

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

Haoran Wang<sup>1</sup>, Dongliang He<sup>1\*</sup>, Wenhao Wu<sup>1,5</sup>, Boyang Xia<sup>2</sup>, Min Yang<sup>1</sup>, Fu Li<sup>1</sup>, Yunlong Yu<sup>3</sup>, Zhong Ji<sup>4</sup>, Errui Ding<sup>1</sup>, and Jingdong Wang<sup>1</sup>

<sup>1</sup> Department of Computer Vision Technology (VIS), Baidu Inc., Beijing, China

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

<sup>3</sup> College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou, China

<sup>4</sup> School of Electrical & Information Engineering, Tianjin University, Tianjin, China

<sup>5</sup> 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

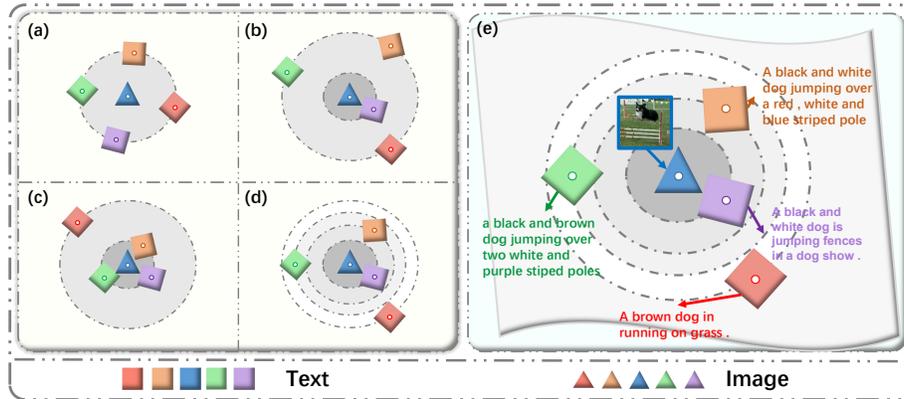
**Abstract.** Image-Text Retrieval (ITR) is challenging in bridging visual and lingual modalities. Contrastive learning has been adopted by most prior arts. Except for limited amount of negative image-text pairs, the capability of constrastive learning is restricted by manually weighting negative pairs 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, 16, 21, 22, 37, 47, 48, 57–60, 62]. In the past decade, tremendous progresses have been made with the prevalence of deep learning [27]. Early works typically associate image with text via learning global [10, 26, 50] or local cross-modal alignment [4, 28]. Follow-up studies attempt to introduce external knowledge information, including commonsense knowledge [45, 49] or scene graph [51] 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 [10, 11] can be regarded as one special case of contrastive loss [14], where only one negative sample is considered. Then, bidirectional Info-NCE loss [39] (BIN), as a typical contrastive loss, has been widely adopted in many tasks [7, 35, 43]. 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 [15], 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 [46]. 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 [15], 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 [31] and Flickr30K [41] 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, 13, 15, 39, 43] has made remarkable progress in unsupervised representation learning. Chen *at el.* [7] shows that contrastive learning in unsupervised visual representation learning benefits from large batch size negatives and stronger data augmentation. He *at el.* [15] 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 [18, 30, 43, 61] are emerging. For video-text retrieval, Liu *at el.* [34] 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 [11, 36, 38, 50] typically employ global features to represent both image and text in a common semantic space. For instance, Kiros *at el.* [26] encoded image and text by CNN and RNN respectively, utilizing BTR loss to train the model. Afterwards, another line of research [4, 9, 28, 53, 55]

employed multi-modal attention mechanism [4, 20, 28, 56] or knowledge aided representation learning [12, 17, 32, 45, 49] to achieve cross-modal alignment by exploiting more fine-grained associations. For instance, Lee *et al.* [28] developed Stacked Cross Attention Network that aligns image regions and textual words.

Except for focusing on representation architecture designing, some studies [6, 10, 33, 54] endeavored to improve the learning objectives. As a seminal work, Faghri [10] *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.* [33] 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

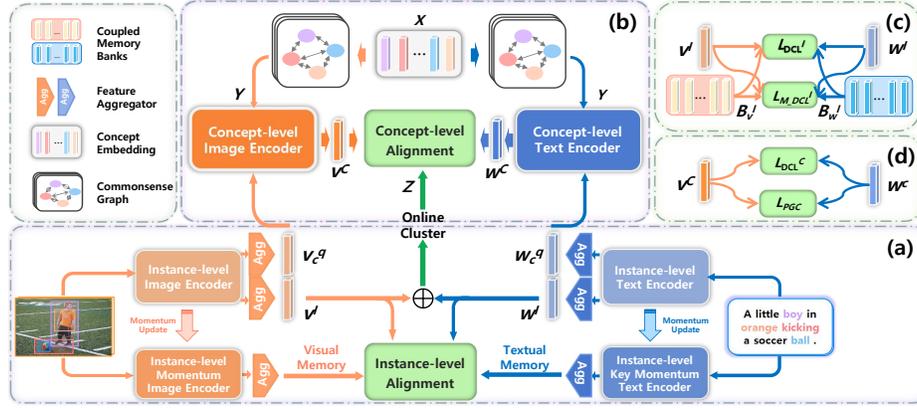
The overall framework of our proposed CODER model is illustrated in Figure 2. In our model, two branches are designed for instance-level and concept-level representation learning. In the instance-level branch (Fig.2(a)), image and text features are encoded and aggregated to be  $\mathbf{v}^I$  and  $\mathbf{w}^I$ , 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) [49] 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  $\mathbf{v}^C$  and  $\mathbf{w}^C$  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, 44] 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 [8] is our textual encoder and  $N$  word-level embeddings  $\{\mathbf{e}_t\}_{t=1}^T$  are also aggregated to instance-level textual embedding  $\mathbf{w}^I \in \mathbb{R}^F$ .

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

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 a 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 [40] is employed to instantiate these concepts as  $\mathbf{X}$ . Following [49], graph convolution network (GCN) [25] 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  $\{\mathbf{o}_1\}_{l=1}^L$  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:

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

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.  $\lambda$  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 [15] 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  $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_Q\}$  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  $\cosine(\mathbf{v}_i, \mathbf{w}_j)$  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  $L_{DCLI}$ . Given  $\mathbf{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_N\}$  and  $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_Q\}$ ,  $L_{DCLI}(\mathbf{V}, \mathbf{W})$  can be formulated as:

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

where  $\mu$  is a temperature scalar;  $\gamma$  is a margin parameter;  $N$  is the number of samples within the mini-batch;  $S_{nq} = \text{cosine}(\mathbf{v}_n, \mathbf{w}_q)$ ,  $S_{qn} = \text{cosine}(\mathbf{w}_q, \mathbf{v}_n)$ ,  $S_{nn} = \text{cosine}(\mathbf{v}_n, \mathbf{v}_n)$  and  $S_{qq} = \text{cosine}(\mathbf{w}_q, \mathbf{w}_q)$  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 [19] 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:

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

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  $\beta$  is a balancing parameter.

## 4 Experiments

### 4.1 Datasets & Evaluation Metrics

**Datasets.** Flickr30K [41] is an image-caption dataset containing 31,783 images, where each image annotated with five sentences. Following [38], we split the dataset into 29,783 training, 1000 validation, and 1000 testing images. The performance evaluation is reported on 1000 testing set. MSCOCO [31] is another

**Table 1.** Comparisons of experimental results on MSCOCO 1K test set and Flickr30K test set. Note that DSRAN [56], GPO [5] and DIME [42] 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@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			
DVSA [23] (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	235.2
m-CNN [36] (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	324.7
DSPE [50] (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	351.0
VSE++ [10] (2018)	ResNet-152	64.7	-	95.9	52.0	-	92.0	-	52.9	-	87.2	39.6	-	79.5	-
SCAN [28] (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
PVSE [46] (2019)	Faster-RCNN	69.2	91.6	96.6	55.2	86.5	93.7	492.8	-	-	-	-	-	-	-
VSRR [29] (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 [49] (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
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	484.2
WCGL [52] (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
SHAN [20] (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	490.0
DSRAN [56] (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 [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) [42] (2021)	Faster-RCNN	77.9	95.9	98.3	63.0	90.5	<b>96.2</b>	521.8	77.4	95.0	97.4	60.1	85.5	91.8	507.2
SGRAF [9] (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	499.6
CODER	Faster-RCNN	<b>82.1</b>	<b>96.6</b>	<b>98.8</b>	<b>65.5</b>	<b>91.5</b>	<b>96.2</b>	<b>530.6</b>	<b>83.2</b>	<b>96.5</b>	<b>98.0</b>	<b>63.1</b>	<b>87.1</b>	<b>93.0</b>	<b>520.9</b>

image-caption dataset, totally including 123,287 images with each image roughly annotated with five textual descriptions. We follow the public split of [23], 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 [8] 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 [40] 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. (5). Our CODER model is trained by Adam optimizer [24] with mini-batch size of 128. The learning rate is set to be 0.0002 for the first 15 epochs and 0.00002 for the next 15 epochs for both datasets. The balancing parameter in Eq. (8) is set to  $\lambda = 3$ . For inference, the controlling parameter  $\beta$  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 1 <sup>6</sup>. 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 [10] 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

<sup>6</sup>We 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”.

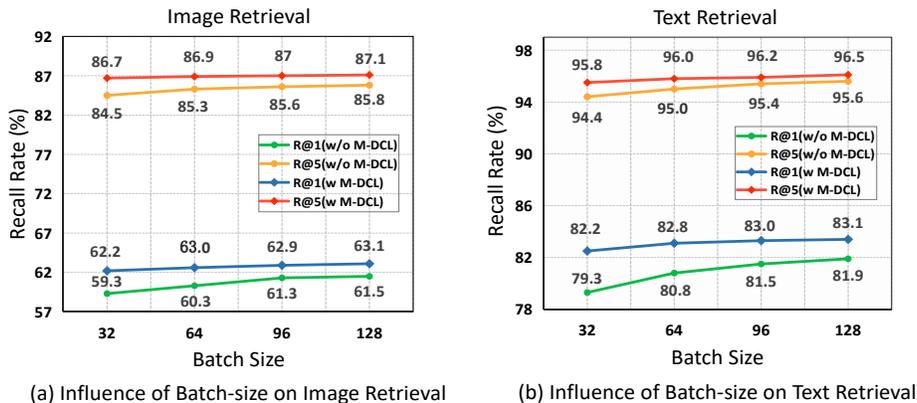
Models	IA		CA		Text Retrieval			Image Retrieval		
	$L_{M\_DCL}^I$	$L_{DCL}^I$	$L_{PGC}$	$L_{DCL}^C$	R@1	R@5	R@10	R@1	R@5	R@10
1					78.7	94.5	97.0	58.6	84.8	90.1
2		✓			81.9	95.6	97.9	61.5	85.8	91.8
3	✓	✓			82.6	96.1	98.0	62.6	86.7	92.3
4	✓	✓		✓	82.8	96.1	98.0	62.8	86.8	92.6
5	✓	✓	✓	✓	<b>83.2</b>	<b>96.5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

**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	$L_{M\_DCL}^I$	$L_{DCL}^I$			Text Retrieval			Image Retrieval		
		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	✓	✓	✓	✓	<b>83.2</b>	<b>96.5</b>	<b>98.0</b>	<b>63.1</b>	<b>87.1</b>	<b>93.0</b>

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”.

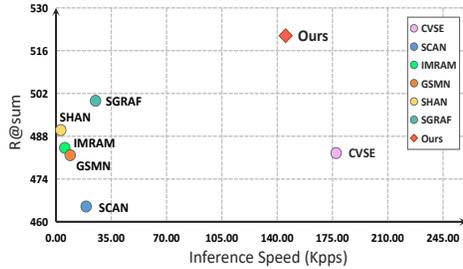
**Table 4.** Impact of different clustering number  $K$  in PGC loss on Flickr30K test set.

$K$	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
5000	82.9	96.3	98.0	<b>63.1</b>	87.0	92.7
10000	<b>83.2</b>	<b>96.5</b>	98.0	<b>63.1</b>	<b>87.1</b>	<b>93.0</b>
15000	<b>83.2</b>	96.3	<b>98.2</b>	63.0	<b>87.1</b>	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 decreasing 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, 9, 20, 28, 32, 49]. 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 [9], 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. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
9. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: AAAI (2021)
10. Faghri, F., Fleet, D.J., Kiros, J., Fidler, S.: Vse++: improved visual-semantic embeddings. In: BMVC (2018)
11. 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)
12. 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)
13. Grill, J.B., Strub, F., Althé, 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)
14. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR (2006)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
16. Hua, T., Zheng, H., Bai, Y., Zhang, W., Zhang, X.P., Mei, T.: Exploiting relationship for complex-scene image generation. In: AAAI (2021)
17. Huang, Y., Wu, Q., Song, C., Wang, L.: Learning semantic concepts and order for image and sentence matching. In: CVPR (2018)
18. 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)
19. Jain, A.: Data clustering: 50 years beyond k-means. *Pattern Recognition Letters* **31**(8), 651–666 (2010)
20. Ji, Z., Chen, K., Wang, H.: Step-wise hierarchical alignment network for image-text matching. In: IJCAI (2021)
21. 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)
22. 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)

23. Karpathy, A., Li, F.F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
24. Kingma, D., Ba, J.: Diederik p. kingma and jimmy ba. In: ICLR (2014)
25. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2016)
26. Kiros, R., Salakhutdinov, R., Zemel, R.: Unifying visual-semantic embeddings with multimodal neural language models. In: NeurIPS Workshop (2014)
27. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *nature* **521**(7553), 436–444 (2015)
28. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: ECCV (2018)
29. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: ICCV (2019)
30. 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)
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
32. Liu, C., Mao, Z., Zhang, T., Xie, H., Wang, B., Zhang, Y.: Graph structured network for image-text matching. In: CVPR (2020)
33. Liu, F., Ye, R., Wang, X., Li, S.: Hal: Improved text-image matching by mitigating visual semantic hubs. In: AAAI (2020)
34. 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)
35. 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)
36. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: ICCV (2015)
37. Ma, L., Lu, Z., Li, H.: Learning to answer questions from image using convolutional neural network. In: AAAI (2016)
38. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). In: ICLR (2015)
39. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
40. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014)
41. 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)
42. Qu, L., Liu, M., Wu, J., Gao, Z., Nie, L.: Dynamic modality interaction modeling for image-text retrieval. In: SIGIR (2021)
43. 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)
44. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
45. Shi, B., Ji, L., Lu, P., Niu, Z., Duan, N.: Knowledge aware semantic concept expansion for image-text matching. In: IJCAI (2019)

46. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval (2019)
47. Tanmay, G., Alexander, G.S., Derek, H.: Vico: Word embeddings from visual co-occurrences. In: ICCV (2019)
48. Wang, B., Ma, L., Zhang, W., Liu, W.: Reconstruction network for video captioning. In: CVPR (2018)
49. Wang, H., Zhang, Y., Ji, Z., Pang, Y., Ma, L.: Consensus-aware visual-semantic embedding for image-text matching. In: ECCV (2020)
50. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: CVPR (2016)
51. Wang, S., Wang, R., Yao, Z., Shan, S., Chen, X.: Cross-modal scene graph matching for relationship-aware image-text retrieval. In: WACV (2020)
52. 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)
53. Wehrmann, J., Kolling, C., Barros, R.C.: Adaptive cross-modal embeddings for image-text alignment. In: AAAI (2020)
54. Wei, J., Xu, X., Yang, Y., Ji, Y., Wang, Z., Shen, H.T.: Universal weighting metric learning for cross-modal matching. In: CVPR (2020)
55. Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F.: Multi-modality cross attention network for image and sentence matching. In: CVPR (2020)
56. 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)
57. Wu, W., Sun, Z., Ouyang, W.: Transferring textual knowledge for visual recognition. *ArXiv:2207.01297* (2022)
58. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: CVPR (2017)
59. 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)
60. 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)
61. 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)
62. Zhao, D., Wang, A., Russakovsky, O.: Understanding and evaluating racial biases in image captioning (2021)