Expanding the Scope of Negatives: Boosting Image-Text Matching with Negatives Distribution Guided Learning

Zhao Zhou[1,2], Weizhong Zhang[1,3*], Xiangcheng Du[1], Yingbin Zheng[4], Cheng Jin[1,2*]

[1]Fudan University, Shanghai, China
[2]Innovation Center of Calligraphy and Painting Creation Technology, MCT, China
[3]Shanghai Collaborative Innovation Center of Intelligent Visual Computing, China
[4]Videt Lab, Shanghai, China
{zzhou21, xcdu22}@m.fudan.edu.cn, zyb@videt.cn, {weizhongzhang, jc}@fudan.edu.cn

Abstract

Image-text matching is a crucial task that bridges visual and linguistic modalities. Recent research typically formulates it into the problem of maximizing the margin with the truly hardest negatives to enhance the learning efficiency and avoid the poor local optima. We argue that such formulation can lead to a serious limitation, i.e., under this formulation, conventional trainers would confine their horizon within the hardest negative examples, while other negative examples offer a range of semantic differences not present in the hardest negatives. In this paper, we propose an efficient negative distribution guided training framework for image-text matching to unlock the substantial promotion space left by the above limitation. Rather than simply incorporating additional negative examples into the training objective, which could diminish both the leading role of the hardest negatives in training and the effect of a large margin learning in producing a robust matching model, our central idea is to supply the objective with distributional information on the entire set of negative examples. To be precise, we first construct the sample similarity matrix based on several pretrained models to extract the distributional information of the entire negative sample dataset. Then we encode it into a margin regularization module to smooth the similarities differences of all negatives. This enhancement facilitates the capture of fine-grained semantic differences and guides the main learning process by maximizing the margin with hard negative examples. Furthermore, we propose a hardest negative rectification module to address the instability in hardest negative selection based on predicted similarity and to correct erroneous hardest negatives. We evaluate our method in combination with several state-of-the-art image-text matching methods, and our quantitative and qualitative experiments demonstrate its significant generalizability and effectiveness.

Introduction

Image-text matching is a key task in multi-modal learning, aiming to determine the relationship between visual content and textual descriptions. This task is fundamental to numerous applications, including multimedia retrieval (Li et al. 2022c) and visual question answering (Anderson et al. 2018). Current image-text matching methods commonly employ triplet loss (Frome et al. 2013) as their primary optimization objective. In image-to-text matching, a triplet includes an anchor query image, a positive text, and a negative text, forming both positive and negative pairs. Triplet loss aims to minimize the distance between anchor and positive samples while maximizing the distance between anchor and negative samples in a common embedding space. Therefore, selecting negative samples is crucial for the effectiveness of triplet loss, attracting significant research attention.

Inspired by the large margin learning theory in machine learning (Cortes and Vapnik 1995), the hard negative mining strategy was introduced in VSE++ (Faghri et al. 2018). This strategy compels the model to focus on negative samples that are close to positive samples, refining the decision boundary to achieve a larger margin over these instances. It is important to note that image-text matching falls under the category of contrastive learning. In contrast to traditional large margin learning, this hard negative mining strategy offers a distinct advantage: it significantly enhances training efficiency by discarding numerous redundant and easily distinguishable positive-negative sample pairs. Some works (Xuan et al. 2020; Yu et al. 2018) point out that inaccurate hardest negative selection may cause distance metrics to fail in capturing semantics, leading to poor local minima. To address this issue, recent methods (Chen, Deng, and Luo 2020; Zhang et al. 2022; Li et al. 2023a) are proposed to refine the selection of the hardest negative samples or to generate the truly hardest negatives.

Despite the promising results reported in the existing studies (Zhang et al. 2022; Li et al. 2023a), we argue that solely considering hardest negatives can lead to a serious limitation. Firstly, compared to the hardest negative in each minibatch, there is a much larger number of other negative samples that offer a range of semantic differences not present in the hardest negatives. These differences are crucial for improving the ability of model to perform fine-grained matching of images and text. Secondly, selecting the hardest negative is unstable, as it requires the model to identify the most similar negative sample among all options. This process is particularly prone to errors in the early stages of training, which can adversely impact model optimization.

We demonstrate the importance of other negative samples for model training from three perspectives: case analysis,

*Corresponding author.

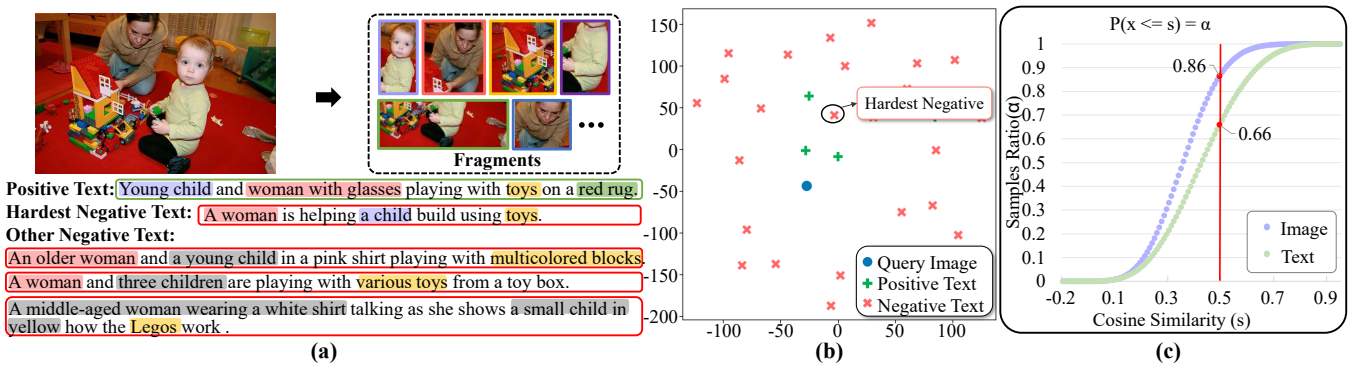

Figure 1: Illustration of our motivation. Negative examples provide a range of semantic differences not captured by the hardest negatives. (a) Matching of an image with corresponding positive and negative texts, where colorful boxes highlight fine-grained correspondences and gray boxes indicate mismatches. (b) t-SNE projections showing features extracted by the trained model. (c) Quantile plot comparing similarity among the hardest negative samples to that of the top-30 most similar negative samples.

feature visualization and statistical analysis. Specifically, as illustrated in Figure 1(a), the hardest negative may mismatch with fragments like "red rug" compared to the positive text, while other negatives offer finer-grained distinctions, such as color or numerical features. This helps model understand the fine-grained characteristics of fragments. Furthermore, Figure 1(b) visualizes the image and text embeddings with t-SNE (Van der Maaten and Hinton 2008), which indicates that the differences between negative samples are significant even among those close to the query. It reveals that while many negative texts have similar semantics to the query, they are dispersed, highlighting semantic variations among the negatives. Figure 1(c) presents the quantile function of similarity scores between the hardest negative samples and the top-30 most similar negative samples. The similarity scores are generally low, with 86% (resp. 66%) of the top-30 most similar negative image (resp. text) samples falling below 0.5, highlighting other negative examples offer a range of semantic differences not present in the hardest negatives.

To unlock the substantial promotion space left by the limiatation above, in this paper, we propose a novel Negatives Distribution Guided Learning (NDGL) framework for image-text matching. Note that simply incorporating additional negative examples into the training objective can diminish both the leading role of the hardest negatives in training and the effect of a large margin learning in producing a robust matching model. To leverage the diverse semantic differences provided by all negative samples to enhance model training, rather than strictly maximizing margins for those negative examples, our key idea is to adopt a soft supervision for training, i.e., guide the model so that its similarity score distribution over negative examples closely aligns with a reference model's distribution. Specifically, we generate a target similarity matrix for all samples by integrating outputs from multiple base models trained with hard negative mining strategy. Considering this matrix as an instantiation of the target distribution, we implement a margin regularization module designed to smooth the similarity differences among all negatives according to the target distribution. Furthermore, a hardest negative rectification module is proposed to address the instability in hardest negative selection based on predicted similarity and to correct erroneous hardest negatives. Based on these designs, in each training iteration, we can fully utilize the semantic information from all negatives from the distributional guidance, and maintain the leading role of the hardest negatives to aid in training a robust model. It is worth noting that our framework is general and can be directly applied to image-text matching methods without requiring modifications to network structures.

The main contributions can be summarized as follows:

- We demoonstrate that other negative examples offer a range of semantic differences not captured by the hardest negatives via three perspectives: case analysis, feature visualization and statistical analysis. This analysis enables us to identify the deficiencies of previous methods and the substantial promotion space that can be unlocked.
- We propose a novel negatives distribution guided image-text semantic learning framework that leverages the diverse semantic differences presented by all negative samples to enhance model training. This framework not only preserves the benefits of large-margin learning but also unlocks substantial promotion space left by relying solely on the hardest negatives.
- We apply NDGL to five state-of-the-art baseline models. Extensive quantitative and qualitative experiments demonstrate the strong generalizability and effectiveness of our approach.

Related Work

Image-Text Matching Methods. Recently, image-text matching has seen significant advancements, primarily following two research directions: global-level matching and local-level matching. The global-level matching focuses on understanding the holistic semantic content of images and texts, assessing the degree of semantic similarity between them. For instance, GPO (Chen et al. 2021) employs a Generalized Pooling Operator to adaptively select the optimal pooling strategy for different features, while

HREM (Fu et al. 2023) explicitly captures both fragment-level and instance-level relations to learn discriminative and robust cross-modal embeddings. Unlike global-level methods, local-level matching methods focus on fine-grained relationships between specific visual and textual elements within images and texts. For instance, DivE (Kim, Kim, and Kwak 2023) employs a slot attention mechanism to capture diverse semantics of the input and introduces a new similarity function called smooth-chamfer similarity to avoid sparse supervision and set collapsing. Similarly, CHAN (Pan, Wu, and Zhang 2023) focuses on the most relevant region-word pairs, disregarding all other alignments as redundant or irrelevant. In this study, to comprehensively evaluate the effectiveness of proposed framework, we have chosen five global-level and local-level matching methods as baselines for comparison.

Negative Samples Mining. The selection of negative samples is critical for training robust image-text matching models. Early approaches (Frome et al. 2013) employed all negative samples using triplet loss for training. VSE++ (Faghri et al. 2018) introduced triplet loss combined with an online hard negative mining strategy, which eliminates redundant, easily distinguishable samples and helps the model learn more distinct representations. Subsequently, the CFM framework (Wei et al. 2022) proposed synthesizing counterfactual samples for more effective image-text matching. LSEH (Gong and Cosma 2023) introduced a semantically-enhanced hard negatives loss function that dynamically modifies the learning objective based on the semantic similarities between unrelated image-description pairs. In contrast to methods that focus on selecting the hardest negative samples to learn more distinct feature representations, our approach leverages the diverse semantic differences presented by all negative samples to enhance model training.

Distillation Learning. Distillation Learning is widely used for model speedup and lightweight optimization (Jiao et al. 2020; Wu et al. 2022). Several methods (Abbasi Koohpayegani, Tejankar, and Pirsiavash 2020; Tejankar et al. 2021) propose utilizing complex models as teachers to guide student models by minimizing the Kullback-Leibler (KL) divergence of their predictions. While our framework is similar, these methods focus on single-modality distillation, whereas our method targets cross-modality distillation. Furthermore, instead of using similarity distributions, we employ margin regularization to constrain predictions. The benefits of these regularization methods will be analyzed in the following section.

Method

In this section, we formally introduce our framework for image-text matching. We begin by discussing the standard triplet loss and hardest negative mining techniques used in prior work. Next, we present the pipeline of our framework and introduce the margin regularization and hardest negative rectification modules. Finally, we discuss the effect of the proposed framework.

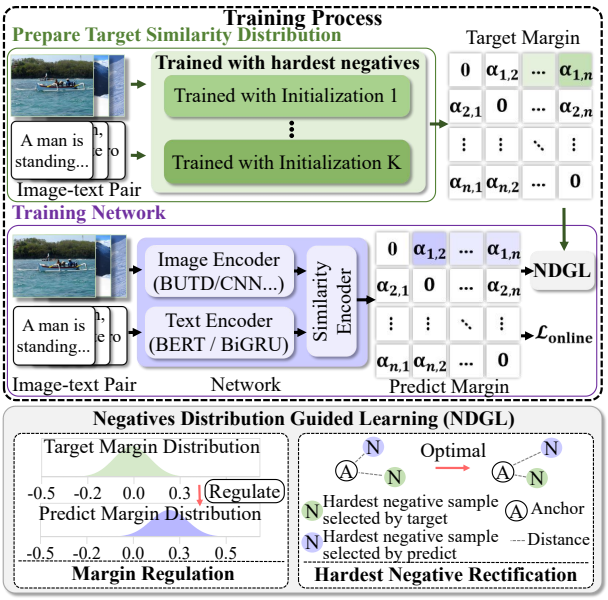

Figure 2: The pipeline of negatives distribution learning.

Preliminaries

Triplet loss is a fundamental concept in the domain of image-text matching. The core idea is to ensure that the distance between a positive pair (an image and its corresponding text) is smaller than the distance between a negative pair (an image and a non-matching text) by a fixed margin. This approach facilitates effective retrieval and matching between the two modalities. Formally, given a training mini-batch containing a set of positive pairs, the standard triplet loss is defined as follows:

$$\mathcal{L}_{\text{triplet}} = \sum_{(i,t)\in P} \Big(\sum_{\bar{t}\in T/t} [\alpha - S(i,t) + S(i,\bar{t})]_+ + \sum_{\bar{i}\in I/i} [\alpha - S(i,t) + S(\bar{i},t)]_+ \Big). \tag{1}$$

Here, α denotes the margin of the triplet loss, and $[x]_+ \equiv max(x, 0)$. The sets I, T, and P represent the images, texts, and positive pairs within the mini-batch, respectively. The variables i and t are the anchor for the image and text terms. The pair (i, t) indicates a positive pair, while $(\bar{i}, t)$ and $(i, \bar{t})$ indicate negative pairs in the mini-batch. The similarity S is used to measure the distance between the image and text.

To avoid redundant and easily distinguishable samples, the online hard negative mining strategy (Faghri et al. 2018) focuses on distinguishing between positive pairs and the most confusing negative pairs. Specifically, for a positive pair (i, t) in a mini-batch, the hard negatives $\hat{t}$ and $\hat{i}$ are defined as $\hat{t} = \arg\max_{c\in T/t} S(i, c)$ and $\hat{i} = \arg\max_{c\in I/i} S(c, t)$, respectively. The resulting triplet loss is then defined as:

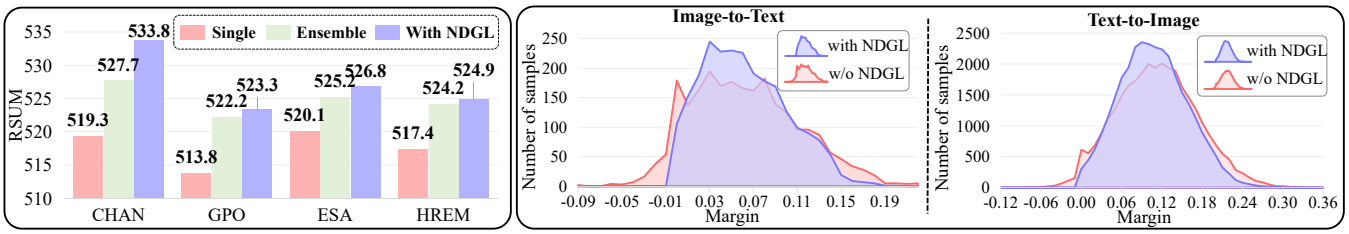

Figure 3: (a) Image-text retrieval performance of recent models, their ensemble results, and their improvement with the NDGL on the Flickr30K dataset. (b) The margin distribution changes when applying NDGL.

$$\mathcal{L}_{\text{online}} = \sum_{(i,t)\in P} \Big([\alpha - S(i,t) + S(i,\hat{t})]_+ + [\alpha - S(i,t) + S(\hat{i},t)]_+ \Big). \tag{2}$$

Negatives Distribution Guided Learning

To avoid the limitations of focusing solely on the hardest negative samples, we propose a novel image-text matching framework called Negative Distribution Guided Learning (NDGL). Our key idea is to leverage diverse semantic differences provided by all negative samples while maintaining the leading role of the hardest negatives and the benefits of large-margin learning during training. The pipeline of NDGL as illustrated in Figure 2.

Preparation of Target Similarity Distribution. Obtaining fine-grained similarity labels from only positive pair labels is difficult. We observe that the similarity distribution of all samples predicted by the ensemble outputs outperforms that of the individual model, even when both use the same underlying model structure. Thus, we use the reference similarity distribution output by model ensembling as the target distribution for soft supervision, guiding the learning of semantic differences among all samples. Specifically, we aggregate the similarity outputs from K models, each trained with a hard negative mining strategy and different initializations, across the entire training dataset to generate a target similarity matrix for extracting distributional information. This matrix, denoted as $\hat{S}$, is formatted as follows:

$$\hat{S}(i,t) = \frac{1}{K}\sum_{k}^{K} S_k(i,t). \tag{3}$$

Here, S_k denote the k-th model outputs similarity. This matrix is then transformed into a target margin distribution to regularize the predicted similarity differences in the margin regularization module and to guide the learning process for selecting truly negative samples in the hardest negative rectification modules.

Margin Regularization. To learn the semantic differences between negative pairs and ensure that the similarity difference distribution predicted by the model closely aligns with the reference model's distribution, we propose a margin regularization module. This module uses the target distribution to regularize the predicted semantic differences among all samples. Formally, given the predicted similarities S and the target similarities $\hat{S}$, the loss is defined as follows:

$$\mathcal{L}_{\text{MR}} = \sum_{(i,t)\in P} \Big(\sum_{\bar{t}\in T/t} \text{KL}(S(i,t) - S(i,\bar{t}) \,\|\, \hat{S}(i,t) - \hat{S}(i,\bar{t})) + \sum_{\bar{i}\in I/i} \text{KL}(S(i,t) - S(\bar{i},t) \,\|\, \hat{S}(i,t) - \hat{S}(\bar{i},t)) \Big). \tag{4}$$

Here, KL denotes the Kullback-Leibler divergence.

Hardest Negative Rectification. According to large margin learning theory, the hardest negatives are crucial for training a robust model. Due to the instability of hardest negative selection based on predicted similarity during training, we believe that the hardest negatives selected based on reference similarity are closer to the true hardest negative samples. Based on this, we propose a hardest negative rectification module that focuses on ranking two hardest negatives from different distributions and to correct erroneous hardest negatives. Formally, for a positive pair (i,t) in a mini-batch, the hard negatives from target similarity $\tilde{t}$ and $\tilde{i}$ are defined as $\tilde{t} = \arg\max_{c\in T/t} \hat{S}(i,c)$ and $\tilde{i} = \arg\max_{c\in I/i} \hat{S}(c,t)$, respectively. The main idea is, for instance in image-to-text matching, to ensure that $S(i,\tilde{t})$ is not lower than $S(i,\hat{t})$. The loss is defined as follows:

$$\mathcal{L}_{\text{HNR}} = \sum_{(i,t)\in P} \Big([\gamma - S(i,\tilde{t}) + S(i,\hat{t})]_+ + [\gamma - S(\tilde{i},t) + S(\hat{i},t)]_+ \Big). \tag{5}$$

Here, γ denotes the margin of the loss. Since the distance between two challenging negatives is small, we set the margin γ to 0.01. When the hardest negative is the same for two similarity scores, we set γ to 0.0.

Training Objective. Following the baseline methods (Pan, Wu, and Zhang 2023; Chen et al. 2021), we implement the triplet loss as described in Equation 2, supplemented by the losses in Equations 4 and 5. The final loss function is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{online}} + \lambda_1 \mathcal{L}_{\text{MR}} + \lambda_2 \mathcal{L}_{\text{HNR}}, \tag{6}$$

where λ_1 and λ_2 are the coefficient balancing the losses.

Effect of Negatives Distribution Learning

Advantages of NDGL. The proposed NDGL framework is arguably the simplest approach to enhancing the perfor-

Data Split	MS-COCO							Flickr30K						
Eval Task	IMG → TEXT			TEXT → IMG				IMG → TEXT			TEXT → IMG			
Method	R@1	R@5	R@10	R@1	R@5	R@10	RSUM	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
BUTD + BiGRU														
DivE	79.9	95.8	98.5	63.2	90.5	95.6	523.5	77.3	93.6	97.0	56.1	83.0	89.4	496.3
DivE+NDGL	79.6	96.0	98.7	64.0	91.2	96.3	525.9(+2.4)	79.2	94.8	97.3	60.8	86.0	91.3	509.4(+13.1)
ESA	78.9	96.4	98.8	63.1	90.9	96.0	524.1	82.1	95.6	97.6	60.4	85.6	91.6	512.9
ESA+NDGL	80.4	96.5	98.8	63.5	91.2	96.2	526.5(+2.4)	83.2	96.1	98.2	62.7	86.8	92.2	519.2(+6.3)
CHAN	79.7	96.3	98.7	64.1	90.6	95.9	525.3	78.0	94.6	96.7	60.7	84.7	90.6	505.3
CHAN+NDGL	81.5	97.0	98.9	66.5	92.1	96.7	532.7(+7.4)	83.7	95.8	98.2	64.8	88.6	93.1	524.2(+18.9)
BUTD + BERT														
GPO	79.2	96.6	98.9	64.7	91.2	96.2	526.8	80.1	95.5	97.8	62.1	86.4	91.9	513.8
GPO+NDGL	81.3	96.6	98.8	65.9	92.0	96.6	531.2(+4.4)	84.0	96.2	98.5	64.2	87.6	92.9	523.3(+9.5)
ESA	80.1	96.6	99.0	64.8	91.5	96.2	528.2	82.7	96.6	98.6	63.0	87.2	92.0	520.1
ESA+NDGL	81.2	97.0	98.9	65.8	92.0	96.4	531.4(+3.2)	85.0	96.4	98.3	65.5	88.6	93.0	526.8(+6.7)
HREM	81.5	96.5	98.8	65.8	91.3	96.0	530.0	82.1	96.5	98.5	62.1	86.5	91.7	517.4
HREM+NDGL	82.3	97.0	99.0	66.4	91.8	96.4	533.0(+3.0)	84.4	96.7	98.6	64.4	88.1	92.5	524.9(+7.5)
CHAN	80.7	96.8	99.1	66.2	91.9	96.6	531.3	79.8	95.8	97.5	65.3	88.0	92.9	519.3
CHAN+NDGL	83.3	97.2	99.2	68.5	93.0	97.1	538.3(+7.0)	84.8	97.0	98.5	69.0	90.2	94.3	533.8(+14.5)
ResNeXt-101 + BERT														
GPO	84.7	97.9	99.4	71.6	93.9	97.4	544.8	88.1	98.6	99.7	74.3	93.3	96.5	550.4
GPO+NDGL	85.5	98.0	99.5	72.9	94.2	97.6	547.7(+2.9)	87.9	99.1	99.7	74.7	93.8	97.1	552.3(+1.9)

Table 1: Image-text retrieval results on MS-COCO 1K Test and Flickr30K datasets. The results for the original model are obtained from re-training using the code provided in the paper.

mance of baseline models. It achieves this without modifying the network architecture or introducing additional parameters, while effectively avoiding poor local minima through the use of the hard negative mining strategy. Unlike previous methods that focus on identifying the hardest negative samples, our framework mines all negative semantic similarities to guide the model in understanding the fine-grained correspondences between images and text. In contrast to methods that optimize using all negatives without accounting for their differences, our approach preserves the critical role of the hardest negatives while also helping the baseline model unlock the significant potential that is otherwise left. As shown in Figure 3(a), the performance of a single model trained with NDGL is comparable to, or even surpasses, that of the ensemble results, which are used as a reference distribution during the training process.

Explanation of Results. The result of NDGL can be explained through the perspective of label smoothing (Szegedy et al. 2016), a technique commonly used to replace "hard labels" with smoothed labels, thereby improving performance across various tasks. Our framework can be viewed as a form of label smoothing, offering smooth semantic differences for sample pairs, and consequently achieving better performance than the "teacher" model. To further validate the regularizing effect of the NDGL method on the prediction results, we calculate the margin distribution for both the original model and the model trained with NDGL. As shown in Figure 3(b), the model trained with NDGL avoiding overfitting to noise or outliers and resulting in a more realistic distribution. This regularization enhances the robustness of predictions, allowing the model to fully leverage its potential and significantly improve overall performance.

Experiments

Experimental Settings

Datasets. We selected two widely-used datasets for our experimentation: Flickr30K (Young et al. 2014) and MS-COCO (Chen et al. 2015). The MS-COCO dataset comprises 123,287 images, each accompanied by 5 annotated captions. Our data partitioning adheres to established practices (Faghri et al. 2018), allocating 113,287 images for training, 5,000 for validation, and 5,000 for testing. We ensure robustness by reporting results averaged over 5 folds of 1,000 test images and validated on the entire 5,000-image test set. The Flickr30K dataset comprises 31,783 images obtained from the Flickr platform, with each image meticulously paired with five corresponding captions. Within the Flickr30K dataset, 29,000 images are allocated for training, 1,000 for testing, and 1,014 for validation purposes.

Evaluation Metrics. Following traditional information retrieval standards, we assess performance using $R@K$, representing the proportion of correctly matched queries among the top-K retrieved instances. Higher $R@K$ values indicate better performance. To provide a thorough evaluation of matching effectiveness and follow by baseline methods, we consolidate all recall values into RSUM, which accounts for both image-to-text and text-to-image matching directions.

Implementation Details. To ensure a comprehensive evaluation, we maintain the network architectures and configurations of all baseline methods exactly as detailed in their respective papers. For each input image, we use bottom-up and top-down attention (BUTD) (Anderson et al. 2018) to extract top-K region-level features or ResNeXt (Mahajan et al. 2018) to obtain image features. For each input text, we use two formulations for text representation: bi-directional gated recurrent unit (BiGRU) or pre-trained BERT (Devlin

Data Split	MS-COCO							Flickr30K						
Eval Task	IMG → TEXT			TEXT → IMG				IMG → TEXT			TEXT → IMG			
Method	R@1	R@5	R@10	R@1	R@5	R@10	RSUM	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
BUTD + BiGRU														
CGMN (Cheng et al. 2022)	76.8	95.4	98.3	63.8	90.7	95.7	520.7	77.9	93.8	96.8	59.9	85.1	90.6	504.1
NAAF† (Zhang et al. 2022)	80.5	96.5	98.8	64.1	90.7	96.5	527.2	81.9	96.1	98.3	61.0	85.3	90.6	513.2
DivE (Kim, Kim, and Kwak 2023)	79.8	96.2	98.6	63.6	90.7	95.7	524.6	77.8	94.0	97.5	57.5	84.0	90.0	500.8
HREM (Fu et al. 2023)	80.0	96.0	98.7	62.7	90.1	95.4	522.8	79.5	94.3	97.4	59.3	85.1	91.2	506.8
ESA (Zhu et al. 2023)	79.6	96.5	98.7	63.5	90.9	96.1	525.3	82.6	95.9	98.1	61.1	85.9	91.1	514.7
X-Dim (Zhang et al. 2023)	80.9	96.9	98.9	64.7	90.9	96.5	528.8	83.1	**96.3**	**98.4**	61.7	86.1	91.4	517.0
NUIF (Zhang et al. 2024a)	79.9	96.7	**99.0**	63.9	90.4	95.8	525.7	81.8	95.7	98.0	59.0	83.9	89.9	508.3
CHAN+NDGL	**81.5**	**97.0**	98.9	**66.5**	**92.1**	**96.7**	**532.7**	**83.7**	95.8	98.2	**64.8**	**88.6**	**93.1**	**524.2**
BUTD + BERT														
VSRN++† (Li et al. 2022b)	77.9	96.0	98.5	64.1	91.0	96.1	523.6	79.2	94.6	97.5	60.6	85.6	91.4	508.9
HREM (Fu et al. 2023)	81.1	96.6	98.9	66.1	91.6	96.5	530.7	83.3	96.0	98.1	63.5	87.1	92.4	520.4
USER (Zhang et al. 2024b)	82.8	96.8	98.8	66.1	90.6	95.6	530.5	82.7	**97.0**	98.3	63.1	86.7	92.1	519.9
DCIN (Li et al. 2023b)	80.9	96.5	98.8	65.1	91.5	96.3	529.1	83.0	96.4	98.6	63.3	87.8	92.4	521.5
ESA (Zhu et al. 2023)	80.3	96.5	99.0	65.2	91.6	96.3	528.9	84.0	96.3	**98.7**	64.7	87.8	92.3	523.8
X-Dim (Zhang et al. 2023)	82.2	97.2	99.1	66.9	92.0	96.6	534.0	83.1	96.3	98.4	61.7	86.1	91.4	517.0
NUIF (Zhang et al. 2024a)	**83.3**	**97.3**	98.9	**69.2**	92.7	96.9	538.2	83.9	96.5	98.2	67.9	89.2	93.6	529.4
CHAN+NDGL	**83.3**	97.2	**99.2**	68.5	**93.0**	**97.1**	**538.3**	**84.8**	**97.0**	98.5	**69.0**	**90.2**	**94.3**	**533.8**
ResNeXt-101 + BERT														
DivE (Kim, Kim, and Kwak 2023)	**86.3**	97.8	99.4	72.4	94.0	**97.6**	547.5	**88.8**	98.5	99.6	74.3	**94.0**	96.7	551.9
GPO+NDGL	85.5	98.0	**99.5**	**72.9**	**94.2**	**97.6**	**547.7**	87.9	**99.1**	**99.7**	**74.7**	93.8	**97.1**	**552.3**

Table 2: Image-text retrieval results on MS-COCO 1K and Flickr30K. Bold and underlined texts denote the top and the runner-up, respectively. †Ensemble models of two hypotheses.

et al. 2019). The default settings for λ_1 and λ_2 are 100.0 and 0.5, respectively.

Main Results

In this section, we empirically analyze the effectiveness and generalization ability of NDGL using several image-text matching methods on the Flickr30K and MS-COCO datasets. We selected both local-level (CHAN, DivE) and global-level (ESA, GPO, HREM) matching methods, which have publicly available code, as the baselines. Table 1 reports the experimental results on both datasets, with all results generated by a single model to ensure a fair comparison. The findings demonstrate that the performance of both basic and state-of-the-art baseline methods is significantly improved by employing our framework, verifying the generality of NDGL across different methods. The local-level methods show greater improvement when applying our framework compared to the global-level methods. Notably, the CHAN method exhibits substantial improvement with NDGL, especially on the Flickr30K dataset. The retrieval performance is significantly boosted from 519.3 to 533.8 RSUM with the BERT text encoder, and there is a 18.9% improvement with the BiGRU text encoder, verifying the effectiveness of NDGL. Moreover, the improvement on the Flickr30K dataset is greater than on MS-COCO, indicating that fully utilizing negative samples is particularly important for training robust models on smaller datasets.

Comparisons with the State-of-the-arts

We apply our method to CHAN and GPO to compare with recent state-of-the-art methods, verifying the effectiveness of our framework. The comparisons on the MS-COCO and Flickr30K datasets are reported in Table 2, showing that our method outperforms all image-text matching methods with the same image encoder and text encoder. Compared with other competitive methods, our method performs better with simpler network structures and smaller datasets. For example, using the BiGRU text encoder, our method achieves RSUM scores of 532.7 and 524.2 on the two datasets, respectively. In the smaller Flickr30K dataset, we achieve superior performance, establishing a significant gap over the second place with the BUTD image encoder. To fully validate the performance of NDGL with different image encoders, we choose GPO as a baseline and training with our framework, finding that the optimized results outperformed current state-of-the-art methods. This result demonstrates that our method significantly enhances performance, enabling the model to achieve a new state-of-the-art results.

Ablation Study

Unless otherwise specified, all ablation experiments are conducted using CHAN with BUTD as image encoder and BERT as text encoder on Flickr30K dataset.

Number of Ensemble Models. We further explore the criteria for determining the optimal number of models to use in the ensemble process. As shown in Table 3, using just one model to regularize the similarity distribution yields a significant improvement, achieving a 9.3% increase compared to the baseline. This underscores the importance of mining all negative samples during the training process. When two models are used for ensemble outputs, performance continues to improve, reaching an RSUM of 533.8. However, as the distribution of similarity changes only slightly with additional ensemble models, further increasing the number of

#Ensemble	IMG → TEXT			TEXT → IMG			
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
1	83.4	96.5	98.1	67.2	89.6	93.8	528.6
2	84.8	97.0	98.5	69.0	90.2	94.3	533.8
3	84.4	96.6	99.0	69.3	90.6	94.5	534.3
4	85.2	96.8	98.8	69.2	90.3	94.3	534.6
5	85.7	96.8	98.6	69.4	90.4	94.3	535.3

Table 3: NDGL with different ensemble number.

Component	IMG → TEXT			TEXT → IMG			
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
Baseline	79.8	95.8	97.5	65.3	88.0	92.9	519.3
+ MR	84.1	96.5	98.5	68.7	90.0	94.4	532.2
+ HNR	84.8	97.0	98.5	69.0	90.2	94.3	533.8
MR→SR	82.9	96.5	97.9	68.2	90.1	94.0	529.7

Table 4: Ablation study of model components. SR denote the similarity regularization.

models results in only marginal performance gains and a trend toward stabilization. Therefore, we selected an ensemble of two models as our default configuration.

Model Components. We verify the efficacy of the proposed Margin Regularization (MR) and Hardest Negatives Rectification (HNR) in Table 4. Compared to the model without these modules, MR results in a 12.9% RSUM improvement. HNR further enhances performance, adding an additional 1.6% to the RSUM. Furthermore, we compared our regularization method with direct similarity regularization (SR), which uses similarity measurements rather than differences. When MR is replaced with SR, the performance drops by approximately 4.1% RSUM but still shows a 10.4% increase compared to the baseline, indicating that regularization is crucial and that margin-level regularization is more suitable for image-text matching methods.

Discussion

Compare with Hard Negative Mining methods. We verify the efficacy of NDGL when applied with other hard negative mining methods using two baseline models, GPO and VSRN (Li et al. 2022b), on the Flickr30K dataset. As shown in Table 5, our framework achieves an 11.5% improvement in RSUM with the VSRN baseline, outperforming the second-best mining method by 6.1%, particularly in text-to-image retrieval performance. For the GPO baseline, we achieve an 8.8% improvement, surpassing other methods. Our method not only considers the hardest negative samples but also fully leverages the variations in similarity with different negatives. This highlights the importance of mining all samples to train a robust model.

The Influence of Batch Size. The efficacy of hard negative mining strategies is highly dependent on batch size (He et al. 2020), as the quality of negative samples improves with the number of negative pairs. Consequently, when the batch size is small, methods based on hard negative mining tend

Baseline	Mining Method	IMG → TEXT			TEXT → IMG			
		R@1	R@5	R@10	R@1	R@5	R@10	RSUM
GPO	-	81.7	95.4	97.6	61.4	85.9	91.5	513.5
GPO	CFM	82.5	95.7	98.1	62.9	86.2	91.8	517.2
GPO	LSEH	82.4	96.0	98.6	63.7	87.1	92.5	520.3
GPO	NDGL	84.0	96.2	98.5	64.2	87.6	92.9	523.3
VSRN	-	71.3	90.6	96.0	54.7	81.8	88.2	482.6
VSRN	AOQ	72.8	91.8	95.8	55.3	82.2	88.4	486.3
VSRN	CFM	72.6	92.8	96.0	55.2	82.0	89.2	487.8
VSRN	LSEH	73.0	92.8	95.7	55.8	81.9	88.8	488.0
VSRN	NDGL	73.7	93.0	96.8	57.6	83.1	89.9	494.1

Table 5: Compared with previous hard negative mining methods on Flickr30K.

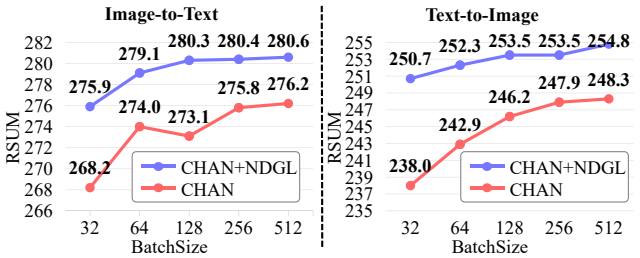

Figure 4: RSUM changes with varying batch sizes.

to perform poorly. Our NDGL framework enhances model performance by mining the hierarchical similarity of all negative samples across datasets, rather than restricting the selection to hardest negative samples within a mini-batch. This approach mitigates the dependency between hard negative mining and batch size. As shown in Figure 4, experimental results indicate that the model trained with NDGL using a small batch size achieves performance comparable to that of the baseline with a larger batch size. The improvements are particularly notable in text-to-image retrieval tasks. Additional discussion are provided in supplementary materials.

Conclusions

In this paper, we demonstrate that other negative examples offer a range of semantic differences not captured by the hardest negatives through case analysis, feature visualization, and statistical analysis, highlighting the importance of these negative samples for model training. We then introduce a novel framework for image-text matching, called negative distribution guided learning. Our framework leverages the semantic similarity differences of all negative samples to enhance model training. This approach not only preserves the benefits of large-margin learning but also exploits the substantial improvement potential left by relying solely on the hardest negatives. Extensive experiments on the Flickr30K and MS-COCO datasets validate the effectiveness of our proposed framework. As future work, we will explore the NDGL with the unified vision-language models such as (Radford et al. 2021; Li et al. 2022a; Wu et al. 2024).

Acknowledgments

This work was supported by National Natural Science Foundation of China (62472097) and Shanxi Archives Research Program (2024-SX-008). The computations in this research were performed using the CFFF platform of Fudan University.

References

Abbasi Koohpayegani, S.; Tejankar, A.; and Pirsiavash, H. 2020. Compress: Self-supervised learning by compressing representations. *Advances in Neural Information Processing Systems*, 33: 12980–12992.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the best pooling strategy for visual semantic embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15789–15798.

Chen, T.; Deng, J.; and Luo, J. 2020. Adaptive offline quintuplet loss for image-text matching. In *European Conference on Computer Vision*, 549–565.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Cheng, Y.; Zhu, X.; Qian, J.; Wen, F.; and Liu, P. 2022. Cross-modal graph matching network for image-text retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(4): 1–23.

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20: 273–297.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, 26.

Fu, Z.; Mao, Z.; Song, Y.; and Zhang, Y. 2023. Learning semantic relationship among instances for image-text matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15159–15168.

Gong, Y.; and Cosma, G. 2023. Improving visual-semantic embeddings by learning semantically-enhanced hard negatives for cross-modal information retrieval. *Pattern Recognition*, 137: 109272.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *Conference on Empirical Methods in Natural Language Processing*, 4163–4174.

Kim, D.; Kim, N.; and Kwak, S. 2023. Improving cross-modal retrieval with set of diverse embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23422–23431.

Li, H.; Bin, Y.; Liao, J.; Yang, Y.; and Shen, H. T. 2023a. Your negative may not be true negative: Boosting image-text matching with false negative elimination. In *ACM International Conference on Multimedia*, 924–934.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2022b. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 641–656.

Li, P.; Xie, H.; Ge, J.; Zhang, L.; Min, S.; and Zhang, Y. 2022c. Dual-stream knowledge-preserving hashing for unsupervised video retrieval. In *European Conference on Computer Vision*, 181–197. Springer.

Li, W.; Su, X.; Song, D.; Wang, L.; Zhang, K.; and Liu, A.-A. 2023b. Towards Deconfounded Image-Text Matching with Causal Inference. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6264–6273.

Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; and Van Der Maaten, L. 2018. Exploring the limits of weakly supervised pretraining. In *European Conference on Computer Vision*, 181–196.

Pan, Z.; Wu, F.; and Zhang, B. 2023. Fine-grained image-text matching by cross-modal hard aligning network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19275–19284.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2826.

Tejankar, A.; Koohpayegani, S. A.; Pillai, V.; Favaro, P.; and Pirsiavash, H. 2021. Isd: Self-supervised learning by iterative similarity distillation. In *IEEE/CVF International Conference on Computer Vision*, 9609–9618.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Wei, H.; Wang, S.; Han, X.; Xue, Z.; Ma, B.; Wei, X.; and Wei, X. 2022. Synthesizing counterfactual samples for effective image-text matching. In *ACM International Conference on Multimedia*, 4355–4364.

Wu, K.; Zhang, J.; Peng, H.; Liu, M.; Xiao, B.; Fu, J.; and Yuan, L. 2022. Tinyvit: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision*, 68–85. Springer.

Wu, Z.; Weng, Z.; Peng, W.; Yang, X.; Li, A.; Davis, L. S.; and Jiang, Y.-G. 2024. Building an open-vocabulary video CLIP model with better architectures, optimization and data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7): 4747–4762.

Xuan, H.; Stylianou, A.; Liu, X.; and Pless, R. 2020. Hard negative examples are hard, but useful. In *European Conference on Computer Vision*, 126–142.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Yu, B.; Liu, T.; Gong, M.; Ding, C.; and Tao, D. 2018. Correcting the triplet selection bias for triplet loss. In *European Conference on Computer Vision*, 71–87.

Zhang, H.; Zhang, L.; Zhang, K.; and Mao, Z. 2024a. Identification of Necessary Semantic Undertakers in the Causal View for Image-Text Matching. In *AAAI Conference on Artificial Intelligence*, volume 38, 7105–7114.

Zhang, K.; Mao, Z.; Wang, Q.; and Zhang, Y. 2022. Negative-aware attention framework for image-text matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15661–15670.

Zhang, K.; Zhang, L.; Hu, B.; Zhu, M.; and Mao, Z. 2023. Unlocking the Power of Cross-Dimensional Semantic Dependency for Image-Text Matching. In *ACM International Conference on Multimedia*, 4828–4837.

Zhang, Y.; Ji, Z.; Wang, D.; Pang, Y.; and Li, X. 2024b. USER: Unified semantic enhancement with momentum contrast for image-text retrieval. *IEEE Transactions on Image Processing*.

Zhu, H.; Zhang, C.; Wei, Y.; Huang, S.; and Zhao, Y. 2023. Esa: External space attention aggregation for image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10): 6131–6143.