A Differentiable Semantic Metric Approximation in Probabilistic Embedding for Cross-Modal Retrieval

Anonymous Author(s)
Affiliation
Address
email

Abstract

Cross-modal retrieval aims to build correspondence between multiple modalities by learning a common representation space. The multiplicity of correspondences between images and texts greatly increases the difficulty of this task. To tackle this problem, probabilistic embeddings are proposed to quantify many-to-many relationships. However, existing datasets (*e.g.*, MS-COCO) and metrics (*e.g.*, Recall@K) are hard to fully represent these diversity correspondences due to non-exhaustive annotations. Based on this observation, we utilize semantic correlation computed by CIDEr to find the potential correspondence. Then we present an effective metric, named Average Semantic Precision (ASP), which can measure the ranking precision of semantic correlation for retrieval sets. Additionally, we introduce a novel and concise objective, coined Differentiable ASP Approximation (DAA). Concretely, DAA can optimize ASP directly by making the ranking function of ASP differentiable through a sigmoid function. To verify the effectiveness of our approach, extensive experiments are conducted on MS-COCO and CUB Captions, which are commonly used in probabilistic embedding for cross-modal retrieval. The results show that our approach obtains superior performance over the state-of-the-art approaches on all metrics. The code and trained models are released at `https://anonymous.4open.science/r/2022-NeurIPS-DAA-4F1F`.

1 Introduction

As one of the most fundamental tasks in cross-modal retrieval, image-text retrieval is in great demand due to the explosive increase of multimedia data in social activities. Given an image or a text as a query, it aims to retrieve the most relevant items in the database of other modalities. Although significant progress has been made [20, 11, 39, 35, 23, 22], this task remains challenging since the heterogeneity across different modalities. To tackle this problem, most existing methods map the query and retrieval items from distinct modalities into a common embedding space, which allows computing the similarity between different modalities.

A line of methods [20, 37, 11] extracts representations of images and texts by constructing a two-stream deep neural network, which can learn the correspondence of inter-modalities well. Another thread of work [39, 35, 47] focuses on the relationship within the intra-modality. Recent studies [23, 22, 26] have shown that vision-language pre-training is effective to learn generic representations from massive image-text pairs. These methods have achieved impressive results on existing benchmarks. However, most of the prior works [20, 37, 11, 39, 35, 47] use projection functions to map a sample as a deterministic embedding, which can only quantify one-to-one relationships. It has no ability to represent the semantic diversity in images or texts, *e.g.*, given a text query, the first four images in the second row of Figure 1 (c) can match it semantically, but these works using deterministic embedding only regard the ground-truth image (green) as positive. To tackle this problem, Chun *et*

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

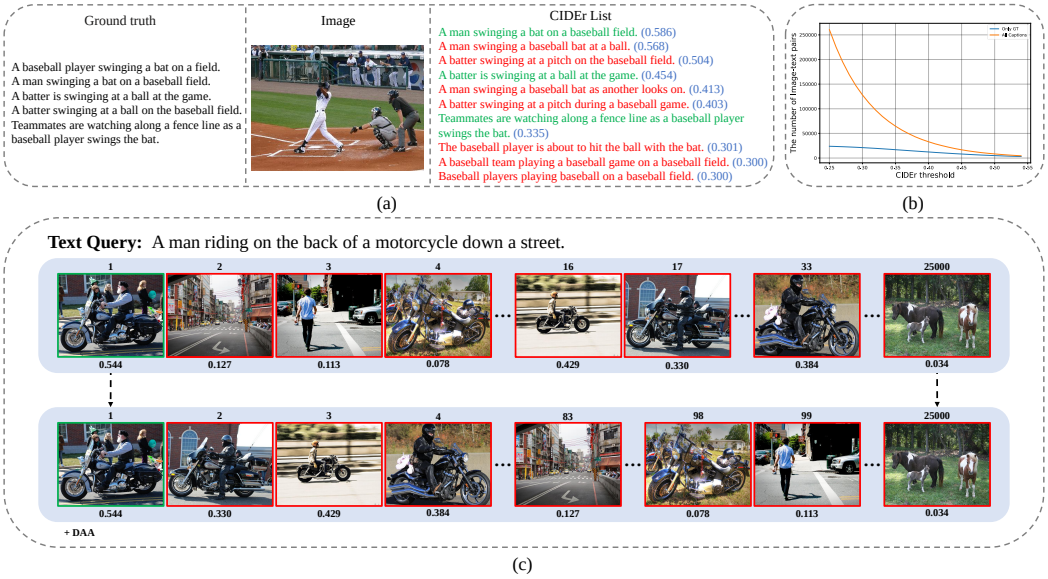

Figure 1: (a) An example of the image and its five ground truth captions. The CIDEr list gives 10 captions, followed by their CIDEr scores in the bracket with ground truth marked green, which are computed with all ground truth shown on the left. It can be observed that the captions with a CIDEr score greater than 0.3 can describe the image correctly (the last three captions in the CIDEr list). (b) The number of image-text pairs greater than the CIDEr threshold shows that there are a large number of image-text pairs with high semantic correspondence besides the ground truth. (c) Both retrieved sets obtain the same score by AP or Recall@K. Nonetheless, compared with the first retrieved set, the second one predicted by the model using DAA is more semantically related to the text query. Below each image is the semantic similarity computed by CIDEr. It shows that the similarities computed by CIDEr are generally consistent with the real semantic similarities between the text queries and the images.

al. [7] introduce probabilistic embedding for cross-modal retrieval, which is powerful to represent many-to-many relationships, by converting the samples into diversified representations according to a distribution.

The multiplicity of probabilistic embedding heavily depends on the well-matched datasets, which are expensive to annotate manually. However, commonly used datasets for the cross-modal retrieval task suffer from non-exhaustive annotations, which miss several positive pairs. These missing positive pairs (MPPs) enormously limit the diversity expressed in training and evaluation processes. To find out all potential positives, we utilize CIDEr [41], a metric to evaluate the similarity of two captions in image captioning, *i.e*., we regard all image-text pairs with CIDEr scores greater than the threshold as positives. Then, we observe that a caption with a CIDEr score greater than 0.3 can typically describe images correctly (see Figure 1 (a)). We measure the data scale of potential positives with different CIDEr thresholds and find that there are 128,050 image-text pairs with the CIDEr scores greater than 0.3 (see Figure 1 (b)), which means that each image equally matches 25.61 captions (only 5 ground-truth captions) on the semantic. The results strongly demonstrate that a large number of potential positives are wrongly regarded as negatives, which severely impede diversity in training and evaluation processes.

Nonetheless, most optimization methods [11, 14, 12, 43] and metrics revolve around the accurate prediction of positives in the ground truth, ignoring the correlations between other retrieved items and the query. As the situation in Figure 1 (c), the retrieved set at the bottom is more in line with our expectations than the one above, even though they obtain the same score by the metrics only focusing on the ground truth, such as AP or Recall@K. It means that ground-truth-based metrics are hard to fully represent the multiplicity for cross-modal retrieval. To address these problems, we carry out the following work.

In this work, we give a solution to alleviate the hindrance of MPPs to multiplicity for probabilistic representations in the training process. Firstly, to evaluate with non-ground-truth relevant items, we propose a new evaluation metric, named Average Semantic Precision (ASP). ASP computes the average deviation between prediction ranking and semantic ranking. It indicates the semantic ranking precision of each retrieved item. We utilize CIDEr as the criterion for judging the semantic correlations of captions. Secondly, we propose a concise and novel Differentiable ASP Approximation (DAA) method, which optimizes ASP directly by relaxing the discrete ranking function in the non-differentiable ASP with a sigmoid function. Eventually, our approach achieves state-of-the-art performance on several multiplicity metrics.

The main contributions and novelties of this paper could be summarized below. i) We count the number of image-text pairs with strong semantic correlations on the most widely used dataset MS-COCO. The result shows that incompletely annotating MPPs is a crucial factor restricting many-to-many mapping for probabilistic embedding. Thus, we propose an effective evaluation metric named ASP, instead of the existing ones, which are powerless to reflect the correlations between each retrieved item and the query. ii) To eliminate the adverse effect of MPPs in the training process, we propose a novel and concise metric learning method DAA, which approximates and optimizes ASP directly. iii) To verify the effectiveness of our method, we conduct experiments on two challenging benchmarks and multiple sensible evaluation metrics.

2 Related Work

In this section, we briefly introduce some recent developments in probabilistic embedding for cross-modal retrieval and metric learning.

2.1 Probabilistic Embedding for Cross-modal Retrieval

Cross-modal retrieval requires a coordinated representation [3] that allows computing the similarity between the query and the retrieved item from distinct modalities. Early works [11, 14] focused on one-to-one mapping, which utilizes multiple projection functions to map the sample as a deterministic embedding. For example, canonical correlation analysis (CCA) is proposed to learn a linear projection function and maximize the correspondences between positive images and texts. Faghri *et al*. [11] fully utilize the hardest negative to boost the matching performance. Some prior works attempt to improve the multiplicity of representations. For example, Song and Soleymani [38] use a multi-head self-attention module to generate multiple candidate representations for a given input. VSRN [21] adopts a graph neural network for semantic reasoning on the region proposals generated by a Faster-RCNN [34] detector. It gains significant performance improvements with the large increase in computing expenses. Recently, Chun *et al*. [7] use probabilistic embedding in cross-modal retrieval to fix the defect of deterministic embedding. Probabilistic embedding has shown strong performance in representing many-to-many correlations. Although there have been feasible approaches [38, 21] to represent the multiplicity of samples, and promising results have been achieved in recent years, the existing not well-aligned datasets (NWD) are hard to support diversity learning. To address this problem, Chun *et al*. [7] propose a many-to-many metric Plausible Match R-Precision (PMRP), which computes similarities between different samples by the multiple class labels in MS-COCO. It regards two samples in MS-COCO whose class label difference is less than 2 as positive, which will make a lot of false matches. Meanwhile, while labeling well-aligned datasets which is as large as MS-COCO, the cost is enormous. Moreover, limited by the incomplete annotations of datasets, there is no appropriate metric to represent the diversity performance of the model trained in NWD. ASP is a metric considering the semantic precision of retrieved items. The diversity performance of the model can be accurately evaluated by ASP even if tested in NWD.

2.2 Metric Learning

Metric Learning maps input data into an embedding space, such that distance metrics can be applied easily. A great body of work has focused on metric learning [2, 1, 44, 40, 29, 6, 45]. For instance, contrastive [6] loss and triplet [45] loss are two of the most widely used methods. The former forces all positive instances closer, while negative pairs are only separated by a fixed margin. In the latter, positive pairs are pushed closer than negative pairs. Despite impressive success has been achieved, the motivation to only minimize distance is limited. The aforementioned methods ignore the ranking

orders, which are essential for rank-based metrics such as Average Precision (AP). Brown *et al*. [5] propose a differentiable AP approximation loss that can optimize AP directly in image retrieval. Compared with triplet loss, it shows better performance in AP, a ranking-based metric. Impressive progress in optimizing AP has been achieved [40, 29, 10, 17, 16, 31, 36], but AP is a limited metric, which only focuses on the ground truth. In our work, a ranking-based metric ASP is proposed to exploit potential semantic information. Besides, we propose a Differentiable ASP Approximation (DAA) to optimize ASP directly. To our knowledge, DAA is the first work that uses direct metric optimization for cross-modal retrieval.

3 Preliminaries

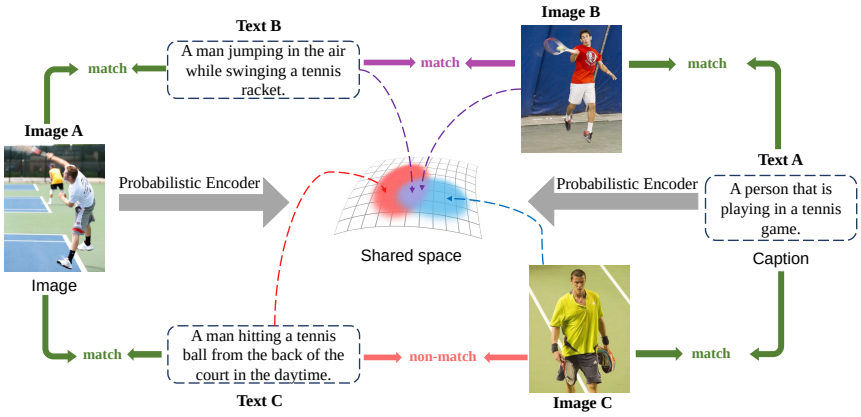

Figure 2: An example of probabilistic embedding to represent multiplicity. For each positive pair (*e.g*., image A and text A), it will generate two distributions through a probabilistic encoder. For the distribution coming from image A (marked red in shared space), captions within the distribution (*e.g*., Text B and C) all match image A. Similarly, Image B and C can match Text A. The image and text on the intersection of two distributions (marked purple in shared space) will match each other (*e.g*., Image B and Text B), while non-match on the outside.

Probabilistic embedding. Chun *et al*. [7] have a previous attempt at Probabilistic Cross-Modal Embedding (PCME) to address the multiplicity of cross-modal correlation. PCME models each sample as a normal distribution (See Figure 2), which could be described by mean and variance together. Therefore, the key to generating probabilistic embedding is determining the mean embedding and the variance matrix. Given an image $v \in \mathcal{V}$ and a caption $t \in \mathcal{T}$ as inputs, probability-based model will learn their normal distributions as:

$$\begin{aligned} p(E(v)|v) &\sim N(h^{\mu}_{\mathcal{V}}(f_v), h^{\sigma}_{\mathcal{V}}(\text{diag}(f_v))) \\ p(E(t)|t) &\sim N(h^{\mu}_{\mathcal{T}}(f_t), h^{\sigma}_{\mathcal{T}}(\text{diag}(f_t))) \end{aligned} \tag{1}$$

where f_v is the feature map extracted from the visual encoder and f_t is the feature sequence extracted from the textual encoder. $E(\cdot)$ is a function mapping input to probabilistic embedding. h^{μ} and h^{σ} are two head modules used to compute the mean embedding and the variance matrix in PCME [7], respectively. To learn the joint probabilistic embedding, the soft cross-modal contrastive loss is used [13], which can be defined as:

$$\mathcal{L}_{v,t}(\theta) = \begin{cases} -\log p_\theta(m|v,t) & \textit{if } v,t \textit{ is a match} \\ -\log(1 - p_\theta(m|v,t)) & \textit{otherwise} \end{cases}, \tag{2}$$

where $\theta = (\theta_V, \theta_T)$ are the parameters for the visual and the textual encoders, respectively. Given an image-text pair (v, t) as input, $p_\theta(m|v,t)$ is the match probability from Euclidean distances can be similarly denoted as:

$$p_\theta(m|v,t) \approx \frac{1}{J^2} \Sigma_j^J \Sigma_{j'}^J \phi(-a||E(v)^j - E(t)^{j'}||_2 + b), \tag{3}$$

where $E(v)^j$ and $E(t)^{j'}$ are samples from the embedding distribution $p(E(v)|v)$ and $p(E(t)|t)$ via a Monte-Carlo estimator [28], respectively. (a, b) are learnable scalars and $\phi(\cdot)$ is a sigmoid function.

4 Approach

In this section, we elaborate on the proposed method. In Section 4.1, we first delineate our novel evaluation metric, *i.e.*, Average Semantic Precision (ASP), which can calculate the semantic correlation scores of the retrieved set easily and precisely even in noisy datasets. In Section 4.2, we detail how to optimize ASP directly in probabilistic embedding by using Differentiable ASP Approximation (DAA).

4.1 Average Semantic Precision (ASP)

Existing diversity metrics [7, 30] for not well-aligned datasets (NWD) are hard to fully represent the semantic ranking accuracy of retrieved sets limited by incomplete correlation labels, such as PMRP [7] or CxC [30]. The former treats samples in MS-COCO [24] whose class label difference numbers are less than 2 as positives, but it will make a lot of false matches. The latter manually annotates the testing set of MS-COCO, but still misses many potential positive pairs. Meanwhile, both of them are merely applicable to MS-COCO, *i.e.*, they are not universal. To solve these problems, we propose Average Semantic Precision (ASP). Given a query set $\Psi = \{Q_i, i = 0, ..., n\}$, a query $q \in \Psi$ can simply calculate the semantic correlation scores of the retrieved set $\Omega = \{G_i, i = 0, ..., m\}$ in various image-text retrieval datasets. Mathematically,

$$ASP_q = \frac{1}{|\Omega|} \sum_{i \in \Omega} \frac{\mathcal{M}in(\mathcal{R}(i, C), \mathcal{R}(i, S))}{\mathcal{M}ax(\mathcal{R}(i, C), \mathcal{R}(i, S))}, \tag{4}$$

where $\mathcal{R}(i, C)$ and $\mathcal{R}(i, S)$ refer to the rankings of the instance i in semantic list C (defined as Eq. 6) computed by CIDEr [41], and similarity list S (defined as Eq. 5) predicted by the model. In Eq. 4, ASP utilizes the overall gap between the semantic ranking and the similarity ranking to represent the precision of a retrieved set.

$$S = \{s_i = \left\langle \frac{E(q)}{\|E(q)\|} \cdot \frac{E(i)}{\|E(i)\|} \right\rangle, \quad i = 0, 1, ..., m\}, \tag{5}$$

$$C = \{c_i = SEM(i, q), \quad i = 0, 1, ..., m\}, \tag{6}$$

The $SEM(i, q)$ is the semantic correlation score between query q and retrieved item i. When q and i refer to an image and a text, we will select all positive texts (5 texts in MS-COCO) corresponding to the image, and calculate their semantic similarities with i through the CIDEr metric, respectively. Eventually, the mean of these semantic similarities will represent the semantic similarity between q and i. Conversely, the positive texts corresponding to image i will be selected to calculate the semantic correlation scores with text q. The specific equation is as follows:

$$SEM(i, q) = \begin{cases} \frac{1}{P} \sum_{j=1}^{P} \psi(i, \mathcal{P}os(q)_j), & \textit{if q is an image} \\ \frac{1}{P} \sum_{j=1}^{P} \psi(\mathcal{P}os(i)_j, q), & \textit{if q is a text} \end{cases} \tag{7}$$

where $\mathcal{P}os(q)$ is the positive texts of q, P is the number of texts in $\mathcal{P}os(q)$, ψ stands for the CIDEr metric.

4.2 Differentiable ASP Approximation (DAA)

Contrastive [6] and triplet [45] losses are two widely used metric learning methods, which are driven by minimizing distance. They both ignore the ranking order, which is important for ranking-based metrics. Thus, the ranking-based metrics will be indirectly optimized in these methods. ASP is a ranking-based metric that is neither differentiable nor decomposable. Consequently, we propose a Differentiable ASP Approximation (DAA) to optimize ASP directly.

Given the ranking-based metric ASP, the reason why it is non-differentiable is that the derivative of a ranking function $\mathcal{R}$ for instance i defined in Eq. 8 is either gradient zero or discontinuous [32], resulting in the inability to be optimized with gradient-based approaches.

$$\mathcal{R}(i, U) = 1 + \sum_{j \in U, j \neq i} \mathbb{1}\{(u_i - u_j) < 0\}, \tag{8}$$

where $\mathbb{1}\{\cdot\}$ serves as an Indicator function, which select 1 while $(u_i - u_j) < 0$, otherwise select 0 [32]. U is the CIDEr score set C or the similarity score set S. The difference matrix of $D^U \in \mathbb{R}^{m \times m}$ can be computed as:

$$D^U = \begin{bmatrix} u_1 & \cdots & u_m \\ \vdots & \ddots & \vdots \\ u_1 & \cdots & u_m \end{bmatrix} - \begin{bmatrix} u_1 & \cdots & u_1 \\ \vdots & \ddots & \vdots \\ u_m & \cdots & u_m \end{bmatrix}, \tag{9}$$

The ASP for the query q can be converted from Eq. 4 to:

$$ASP_q = \frac{1}{|\Omega|} \sum_{i \in \Omega} \frac{1 + \mathcal{M}in(\sum_{j \in \Omega, j \neq i} \mathbb{1}\{D^C_{ij} > 0\}, \sum_{j \in \Omega, j \neq i} \mathbb{1}\{D^S_{ij} > 0\})}{1 + \mathcal{M}ax(\sum_{j \in \Omega, j \neq i} \mathbb{1}\{D^C_{ij} > 0\}, \sum_{j \in \Omega, j \neq i} \mathbb{1}\{D^S_{ij} > 0\})}, \tag{10}$$

To make ASP differentiable, which could achieve direct optimization of ASP, we utilize a sigmoid function $\phi(\cdot)$ to replace the Indicator function $\mathbb{1}\{\cdot\}$. After substituting $\phi(\cdot)$ into Eq. 8, the approximation of ASP is denoted as:

$$ASP_q = \frac{1}{|\Omega|} \sum_{i \in \Omega} \frac{1 + \mathcal{M}in(\sum_{j \in \Omega} \phi(D^C_{ij} > 0), \sum_{j \in \Omega} \phi(D^S_{ij} > 0))}{1 + \mathcal{M}ax(\sum_{j \in \Omega} \phi(D^C_{ij} > 0), \sum_{j \in \Omega} \phi(D^S_{ij} > 0))}, \tag{11}$$

which is differentiable and could be optimized with gradient-based methods. Therefore, the final loss function is defined as:

$$\mathcal{L}_{ASP} = \frac{1}{|\Psi|} \sum_{k \in \Psi} (1 - ASP_k), \tag{12}$$

5 Experiment

In this section, we conduct experiments to verify the effectiveness of DAA, which could optimize ASP directly. We describe the evaluation protocol including datasets and metrics, as well as the implementation details. Then we report experimental results on MS-COCO [24], and CUB Captions [7] for the cross-modal retrieval task. After that, we carry out a series of ablation studies to verify the effectiveness of DAA in improving multiplicity. Meanwhile, we present an experimental analysis of the effectiveness of DAA.

5.1 Evaluation Protocol

5.1.1 Datasets

We conduct the experiments on MS-COCO [24] and CUB Captions [7] to show the performance of probabilistic embedding using DAA for cross-modal retrieval.

MS-COCO is a widely used dataset for the cross-modal retrieval task. It contains 123,287 images with five captions per image. We follow the data partition in [19], which consists of 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. We report our results on both 5K and 1K testing sets (the average over 5-folds of 1K testing images).

CUB Captions [7, 33, 42] is a benchmark proposed to reduce the impact of false matching for probabilistic embedding in MS-COCO. It contains 11,788 images of 200 fine-grained bird categories with 10 captions per image. All the image-text pairs which belong to the same class are considered as positive pairs, hence false negatives rarely exist. Meanwhile, the homogeneity of image-text pairs in the same class could largely suppress false positives. We follow the class splits in [46], which consists of 150 classes for training and validation, and the remaining 50 classes for testing.

5.1.2 Metrics

We take **Recall@K** (R@K) as a measurement in both benchmarks. R@K is a widely used metric in cross-modal retrieval. We report R@1, R@5, and R@10 for a comprehensive evaluation. However, R@K only focuses on the position of the first relevant item according to the ground truth, reflecting limited information.

R-Precision (R-P) is an alternative proposed by Musgrave *et al*. [27]. For each query, R-P computes the ratio of matched items in the top-r retrieved items, where r is the number of ground-truth matches.

In short, for all matched items, the fewer negatives before, the higher the R-P score is. Compared with R@K, R-P could better evaluate the one-to-many matching for cross-modal retrieval. We use R-P to evaluate the performance of DAA on CUB Captions.

As MS-COCO, we take **Plausible Match R-Precision** (PMRP) [7] metric as a measurement, which utilizes the class labels for MS-COCO to find out all hidden positive image-text pairs. Concretely, PMRP regards the pair (v, t), whose binary label vectors $y^v, y^t \in \{0,1\}^{d_{label}}$, differ at most at ζ positions ($\zeta \in \{0, 1, 2\}$), as positive matches.

The above metrics can only reflect the precision of ground-truth items, ignoring the correlations between other non-matched retrieved items and queries. Therefore, we refer to **ASP** as an evaluation measurement for both benchmarks.

In most cases, we pay more attention to the relevance of header retrieved items. However, the top-k retrieved items predicted by current high-performance models are typically highly correlated, only with a small CIDEr difference. Concretely, as the second row in Figure 1 (c), the first four images can match with the query semantically, they should be equally treated as good matches even if they are not ranked according to CIDEr score. Thus, we apply a variant of ASP to evaluate the semantic correlations in the top-K retrieved items, named **ASP@K** (A@K). A@K computes the ratio between the total CIDEr scores of top-K retrieved items predicted by the model, and the sum of the largest K in CIDEr list C (defined as Eq. 6). We compute A@5, A@10, A@20 in the ablation studies on MS-COCO (5K).

5.2 Implementation Details

We use ResNet [15] and the basic version of the pre-trained BERT [8] as visual and textual encoders. There are three phases in training proceeds: The first is a pre-training phase, where only the head modules in the visual encoder and all parameters in the textual encoder are trained. A warm-up phase lasting three epochs is used in the pre-training phase. Finally, all parameters will be trained in an end-to-end fine-tuning phase. Concretely, for MS-COCO, we use ResNet-152 backbone with embedding dimension $D = 1024$. For CUB, we use ResNet-50 backbone with embedding dimension $D = 512$. Meanwhile, We utilize AdamP Optimizer [18] for model training with the batch size of 128 and 150 for MS-COCO and CUB, respectively. For both datasets, we set the initial learning rate as 0.00015 with the cosine learning rate scheduler [25]. Besides, models are always trained on a single RTX A6000 GPU with Cutout [9] and random caption dropping [4] augmentation strategies with 0.2 and 0.1 erasing ratios, respectively.

Table 1: Comparison on MS-COCO

	1K Test Images				5K Test Images					
Method	i2t		t2i		i2t			t2i		
	PMRP	R@1	PMRP	R@1	PMRP	R@1	ASP	PMRP	R@1	ASP
VSE++	-	64.6	-	52.0	-	41.3	-	-	31.3	-
PVSE K=1	40.3	66.7	41.8	53.5	29.3	41.7	-	30.1	30.6	-
PVSE K=2	42.8	69.2	43.6	55.2	31.8	45.2	-	32.0	32.4	-
PCME	45.0	68.8	46.0	54.6	34.1	44.2	54.5	34.4	31.9	55.0
PCME-bert	44.6	68.8	45.7	56.5	33.3	44.8	54.4	33.8	33.4	54.4
Ours	**45.9**	**71.2**	**46.7**	**57.8**	**34.6**	**48.2**	**58.3**	**34.9**	**35.1**	**57.0**

5.3 Comparisions with State of The Arts

In this section, we compare our method with previous works using ResNet as the visual encoder on the two datasets. The comparison methods include VSE0 [11], VSE++ [11], PVSE [38], and PCME [7]. Our model uses the basic version of BERT as the textual encoder. To be fair, PCME is also retrained with BERT as the textual encoder and compared with our approach.

Results on MS-COCO. Table 1 shows the quantitative results of MS-COCO 1K and 5K testing sets for image-to-text (i2t) and text-to-image (t2i) retrieval. We can observe that our approach obtains a consistent improvement in performance on all metrics, compared with previous state-of-the-art

models. Exactly, DAA outperforms PCME [7], which is the state-of-the-art model using probabilistic embedding for cross-modal retrieval, by significant margins in R@1, PMRP, and ASP. As for MS-COCO (1K), compared with PCME, we achieve absolute boosts on PMRP and R@1 in both i2t and t2i, which are 0.9%, 2.4% and 0.7%, 3.2%, respectively. As for MS-COCO (5K), compared with PCME, the boosts on PMRP and R@1 are 0.5%, 4.0% and 0.5%, 3.2%. Besides, we evaluate the experimental results with newly proposed metrics ASP on MS-COCO (5K). The results show that DAA could improve ASP in both directions, especially i2t with a 3.8% increase compared with PCME, which indicates that DAA is effective in ranking retrieval sets according to semantics. Moreover, DAA also has a better performance than PCME retrained with BERT in the metrics mentioned before.

Results on CUB Captions. To further demonstrate the robustness of the proposed model, we evaluate our model on CUB Captions. Compared with MS-COCO, R-P is more suitable for CUB due to rare false pairs, which is acceptable for evaluating the multiplicity of probabilistic embedding. The results on CUB are presented in Table 2. It can be seen that DAA outperforms all the state-of-the-art methods. Especially, for image-to-text retrieval, DAA surpasses the previous best method on R-P, R@1, and ASP with an increase of 1.9%, 6.3%, and 0.3% respectively. Besides, for text-to-image retrieval, DAA also outperforms the best counterpart by a significant margin (1.7% for R-P, 2.5% for R@1, and 0.8% for ASP). The above results demonstrate the effectiveness and importance of exploiting the semantic information in cross-modal retrieval.

Table 2: Comparison on CUB Captions

Method	Image-to-text			Text-to-image		
	R-P	R@1	ASP	R-P	R@1	ASP
VSE0	22.4	44.2	-	22.6	32.7	-
PVSE K=1	22.3	40.9	-	20.5	31.7	-
PVSE K=2	19.7	47.3	-	21.2	28.0	-
PCME	26.3	46.9	53.1	26.8	35.2	56.8
PCME-bert	27.0	49.7	52.8	27.2	35.1	55.8
Ours	**28.2**	**53.2**	**53.4**	**28.5**	**37.7**	**57.6**

5.4 Ablation Studies

In this section, we perform several ablation studies on MS-COCO 5K to explore the superiority of DAA over the optimization method only focusing on ground truth (*e.g.*, Smooth-AP [5]), on the probability-based model.

To explore the effect of DAA and Smooth-AP, we test the impact of using these two methods on multiple metrics. The results are shown in Table 3. To start with, we first evaluate the model using BERT as the baseline, where the sample number is 5. Subsequently, we train our model with Smooth-AP and find that there are only few improvements on R@1 and A@K (Section 5.1.2), even a decline in PMRP. Finally, we add the DAA to the baseline, resulting in considerable boosts on all metrics, especially R@1 (2.9% for i2t, 1.8% for t2i) and ASP (4.1% for i2t, 2.8% for t2i). The significant improvement of ASP shows that DAA shifts retrieved items with high relevance forward overall. Besides, DAA obtain impressive improvements on A@K (K=5, 10, 20), which reflects that DAA can improve the semantic ranking precision even for the retrieval set with high semantic correlation. These results implicate excavating semantic correlation can not only improve the multiplicity of probability-based models, but also improve the traditional ground-truth-based metrics, such as R@K. The visualization of retrieval results is shown in the second row of Figure 1 (c).

Table 3: Effect of Smooth-AP and DAA on MS-COCO 5K

Method	Image-to-text						Text-to-image					
	PMRP	R@1	ASP	A@5	A@10	A@20	PMRP	R@1	ASP	A@5	A@10	A@20
Baseline (B)	34.3	45.3	54.2	61.6	62.2	62.3	34.4	33.3	54.2	63.8	63.3	63.0
B+Smooth-AP	33.0	45.4	54.2	61.8	62.3	62.2	33.5	33.3	54.2	63.9	63.4	63.1
B+DAA	**34.6**	**48.2**	**58.3**	**62.7**	**63.2**	**63.4**	**34.9**	**35.1**	**57.0**	**65.2**	**64.9**	**65.0**

The results in Table 3 prove that DAA can increase the semantic similarity in the whole retrieval set and top-K ranked retrieval subset, which can be regarded as two sets with large and small CIDEr score variance. To validate the effect of DAA for subsets with different CIDEr score variance, we rank the retrieved set according to the CIDEr score in descending order and sample the top N to compute ASP. We draw the contrastive curves of ASP in Baseline, DAA, and Smooth-AP [5] with N

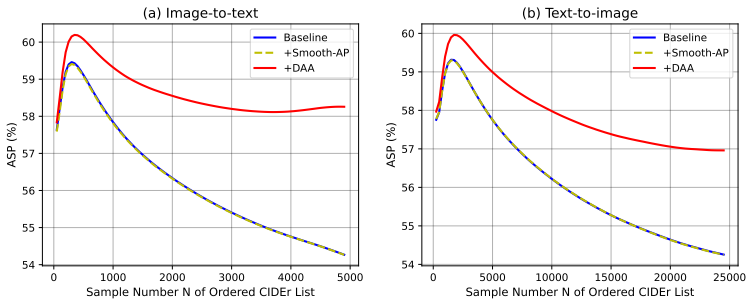

Figure 3: Ablation studies of the retrieval subsets with different CIDEr score variance on ASP metric. Given a retrieval set in descending order according to CIDEr score, we compute the ASP of top-N items. The variance will increase with the increase of N.

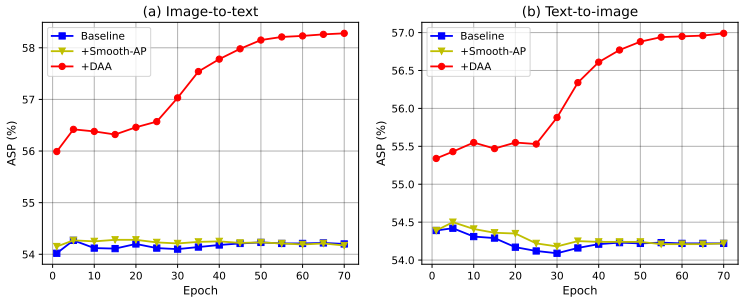

Figure 4: Ablation studies of different training strategies during training on ASP metric. The training process consists of 70 epochs. The first 3 epochs are the warm-up phase, the next 27 epochs are the pre-training phase, and the last 40 epochs are the fine-tuning phase.

from 0 to the length of the whole retrieval set in Figure 3. We can observe that DAA always performs better than others and the gap becomes larger along with the increase of N. The results indicate that DAA could enhance diversity performance in any retrieval set. Meanwhile, The curves of Smooth-AP and baseline are almost consistent, which shows that only focusing on ground truth is limited to improving multiplicity.

To make clear the impacts of different methods on multiplicity in the training process, we test ASP in each epoch for baseline, Smooth-AP, and DAA. The results are shown in Figure 4, which provides a quantitative analysis of ASP against epochs. We can observe that with the increase of training rounds, the curve of the training strategy using DAA grows constantly. It substantially verifies that DAA effectively captures complex semantic correspondences and enhances multiplicity in the training process. The ASP curves of baseline and Smooth-AP strategies are almost invariant, while Recall@K improves constantly. This means that boosting the prediction accuracy of the items in ground truth may not lead to the improvement of multiplicity. All of these ablation studies strongly demonstrate that DAA is superior to the optimization method focusing only on ground truth in improving multiplicity.

6 Conclusion

This paper attempts to address a critical problem in probabilistic embedding for cross-modal retrieval, *i.e.*, the not well-aligned datasets and existing metrics can hardly train and evaluate the multiplicity performance of the model. To solve this problem, we introduce Average Semantic Precision (ASP) to evaluate the performance of multiplicity in these not well-aligned datasets. Moreover, we propose a novel Differentiable ASP Approximation (DAA), which could make the model diversified even if training in these noisy datasets. We have demonstrated the effectiveness of our DAA for boosting the multiplicity of probabilistic embedding for cross-modal retrieval by outperforming the state-of-the-art models.

References

[1] Relja Arandjelovic and Andrew Zisserman. "All about VLAD". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2013, pp. 1578–1585.

[2] Relja Arandjelovic et al. "NetVLAD: CNN architecture for weakly supervised place recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5297–5307.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy". In: *IEEE transactions on pattern analysis and machine intelligence* 41.2 (2018), pp. 423–443.

[4] Samuel R Bowman et al. "Generating sentences from a continuous space". In: *arXiv preprint arXiv:1511.06349* (2015).

[5] Andrew Brown et al. "Smooth-ap: Smoothing the path towards large-scale image retrieval". In: *European Conference on Computer Vision*. Springer. 2020, pp. 677–694.

[6] Sumit Chopra, Raia Hadsell, and Yann LeCun. "Learning a similarity metric discriminatively, with application to face verification". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 539–546.

[7] Sanghyuk Chun et al. "Probabilistic embeddings for cross-modal retrieval". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 8415–8424.

[8] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[9] Terrance DeVries and Graham W Taylor. "Improved regularization of convolutional neural networks with cutout". In: *arXiv preprint arXiv:1708.04552* (2017).

[10] Martin Engilberge et al. "Sodeep: a sorting deep net to learn ranking loss surrogates". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10792–10801.

[11] Fartash Faghri et al. "Vse++: Improving visual-semantic embeddings with hard negatives". In: *arXiv preprint arXiv:1707.05612* (2017).

[12] Andrea Frome et al. "Devise: A deep visual-semantic embedding model". In: *Advances in neural information processing systems* 26 (2013).

[13] Raia Hadsell, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. IEEE. 2006, pp. 1735–1742.

[14] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. "Canonical correlation analysis: An overview with application to learning methods". In: *Neural computation* 16.12 (2004), pp. 2639–2664.

[15] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[16] Kun He, Yan Lu, and Stan Sclaroff. "Local descriptors optimized for average precision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 596–605.

[17] Kun He et al. "Hashing as tie-aware learning to rank". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4023–4032.

[18] Byeongho Heo et al. "Adamp: Slowing down the slowdown for momentum optimizers on scale-invariant weights". In: *arXiv preprint arXiv:2006.08217* (2020).

[19] Andrej Karpathy and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3128–3137.

[20] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. "Unifying visual-semantic embeddings with multimodal neural language models". In: *arXiv preprint arXiv:1411.2539* (2014).

[21] Kunpeng Li et al. "Visual semantic reasoning for image-text matching". In: *Proceedings of the IEEE/CVF International conference on computer vision*. 2019, pp. 4654–4662.

[22] Wei Li et al. "Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning". In: *arXiv preprint arXiv:2012.15409* (2020).

[23] Xiujun Li et al. “Oscar: Object-semantics aligned pre-training for vision-language tasks”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 121–137.

[24] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[25] Ilya Loshchilov and Frank Hutter. “Sgdr: Stochastic gradient descent with warm restarts”. In: *arXiv preprint arXiv:1608.03983* (2016).

[26] Xiaopeng Lu, Tiancheng Zhao, and Kyusong Lee. “VisualSparta: An Embarrassingly Simple Approach to Large-scale Text-to-Image Search with Weighted Bag-of-words”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 5020–5029.

[27] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. “A metric learning reality check”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 681–699.

[28] Seong Joon Oh et al. “Modeling uncertainty with hedged instance embedding”. In: *arXiv preprint arXiv:1810.00319* (2018).

[29] Hyun Oh Song et al. “Deep metric learning via lifted structured feature embedding”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4004–4012.

[30] Zarana Parekh et al. “Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO”. In: *arXiv preprint arXiv:2004.15020* (2020).

[31] Sebastian Prillo and Julian Eisenschlos. “Softsort: A continuous relaxation for the argsort operator”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7793–7802.

[32] Tao Qin, Tie-Yan Liu, and Hang Li. “A general approximation framework for direct optimization of information retrieval measures”. In: *Information retrieval* 13.4 (2010), pp. 375–397.

[33] Scott Reed et al. “Learning deep representations of fine-grained visual descriptions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 49–58.

[34] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015).

[35] Shuhuai Ren et al. “Learning Relation Alignment for Calibrated Cross-modal Retrieval”. In: *arXiv preprint arXiv:2105.13868* (2021).

[36] Jerome Revaud et al. “Learning with average precision: Training image retrieval with a listwise loss”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 5107–5116.

[37] Hwanjun Song et al. “Learning from noisy labels with deep neural networks: A survey”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[38] Yale Song and Mohammad Soleymani. “Polysemous visual-semantic embedding for cross-modal retrieval”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 1979–1988.

[39] Christopher Thomas and Adriana Kovashka. “Preserving semantic neighborhoods for robust cross-modal retrieval”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 317–335.

[40] Evgeniya Ustinova and Victor Lempitsky. “Learning deep embeddings with histogram loss”. In: *Advances in Neural Information Processing Systems* 29 (2016).

[41] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. “Cider: Consensus-based image description evaluation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4566–4575.

[42] Catherine Wah et al. “The caltech-ucsd birds-200-2011 dataset”. In: (2011).

[43] Liwei Wang, Yin Li, and Svetlana Lazebnik. “Learning deep structure-preserving image-text embeddings”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5005–5013.

[44] Xinshao Wang et al. “Ranked list loss for deep metric learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5207–5216.

[45] Kilian Q Weinberger and Lawrence K Saul. “Distance metric learning for large margin nearest neighbor classification.” In: *Journal of machine learning research* 10.2 (2009).

[46] Yongqin Xian, Bernt Schiele, and Zeynep Akata. "Zero-shot learning-the good, the bad and the ugly". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4582–4591.

[47] Xin Yuan et al. "Multimodal contrastive training for visual representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 6995–7004.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1
   (b) Did you describe the limitations of your work? [No] It need to be explored.
   (c) Did you discuss any potential negative societal impacts of your work? [No] No explicit negative impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 4
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 4
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See abstract, as a URL
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 5
   (b) Did you mention the license of the assets? [Yes] See Section 5
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See abstract, as a URL
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]