NAPReg: Nouns As Proxies Regularization for Semantically Aware Cross-Modal Embeddings

Bhavin Jawade*, Deen Dayal Mohan*, Naji Mohamed Ali, Srirangaraj Setlur, Venu Govindaraju
Department of Computer Science and Engineering
University at Buffalo, SUNY
{bhavinja, dmohan, najimoha, setlur, govind}@buffalo.edu

Abstract

Cross-modal retrieval is a fundamental vision-language task with a broad range of practical applications. Text-to-image matching is the most common form of cross-modal retrieval where, given a large database of images and a textual query, the task is to retrieve the most relevant set of images. Existing methods utilize dual encoders with an attention mechanism and a ranking loss for learning embeddings that can be used for retrieval based on cosine similarity. Despite the fact that these methods attempt to perform semantic alignment across visual regions and textual words using tailored attention mechanisms, there is no explicit supervision from the training objective to enforce such alignment. To address this, we propose NAPReg, a novel regularization formulation that projects high-level semantic entities i.e Nouns into the embedding space as shared learnable proxies. We show that using such a formulation allows the attention mechanism to learn better word-region alignment while also utilizing region information from other samples to build a more generalized latent representation for semantic concepts. Experiments on three benchmark datasets i.e. MS-COCO, Flickr30k and Flickr8k demonstrate that our method achieves state-of-the-art results in cross-modal metric learning for text-image and image-text retrieval tasks. Code: `https://github.com/bhavinjawade/NAPReq`

1. Introduction

Learning robust embeddings for text-image matching or cross-modality retrieval is an essential goal for vision-language understanding. Cross-modal retrieval research is motivated by the need for solutions for a variety of practical challenges, such as product retrieval, person search [18], and compositional retrieval [30]. In contrast to uni-modal tasks such as image-to-image search, cross-modal retrieval requires models to learn exhaustive correspondences between modalities to model intricate relationships among entities.

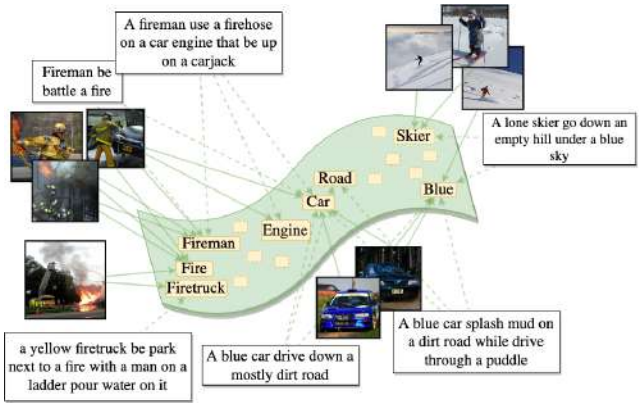
A fireman use a firehose on a car engine that be up on a carjack
Fireman be battle a fire
A lone skier go down an empty hill under a blue sky
Skier
Road
Car
Blue
Engine
Fireman
Fire
Firetruck
a yellow firetruck be park next to a fire with a man on a ladder pour water on it
A blue car drive down a mostly dirt road
A blue car splash mud on a dirt road while drive through a puddle

Figure 1: An illustration depicting the interaction of text and image features with the shared semantic entities to learn more robust visual representations while refining region-text alignment to bring contextually relevant pairs closer in the embedding space. (Best viewed in digital.)

Early works [8, 13] that project image and text representations into a shared embedding space for retrieval are unable to capture the fine-grained interactions between high-level and coarse features over modalities. Capturing the alignment between features across text and images is essential for determining similarity that could discriminate across seemingly alike samples. Recent works have tried to capture these relationships between visual regions and textual words using tailored attention mechanisms, feature alignment methods, and feature aggregation modules. Lee et al. [15] proposed refinement of the region-to-word alignment by utilizing stacked cross attention to compute aggregated image-to-sentence similarity. Liu et al. [20] aimed to learn the correspondence between textual and visual graphs by modeling relationships among attributes and objects. Diao et al. [7] improved upon [15] by modeling the local and the global feature similarity as graph nodes connected through directed edges to iteratively compute final similarity.

[*] Equal contribution authors in alphabetical order

Though these approaches have explored different representation learning strategies with incremental boosts in performance, ranking loss has remained a de-facto training objective for cross-modal retrieval. While most previous works have utilized triplet loss with a random sample mining strategy, Chen et al.[6] proposed an offline hard negative sampling method and Wei et al.[33] proposed a polynomial function for weighting positive and negative informative pairs. Although triplet based losses have worked well for uni-modal/multi-modal retrieval in the past, an additional supervision to improve image-to-text alignment that could complement the newer cross-attention mechanisms is desirable.

In this work, we propose NAPReg, a regularization objective that provides direct supervision to improve the cross-modal alignment by projecting the high-level semantic notions that are captured by nouns in a sentence as shared learnable proxies in the embedding space. This regularization term aids the existing attention mechanisms to learn a better region-to-word alignment. NAPReg can be easily integrated into existing triplet based formulations and can complement a variety of existing cross-attention and feature alignment modules to learn more robust feature representations.
The main contributions of this paper are summarized as follows:

1. We identify the need for direct supervision from the training objective to learn better region-to-word alignment for text-image retrieval.
2. We propose NAPReg, a proxy based formulation that maximizes the similarity between the aggregated context vector and the shared semantic proxies to achieve better region-word alignment and learn more generalized visual latent representations.
3. We specifically design NAPReg to complement existing cross-attention techniques and provide them with the required supervision to achieve robust feature alignment. We conduct extensive experiments on three benchmark datasets: MS-COCO, Flickr30k, and Flickr8k, and demonstrate the effectiveness of NAPReg with multiple feature alignment methods. NAPReg consistently achieves superior results over the state-of-the-art. We also perform a rigorous empirical study and qualitative analysis to evaluate the role of different hyper-parameters involved in the regularization term.

2. Related Works

Current methods for cross-modal retrieval have broadly focused on two approaches: i) improving the backbone architectures for feature extraction and alignment, and ii) improving the training objective and loss function formulation to learn more discriminative features. The latter, also known as cross-modal metric learning, explores novel approaches in a) modality interaction and feature aggregation using attention, and b) deep metric learning using sampling and hard mining strategies, and custom loss formulations.

Cross-Modal Feature Extraction and Aggregation - The representation learning backbone for cross-modality retrieval consists of two parts: the feature extractor and the feature aggregator. Lu et al.[21] performed feature extraction using an additional supervision input from the Faster-RCNN object detector to provide a vision transformer encoder with labeled image patches. Other methods [17, 3] utilize pretrained visual representations from a bottom-up attention network [1] and augment it with a novel feature alignment module. Li et al.[17] proposed a region relationship model and a global semantic reasoning model built upon a graph convolutional network and GRUs using image features from bottom-up attention [1] which were jointly optimized using a matching loss (hinge-based triplet ranking loss) and a generation loss (log-likelihood captioning loss). [26, 37] proposed different attention based fusion architectures for multi-modality features. Chen et al. [4] proposed a generalized pooling strategy which computes weights for increasing orders of max pool operator using a BiGRU for cross-modality retrieval on VSE++ features. [32] proposed utilizing the object's position information along with an attention mechanism to learn a region position feature vector for improving cross-modality retrieval. [34] proposed a transformer based intra-modal and inter-modal attention network to learn a multiple sample embedding. The use of triplet loss formulation as a training objective is a key feature shared by all recent feature aggregation methods.

Metric Learning Methods Lecun et al. [10] proposed a contrastive loss formulation that tried to reduce the distance between the feature representations of images if they belonged to the same class and increased it if they belonged to a different class. Triplet loss [28], lifted structure loss [25], and N-Pair loss [29] introduced the notion of negative samples and proposed sampling strategies in which batches are constructed intelligently based on the relative importance of different samples. Furthermore, techniques such as pair weighting [35] and curriculum learning [2] have been proposed to improve the sampling process. Even though these methods were initially proposed for uni-modal data retrieval, they have also been used in various cross-modal data retrieval tasks. Wei et al. [33] proposed a self-similarity and relative-similarity based polynomial formulation of triplet similarities for cross-modal metric learning. Mining for informative samples often becomes a computationally intensive task. Yair et al. [23] proposed a proxy-based method to overcome this computational overhead. Anchor proxy loss [14] further improved the formulation to incorporate the relative distance between samples in the feature space. These proxy-based methods use anchor points (usually one per class) to act as a rallying point for all the positive image features belonging to a specific class. Due to the class-specific nature of these loss formulations and the lack of global categorical information for text-image matching tasks, proxy based methods are not directly applicable to cross-modal retrieval.

Large Scale VL Pretraining Recently, transformer based large scale vision language pretraining (also referred to as *foundational models*) have gained interest. Primarily, the goal in this domain is to train large transformer based vision and text encoders on typically large data sets (millions of image-text pairs). The two avenues of research in this domain are: (i) Contrastive Pretraining and (ii) Cross-Attention based pre-training. [27], [24], [36] demostrate that contrastively pre-training image and text embeddings on a large number of image-text pairs shows robust generalizability for zero-shot classification. [16], and [12], showed that joint contrastive and cross-attention based pretraining improves performance on language based downstream tasks such as VQA and visual grounding.

Majority of the loss formulations used in existing cross-modal retrieval methods are inspired from deep metric learning. These methods lack an explicit objective to enforce fine grained alignment across modalities. In this paper, we present a novel regularization method that augments the capability of existing cross-modal loss formulations. Our regularization method overcomes the class dependency of proxy based methods by synthesizing proxies from tangible entities present in textual content and using them as shared semantic notions.

3. Methodology

In this section, we will first revisit the formulation of a cross-modal image-text retrieval problem and subsequently provide the motivation and design for the proposed regularization (NAPReg).

3.1. Problem Statement

Consider the visual features of an image $V = \{v_1, v_2, ...v_n\}$, where $v_i \in \mathbb{R}^{d_v}$ is the feature representation corresponding to the i^{th} region in the image. Here n is the number of visual regions under consideration. Let $T = \{t_1, t_2, ...t_m\}$ be the text features corresponding to a sentence, where $t_i \in \mathbb{R}^{d_t}$ is the encoded representation of each word. m is the number of words in the sentence.

Given a query belonging to a specific modality, cross-modal retrieval aims to find the best possible mated representation from a gallery of samples belonging to the other modality. The similarity of an image and text pair is given by:

$$S(V,T) = f(\Phi(V;\theta_i), \Psi(T;\theta_j)) \tag{1}$$

where Φ is an MLP or any other non-linear transformation and Ψ is a sequence model (ex. LSTM, BERT, etc.) that projects the corresponding feature representations into a shared embedding space. θ_i and θ_j are the parameters for the corresponding modalities. f is the function to compute similarity between the two representations.

Following [15], an optimal strategy to aggregate different region level and word level features for cross-modal retrieval, is to use a stacked cross attention module. Considering the objective of image-to-text matching, for each visual location, an attended combination of word representation a (i.e. the attended sentence vector a_i^t, with respect to the i-th image region a_t^i) is constructed as defined below:

$$\begin{aligned} s_{ij} &= \frac{v_i^T t_j}{||v_i||.||t_j||}, i \in [1,n], j \in [1,m] \\ w_{ij} &= \frac{exp(\tau.\bar{s}_{ij})}{\sum_{j=1}^{n} exp(\tau.\bar{s}_{ij})} \\ a_i^t &= \sum_{j=1}^{m} w_{ij} * t_j \end{aligned} \tag{2}$$

where s_{ij} is the cosine similarity of each individual L2 normalized word representation j with an image region i and τ is the inverse temperature of the softmax. The overall cosine similarity between the image-text pair is given by:

$$S(V,T) = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_i \cdot \hat{a_i}^t \tag{3}$$

In general, given a training set $\mathcal{X}$ consisting of image-text pairs, the objective can be stated as $S(V_i, T_i) > S(V_i, T_j) \forall i \neq j$. One of the most widely used loss formulations for such retrieval problems is the triplet loss. The matching objective can be stated as:

$$\begin{aligned} \mathcal{L} = \sum_{(i,j \sim \mathcal{X})} & \{S(V_j, T_i) - S(V_i, T_i) + \alpha\}_+ \\ & + \{S(V_i, T_j) - S(V_i, T_i) + \alpha\}_+ \end{aligned} \tag{4}$$

where, $[x]_+ = \max(x, 0)$.

3.2. Nouns as Proxies

Our analysis of existing objective functions used for cross-modal retrieval approaches reveal a shortcoming that we intend to address by improving the optimization criteria. We first motivate the need for such a regularization and then present the formulation.

3.2.1 Motivation

In order to get a better understanding of the problem at hand, let us first consider an example of an image which has three salient regions and a text string that describes this image. Following the discussion in the previous section, we know the visual features that describe the image are given by $V = \{v_1, v_2 ......v_n\}$. Similarly, $T = \{t_1, t_2 ......t_m\}$ is the set of word level features used to describe the image.

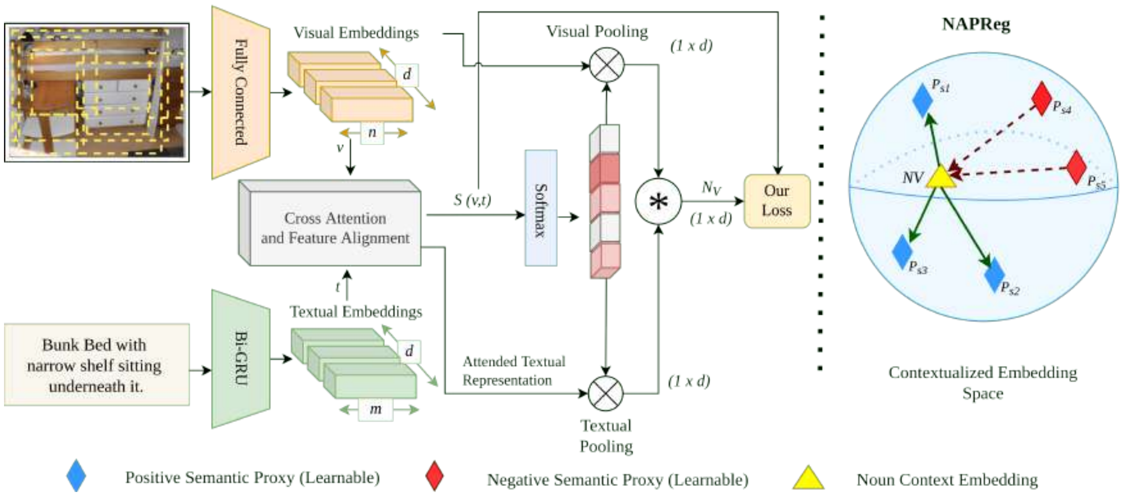
Fully Connected
Visual Embeddings
d
n
v
Visual Pooling
(1 x d)
S (v,t)
Softmax
Cross Attention and Feature Alignment
t
Textual Embeddings
d
m
Bi-GRU
Bunk Bed with narrow shelf sitting underneath it.
Attended Textual Representation
Textual Pooling
(1 x d)
N_V
(1 x d)
Our Loss
NAPReg
P_{s1}
P_{s4}
NV
P_{s5}
P_{s3}
P_{s2}
Contextualized Embedding Space
Positive Semantic Proxy (Learnable)
Negative Semantic Proxy (Learnable)
Noun Context Embedding

Figure 2: Overview of the proposed loss function. For each positive sample (text-image pair) in the training mini-batch, we maximize its similarity to its hard positives in another modality while minimizing similarity to hard negatives. Simultaneously, we compute a noun context vector $\mathcal{N}_{\mathcal{V}}$ by pooling the visual and textual features. In the regularization term, we maximize similarity of this noun context vector with respect to relevant learnable semantic proxies and minimize similarity with respect to irrelevant proxies. The resulting loss is the weighted combination of this pair loss and the regularization term. (Best viewed in digital.)

Consider v_i, v_j,v_k, for i,j,$k \in n$, are the location/region level representation associated with the three salient regions(objects,actions,attributes etc.) and t_a,t_b,t_c for a,b,$c \in m$ are words in the sentence that are associated with these regions. Given the presence of these three salient regions, it seems logical that the similarity between salient regions (v_i, v_j,v_k) and the corresponding attended textual vectors(a_i^t,a_j^t,a_k^t) should contribute significantly more to the global similarity score than the non-salient visual regions. This objective can be written as

$$\{\hat{v}_i \cdot a_i^t + \hat{v}_j \cdot a_j^t + \hat{v}_k \cdot a_k^t\} > \{\sum_{x=1}^{n} \hat{v}_x \cdot \hat{a}_x^t, r \notin i,j,k\} \tag{5}$$

The right hand side of Eq. 5 represents the similarity between irrelevant/non salient visual regions. In order to fulfil this additional constraint, a_i^t should have more weightage (w_{ij}) from the corresponding relevant word (t_j) in the text. Enforcing such a constraint is non trivial since the labels for region level features are either unavailable or the feature extractors' predictions are inaccurate. Hence the current loss formulations fail to provide an explicit supervision to enforce region word alignment for cross modal retrieval.

3.2.2 Defining Proxies

To provide additional supervision, we make use of the parts of speech of the words in the sentence describing the image. Let M = $\{m_1$,m_2,....$m_l\}$ be concepts extracted from a sentence, using a standard part-of-speech (POS) tagger, then C will be the set of all concepts aggregated from all the sentences present in the dataset. Even though the concepts here can be nouns, verbs, adjective etc, which can have possible association with image regions, majority of the cross-modal retrieval methods rely on visual feature extractors like [1] which was trained to detect objects and attributes. Since we also rely on these feature extractors, we restrict our concepts to nouns. Given N as the total number of unique noun entities that occur more than K times in C, we define $\mathcal{P}$ as the set of (N, d) dimensional learnable proxy embeddings, each representing a unique noun entity. These proxy embeddings can be used to provide a notion of shared semantics as additional supervision required for the region level image-to-text alignment. To enable these proxy embeddings to refine the image-text alignment, we need an aggregated representation of relevant image and textual features. In order to achieve this, we introduce the concept of a noun context vector $\mathcal{N}_{\mathcal{V}}$, which is an aggregated representation of visual and textual regions. To compute the noun context vector, we first use the individual region to text alignment scores.

$$S = \{\hat{v}_i \cdot \hat{a}_i^t\} \forall i \in n \tag{6}$$

The relative importance of any visual region can be measured by how well it aligns with the corresponding attended sentence vector compared to all the other visual regions. Formally:

$$s = \text{softmax}(S) \tag{7}$$

Given the relative importance weights s, the final noun context vector is created by pooling the visual features and textual features. This can be written as.

$$\mathcal{N}_{\mathcal{V}} = (\sum_{i=1}^{n} s_i * a_i^t) \odot (\sum_{i=1}^{n} s_i * v_i) \tag{8}$$

Where $\odot$ denotes the Hadamard product (Refer Fig. 2).

Once the $\mathcal{N}_{\mathcal{V}}$ is computed, we explicitly force the noun context vector to better align with noun proxies to enhance the relationship with salient objects, using our proposed regularization constraint which we call Nouns as Proxies (NAPReg). Let D^+ be the set of all positive image-text pairs in the training set. If $NS = \{n_1, n_2....n_l\}$ represent nouns in a Text T_1 then $P^+ = \{p_n^1, p_n^2, ...p_n^l\}$ corresponds to the positive proxies of these nouns and all the other proxies in N are regarded as negatives. If there are multiple descriptions associated with the same image we aggregate the noun entities for all text samples, given by:

$$Nb_i = \cup_{j=1}^{c} NS_j \tag{9}$$

Where Nb_i denotes all the noun entities belonging to the c other descriptions of the i^{th} image. Subsequently, P^+ will also be augmented by adding the corresponding proxies to the set of positives. This is done to prevent the noun context vector $\mathcal{N}_{\mathcal{V}}$ being separated during optimization from proxies associated with nouns that are synonyms of each other. Following standard practice [31], we formulate the regularization as a log sum of exponents term, which is defined as:

$$\mathcal{L}_{nap} = \sum_{\mathcal{X}} \left\{ \frac{1}{\alpha_1} \log\left(1 + \sum_{p \in P^+} e^{-\alpha_1 (S_{np} - \lambda_1)}\right) + \frac{1}{\beta_1} \log\left(1 + \sum_{p \notin P^+} e^{\beta_1 (S_{np} - \lambda_1)}\right) \right\} \tag{10}$$

Where $S_{np} = \hat{P} \cdot \hat{\mathcal{N}_{\mathcal{V}}}$ is the cosine similarity between the noun context vector and the proxies. One can note that in the second log-sum-exponent term in regularization, the noun context vector is pushed away from negative noun proxies. This is advantageous because some of the negative proxies would serve as hard negatives, enhancing the positive region to noun association, and therefore, representation. Since the regularization is a secondary objective, we combine this term with the primary objective of separating positive and negative image-text pairs. We follow a similar formulation which can be written as:

$$\mathcal{L}_{pair} = \sum_{\mathcal{X}} \left\{ \frac{1}{\alpha_2} \log\left(1 + \sum_{(v,t) \in D^+} e^{-\alpha_2 (\bar{S} - \lambda_2)}\right) + \frac{1}{\beta_2} \log\left(1 + \sum_{(v,t) \notin D^+} e^{\beta_2 (\bar{S} - \lambda_2)}\right) \right\} \tag{11}$$

where $(v, t) \in D^+$ denotes a positive image-text pair while $(v, t) \notin D^+$ denotes an image-text pair that is unrelated. $\bar{S}$ denotes the mean aggregated similarity score on S for each image-text pair. The final loss formulation is defined by

$$\mathcal{L} = \mathcal{L}_{pair} + \gamma \mathcal{L}_{nap} \tag{12}$$

4. Experiments

4.1. Datasets

We perform several experiments and ablation studies on three image-text benchmark datasets: Flickr8k, Flickr30k, and MSCOCO following the standard protocol used in [15, 32, 7]. Flickr8k dataset contains 6000 images in the train, 1000 in the validation and 1000 in the test set. Each of these images has 5 captions associated with it. Flickr30k dataset contains 31000 images, with 5 captions per image, out of which 1000 images are used for testing, 1000 for validation and 29000 for training. We demonstrate the scalability of our loss function on MSCOCO which is a large-scale benchmark with 123,287 images with five captions each. We utilize 5000 images for validation, 5000 for testing and 113,287 for training. Results are reported on both the full 5000 image test set, and the 1000 image test set averaged over 5 folds. The performance is evaluated using $Recall@K$ metric where $K \in \{1, 5, 10\}$. We report results on both text-to-image and image-to-text retrieval tasks.

4.2. Implementation details

For fair comparison, we follow the experimental setup used by other methods [15, 7, 33]. Following the conventional practice, we extract $(36, 2048)$ dimensional visual features from the bottom-up attention network pretrained on the visual genome dataset. A set of word features which are encoded by a bi-directional Gated Recurrent Unit (GRU) are used as textual features. For all experiments, we fix the embedding dimension of the proxies and the feature vectors to 1024. For extracting noun entities from text captions, we use the *nltk* part-of-speech tagger before performing lemmatization and stemming on words. We extract N proxies from the training dataset by thresholding on the frequency of the particular word in the whole dataset. For the results reported in the tables we use $N = 1551$ for Flickr30k, $N = 2275$ for MSCOCO and $N = 2444$ for Flickr8k as the number of proxies. We use a higher learning rate for the proxies than the features. The proxy learning rate for all experiments is 0.08 and the learning rate for model parameters is 0.0002. For flickr8k, we use $\gamma = 0.30$ and for MSCOCO and Flickr30k we use $\gamma = 0.15$. For the other parameters, we used the default settings mentioned in the implementation of [31](refer supplementary material for more details). In order to augment SCAN [15] attention with our regularization, we compute the noun context vector as described in Eq.8 to calculate L_{nap}, and utilize pair wise image-to-text similarity score $S(v, t)$ computed by [15] to calculate L_{pair}. Similarly, for augmenting SGRAF's [7] similarity graph reasoning module, we utilize pair wise image-to-text similarity score $S(v, t)$ computed by [7] to calculate L_{pair} and extract noun context vector through Eq.8 that uses attention mechanism as described in

Table 1: Recall@K(%) performance on Flickr30K dataset

Method	Reference	Loss	Text-to-Image			Image-to-Text		
			R@1	R@5	R@10	R@1	R@5	R@10
BFAN[19]	MMM'19	Triplet	50.8	78.4	85.8	68.1	91.4	95.9
IMRAM$_{\text{Full}}$[3]	CVPR'20	Triplet	53.9	79.4	87.2	74.1	93	96.6
GSMN$_{\text{Sparse}}$[20]	CVPR'20	Triplet	53.9	79.7	87.1	71.4	92	96.1
PFAN$_{\text{i2t}}$[32]	IJCAI'21	Triplet	45.7	74.7	83.6	67.6	90.0	93.8
SCAN$_{\text{i2t}}$[15]	ECCV'18	Triplet	43.9	74.2	82.8	67.9	89	94.4
SMFEA[9]	MMM'21	Triplet	54.7	82.1	88.4	73.7	92.5	96.1
SHAN[11]	IJCAI'21	Triplet	55.3	81.3	88.4	74.6	93.5	96.9
VSE∞[5]	CVPR'21	Triplet	56.4	83.4	89.9	76.7	94.2	97.7
UWML$_{\text{i2t}}$[33]	CVPR'21	Polyloss	47.5	75.5	83.1	69.4	89.4	95.4
NAAF$_{\text{BiGRU}}$[38]	CVPR'22	Polyloss	55.5	81.0	87.9	75.9	93.6	97.7
SGRAF$_{\text{SGR}}$[7]	AAAI'21	Triplet	56.2	81	86.5	75.2	93.3	96.6
SCAN$_{\text{i2t}}$	**Ours**	**Ours**	**51.4**	**77.6**	**85.7**	**70.8**	**90.9**	**95.3**
SGRAF$_{\text{SGR}}$	**Ours**	**Ours**	**58.3**	**83.1**	**89.2**	**79.2**	**95.3**	**97.7**
SGRAF$_{\text{SGR+SAF}}$[7]	AAAI'21	Triplet	58.5	83.0	88.8	77.8	94.1	97.4
SCAN$_{\text{i2t+t2i}}$[15]	ECCV'18	Triplet	48.6	77.7	85.2	67.4	90.3	95.8
SGRAF$_{\text{SGR+SAF}}$	**Ours**	**Ours**	**60.0**	**84.1**	**90.2**	**79.6**	**95.6**	**98.0**

Methods highlighted in same color use exactly same backbone and aggregation method for comparison. For NAAF we report numbers for Bi-GRU textual features for fair comparison with all other works. Best results are in bold.

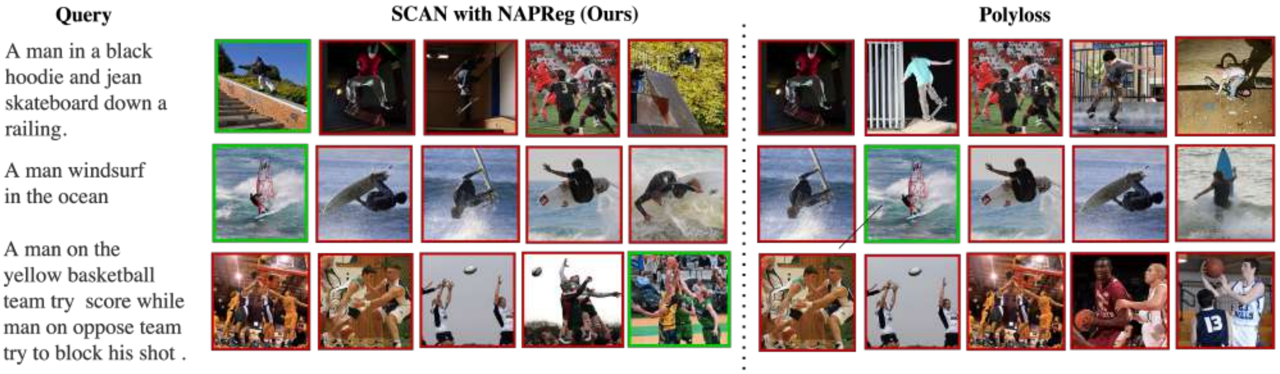
Query
SCAN with NAPReg (Ours)
Polyloss
A man in a black hoodie and jean skateboard down a railing.
A man windsurf in the ocean
A man on the yellow basketball team try score while man on oppose team try to block his shot .

Figure 3: Left Side - Our loss, Right Side - Polyloss. Qualitative results on the Flickr8k. For each query image the top 5 predictions are presented in sorted order. The results with green boundary represent successful retrievals of a correct image, while images with red boundary are incorrect retrievals. Note - there is only one correct retrieval per text query in the dataset.(Best viewed in digital)

Eq. 2.

4.3. Comparison with the state-of-the-art

We compare the performance of our proposed loss function Eq.12 , against the most representative works in the domain of cross-modal metric learning based retrieval. Further, we demonstrate the robustness of the proposed formulation by using it alongside state-of-the-art text-image matching architectures such as SCAN and SGRAF. In Table 1, we observe that SCAN$_{\text{i2t}}$ trained with the NAPReg objective, outperforms all previous SCAN$_{\text{i2t}}$ based approaches by a large margin on Flickr30k. Moreover, we establish a new state-of-the-art result on Flickr30k by training SGRAF$_{\text{SGR}}$ with our regularization. We note that SGRAF$_{\text{SGR}}$ trained with the proposed formulation surpasses the original SGRAF$_{\text{SGR}}$ trained using a triplet-based loss by 2.1% R@1 for the text-to-image retrieval task and 4.0 % R@1 for image-to-text retrieval. Futhermore, the averaging the performance of SGRAF$_{\text{SGR}}$ trained using NAPReg with SGRAF $_{\text{SAF}}$, has 2.0 % improvement in R@1 for text-to-image retrieval . This significant improvement in the cross-modal retrieval task can be attributed to robust alignment of the salient image region to the corresponding text. Table 2 shows results on MSCOCO dataset for 5k (full test set) and 1k (5 fold) evaluation protocols respectively. Here, we again observe that the addition of NAPReg improves performance over all existing methods for cross-modality retrieval. Specifically, using the proposed loss formulation with SGRAF$_{\text{SGR+SAF}}$ provides 1.1% improvement on R@1 for text-to-image matching on the 5K test set and likewise a 3.7% improvement on R@1 for text-to-

Table 2: Recall@K(%) performance on MSCOCO dataset

Method	Reference	Loss	Text-to-Image			Image-to-Text		
			R@1	R@5	R@10	R@1	R@5	R@10
MSCOCO - 1K Evaluation								
IMRAM$_{\text{Full}}$[3]	CVPR'20	Triplet	61.7	89.1	95	76.7	95.6	98.5
GSMN$_{\text{Sparse}}$[20]	CVPR'20	Triplet	60.4	88.7	95	76.1	95.6	98.3
PFAN$_{\text{i2t}}$[32]	IJCAI'21	Triplet	53.0	84.5	92.6	70.7	94.1	97.8
SCAN$_{\text{i2t}}$[15]	ECCV'18	Triplet	54.4	86	93.6	69.2	93.2	97.5
SHAN [11]	IJCAI'21	Triplet	62.6	89.6	95.8	76.8	96.3	98.7
VSE ∞ [5]	CVPR'21	Triplet	61.7	90.3	95.6	78.5	96.0	98.7
UWML$_{\text{i2t}}$[33]	CVPR'21	Polyloss	56.8	86.7	93	71.1	93.7	98.2
NAAF$_{\text{BiGRU}}$[39]	CVPR'22	Triplet	61.3	90.6	96.0	76.8	95.2	98.2
SGRAF$_{\text{SGR}}$[7]	AAAI'21	Triplet	61.4	89.3	95.4	78	95.8	98.2
SCAN$_{\text{i2t}}$	**Ours**	**Ours**	**58.6**	**87.5**	**93.8**	**71.6**	**94.5**	**98.2**
SGRAF$_{\text{SGR}}$	**Ours**	**Ours**	**63.3**	**90.0**	**95.6**	**78.7**	**96.2**	**98.8**
SCAN$_{\text{i2t+t2i}}$	ECCV'18	Triplet	58.8	88.4	94.8	72.7	94.8	98.4
SGRAF$_{\text{SGR+SAF}}$	AAAI'21	Triplet	63.2	90.7	96.1	79.6	96.2	98.5
SGRAF$_{\text{SGR+SAF}}$	**Ours**	**Ours**	**66.9**	**91.6**	**96.5**	**81.9**	**97.5**	**99.2**
MSCOCO-5K Evaluation								
IMRAM$_{\text{Full}}$[3]	CVPR'20	Triplet	39.7	69.1	79.8	53.7	83.2	91
SCAN$_{\text{i2t}}$[15]	ECCV'18	Triplet	34.4	64.2	75.9	46.4	77.4	87.6
UWML$_{\text{i2t}}$[33]	CVPR'21	Polyloss	34.4	64.2	75.9	46.9	77.7	87.6
SGRAF$_{\text{SGR}}$[7]	AAAI'21	Triplet	40.2	-	79.8	56.9	-	90.5
SCAN$_{\text{i2t}}$	**Ours**	**Ours**	**36.5**	**66.0**	**77.6**	**48.0**	**78.6**	**88.3**
SGRAF$_{\text{SGR}}$	**Ours**	**Ours**	**41.7**	**71.2**	**81.5**	**58.0**	**85.1**	**91.6**
SCAN$_{\text{i2t+t2i}}$	ECCV'18	Triplet	38.6	69.3	80.4	50.4	82.2	90.0
SGRAF$_{\text{SGR+SAF}}$	AAAI'21	Triplet	41.9	-	79.8	57.8	-	91.6
SGRAF$_{\text{SGR+SAF}}$	**Ours**	**Ours**	**43.0**	**72.1**	**82.4**	**59.8**	**86.0**	**92.6**

Methods highlighted in same color use exactly same backbone and aggregation method for comparison. For NAAF we report numbers for Bi-GRU textual features for fair comparison with all other works. Best results are in bold.

Table 3: Recall@K(%) performance on Flickr8K dataset

Method	Reference	Loss	Text-to-Image			Image-to-Text		
			R@1	R@5	R@10	R@1	R@5	R@10
DeViSE[8]	NIPS'13	Hinge	5.9	20.1	29.6	4.8	16.5	27.3
DVSA[13]	PAMI'16	Triplet	11.8	32.1	44.7	16.5	40.6	54.2
m-CNN [22]	CVPR'15	Triplet	20.3	47.6	61.7	24.8	53.7	67.1
IMRAM$_{\text{Image}}$[3]	CVPR'20	Triplet	32	61.4	73.9	48.5	78.1	85.3
SCAN*$_{\text{i2t}}$[15]	ECCV'18	Triplet	32.3	62.3	74.3	51.2	77.6	87.2
UWML*$_{\text{i2t}}$[33]	CVPR'21	Polyloss	33.3	63.8	75.5	50.7	78.9	88.4
SCAN$_{\text{i2t}}$	**Ours**	**Ours**	**39.2**	**69.1**	**79.7**	**56.2**	**82.8**	**90.7**

{*} Papers did not report numbers on Flickr8k. We produce the experimental results using the code provided by the authors. Methods highlighted in same color use exactly same backbone and aggregation method for comparison. Best results are in bold.

image matching on the 1K test set. Evaluation on the large scale MSCOCO dataset establishes the scalability of our loss function for bigger training sets.

4.4. Discussion

Effect of γ on Performance - To understand the behaviour of our final loss formulation with varying influence from the regularization term, we conduct experiments on Flickr8k for different values of γ. For the experiment, we use the SCAN[15] architecture. As seen in table 4, we find that the best performance is obtained when γ = 0.3. After γ = 0.3, we observe that performance gradually declines because the regularization in loss begins to overpower the discriminative term. Based on this analysis, we may conclude

Table 4: Ablation to evaluate effect of Gamma γ on matching performance on Flickr8k dataset.

Gamma γ	Text-to-Image		Image-to-Text	
	R@1	Rsum	R@1	Rsum
0.0	37.7	184.3	52.1	226.4
0.1	37.6	184.9	54.4	227.0
0.2	38.1	186.4	54.5	228.4
0.3	**39.2**	**188.0**	**56.2**	**229.7**
0.4	38.3	186.5	54.8	228.7

Rsum denotes aggregation of R@1, R@5 and R@10

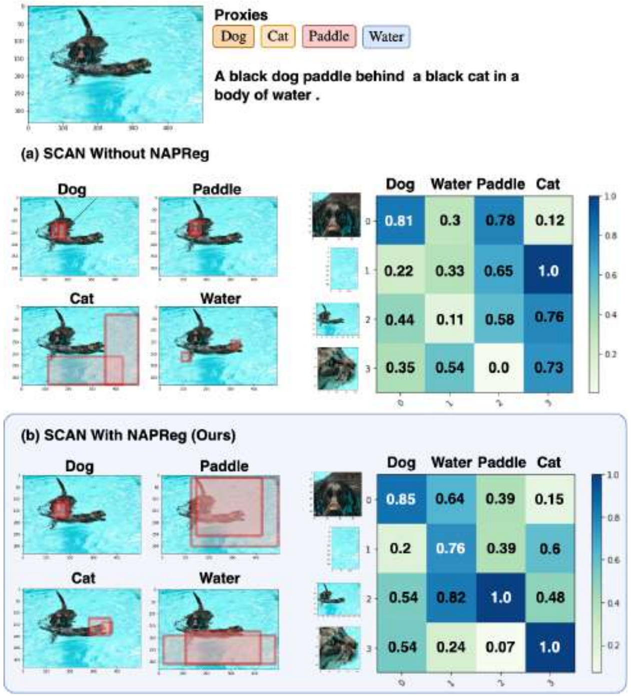

Figure 4: Qualitative analysis of region to word alignment. (a) SCAN without NAPReg - Shows the top 2 regions attended by each proxy word in the image on Left and heatmap between the similarity of selected visually relevant regions and the word proxies on the right. (b) SCAN with NAPReg. Here the similarity scores are min-max normalized for visualization. 1.0 denotes highly similar and 0.0 denotes highly dissimilar. (Best viewed in digital)

that for the current experimental setup, a positive value of γ in the range of [0.1-0.3] produces the best results. Further, from our experiments on other datasets we conclude that the optimal value of γ is greater for smaller datasets compared to the larger ones. This can be attributed to the fact that in the smaller datasets, there are fewer image-text instances available for the network to learn how to align the salient region to the noun text correctly, when compared with larger datasets. A larger value of γ provides more weightage to the above mentioned alignment process for each image-text pair.

Qualitative results - Qualitative results for top-5 recall on Flickr8k dataset can be seen in Figure 2. The examples shown demonstrate that our method produces better retrieval results. In the first row, we can see that a better alignment of the nouns to the salient regions helps to retrieve an image that best matches the query. Another interesting observation (second row) is that the model trained with our loss formulation can distinguish between a windsurf and a surfboard. Since both windsurf and surfboard mostly occur in a similar image setting, the alignment of corresponding region-text is challenging. However, when utilizing NAPReg regularization, developing a robust representation is easier because both windsurf and surfboard have independent proxies that behave as negatives to each other. This result provides qualitative validation of the theoretical analysis presented for the proposed method.

Figure 4a shows the attention map generated for each salient object in the image, with and without our regularization term. As one can see, the alignment between the text and the image region is much more refined when using our regularization. Furthermore, it is able to distinguish between the image regions corresponding to cat and dog, even when there is only a subtle difference between the two. Another interesting observation is that when using regularization, the model is able to identify various relevant larger regions that correlate to the term water. Figure 4b shows the similarity score (min-max normalized) of selected visually relevant regions with semantically dominant words in the description. It can be seen that the similarity score of the visual region containing the cat and the dog is highest for the corresponding word in the text. Furthermore, the magnitude of the scores has also increased in comparison to the model without the proposed regularization. This shows that the alignment generated by our proposed loss function is superior to that of prior loss formulations.

5. Conclusion

Cross-Modal image-text retrieval finds application in a variety of challenging domains. Further, developing feature representations for both modalities that can map semantic relationships between visual and text elements is critical. As can be seen from different attention methods, the objective function used for creating these representations also plays a crucial role. In this work, we have identified an inadequacy in existing loss formulations where they lack the much needed emphasis on alignment of salient regions in an image-text pair. To address this limitation, we have proposed a novel regularization. We have provided a theoretical basis for the proposed proxy-based regularization and show, using both qualitative and quantitative results, that this novel formulation aids in the creation of more generalizable representations. The proposed method achieves state-of-the-art results on all three standard image-text retrieval datasets.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[2] Srikar Appalaraju and Vineet Chaoji. Image similarity using deep cnn and curriculum learning. *arXiv preprint arXiv:1709.08761*, 2017.

[3] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. *CoRR*, abs/2011.04305, 2020.

[5] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15789–15798, June 2021.

[6] Tianlang Chen, Jiajun Deng, and Jiebo Luo. Adaptive offline quintuplet loss for image-text matching. In *European Conference on Computer Vision*, pages 549–565. Springer, 2020.

[7] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. *CoRR*, abs/2101.01368, 2021.

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[9] Xuri Ge, Fuhai Chen, Joemon M. Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. Structured multi-modal feature embedding and alignment for image-sentence retrieval. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[10] 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)*, volume 2, pages 1735–1742. IEEE, 2006.

[11] Zhong Ji, Kexin Chen, and Haoran Wang. Step-wise hierarchical alignment network for image-text matching. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 765–771. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *CoRR*, abs/2102.05918, 2021.

[13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, 2017.

[14] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3238–3247, 2020.

[15] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[16] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651, 2021.

[17] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching, 2019.

[18] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[19] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 3–11, New York, NY, USA, 2019. Association for Computing Machinery.

[20] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[21] 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)*, pages 5020–5029, Online, Aug. 2021. Association for Computational Linguistics.

[22] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[23] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 360–368, 2017.

[24] Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. SLIP: self-supervision meets language-image pre-training. *CoRR*, abs/2112.12750, 2021.

[25] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

[26] Juan-Manuel Perez-Rua, Valentin Vielzeuf, Stephane Pateux, Moez Baccouche, and Frederic Jurie. Mfas: Multimodal fusion architecture search. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6959–6968, 2019.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[29] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[30] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. *CoRR*, abs/1812.07119, 2018.

[31] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019.

[32] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. *CoRR*, abs/1907.09748, 2019.

[33] Jiwei Wei, Yang Yang, Xing Xu, Xiaofeng Zhu, and Heng Tao Shen. Universal weighting metric learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[34] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10938–10947, 2020.

[35] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017.

[36] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19163–19173, June 2022.

[37] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *CoRR*, abs/1708.01471, 2017.

[38] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15661–15670, June 2022.

[39] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15661–15670, June 2022.