common character of above works is designing constraint strategy for pairwise multi-modal data, whilst our DCL additionally performs diversity estimation especially for each sample. Moreover, we introduce MCL to promote large-scale negative interaction, which leads to more comprehensive diversity estimation in DCL.

3 Methodology

3.1 Overall Framework

The overall framework of our proposed CODER model is illustrated in Figure 2. In our model, two branches are designed for instance-level and concept-level representation learning. In the instance-level branch (Fig.2(a)), image and text features are encoded and aggregated to be \mathbf{v}^I and \mathbf{w}^I , momentum encodes are used for the two modalities to serve as coupled memory banks. Instance-level alignment is achieved via employing our proposed diversity-sensitive contrastive loss L_{DCL}^I as well as memory-aided DCL loss $L_{M_DCL}^I$ (Fig.2(c)). As for the concept-level branch (Fig.2(b)), statistical commonsense representation (SCC) [54] denoted as \mathbf{Y} , is adopted as homogeneous feature basis. Query features \mathbf{v}_C^q and \mathbf{w}_C^q are obtained from image and text, respectively. Then concept-level features \mathbf{v}_C^C and \mathbf{w}_C^C are obtained by learning to query from feature basis \mathbf{Y} . For concept-level alignment (Fig.2(d)), except for DCL loss L_{DCL}^C , an online-clustering based prototype-guided classification loss L_{PGC} is additionally leveraged.

3.2 Instance and Concept Level Representations

Instance-level Representation For image encoding, we adopt Faster-RCNN [1, 48] to obtain L region-level features $\{\mathbf{o}_l\}_{l=1}^L$ and then aggregate these features to be a instance-level visual embedding $\mathbf{v}^I \in \mathbb{R}^F$. Pre-trained BERT [9] is our textual encoder and N word-level embeddings $\{\mathbf{e}_t\}_{t=1}^T$ are also aggregated to instance-level textual embedding $\mathbf{w}^I \in \mathbb{R}^F$.

$$\mathbf{v}^I = g_{vis}(\{\mathbf{o}_l\}_{n=1}^L), \quad \mathbf{w}^I = g_{text}(\{\mathbf{e}_t\}_{t=1}^T), \quad (1)$$

where $g_{vis}(\cdot)$ and $g_{text}(\cdot)$ are visual and textual aggregators.

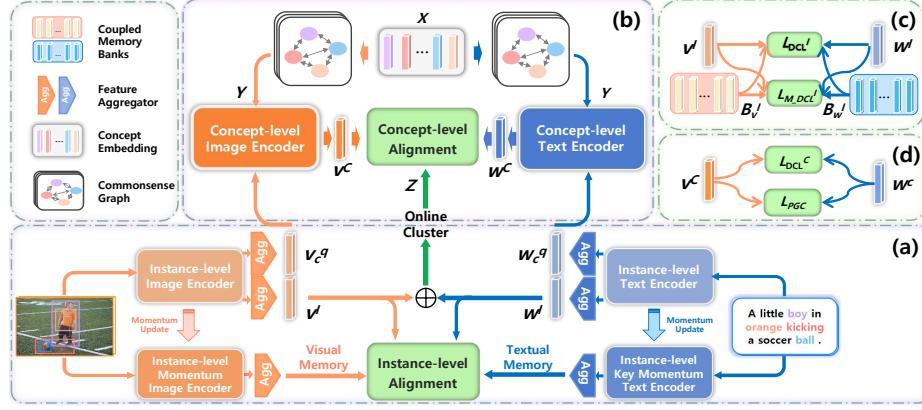


Fig. 2. The overall architecture of our proposed CODER model for image-text retrieval. It is composed of an instance-level representation branch (a) and an concept-level one which leverages external knowledge (b). The former branch is optimized by minimizing instance-level DCL loss and memory-based DCL loss (denoted as L_{DCL}^I and $L_{M_DCL}^I$, respectively) (c). The other one is learned by employing concept-level DCL loss L_{DCL}^C and online clustering based prototype-guided classification L_{PGC} as objectives (d).

Concept-level Representation The concept-level representations for both modalities are built based on a group of *concepts*. Firstly, we extract g representative concepts from the texts over the whole image-caption dataset. Afterwards, the GloVe [44] is employed to instantiate these concepts as \mathbf{X} . Following [54], graph convolution network (GCN) [28] is utilized to process to produce the statistical commonsense aided concept (SCC) representations $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_g\}$. Please refer to the supplementary materials for more details.

To generate concept-level representations, we generate representations (\mathbf{v}_C^q and \mathbf{w}_C^q) by using another group of feature aggregators ($g_{vis}(\cdot)$ and $g_{text}(\cdot)$) to combine local features $\{\mathbf{o}_l\}_{l=1}^L$ and $\{\mathbf{e}_t\}_{t=1}^T$, respectively. Then, as depicted in Figure 2, \mathbf{v}_C^q and \mathbf{w}_C^q are fed into concept-level feature encoders, which are taken as input vectors to query from the SCC representations \mathbf{Y} . The output scores for different concepts allow us to uniformly utilize the linear combination of the SCC representations to represent both modalities. Mathematically, the concept-level representation \mathbf{v}^C and \mathbf{w}^C can be calculated as:

$$\begin{aligned} \mathbf{v}^C &= \sum_{i=1}^g a_i^v \mathbf{y}_i; \quad a_i^v = \frac{e^{\lambda \mathbf{v}_C^I \mathbf{W}^v \mathbf{y}_i^\top}}{\sum_{i=1}^g e^{\lambda \mathbf{v}_C^I \mathbf{W}^v \mathbf{y}_i^\top}}. \\ \mathbf{w}^C &= \sum_{j=1}^g a_j^w \mathbf{y}_j; \quad a_j^w = \frac{e^{\lambda \mathbf{w}_C^I \mathbf{W}^w \mathbf{y}_j^\top}}{\sum_{j=1}^g e^{\lambda \mathbf{w}_C^I \mathbf{W}^w \mathbf{y}_j^\top}} \end{aligned} \quad (2)$$

where $\mathbf{W}^v \in \mathbb{R}^{F \times F}$ and $\mathbf{W}^w \in \mathbb{R}^{F \times F}$ denote the learnable parameter matrix, \mathbf{a}_i^v and \mathbf{a}_j^w denote the visual and textual score corresponding to the concept \mathbf{z}_i , respectively. λ controls the smoothness of the softmax function.

Coupled Memory Banks Building We propose to leverage a couple of dynamic memory banks B_v^I and B_w^I to restore more visual and textual embeddings to enlarge the scale of negative samples for both modalities. We follow MoCo [18] to obtain instance-level momentum image encoder and text encoder by momentum updating their weights according to the corresponding image and text encoders. Visual or textual instances from the latest training iterations are fed to the momentum encoders to generate visual and textual embeddings, which are restored in coupled memory banks. Such a process can be conveniently implemented via queues.

3.3 Diversity-Sensitive Contrastive Loss

Estimating the semantic Diversity of instance plays important role in enhancing cross-modal discrimination. Specifically, to describe our diversity-sensitive contrastive loss, we start from diversity estimation, and then introduce our *explicit* diversity-sensitive loss.

Diversity Estimation For simplicity, we take as example that visual feature \mathbf{v}_i is an anchor sample and Q text features $\mathbf{W} = \{\mathbf{w}_i, \mathbf{w}_2, \dots, \mathbf{w}_Q\}$ are to be compared (among which only \mathbf{w}_i is a matching sample for \mathbf{v}_i), to illustrate how we estimate diversity of an anchor sample. The cosine similarity of $\text{cosine}(\mathbf{v}_i, \mathbf{w}_j)$ is defined as S_{ij} . We propose a simple but effective metric to estimate the semantic diversity explicitly.

In joint embedding space, if an anchor sample with low diversity indicates the close similarities between it and numerous negatives, this case is undesired. By contrast, an ideal data distribution space should be more structured and consistent with text-image pair annotations. Intuitively, we propose to quantify the diversity of anchor sample via employing one statistical variable, *i.e.* standard deviation (SD). Concretely, a low-diversity anchor sample has multiple negatives with close distances to it, implying the SD value of cross-modal similarities between it and them will be small. Conversely, the high SD value means an anchor sample has high diversity. Since the SD value between negative cross-modal similarities are proportional to the diversity of anchor, we propose to estimate the semantic diversity explicitly based on SD value. Taking image sample \mathbf{v}_i for instance, the computation process of its diversity value is defined as:

$$\begin{aligned} SD(\mathbf{v}_i) &= \sqrt{E(S_{ij}^2) - [E(S_{ij})]^2}, i \neq j; \\ div(\mathbf{v}_i) &= 1/\sigma(\epsilon/SD); \\ div(\mathbf{v}_i) &= div(\mathbf{v}_i)/\max\{div(\mathbf{v}_1), \dots, div(\mathbf{v}_Q)\}, \end{aligned} \tag{3}$$

where $E(\cdot)$ is the mathematical expectation function and $\sigma(\cdot)$ denotes the Sigmoid function that normalizes the reciprocal of SD value to a uniform scale, assuring it vary in a relatively stable range. $div(\mathbf{v}_i)$ denotes the diversity score of \mathbf{v}_i calculated from the candidate textual samples to be compared with. $\epsilon = 0.1$ is a tuning parameter. Finally, we divide each diversity score $div_{std}(\mathbf{v}_i)$ by the maximum value of them in mini-batch for normalization. Likewise, the diversity of text sample can be calculated in similar manner.

Diversity-Sensitive Loss As mentioned in Section.1, we aim to highlight the discrepancy among the anchor sample with low-diversity and its negatives. To achieve it, we need to allocate more attention to such cases in order for an optimal alignment model. To begin with, let us term the contrastive objective that insensitive to diversity as L_{DCL_I} . Given $\mathbf{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_N\}$ and $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_Q\}$, $L_{DCL_I}(\mathbf{V}, \mathbf{W})$ can be formulated as:

$$\begin{aligned} l_{DCL_I}(\mathbf{V}, \mathbf{W}) &= \frac{\mu}{N} \sum_{n=1}^N [\log(\sum_{q \neq n} \exp(\frac{(S_{nq}-\gamma)}{\mu})) + 1) - \log(S_{nn} + 1)]; \\ l_{DCL_I}(\mathbf{W}, \mathbf{V}) &= \frac{\mu}{Q} \sum_{q=1}^Q [\log(\sum_{n \neq q} \exp(\frac{(S_{qn}-\gamma)}{\mu})) + 1) - \log(S_{qq} + 1)]; \\ L_{DCL_I}(\mathbf{W}, \mathbf{V}) &= l_{DCL_I}(\mathbf{W}, \mathbf{V}) + l_{DCL_I}(\mathbf{V}, \mathbf{W}) \end{aligned} \quad (4)$$

where μ is a temperature scalar; γ is a margin parameter; N is the number of samples within the mini-batch; $S_{nq} = \text{cosine}(\mathbf{v}_n, \mathbf{w}_q)$, $S_{qn} = \text{cosine}(\mathbf{w}_q, \mathbf{v}_n)$, $S_{nn} = \text{cosine}(\mathbf{v}_n, \mathbf{w}_n)$ and $S_{qq} = \text{cosine}(\mathbf{w}_q, \mathbf{v}_q)$ denote the cosine similarities.

To explicitly introduce diversity awareness, we extend the above loss to DCL loss L_{DCL} . Mathematically,

$$\begin{aligned} L_{DCL}(\mathbf{V}, \mathbf{W}) &= l_{DCL}(\mathbf{W}, \mathbf{V}) + l_{DCL}(\mathbf{V}, \mathbf{W}) \\ l_{DCL}(\mathbf{V}, \mathbf{W}) &= \frac{\mu}{N} \sum_{n=1}^N [\log(\sum_{q \neq n} \exp(\frac{(S_{nq}-\gamma)}{\mu \cdot div(\mathbf{v}_n)})) + 1) - \log(S_{nn} + 1)]; \\ l_{DCL}(\mathbf{W}, \mathbf{V}) &= \frac{\mu}{Q} \sum_{q=1}^Q [\log(\sum_{n \neq q} \exp(\frac{(S_{qn}-\gamma)}{\mu \cdot div(\mathbf{w}_q)})) + 1) - \log(S_{qq} + 1)]; \end{aligned} \quad (5)$$

where $div(\mathbf{v}_n)$ and $div(\mathbf{w}_q)$ denotes the diversity of \mathbf{v}_n and \mathbf{w}_q , respectively and they are used to adaptively weight the negative samples.

DCL Loss Based Cross-Modal Alignment Instance-level DCL Loss. For instance-level representation, two items of DCL loss is employed. First, it is imposed on data pairs in mini-batch, named as L_{DCL}^I . Secondly, it is imposed on anchor sample in mini-batch and items from coupled memory banks, namely Memory-aided Diversity-sensitive Contrastive Learning (M-DCL) and abbreviated as $L_{M_DCL}^I$. Formally, using \mathbf{V}^I and \mathbf{W}^I to denote a mini-batch of embeddings \mathbf{v}^I and \mathbf{w}^I , these loss items are defined as:

$$\begin{aligned} L_{DCL}^I &= L_{DCL}(\mathbf{V}^I, \mathbf{W}^I), \\ L_{M_DCL}^I &= L_{DCL}(\mathbf{V}^I, B_w^I) + L_{DCL}(\mathbf{W}^I, B_v^I). \end{aligned} \quad (6)$$

Please note that in Eq.6, because the presence of memory banks, diversity estimation is processed as the average of diversity values at mini-batch level and memory bank level.

Concept-level DCL Loss. For concept-level representation, we only impose DCL Loss on data pairs in a mini-batch, the concept-level DCL loss is represented as: $L_{DCL}^C = L_{DCL}(\mathbf{V}^C, \mathbf{W}^C) + L_{DCL}(\mathbf{W}^C, \mathbf{V}^C)$.

3.4 Prototype-guided Classification Loss

In this section, we present a novel Prototype-guided Classification (PGC) loss, which aims to enhance cross-modal discrimination by leveraging the complementary semantics between instance-level and concept-level representations. In particular, we perform K-means [22] clustering in an on-line manner during training based on the summation of instance-level representations \mathbf{v}^I and \mathbf{w}^I , which contains more individual information. We name the output clusters as *prototypes* that are able to capture the shared semantic information between semantically related samples. Accordingly, The prototype ids of image/text instances serve as the pseudo class ids and are taken as supervision $\mathbf{Z} = \{z_1, \dots, z_K\}$ for concept-level representation learning. Specifically, the PGC loss is formally defined as:

$$\begin{aligned} \mathbf{P}_v &= softmax(\mathbf{P}^C \mathbf{v}^C), \mathbf{P}_w = softmax(\mathbf{P}^C \mathbf{w}^C), \\ L_{PGC} &= L_{PGC}^v + L_{PGC}^w = L_{cls}(\mathbf{P}_w, \mathbf{Z}) + L_{cls}(\mathbf{P}_v, \mathbf{Z}) \end{aligned} \quad (7)$$

where $\mathbf{P}^C \in \mathbb{R}^{K \times F}$ is one learnable parameter matrix that outputs distributions over the K category labels for both \mathbf{v}^C and \mathbf{w}^C . $\mathbf{P}_v \in \mathbb{R}^K$ and $\mathbf{P}_w \in \mathbb{R}^K$ denote probabilities over all labels. L_{cls} denotes the cross-entropy classification loss.

3.5 Training and Inference

Training Objective. We deploy the summation of instance-level and concept-level aligning losses as overall training objectives:

$$L = \lambda L_{DCL}^I + L_{M-DCL}^I + L_{DCL}^C + L_{PGC}, \quad (8)$$

Inference Scheme. For inference, we use the weighted summation of instance-level and concept-level cosine similarities to measure the overall cross-modal similarity $S = \beta S(\mathbf{v}^I, \mathbf{w}^I) + (1 - \beta)S(\mathbf{v}^C, \mathbf{w}^C)$, where β is a balancing parameter.

4 Experiments

4.1 Datasets & Evaluation Metrics

Datasets. Flickr30K [45] is an image-caption dataset containing 31,783 images, where each image annotated with five sentences. Following [42], we split the dataset into 29,783 training, 1000 validation, and 1000 testing images. The performance evaluation is reported on 1000 testing set. MSCOCO [34] is another

Table 1. Comparisons of experimental results on MSCOCO 1K test set and Flickr30K test set. Note that DSRAN [61], GPO [5] and DIME [46] employ BERT as we use, the rest use inferior text encoders.

Methods	Image Encoder	MSCOCO						Flickr30K						
		Text retrieval			Image Retrieval			R@sum	Text 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
DVSA [26] (2015)	R-CNN	38.4	69.9	80.5	27.4	60.2	74.8	351.2	22.2	48.2	61.4	15.2	37.7	50.5
m-CNN [39] (2015)	VGG-19	42.8	73.1	84.1	32.6	68.6	82.8	384.0	33.6	64.1	74.9	26.2	56.3	69.6
DSPE [55] (2016)	VGG-19	50.1	79.7	89.2	39.6	75.2	86.9	420.7	40.3	68.9	79.9	29.7	60.1	72.1
VSE++ [11] (2018)	ResNet-152	64.7	-	95.9	52.0	-	92.0	-	52.9	-	87.2	39.6	-	79.5
SCAN [31] (2018)	Faster-RCNN	72.7	94.8	98.4	58.8	88.4	94.8	507.9	67.4	90.3	95.8	48.6	77.7	85.2
PVSE [50] (2019)	Faster-RCNN	69.2	91.6	96.6	55.2	86.5	93.7	492.8	-	-	-	-	-	-
VSRN [32] (2019)	Faster-RCNN	76.2	94.8	98.2	62.8	89.7	95.1	516.8	71.3	90.6	96.0	54.7	81.8	88.2
CVSE [54] (2020)	Faster-RCNN	74.8	95.1	98.3	59.9	89.4	95.2	512.7	73.5	92.1	95.8	52.9	80.4	87.8
IMRAM [4] (2020)	Faster-RCNN	76.7	95.6	98.5	61.7	89.1	95.0	516.6	74.1	93.0	96.6	53.9	79.4	87.2
WCGL [57] (2021)	Faster-RCNN	75.4	95.5	98.6	60.8	89.3	95.3	514.9	74.8	93.3	96.8	54.8	80.6	87.5
SHAN [23] (2021)	Faster-RCNN	76.8	96.3	98.7	62.6	89.6	95.8	519.5	74.6	93.5	96.9	55.3	81.3	88.4
DSRAN [61] (2021)	Faster-RCNN	77.1	95.3	98.1	62.9	89.9	95.3	518.6	75.3	94.4	97.6	57.3	84.8	90.9
GPO [5] (2021)	Faster-RCNN	78.6	96.2	98.7	62.9	90.8	96.1	523.3	78.1	94.1	97.8	57.4	84.5	90.4
DIME (i-t) [46] (2021)	Faster-RCNN	77.9	95.9	98.3	63.0	90.5	96.2	521.8	77.4	95.0	97.4	60.1	85.5	91.8
SGRAF [10] (2021)	Faster-RCNN	79.6	96.2	98.5	63.2	90.7	96.1	524.3	77.8	94.1	97.4	58.5	83.0	88.8
CODER	Faster-RCNN	82.1	96.6	98.8	65.5	91.5	96.2	530.6	83.2	96.5	98.0	63.1	87.1	93.0
														520.9

image-caption dataset, totally including 123,287 images with each image roughly annotated with five textual descriptions. We follow the public split of [26], including 113,287 training images, 1000 validation images, and 5000 testing images. The result is reported by averaging the results over 5-folds of 1K testing images.

Evaluation Metrics. We utilize two commonly used evaluation metrics, *i.e.*, R@K and “R@sum”. Specifically, R@K refers to the percentage of queries in which the ground-truth matchings appear in the top K retrieved results. “R@sum” is the summation of all six recall rates of R@K, which provides a more comprehensive evaluation to testify the overall retrieval performance.

4.2 Implementation Details

For visual feature encoding, the amount of regions is $L = 36$ and the dimension of region embeddings is 2048. For text encoding, a BERT-base [9] model is used to extract 768-dimension textual embeddings. The dimension of joint space is set to $F=1024$. For concept-level representation, we adopt 300-dim GloVe [44] trained on the Wikipedia dataset to initialize the semantic concepts. The volume of the concept vocabulary is $g = 400$. The size of couple memory banks is set to 4096 and the momentum coefficient is 0.995. The cluster number K of PGC loss is set to 10000 and 20000 for Flickr30K and MSCOCO dataset, respectively. For the training objective, we empirically set $\mu = 0.1$ and $\gamma = 0.3$ in Eq. (9). Our CODER model is trained by Adam optimizer [27] with mini-batch size of 128. The learning rate is set to be 0.00002 for the first 15 epochs and 0.000002 for the next 15 epochs for both datasets. The balancing parameter in Eq. (8) is set to $\lambda = 3$. For inference, the controlling parameter β is equal to 0.9. All our experiments are conducted on a NVIDIA Tesla P40 GPU.

4.3 Comparisons with state-of-the-art Methods

The experimental results on two benchmark datasets are listed in Table 1⁶. As for MSCOCO, we can observe that our CODER is obviously superior to the competitors in most evaluation metrics, which yields a result of 82.1% and 65.5% on R@1 for text retrieval and image retrieval, respectively. Specifically, compared with the best competitor SGRAF method, it achieves absolute boost (2.5%, 0.4%, 0.3%) on (R@1, R@5, R@10) for text retrieval. For image retrieval, our method also outperforms other algorithms. Moreover, on Flickr30K dataset, as for the most persuasive criteria, the “R@sum” achieved by our model exceeds the second best performance by 13.7%. These results solidly validate the advance of our method.

4.4 Ablation Study

In this section, we perform a series of ablation studies to explore the impact of the main modules in our CODER method. All the comparative experiments are conducted on the Flickr30K dataset.

To begin with, we first investigate the effect of each module for instance-level representation. In Table 2, we employ a framework without adopting coupled memory banks for M-DCL as the baseline (#1), which utilizes the traditionally prevailing BTR loss [11] to perform instance-level alignment instead of our DCL loss. From Table 2, Comparing #1 with #2 based on R@1, the DCL loss brings about 3.2% improvement for text retrieval and 2.9% for image retrieval. Moreover, when the coupled memory banks is exploited for M-DCL, Comparing #3 with #2, we can obtain additional performance improvement. These results confirm the effectiveness of our proposed DCL learning paradigm for enhancing instance-level discrimination.

In addition, we explore how the modules for concept-level representation affects the retrieval performance. As shown in Table 2, comparing #4 with #3 based on R@1, the L_{DCL}^C loss leads to 0.2% improvement for text retrieval and 0.2% for image retrieval. It validates our DCL loss is also effective for concept-level representation learning. Furthermore, when our presented PGC loss is leveraged, comparing #5 with #4, it achieves (0.4%, 0.4%) boost on (R@1, R@5) for text retrieval and (0.3%, 0.3%) boost on (R@1, R@5) for image retrieval. The above results prove our designed concept-level representation learning module can provides more complementary semantics for instance-level one thereby enhancing cross-modal discrimination.

Impact of Different Configurations of DCL In this part, we perform ablation studies to explore the impact of different configurations for the DCL module.

To analyze the impacts of various components in DCL module, we perform a group of experiments and present the results in Table 3. We take the model

⁶We report our replicated results of [5] by using its official code without changing, more discussions are given in the supplementary materials

Table 2. Performance comparison of our CODER with different main components on Flickr30K test set. “Instance-level Alignment” is abbreviated as “IA”. “Concept-level Alignment” is abbreviated as “CA”.

Models	IA		CA		Text Retrieval			Image Retrieval		
	$L_{M,DCL}^I$	L_{DCL}^I	L_{PGC}	L_{DCL}^C	R@1	R@5	R@10	R@1	R@5	R@10
1					78.7	94.5	97.0	58.6	84.8	90.1
2		✓			81.9	95.6	97.9	61.5	85.8	91.8
3	✓	✓			82.6	96.1	98.0	62.6	86.7	92.3
4	✓	✓		✓	82.8	96.1	98.0	62.8	86.8	92.6
5	✓	✓	✓	✓	83.2	96.5	98.0	63.1	87.1	93.0

Table 3. Effect of different configurations of DCL module on Flickr30K test set. Implicit Diversity estimation is abbreviated as “IE”. Explicit Diversity estimation is abbreviated as “EE”. “MB” means using memory banks for Explicit Diversity estimation.

Models	L_{DCL}^I			Text Retrieval			Image Retrieval			
	$L_{M,DCL}^I$	IE	EE	MB	R@1	R@5	R@10	R@1	R@5	R@10
1		✓			80.3	94.8	97.3	60.2	85.3	91.1
2		✓	✓		81.5	95.6	97.5	61.2	85.6	91.2
3		✓	✓	✓	81.9	95.6	97.7	61.5	85.8	91.4
4	✓	✓			82.0	95.8	97.7	61.6	86.2	92.2
5	✓	✓	✓		82.8	96.3	97.9	62.6	87.0	92.7
6	✓	✓	✓	✓	83.2	96.5	98.0	63.1	87.1	93.0

adopting $L_{DCL,I}$ loss in Eq. 4 as baseline, named implicit Diversity contrastive loss and abbreviated as “IE”. As shown in Table 3, comparing #2 with #1 based on R@1, the explicit Diversity estimation additionally leads to 1.2% improvement for text retrieval and 1.0% for image retrieval. Moreover, the comparison between #3 and #2 validates the introduce of memory bank items in Diversity estimation really matters for alleviating semantic ambiguity. Furthermore, comparing (#4,#5,#6) with (#1,#2,#3), we find the combination of $L_{M,DCL}$ and L_{DCL}^I loss can lead to significant retrieval performance boost, which validates they are mutually beneficial to each other and collaborate to promote discriminative cross-modal embedding learning.

Impact of size in Mini-Batch. Then, we investigate the impact of size in mini-batch on performance. From Figure 3, we can see that when mini-batch size decreases from 128 to 32, the R@1 metric of the model “w/o M-DCL” falls from 61.5% to 59.3% for image retrieval, meanwhile falls from 81.9% to 79.3% for text retrieval. By contrast, in the same setting, the R@1 metric of the model “w M-DCL” only degrades by 0.9% and 0.9% for image retrieval and text retrieval, respectively. These results reveal that, even though the mini-batch size decreases sharply, our CODER with M-DCL can still achieve stable and superior performance, which is achieved by leveraging the coupled memory banks to enlarge interaction with negative samples. Additionally, the insensitivity to mini-batch size indicates our method is able to remain competitive even if the available computation resource is limited.

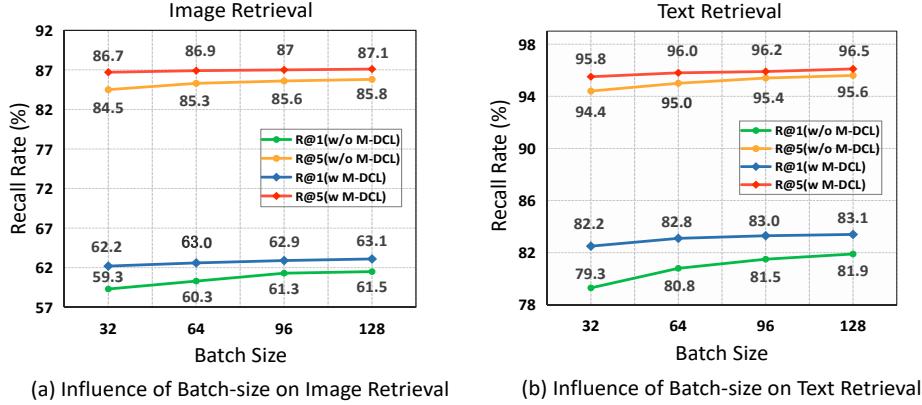


Fig. 3. Performance comparison of CODER model with M-DCL and without M-DCL. The model with M-DCL is abbreviated as “w M-DCL” and that without M-DCL is abbreviated as “w/o M-DCL”.

Table 4. Impact of different clustering number K in PGC loss on Flickr30K test set.

K	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
5000	82.9	96.3	98.0	63.1	87.0	92.7
10000	83.2	96.5	98.0	63.1	87.1	93.0
15000	83.2	96.3	98.2	63.0	87.1	92.8
20000	82.8	96.2	97.9	62.8	87.0	92.5

Impact of Different Configurations of PGC Loss In this part, we explore the influence of the clustering number K in PGC loss. The corresponding experimental results are listed in Table 4. It can be seen that the performance is not obviously affected by clustering number, archiving best results at $K = 10000$. Afterwards, the performance degrades slightly accompanied by the increase of clustering number, which implies the deceasing samples of one prototypical class may weaken the general semantics conveyed by concept-level representations.

4.5 Analysis on Accuracy and Efficiency of Model

The retrieval latency is also very important in real application scenario, whereas was seldom investigated in previous works. Thus, we report both retrieval recall and consuming time for more comprehensive performance comparisons. To achieve that, we compare our CODER with six leading methods [4, 10, 23, 31, 35, 54]. Note that the inference time of them are reported by re-implementing their open-sourced codes in the same environment. As shown in Figure 4, we can see that the inference speed of our method is comparable to CVSE, but its retrieval recall surpasses the latter by a large margin. Besides, in comparison to the best competitor SGRAF [10], our method surpasses it up to nearly $6 \times$ faster

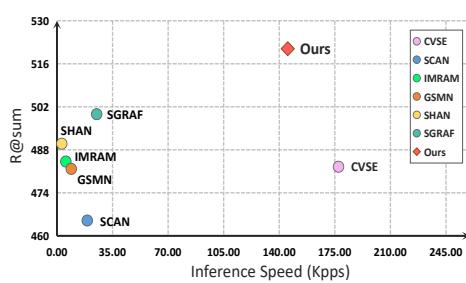


Fig. 4. Performance comparison of inference speed and recall between different methods. The Kpps on the horizontal axis denote the similarities of how many image-text pairs are calculated per second, the higher the better.



Fig. 5. The qualitative bi-directional retrieval results on Flickr30K dataset. For text retrieval, the ground-truth and non ground-truth descriptions are marked in red and black, respectively. For image retrieval, the number in the upper left corner denotes the ranking order, and the ground-truth images are annotated with green check mark.

for inference, meanwhile achieves considerable advantage over it on “R@sum” recall metric. Therefore, our method is superior to these approaches from both perspectives of effectiveness and efficiency.

4.6 Retrieval Result Visualization

To further qualitatively show the performance of our model, in Figure 5, we select several images and texts as queries to display their retrieval results on Flickr30K dataset. The bidirectional ITR results demonstrate our CODER model can return reasonable retrieval results.

5 Conclusions

In this paper, we proposed a Coupled Diversity-Sensitive Momentum Contrastive Learning (CODER) model for image-text retrieval. Specifically, Momentum Contrastive Learning (MCL) is extended to coupled form with dual dynamic modality-specific memory banks to enlarge interactions among instance pairs for cross-modal representation learning. Meanwhile, a novel diversity-sensitive contrastive loss is designed to take semantic ambiguity of sample embedding into account, which flexibly and dynamically allocate attention weights to negative pairs. In parallel, we devise an on-line clustering based strategy to exploit complementary knowledge between hierarchical semantics to promote discriminative feature learning. Furthermore, we systematically studied the impact of multiple components in our model, and its superiority is validated via substantially surpassing state-of-the-art approaches on two benchmarks with very low latency. In the near future, we plan to integrate our proposed learning paradigm into more large-scale vision-language pre-training models.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and vqa. In: CVPR (2018)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: ICCV (2015)
3. Bai, Y., Fu, J., Zhao, T., Mei, T.: Deep attention neural tensor network for visual question answering. In: ECCV (2018)
4. Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., Han, J.: Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: CVPR (2020)
5. Chen, J., Hu, H., Wu, H., Jiang, Y., Wang, C.: Learning the best pooling strategy for visual semantic embedding. In: CVPR (2021)
6. Chen, T., Deng, J., Luo, J.: Adaptive offline quintuplet loss for image-text matching. In: ECCV (2020)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
8. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: CVPR (2019)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
10. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: AAAI (2021)
11. Faghri, F., Fleet, D.J., Kiros, J., Fidler, S.: Vse++: improved visual-semantic embeddings. In: BMVC (2018)
12. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: CVPR (2015)
13. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: NeurIPS (2013)
14. Ge, X., Chen, F., Jose, J.M., Ji, Z., Wu, Z., Liu, X.: Structured multi-modal feature embedding and alignment for image-sentence retrieval. In: ACMMM (2021)
15. Ge, X., Chen, F., Jose, J.M., Ji, Z., Wu, Z., Liu, X.: Structured multi-modal feature embedding and alignment for image-sentence retrieval. ACMMM (2021)
16. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: NeurIPS (2020)
17. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR (2006)
18. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
19. Hua, T., Zheng, H., Bai, Y., Zhang, W., Zhang, X.P., Mei, T.: Exploiting relationship for complex-scene image generation. In: AAAI (2021)
20. Huang, Y., Wu, Q., Song, C., Wang, L.: Learning semantic concepts and order for image and sentence matching. In: CVPR (2018)
21. Huo, Y., Zhang, M., Liu, G., Lu, H., Gao, Y., Yang, G., Wen, J., Zhang, H., Xu, B., Zheng, W., et al.: Wenlan: Bridging vision and language by large-scale multi-modal pre-training. arXiv preprint arXiv:2103.06561 (2021)
22. Jain, A.: Data clustering: 50 years beyond k-means. Pattern Recognition Letters **31**(8), 651–666 (2010)

23. Ji, Z., Chen, K., Wang, H.: Step-wise hierarchical alignment network for image-text matching. In: IJCAI (2021)
24. Jiao, Y., Jie, Z., Chen, J., Ma, L., Jiang, Y.G.: Suspected object matters: Rethinking model's prediction for one-stage visual grounding. ArXiv:2203.05186 (2022)
25. Jiao, Y., Jie, Z., Luo, W., Chen, J., Jiang, Y.G., Wei, X., Ma, L.: Two-stage visual cues enhancement network for referring image segmentation. In: ACM MM (2021)
26. Karpathy, A., Li, F.F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
27. Kingma, D., Ba, J.: Diederik p. kingma and jimmy ba. In: ICLR (2014)
28. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2016)
29. Kiros, R., Salakhutdinov, R., Zemel, R.: Unifying visual-semantic embeddings with multimodal neural language models. In: NeurIPS Workshop (2014)
30. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
31. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: ECCV (2018)
32. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: ICCV (2019)
33. Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., Wang, H.: Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In: ACL (2021)
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
35. Liu, C., Mao, Z., Zhang, T., Xie, H., Wang, B., Zhang, Y.: Graph structured network for image-text matching. In: CVPR (2020)
36. Liu, F., Ye, R., Wang, X., Li, S.: Hal: Improved text-image matching by mitigating visual semantic hubs. In: AAAI (2020)
37. Liu, S., Fan, H., Qian, S., Chen, Y., Ding, W., Wang, Z.: Hit: Hierarchical transformer with momentum contrast for video-text retrieval. arXiv preprint arXiv:2103.15049 (2021)
38. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. ArXiv **abs/2104.08860** (2021)
39. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: ICCV (2015)
40. Ma, L., Lu, Z., Li, H.: Learning to answer questions from image using convolutional neural network. In: AAAI (2016)
41. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**, 2579–2605 (2008)
42. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). In: ICLR (2015)
43. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
44. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014)
45. Plummer, B.A., Wang, L., Cervantes, C., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV (2015)
46. Qu, L., Liu, M., Wu, J., Gao, Z., Nie, L.: Dynamic modality interaction modeling for image-text retrieval. In: SIGIR (2021)

47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
48. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
49. Shi, B., Ji, L., Lu, P., Niu, Z., Duan, N.: Knowledge aware semantic concept expansion for image-text matching. In: IJCAI (2019)
50. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval (2019)
51. Tanmay, G., Alexander, G.S., Derek, H.: Vico: Word embeddings from visual co-occurrences. In: ICCV (2019)
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
53. Wang, B., Ma, L., Zhang, W., Liu, W.: Reconstruction network for video captioning. In: CVPR (2018)
54. Wang, H., Zhang, Y., Ji, Z., Pang, Y., Ma, L.: Consensus-aware visual-semantic embedding for image-text matching. In: ECCV (2020)
55. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: CVPR (2016)
56. Wang, S., Wang, R., Yao, Z., Shan, S., Chen, X.: Cross-modal scene graph matching for relationship-aware image-text retrieval. In: WACV (2020)
57. Wang, Y., Zhang, T., Zhang, X., Cui, Z., Huang, Y., Shen, P., Li, S., Yang, J.: Wasserstein coupled graph learning for cross-modal retrieval. In: ICCV (2021)
58. Wehrmann, J., Kolling, C., Barros, R.C.: Adaptive cross-modal embeddings for image-text alignment. In: AAAI (2020)
59. Wei, J., Xu, X., Yang, Y., Ji, Y., Wang, Z., Shen, H.T.: Universal weighting metric learning for cross-modal matching. In: CVPR (2020)
60. Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F.: Multi-modality cross attention network for image and sentence matching. In: CVPR (2020)
61. Wen, K., Gu, X., Cheng, Q.: Learning dual semantic relations with graph attention for image-text matching. IEEE Transactions on Circuits and Systems for Video Technology **31**, 2866–2879 (2021)
62. Wu, W., Sun, Z., Ouyang, W.: Transferring textual knowledge for visual recognition. ArXiv:2207.01297 (2022)
63. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: CVPR (2017)
64. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)
65. Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., Gao, J.: Unified contrastive learning in image-text-label space. ArXiv **abs/2204.03610** (2022)
66. Zhang, L., Wu, H., Chen, Q., Deng, Y., Li, Z., Kong, D., Cao, Z., Siebert, J., Han, Y.: Vldeformer: Learning visual-semantic embeddings by vision-language transformer decomposing. ArXiv **abs/2110.11338** (2021)
67. Zhao, D., Wang, A., Russakovsky, O.: Understanding and evaluating racial biases in image captioning (2021)

– Supplementary Material –

**CODER: Coupled Diversity-Sensitive
Momentum Contrastive Learning
for Image-Text Retrieval**

In this appendix, we provide additional details which were omitted in the main manuscript owing to the limited space. From the perspective of algorithm details, we first give more technical details of representation modules. For instance-level representation, the feature aggregators will be introduced. For concept-level representation, the details of concept selection and the creation of statistical commonsense graph are described. Then, more designed motivation and illustration of our proposed Diversity-sensitive Contrastive Learning (DCL) loss will be given. After that, we also report more experimental results, including the performance of models with different data encoder, influence of different diversity estimation functions, impact of hyper-parameters, data distribution visualization of joint embedding space, performance comparison with different contrastive objectives, and bidirectional image-text retrieval results.

A Methodology

A.1 Aggregator for Instance-level Representation

For simplicity, here we only describe the image feature aggregator for visual modality, since the same goes for the textual branch. Specifically, we employ the Generalized Pooling Operator proposed in [5], which leverages the encoder-decoder architecture to build the image feature aggregator $g_{vis}(\cdot)$: (1) A positional encoding function that turns position index of local features into a vector. (2) A decoding module that takes the positional encoding output to produce pooling weights.

Position Encoder. To represent each position index l by a dense vector, the positional encoding strategy in Transformer [52] is adopted:

$$\mathbf{p}_l^i = \begin{cases} \sin(u_j, l), & \text{if } i = 2j, \forall i, \\ \cos(u_j, l), & \text{if } i = 2j + 1, \forall i. \end{cases} \quad (1)$$

where $u_j = \frac{1}{10000^{2j/d_p}}$ and d_p denotes the dimension for positional encoding.

Position Decoder. Given the dense vector $\mathbf{p}_l \in \mathbb{R}^{d_p}$, we feed them into a sequence model, which outputs the corresponding pooling weights $\theta = \{\theta\}_{l=1}^L$. The decoder function contains a bidirectional-GRU (BiGRU) and a two-layer perceptron (MLP):

$$\{\mathbf{h}\}_{l=1}^L = BiGRU(\{\mathbf{p}_l\}_{l=1}^L), \theta_k = MLP(\mathbf{h}_l) \quad (2)$$

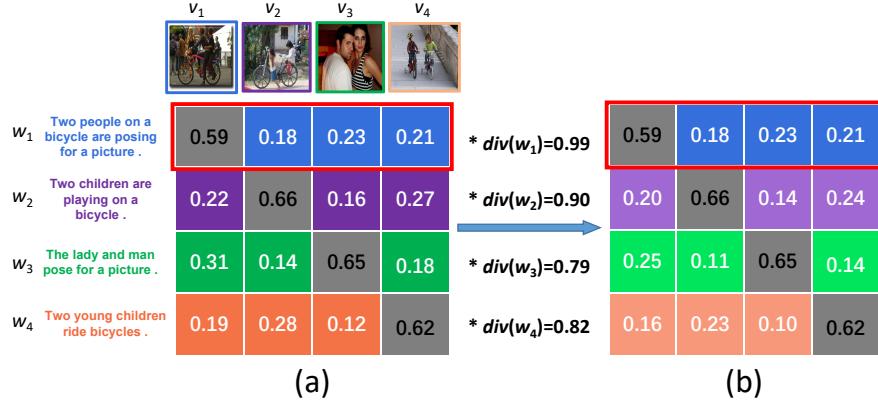


Fig. 1. Conceptual illustration of how diversity affects the cross-modal alignment. Sub-figure (a) depicts the image-text similarity matrix and Sub-figure (b) illustrates the similarity matrix being sensitive to the semantic diversity. After the incorporation of diversity score, our DCL loss will focus on handling the samples with high diversity.

where \mathbf{h}_l is the hidden states output by Bi-GRU at the position index k . Then, we can aggregate the local image features into a holistic instance-level image representation \mathbf{v}^I :

$$\mathbf{v}^I = g_{vis}(\{\mathbf{o}_l\}_{n=1}^L) = \sum_{l=1}^L \theta_l * \mathbf{o}_l \quad (3)$$

Similarly, we can obtain the textual feature aggregator $g_{text}(\cdot)$ and global instance-level text representation \mathbf{w}^I .

A.2 Concept-level Representation

Concept Initialization Our statistical commonsense knowledge is extracted from certain meaningful concepts and their semantic correlations, which are collected from the texts of the whole image-caption dataset. In order to filter out most meaningless and infrequent concepts, we follow [12, 20, 54] to select the representative words with top- q appearing frequencies in the concept vocabulary, which are roughly categorized into three types, *i.e.*, *Object*, *Motion*, and *Property*. Then, following [54], according to the appearance frequency over dataset, the ratio of the concepts with type of (*Object*, *Motion*, *Property*) is set to (7:2:1). Afterwards, we adopt the glove [44] to initialize them and denote them as \mathbf{X} .

Commonsense Aided Concept Representation. To model the statistical commonsense knowledge, we follow [54] to utilize the co-occurrence relationship between concepts to build one correlation graph. To be more specific, we construct a conditional probability matrix \mathbf{P} to model the relation between different

concepts, with each element \mathbf{P}_{ij} denoting the appearance probability of concept C_i when concept C_j appears: $\mathbf{P}_{ij} = \mathbf{B}_{ij}/N_i$, where $\mathbf{B} \in \mathbb{R}^{q \times q}$ is the concept co-occurrence matrix, \mathbf{B}_{ij} represents the co-occurrence times of C_i and C_j , and N_i is the appearance times of C_i in the corpus.

Afterwards, to further prevent the correlation matrix from being over-fitted and improve its generalization ability, we follow [8, 54] to apply binary operation to the rescaled matrix \mathbf{P} :

$$\mathbf{H}_{ij}^{sc} = \begin{cases} 0, & \text{if } \mathbf{P}_{ij} < \epsilon, \\ 1, & \text{if } \mathbf{P}_{ij} \geq \epsilon, \end{cases} \quad (4)$$

where ϵ denotes a threshold parameter filters noisy edges. Given the LCC representations \mathbf{X}^l and statistical commonsense graph \mathbf{H}^{ss} , we employ one Graph Convolution Network (GCN) [28] to process them, after one-layer convolution operation, the statistical commonsense aided concept (SCC) representations can be computed as:

$$\mathbf{Y} = \rho(\tilde{\mathbf{A}}_{sc}\mathbf{X}^l\mathbf{W}_{sc}) \quad (5)$$

where $\tilde{\mathbf{A}}_{sc} = \mathbf{D}_{ss}^{-\frac{1}{2}}\mathbf{H}^{sc}\mathbf{D}_{sc}^{-\frac{1}{2}} + \mathbf{I}$ denotes the normalized symmetric matrix and \mathbf{W}_{sc} is the learnable weight matrix.

Commonsense Aided Concept-level Representation. To generate concept-level representations, we generate representations $(\mathbf{v}_C^q$ and $\mathbf{w}_C^q)$ by using another group of feature aggregators ($g_{vis}(\cdot)$ and $g_{text}(\cdot)$) to combine local features $\{\mathbf{o}_i\}_{i=1}^L$ and $\{\mathbf{e}_t\}_{t=1}^T$, respectively. Note that the weights of both visual feature aggregators for \mathbf{v}_I and \mathbf{v}_C^q are shared, and we empirically find this operation helps to make our method converge better. Afterwards, \mathbf{v}_C^q and \mathbf{w}_C^q are taken as input vectors to query from the SCC representations \mathbf{Y} . As consequence, the output scores for different concepts allow us to uniformly utilize the linear combination of the SCC representations to represent both modalities. Mathematically, the concept-level representation \mathbf{v}^C and \mathbf{w}^C can be calculated as:

$$\begin{aligned} \mathbf{v}^C &= \sum_{i=1}^g a_i^v \mathbf{y}_i; \quad a_i^v = \frac{e^{\lambda \mathbf{v}_C^I \mathbf{W}^v \mathbf{y}_i^\top}}{\sum_{i=1}^q e^{\lambda \mathbf{v}_C^I \mathbf{W}^v \mathbf{y}_i^\top}}. \\ \mathbf{w}^C &= \sum_{j=1}^g a_j^w \mathbf{y}_j; \quad a_j^w = \frac{e^{\lambda \mathbf{w}_C^I \mathbf{W}^w \mathbf{y}_j^\top}}{\sum_{j=1}^q e^{\lambda \mathbf{w}_C^I \mathbf{W}^w \mathbf{y}_j^\top}} \end{aligned} \quad (6)$$

where $\mathbf{W}^v \in \mathbb{R}^{F \times F}$ and $\mathbf{W}^w \in \mathbb{R}^{F \times F}$ denote the learnable parameter matrix, \mathbf{a}_i^v and \mathbf{a}_j^w denote the visual and textual score corresponding to the concept \mathbf{z}_i , respectively. λ controls the smoothness of the softmax function.

A.3 Illustration of Diversity in DCL

In this section, we describe how our proposed semantic *diversity* affects the cross-modal alignment. First, we briefly review the mathematical definitions of

diversity and DCL loss defined in the main manuscript. Specifically, we take as example that visual feature \mathbf{v}_i is an anchor sample and Q text features $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_Q\}$ are to be compared (among which only \mathbf{w}_i is a matching sample for \mathbf{v}_i), to illustrate how we estimate diversity of an anchor sample. The cosine similarity of $\text{cosine}(\mathbf{v}_i, \mathbf{w}_j)$ is defined as S_{ij} . Then, the diversity of anchor \mathbf{v}_i (\mathbf{w}_i) is defined as:

$$\begin{aligned} SD(i) &= \sqrt{E(S_{ij}) - [E(S_{ij})]^2}, i \neq j, j \in [1, Q]; \\ div_{std}(\mathbf{v}_i) &= 1/\sigma(\epsilon/SD(\mathbf{v}_i)); \\ div_{std}(\mathbf{v}_i) &= div_{std}(\mathbf{v}_i)/\max\{div_{std}(\mathbf{v}_1), \dots, div_{std}(\mathbf{v}_Q)\}, \end{aligned} \quad (7)$$

where $E(\cdot)$ is the mathematical expectation function and $\sigma(\cdot)$ denotes the Sigmoid function that normalizes the reciprocal of SD value to a uniform scale. $div_{std}(\mathbf{v}_i)$ denotes the diversity score of \mathbf{v}_i calculated from the candidate textual samples to be compared with. $\epsilon = 0.1$ is a tuning parameter. Lastly, we divide each diversity score $div_{std}(\mathbf{v}_i)$ by the maximum value of them in mini-batch for normalization. Similarly, the diversity of \mathbf{w}_j can be obtained.

Except for the definition above, we also explore another method to define diversity, which is built based on employing *information entropy*. It is commonly used to measure the information volume conveyed by variables. To incorporate cross-modal similarity into information entropy computation, we first utilize softmax function to convert similarity score to probability form. Then, we can use information entropy to estimate the diversity of anchor sample. Formally, the information entropy based diversity of anchor \mathbf{v}_i (\mathbf{w}_i) is defined as:

$$\begin{aligned} P_{ij} &= \frac{e^{S_{ij}}}{\sum_{j=1}^Q e^{S_{ij}}}; \\ H(\mathbf{v}_i) &= -P_{ij} \cdot \sum_{j=1}^Q \log_2(P_{ij}), i \neq j; \\ div_{ent}(i) &= 1/\sigma(\epsilon/H(i)); \\ div_{ent}(\mathbf{v}_i) &= div_{ent}(\mathbf{v}_i)/\max\{div_{ent}(\mathbf{v}_1), \dots, div_{ent}(\mathbf{v}_Q)\}, \end{aligned} \quad (8)$$

where $H(\cdot)$ represents the function for calculating information entropy. $div_{ent}(\mathbf{v}_i)$ denotes the information entropy based diversity score of \mathbf{v}_i calculated from the candidate textual samples to be compared with. The effect comparison between two types of diversity will be presented in Section B.2.

Furthermore, given $\mathbf{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_N\}$ and $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_Q\}$, based on the diversity score defined in Eq.7, our proposed DCL loss L_{DCL} is defined as:

$$\begin{aligned} L_{DCL}(\mathbf{V}, \mathbf{W}) &= l_{DCL}(\mathbf{W}, \mathbf{V}) + l_{DCL}(\mathbf{V}, \mathbf{W}) \\ l_{DCL}(\mathbf{V}, \mathbf{W}) &= \frac{\mu}{N} \sum_{n=1}^N [\log(\sum_{q \neq n} \exp(\frac{(S_{nq} - \gamma)}{\mu \cdot div_{std}(\mathbf{v}_n)}) + 1) - \log(S_{nn} + 1)]; \\ l_{DCL}(\mathbf{W}, \mathbf{V}) &= \frac{\mu}{Q} \sum_{q=1}^Q [\log(\sum_{n \neq q} \exp(\frac{(S_{qn} - \gamma)}{\mu \cdot div_{std}(\mathbf{w}_q)}) + 1) - \log(S_{qq} + 1)]; \end{aligned} \quad (9)$$

where $div_{std}(\mathbf{v}_n)$ and $div_{std}(\mathbf{w}_q)$ denotes diversity of \mathbf{v}_n and \mathbf{w}_q , respectively and they are used to adaptively weight each negative sample.

Then, we take the similarity measuring matrix of text-to-image as example. As shown in Figure1(a), for query sample \mathbf{w}_1 with higher diversity, *i.e.* semantic ambiguity, the calculated similarity difference between it and other three mismatched samples is very small. According to Eq.7, the diversity of \mathbf{w}_1 is larger than those of others. As a result, seeing Figure1(b), the similarity ratio between positive pair and negative pairs of \mathbf{w}_1 remain unchanged. By contrast, those of other samples all become larger than before. From Eq.9, we can know that this adaptive weighting strategy will lead to imposing harder punishment on sample \mathbf{w}_1 than others. And it is consistent with our designing principle.

Note that distinct from previous works [6, 11] that focus on mining hard negatives specifically for a single sample, the diversity in DCL is defined from a more holistic view, which is calculated based on the statistical information of data distribution. Consequently, our DCL loss aims at reducing the cross-modal distribution discrepancy, which captures more hierarchical semantic structure in joint space by alleviating the negative influence brought by samples with high diversity. Furthermore, we will display how the diversity in DCL loss affects data distribution in the joint embedding space by t-SNE visualizing in Section B.4.

B Experiments

B.1 More Results and Comparisons for Image-Text Retrieval

The additional experimental results are presented in Table 1. Note that the results of [5] are reported by our replicated number. Since the instance-level representation part of our model is built according to [5], a solid performance of baseline is needed to reasonably evaluate the impact of our contributions. Thus, we report our replicated results by using their open-sourced code with no change, and mark them with * symbol in Table 1 & 2. To further assure the fairness of comparisons, we divide the experiments into two groups. One group of approaches adopt “Faster-RCNN + BiGRU” as image and text encoders, meanwhile another group of methods is uniformly built based on “Faster-RCNN + BERT” as encoders. The experimental results on Flickr30K test set are presented in Table 1. First, in contrast to other methods adopting “Faster-RCNN + BiGRU” architecture, the “R@sum” achieved by our CODER surpasses the second best performance by 13.6%. Secondly, compared with those employing “Faster-RCNN + BERT” for encoding multi-modal data, our method outperforms the best competitor by 13.7% on the “R@sum” metric.

As shown in Table 1, on MSCOCO 1k test set, our CODER also significantly outperforms all other compared methods. Although employing the “Faster-RCNN + BiGRU” as image and text encoders, there is still a performance gap between CODER and best competitor SMFEA [15] on the R@sum metric, *e.g.* 4.0% improvement. Moreover, the retrieval performances on MSCOCO 5K test set are listed in Table 2. Comparing with best competitor GPO (BERT) [5], our

Table 1. Comparisons of experimental results on MSCOCO 1K test set and Flickr30k test set, employing different image and text encoders (denoted by bold section title).

Methods	Image Encoder	MSCOCO 1K										Flickr30K									
		Text Retrieval			Image Retrieval			R@sum	Text Retrieval			Image Retrieval			R@sum						
		R@1	R@5	R@10		R@1	R@5	R@10		R@1	R@5	R@10		R@1	R@5	R@10		R@1	R@5	R@10	
Faster-RCNN + BiGRU																					
SCAN [31] (2018)	Faster-RCNN	72.7	94.8	98.4	58.8	88.4	94.8	507.9	67.4	90.3	95.8	48.6	77.7	85.2	465.0						
VSRN [32] (2019)	Faster-RCNN	76.2	94.8	98.2	62.8	89.7	95.1	516.8	71.3	90.6	96.0	54.7	81.8	88.2	482.6						
CVSE [54] (2020)	Faster-RCNN	74.8	95.1	98.3	59.9	89.4	95.2	512.7	73.5	92.1	95.8	52.9	80.4	87.8	482.5						
MMCA [60] (2020)	Faster-RCNN	74.8	95.6	97.7	61.6	89.8	95.2	514.7	74.2	92.8	96.4	54.8	81.4	87.8	487.4						
GSMN [35] (2020)	Faster-RCNN	76.1	95.6	98.3	60.4	88.7	95.0	514.0	74.4	91.5	95.3	54.1	79.9	86.6	481.8						
SMFEA [15] (2021)	Faster-RCNN	75.1	95.4	98.3	62.5	90.1	96.2	517.6	73.7	92.5	96.1	54.7	82.1	88.4	487.5						
WCGL [57] (2021)	Faster-RCNN	75.4	95.5	98.6	60.8	89.3	95.3	514.9	74.8	93.3	96.8	54.8	80.6	87.5	487.8						
GPO (BiGRU) [5] (2021) *	Faster-RCNN	76.2	95.4	98.5	60.1	89.8	95.2	515.2	74.8	93.5	97.0	55.1	83.8	89.4	493.6						
CODER (BiGRU)	Faster-RCNN	78.9	95.6	98.6	62.5	90.3	95.7	521.6	79.4	94.9	97.7	59.0	85.2	91.0	507.2						
Faster-RCNN + BERT																					
DSRAN [61] (2021)	Faster-RCNN	77.1	95.3	98.1	62.9	89.9	95.3	518.6	75.3	94.4	97.6	57.3	84.8	90.9	500.3						
GPO (BERT) [5] (2021) *	Faster-RCNN	78.6	96.2	98.7	62.9	90.8	96.1	523.3	78.1	94.1	97.8	57.4	84.5	90.4	502.3						
DIME (i-t) [46] (2021)	Faster-RCNN	77.9	95.9	98.3	63.0	90.5	96.2	521.8	77.4	95.0	97.4	60.1	85.5	91.8	507.2						
CODER (BERT)	Faster-RCNN	82.1	96.6	98.8	65.5	91.5	96.2	530.6	83.2	96.5	98.0	63.1	87.1	93.0	520.9						

Table 2. Comparisons of experimental results on MSCOCO 5K test set, employing different image and text encoders (denoted by bold section title).

Methods	Image Encoder	MSCOCO 5K										
		Text retrieval			Image Retrieval			R@sum				
		R@1	R@5	R@10		R@1	R@5	R@10				
Faster-RCNN + BiGRU												
SCAN [31] (2018)	Faster-RCNN	50.4	82.2	90.0	38.6	69.3	80.4	410.9				
VSRN [32] (2019)	Faster-RCNN	53.0	81.1	89.4	40.5	70.6	81.1	415.7				
MMCA [60] (2020)	Faster-RCNN	54.0	82.5	90.7	38.7	69.7	80.8	416.4				
SMFEA [15] (2021)	Faster-RCNN	54.2	-	89.9	41.9	-	83.7	-				
GPO (BiGRU) [5] (2021) *	Faster-RCNN	55.2	83.1	91.0	39.3	69.9	81.1	419.6				
CODER (BiGRU)	Faster-RCNN	58.5	84.3	91.5	40.9	70.8	81.4	427.2				
Faster-RCNN + BERT												
DSRAN [61] (2021)	Faster-RCNN	53.7	82.1	89.9	40.3	70.9	81.3	418.2				
DIME (i-t) [46] (2021)	Faster-RCNN	56.1	83.2	91.1	40.2	70.7	81.4	422.7				
GPO (BERT) [5] (2021) *	Faster-RCNN	57.3	84.5	91.6	41.1	71.9	82.6	429.0				
CODER (BERT)	Faster-RCNN	62.6	86.6	93.1	42.5	73.1	83.3	441.3				

CODER outperforms it by 5.3% improvement for text retrieval and 1.4% for image retrieval on R@1 criteria. The above results are obtained under totally fair conditions with the same data encoders, thus they can solidly validate the superiority of our method for image-text retrieval.

B.2 Impact of Different Functions for Diversity Estimation

In this section, we explore the effect of different diversity estimation functions. As shown in Table 3, the experimental results based on estimation functions of $div_{std}(\cdot)$ and $div_{ent}(\cdot)$ are listed. For comparison, the results without diversity estimating are also presented. From Table 3, we can see our proposed two types of diversity estimation functions can both bring about substantial performance

Table 3. Impact of different diversity estimation functions in DCL loss on Flickr30K test set. Explicit diversity estimation is abbreviated as “EE”.

EE	Diversity Estimation Function	Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
✗	-	81.5	95.8	98.1	61.3	85.7	91.0
✓	$div_{std}(\cdot)$	83.2	96.5	98.0	63.1	87.1	93.0
✓	$div_{ent}(\cdot)$	83.0	96.2	98.1	62.4	86.9	92.5

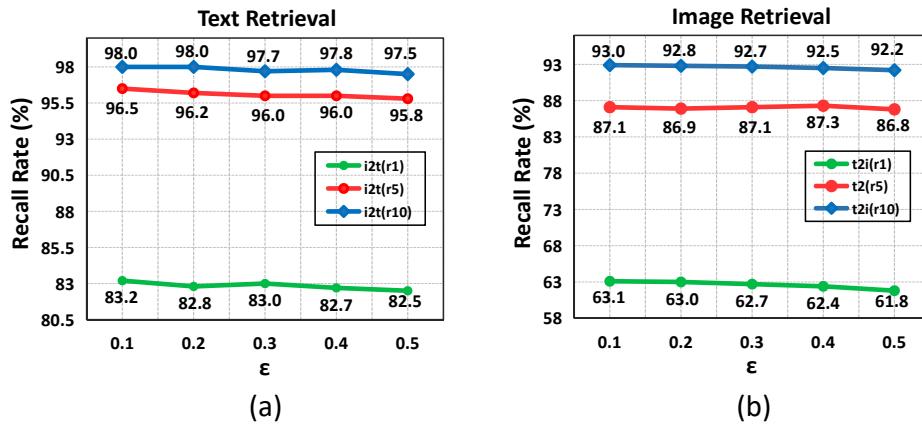


Fig. 2. Impact of varied controlling parameters ϵ on Flickr30K test set. Sub-figure (a) shows image-to-text retrieval performance with different values of ϵ in DCL loss. Sub-figure (b) depicts the corresponding text-to-image retrieval performance.

boost. It further validates our train of thought for diversity estimation is reasonable. Besides, the performance of model using $div_{ent}(\cdot)$ is slightly inferior to that with $div_{std}(\cdot)$. The potential reason may be that the softmax function adopted in $div_{ent}(\cdot)$ will make the original data distribution of cross-modal similarity to be more smooth.

B.3 Hyper-Parameter Analysis for Diversity Estimation

In this part, we investigate the affect of controlling parameter ϵ of diversity in Eq.7 on retrieval performance. As shown in Figure 2, with the variant ϵ , the retrieval results vary moderately, indicating our model is robust to ϵ within a proper range. Additionally, the increase of ϵ value implicates the narrower variation range of diversity score. Thus, from Figure 2, we can infer the proper sensitiveness of DCL loss on parameter ϵ also leads to performance gain. Overall, these results reveal that diversity plays critical role in DCL loss for learning more discriminative cross-modal embeddings.

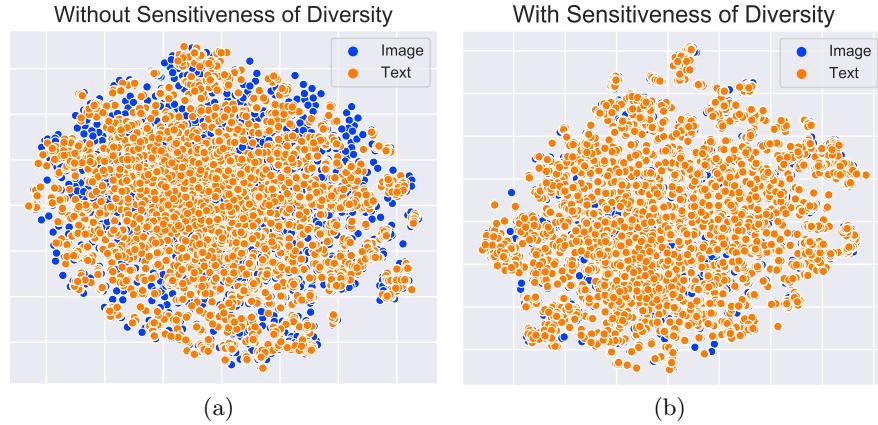


Fig. 3. T-SNE visualization of the image-text representations generated by (a) baseline model with L_{DCL_I} loss and (b) full CODER model on Flickr30K test set (1000 images and 5000 texts).

	Query	Baseline	CODER
Flickr30K		<p>1. A pride softball team member hits the ball and runs towards first base while the umpire and catcher watch the ball. 2. A girl dressed in a red uniform is hitting a softball with a bat while a catcher and home plate umpire look on. 3. A woman runs after making a hit in women's softball , the catcher rises to her feet . 4. A baseball catcher trying to tag a base runner in a baseball game .</p>	<p>1. A pride softball team member hits the ball and runs towards first base while the umpire and catcher watch the ball. 2. A woman runs after making a hit in women's softball , the catcher rises to her feet . 3. Girl hits a ball and the catcher looks on. 4. A girl dressed in a red uniform is hitting a softball with a bat while a catcher and home plate umpire look on.</p>
		<p>1. Three male field hockey players are running onto the field while the goalie is standing in the goal looking on .. 2. The three field hockey players dressed in orange make for the ball . 3. A group of guys are playing roller hockey . 4. A large goalie towers over his opposing teammates .</p>	<p>1. Three male field hockey players are running onto the field while the goalie is standing in the goal looking on .. 2. The three field hockey players dressed in orange make for the ball . 3. A team in orange uniforms are near a goal and a goalkeeper in green . 4. A large goalie towers over his opposing teammates .</p>
		<p>Two large dogs chase after another dog that has a ball in his mouth and runs from them .</p> <p>① ② ③ </p>	<p>① ② ③ </p>
		<p>A bicyclist near town is racing in a race while , wearing yellow and a helmet .</p> <p>① ② ③ </p>	<p>① ② ③ </p>
MSCOCO		<p>1. A man holding a book and a phone in his hands. 2. person lying on ground reading book and holding cell phone. 3. Person laying on ground looking at book and phone. 4. A person holding up a smart phone in a public space.</p>	<p>1. A man holding a book and a phone in his hands. 2. person lying on ground reading book and holding cell phone. 3. A person laying down with a book in one hand and a cell phone in another. 4. Person laying on ground looking at book and phone.</p>
		<p>1. The two giraffes are walking in their pen. 2. Two giraffes in a grassy area with a fence and trees next to them. 3. A couple of giraffes that are standing in a fence. 4. Two giraffes standing in a brush covered area.</p>	<p>1. Two giraffes in a grassy area with a fence and trees next to them. 2. The two giraffes are walking in their pen. 3. Two giraffes roaming around an enclosed area on a sunny day. 4. A couple of giraffes that are standing in a fence.</p>
		<p>A tray topped with two sandwiches, pie and a plate of coleslaw.</p> <p>① ② ③ </p>	<p>① ② ③ </p>
		<p>A row of motorcycles parked in front of a building.</p> <p>① ② ③ </p>	<p>① ② ③ </p>

Fig. 4. The qualitative bi-directional retrieval results on Flickr30k and MSCOCO datasets. For text retrieval, the ground-truth and non ground-truth describing sentences are marked in red and black, respectively. For image retrieval, the number in the upper left corner denotes the ranking order, and the ground-truth images are annotated with green check mark.

B.4 T-SNE Visualization of Cross-Modal Representation

To better understand how our DCL loss affects the cross-modal joint embedding space, we utilize t-SNE [41] to visualize the learned representations from Flickr30K test set, including 1000 images and 5000 texts. Specifically, Figure 3(a) displays the feature distribution of the baseline model (referring to the model #1 in Table 4 defined in the main manuscript, employing the $L_{DCL,I}$ loss as learning objective), and those of full CODER model is illustrated in Figure 3(b). We can see that the data distribution in Figure 3(b) is obviously more desirable than that in Figure 3(a), which lies in two main points: 1) The distribution discrepancy between images and texts is alleviated significantly. 2) The learned joint space is characterized by being structured and hierarchical rather than being irregular and scattered. Considering the unique difference between both models is the varied configuration of DCL loss, we believe the main factor improving the data distribution is the combination between coupled memory banks ($L_{M_DCL}^I$) and diversity estimation. Benefiting from the large-scale negative interactions from the former, we can achieve more accurate diversity estimation for DCL. It is able to regularize the joint embedding space by alleviating the influence of sample with high diversity, such as some visual instances existing on the left side of Figure 3(a).

B.5 Retrieval Result Visualization

To further validate the effectiveness of our method, in Figure 4, we choose several images and texts as queries and exhibit their retrieval results. Note that we take CODER adopting BTR loss [11] instead of our DCL loss as baseline model. As shown in Figure 4, comparing with baseline, the CODER model with the aid of DCL loss is able to return better image-text retrieval results.