# Emphasizing Complementary Samples for Non-literal Cross-modal Retrieval

Christopher Thomas*
Columbia University
New York, NY
christopher.thomas@columbia.edu

Adriana Kovashka
University of Pittsburgh
Pittsburgh, PA
kovashka@cs.pitt.edu

## Abstract

*Existing cross-modal retrieval methods assume a straightforward relationship where images and text contain portrayals or mentions of the same objects. In contrast, real-world image-text pairs (e.g. an image and its caption in a news article) often feature more complex relations. Importantly, not all image-text pairs have the same relationship: in some pairs, image and text may be more closely aligned, while others are more loosely aligned hence complementary. In order to ensure the model learns a semantically robust space which captures nuanced relationships, care must be taken that loosely-aligned image-text pairs have a strong enough impact on learning. In this paper, we propose a novel approach to prioritize loosely-aligned samples. Unlike prior sample weighting methods, ours relies on estimating to what extent* semantic similarity *is preserved in the separate channels (images/text) in the learned multimodal space. In particular, the image-text pair weights in the retrieval loss focus learning towards samples from* diverse *or* discrepant *neighborhoods: samples where images or text that were close in a semantic space, are distant in the cross-modal space (diversity), or where neighbor relations are asymmetric (discrepancy). Experiments on three challenging datasets exhibiting abstract image-text relations, as well as COCO, demonstrate significant performance gains compared to recent state-of-the-art models and sample weighting approaches.*

## 1. Introduction

We live in a multimodal world. Modern media use this multimodality to better convey stories: the same overall topic is expressed in text, images, video, audio, etc. However, the modalities (e.g. images and text) tell different *parts* of the story. For example, the text might describe the atrocities of Russia's attack on Ukraine, while the image *illustrates one aspect,* the suffering of a refugee child separated from his father. Given the enormity of multimodal data on the web, intelligent systems must reason across modalities. The basic step is constructing a shared semantic space such that images and text corresponding to the same concept neighbor each other. However, to model complex multimodal media, cross-modal methods must understand not just the *value* (close/not) but also the *nature of the relationship* between the co-occurring image and text.

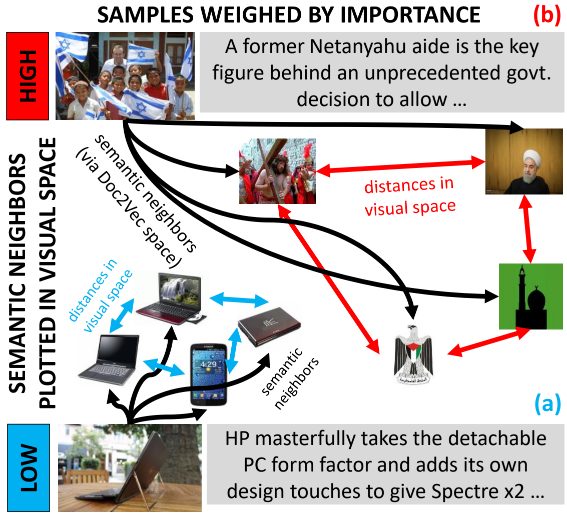

SAMPLES WEIGHED BY IMPORTANCE
(b)
HIGH
A former Netanyahu aide is the key figure behind an unprecedented govt. decision to allow ...
SEMANTIC NEIGHBORS PLOTTED IN VISUAL SPACE
semantic neighbors (via Doc2Vec space)
distances in visual space
distances in visual space
semantic neighbors
(a)
LOW
HP masterfully takes the detachable PC form factor and adds its own design touches to give Spectre x2 ...

Figure 1. Image-text pairs weighed by importance: we prioritize loosely-aligned image-text pairs. One of our metrics emphasizes images corresponding to similar texts (shown with black arrows) which are *not* similar in visual space (distances shown with red/blue arrows). The "Israeli flag" image-text pair will be prioritized because the semantic neighbors are highly diverse visually.

Most cross-modal retrieval approaches assume the alignment between image and text is literal, e.g. the image shows an airplane in the sky and the caption describes that airplane. This makes sense when the *purpose* of a caption is to provide the exact same content as the image, e.g. to serve a visually-impaired user. However, in real-world media (e.g. blog posts or news articles), the *type* of relationship between image and text in a pair will vary. *Some* image-text pairs

[*] Work done while at University of Pittsburgh

will have a direct relation, with objects in the image directly corresponding to words in text, as in Fig. 1 (a) where the caption describes the features of the laptop shown. However, in another pair, the image may illustrate aspects of the text, complement it by providing extra context, or figuratively underscore a textual point. For example, Fig. 1 (b) shows children holding Israeli flags to illustrate a report about a political event in Israel. Simple cases like (a) can be matched relatively easily and provide strong training signal, but samples with a less direct relation (b), may be more challenging and thus informative. Without proper care, the latter samples (Israeli flag) will be drowned out by the easier cases.

To address this problem, we dynamically weigh each image-text pair within a training batch. We model the relationship between images and text, as well as their surrounding samples, by measuring the extent to which the image and text modalities come from a diverse or discrepant neighborhood. We refer to *texts* that are close in text embedding space (and their co-occurring, paired *images*) as "semantic neighbors." We measure the *diversity* of each sample's semantic neighbors. Samples score high on diversity if they have (1) semantic neighbors with dissimilar visual embeddings in the *joint* cross-modal space, or (2) dissimilar textual embeddings in the *joint* space, for texts that are originally semantic neighbors. High diversity in the joint space for semantic neighbors could imply that the same semantic concept shares multiple visual expressions, i.e. these are samples where the relationship between image and text is abstract and they should be prioritized in cross-modal learning. We also measure the sample's neighborhood discrepancy by computing the distance of the sample to the *semantic neighbors of its semantic neighbors*. This tests for symmetry of the sample-to-neighbor relationship, both within and across modalities. If a sample is far from the neighbors-of-its-neighbors, this could indicate the sample is more likely to have multiple senses, much like samples with high diversity. We illustrate our two weighting cues in Fig. 2. We also propose a mechanism to combine the diversity and discrepancy metrics.

Our main contribution is a method of learning visual-semantic embeddings on challenging data consisting of abstract, loosely-aligned image-text pairs with complementary information in each modality. Our approach can be easily integrated into standard ranking losses. We perform detailed experimental analysis on three datasets that exhibit abstractness, namely GoodNews [4], Politics [28], and Conceptual Captions [25]. We also evaluate on a retrieval dataset with well-aligned images and captions, namely COCO [17]. We discover that diversity is very helpful for the abstract datasets, while discrepancy is more helpful for the literal COCO dataset. We outperform six recent, state-of-the-art approaches by a large margin. Importantly, four of these prior approaches are sample weighting strategies.

## 2. Related Work

**Connecting vision and language.** Most visual-semantic embedding (VSE) approaches learn a joint visual-text space where the distance between embedded samples reflects their semantic relationship [34]. Following the early deep VSE models [10, 20], research has focused on improving the learning objectives [13, 30], and explored different ways to fuse image and text representations [23, 37], including through cross-modal attention [22]. We use a traditional, well-understood two-stream visual semantic embedding framework trained via a ranking loss, following recent work [1, 7, 31, 39]. However, our contribution is agnostic to backbone model architecture.

**Retrieval losses.** The most commonly used loss for learning cross-modal embeddings is triplet loss [24], but others have also been proposed [5, 11]. Triplet loss can be challenging to train [15] due to the difficulty of choosing informative dissimilar samples. Many have exploited hard negative mining [9, 38], while others have tackled issues stemming from negative sample choice [36], e.g. by pushing multiple negatives away [26] or facilitating learning using *easy* negatives [35]. Other approaches [8, 40] rely on the use of classification labels or metadata, e.g. to ensure negatives in the triplet belong to different classes than the positive. Unlike these, our approach only uses the supervision of image-text co-occurrence. Our method exploits the structure of the image and text unimodal spaces by comparing their structure. [32] uses a related idea, but relies on the presence of five captions per image, which is not applicable for most datasets we consider.

Most related to our approach is [29] which uses two within-modality triplet losses to constrain texts which are semantic neighbors (those close in a pre-trained text embedding space) and their paired images, to be close in the joint space. [29] explicitly impose structural constraints on the space which may be too strict in some cases. In contrast, we emphasize samples without imposing structural conditions on the learned space. Moreover, our method does not consider all samples equally important, which allows the model to prioritize its efforts towards challenging samples. We significantly outperform this work in all settings, and show large gains on the Conceptual Captions dataset which has both closely- and weakly-aligned image-text pairs, as shown in [2].

**Sample weighting.** Rather than hard negative mining, some methods (including ours) use a soft sample weighting. Some methods, like [33], are general and can be applied to different settings including classification and retrieval. While we compare against [33], we primarily focus on and compare against sample weighting methods specific to cross-modal retrieval, such as [3, 18, 21]. In [21], positive samples which violate the margin but are still correctly retrieved are weighted less, while others incur a larger penalty. [18] use sample weights to address hubness (a phenomenon where a small number of embeddings remain undesirably close to many others), such that samples which are hubs receive more attention. Our weights are designed to improve the semantic properties of the learned space by emphasizing samples where the relation between image and text is abstract, not necessarily "hard" samples. This is an important distinction, because some "hard" samples in large-scale datasets are likely to be noisy; we found using hard negative mining prevented methods from training successfully on several of the challenging datasets we tested on. To isolate the effect of outliers, [3] estimate density by computing the correlation between samples from different modalities. In contrast, we look at the discrepancy between semantic and visual proximity, rather than density.

## 3. Approach

We propose two metrics to capture the degree of abstractness of an image-text pair. Both metrics rely on the sample's neighborhood, both within and across modalities. One metric, DISCREPANCY, relies on the relationship of a sample to its *neighbors-of-neighbors*. Because it is a sample-to-neighbor comparison, we consider this method a simpler, *first-order* technique. Another metric, DIVERSITY, considers the relation of the sample's neighbors to other sample neighbors, and we refer to it as a *second-order* technique. In our experiments, we discover that the simpler first-order technique (DISCREPANCY) is more appropriate for simpler datasets like COCO, while the second-order technique (DIVERSITY) works best for more abstract datasets.

### 3.1. Training objective

Let $\mathcal{D} = \{\mathbf{I}, \mathbf{T}\}$ represent a dataset of $n$ image-text pairs, where $\mathbf{I} = \{x_1, x_2, \ldots, x_n\}$ and $\mathbf{T} = \{y_1, y_2, \ldots, y_n\}$ represent the set of images and text, respectively, and $y_i$ is text co-occurring with image $x_i$ (the two are semantically related). We refer to $(x_i, y_i)$ as positive pairs and either $(x_i, y_{j \neq i})$ or $(x_{j \neq i}, y_i)$ as negative pairs. In order to compare and retrieve across modalities, we seek a common manifold $\mathcal{M}$. A convolutional network $f : \mathbf{I} \rightarrow \mathcal{M}$ is used to project images into the joint space, while a recurrent network $g : \mathbf{T} \rightarrow \mathcal{M}$ projects text. We use the notational shorthand $f(x) = x \in \mathbb{R}^{K \times H}$ and $g(y) = y \in \mathbb{R}^{K \times H}$, where $K$ is the number of embeddings per sample and $H$ is the dimension of the learned manifold. Most prior methods assume $K = 1$ but this may be too stringent when image and text have multiple meanings. Recently [27] propose a polysemous embedding model (PVSE), where every image and text are represented by $K$ embeddings encouraged to be diverse; we adopt this formulation for all methods compared. When comparing two samples, we use the maximum cosine similarity across all $K^2$ pairs: $s(x_i, y_i) = \max_{(k_1,k_2) \in \{1,\ldots,K\} \times \{1,\ldots,K\}} \left\langle \frac{x_{i_{k_1}}}{\|x_{i_{k_1}}\|_2}, \frac{y_{i_{k_2}}}{\|y_{i_{k_2}}\|_2} \right\rangle : \mathbb{R}^{K \times H} \times \mathbb{R}^{K \times H} \rightarrow \mathbb{R}$. For notational simplicity, we omit the reference to the $K$ embeddings in the remaining text.

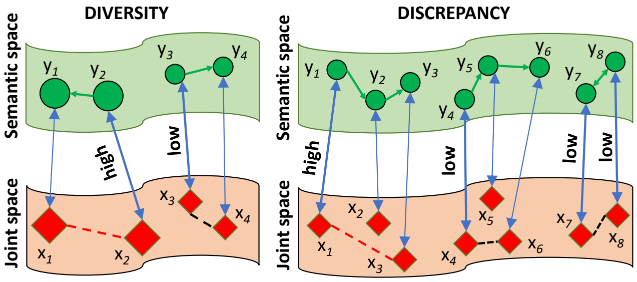


Figure 2. Our two weighting cues. The top shows neighbors in the original semantic space, with green arrows connecting the query to the nearest neighbor. Blue links show co-occurring images and text; thicker links show pairs whose scores we wish to compute. Red links connect images which are semantic neighbors, but their visual embeddings are far in the joint cross-modal space (high *diversity*). Images corresponding to neighbors-of-neighbors in text space, which are far in the joint space, have high *discrepancy*. We use diversity and discrepancy weights to compute the $\boldsymbol{\alpha}$ scores in Eq. 1; this figure shows computation of $\boldsymbol{\alpha_X}$ only.

We assume the same pairwise ranking objective (triplet loss) as other recent VSE methods [3, 9, 18, 27], but introduce a weighting constraint to emphasize semantically informative samples. We optimize a sample-weighted bidirectional n-pairs [26] triplet loss $\mathcal{L}_{\text{RANK}}$ given by:

$$\mathcal{L}_{\text{RANK}} = \frac{1}{2N^2} \Bigg( \sum_{x_i \in I_B} \sum_{y_j \in T_B} \alpha_i [\, s(x_i, y_{j \neq i}) - s(x_i, y_i) + m \,]_+ + \sum_{y_i \in T_B} \sum_{x_j \in I_B} \alpha_i [\, s(x_{j \neq i}, y_i) - s(x_i, y_i) + m \,]_+ \Bigg) \quad (1)$$

where $m$ is the margin, $[\cdot]_+ = \max(0, \cdot)$ and $I_B, T_B$ are images and text, respectively, within a minibatch of samples. We introduce a per-positive-sample weight $\alpha_i$, given by our method. All methods and baselines (except where noted) use this loss to train, but vary in how $\alpha_i$ are computed.

**Limitations of hard negative mining:** Traditional VSE methods give all samples equal weight within a minibatch. To facilitate learning, most recent methods [9, 27, 32] also perform hard negative mining, where only the most challenging negative sample is used (e.g. $\max_j s(x_i, y_{j \neq i})$). While this makes sense in common captioning datasets with strong, literal image-text alignment, we found it prevented models from successfully training on more challenging datasets. When using hard negatives, the problem becomes too hard since many negative image-text matches are plausible, even if technically incorrect. Moreover, relying only on hard negatives makes the model more vulnerable to noise, which is present within the webly-harvested datasets we consider. [33] propose a compromise solution, where negatives are sampled with probability inversely proportional to their distance from the anchor (i.e. hard negatives are more likely, but other samples are also not neglected). We found this approach did not yield competitive results on the datasets we tested. [21] proposes a soft (weighted) semi-hard negative mining approach to enable learning, which we outperform.

### 3.2. Measuring semantic neighborhood discrepancy

In order to emphasize informative, weakly-aligned image-text samples, we first must detect and weight them. We first consider the relationship of the query sample to its neighborhood: a first-order metric. We discover each image-text pair's *semantic neighbors* in text, $\Omega(\mathbf{T})$, following [29]'s implementation. We compute neighbors in text space because the text domain provides the cleanest semantic representation of the image-text pair. Let $\Psi\left(\Omega(y_i)\right) = \left\{\left\langle x'_{i_n}, y'_{i_n}\right\rangle\right\}_{n=1}^{N}$ represent the semantic nearest neighbor function over $\Omega(\mathbf{T})$, where $\left\{\left\langle x'_{i_n}, y'_{i_n}\right\rangle\right\}_{n=1}^{N}$ denotes the set of the $N$ neighbors of $\langle x_i, y_i\rangle$ and $\langle x_i, y_i\rangle \notin \Psi\left(\Omega(y_i)\right)$. Note that *semantically* neighboring images $\left\{x'_{i_n}\right\}^{N}$ are not necessarily *visual* neighbors of $x_i$.

Because our formulation is equivalent for both image/text neighbors, we let $s_i$ represent a sample from either domain but require samples $s_i$ and $s_j$ come from the same domain. We examine the relation: $s_i \in \Psi\left(\Omega(\Psi\left(\Omega(s_i)\right))\right)$, i.e. whether a sample is a semantic neighbor of its semantic neighbors. For images, this amounts to using the ground-truth text paired with the image. This criterion quantifies whether the surrounding space is compact or grid-like, which would result in a symmetric-like neighborhood relation, and high similarity of neighbors in the joint space. See Fig. 2 (right). Images/texts with high discrepancy may have multiple meanings or may be used figuratively.

Formally, let $\Psi\left(\Omega(\Psi\left(\Omega(s_i)\right))\right) = \left\{s''_{i_n}\right\}_{n=1}^{N^2}$ represent the set of the semantic neighbors of $s_i$'s semantic neighbors. Let $\mathbf{s}''_i = [s''_{i_1}, s''_{i_2}, \dots, s''_{i_{N^2}}]^\intercal$ denote the matrix of size $N^2 \times H$ of the embeddings of the neighbors of neighbors, and $\mathbf{V} = \mathbf{s}''_i s_i$ is the matrix-vector product of the sample's neighborhood and the sample (size $N^2 \times 1$). We use the $f : \mathbf{I} \rightarrow \mathcal{M}$ and $g : \mathbf{T} \rightarrow \mathcal{M}$ projections of image and text into the joint space. Then, the *semantic discrepancy score* $\Upsilon_i^{DIS}$ and corresponding scaled score $\alpha_i^{DIS}$ of $s_i$ is:

$$\Upsilon_i^{DIS} = \Gamma^{DIS} \times \frac{1}{N^2} \sum_{r=1}^{N^2} \mathbf{V}_{(r)} \quad (2) \qquad \alpha_i^{DIS} = \lambda \times \frac{e^{\Upsilon_i^{DIS}}}{\sum_{j=1}^{B} e^{\Upsilon_j^{DIS}}} \quad (3)$$

where $r$ indices $\mathbf{V}$'s entries, $B$ is the minibatch size, $\lambda$ is a scaling constant (see implementation details), and $\Gamma^{DIS} \in \{1, -1\}$ is a switching parameter, which controls whether more weight is given to more (1) or less ($-1$) similar samples, respectively (we show $\Gamma^{DIS} = -1$ is superior in experiments). The final attention vector is given by stacking sample weights: $\boldsymbol{\alpha^{DIS}} = [\alpha_1^{DIS}, \alpha_2^{DIS}, \dots, \alpha_B^{DIS}]$.

We compute $\boldsymbol{\alpha^{DIS}}$ for the image and text domains separately (i.e. $s_i = x_i$ or $s_i = y_i$), then combine the two vectors by addition: $\boldsymbol{\alpha^{DIS}} = \text{softmax}\left(\boldsymbol{\alpha_X^{DIS}} + \boldsymbol{\alpha_Y^{DIS}}\right)$, or by taking their absolute difference: $\boldsymbol{\alpha^{DIS}} = \text{softmax}\left(\left|\boldsymbol{\alpha_X^{DIS}} - \boldsymbol{\alpha_Y^{DIS}}\right|\right)$; we show the latter is superior in Tab. 3. The weights can now be directly used in Eq. 1 to weight samples by the semantic discrepancy measure.

In early experiments, we also tested combination via multiplication, or taking the larger (max) score; both performed worse than summation. We also experimented with measuring the relation between a sample and its direct neighbors, not neighbors-of-neighbors, and obtained inferior results. The cost for computing neighborhoods is small because we dynamically cache sample embeddings into a memory bank while training. Finding semantic neighbors occurs once using a pre-trained text embedding model. Thus, computing $\boldsymbol{\alpha}$ weights is efficient as it only requires multiplication.

### 3.3. Measuring semantic neighborhood diversity

We next present our second-order method. We observe that semantic concepts with non-literal portrayals are likely to be visually diverse. For example, a piece of text about "patriotism" could be paired with an image of a flag, an eagle, army servicepeople, a crowd of protesters, etc. In contrast, images paired with the caption "a bowl of apples on a table" would likely be much more visually similar.

We measure the diversity of the semantic neighbors in both the image and text domains, Let $\mathbf{s}'_i = [s'_{i_1}, s'_{i_2}, \dots, s'_{i_N}]^\intercal$ denote the $N \times H$ matrix of embeddings of the neighbors of $s_i$ found via $\Psi$, and $\mathbf{U} = \mathbf{s}'_i \mathbf{s}'^{\intercal}_i$ compute the pairwise similarities between all semantic neighbors through cross-product. We compute the *semantic diversity score* $\Upsilon_i^{DIV}$ for $s_i$ as follows:

$$\Upsilon_i^{DIV} = \Gamma^{DIV} \times \frac{1}{N^2} \sum_{r=1}^{N} \sum_{c=1}^{N} \mathbf{U}_{(r,c)} \tag{4}$$

where $r, c$ index over the rows and columns of $\mathbf{U} = \mathbf{s}'_i \mathbf{s}'^{\intercal}_i$ and $\Gamma^{DIV} \in \{1, -1\}$. We finally enforce that all $\Upsilon_i^{DIV}$ in a minibatch form an attention-like vector $\boldsymbol{\alpha^{DIV}}$ as follows:

$$\boldsymbol{\alpha^{DIV}} = \left[\alpha_1^{DIV}, \alpha_2^{DIV}, \dots, \alpha_B^{DIV}\right], \tag{5}$$

$$\alpha_i^{DIV} = \lambda \times \frac{e^{\Upsilon_i^{DIV}}}{\sum_{j=1}^{B} e^{\Upsilon_j^{DIV}}} \tag{6}$$

### 3.4. Combination methods

We explore two methods for combining our two proposed weighting strategies. The first combination approach is standard, and can be expressed as follows:

$$\alpha_i^{COMB-VAL} = \lambda \times \text{softmax}\left(\hat{\beta} * \Upsilon_i^{DIV} + \hat{\gamma} * \Upsilon_i^{DIS}\right) \tag{7}$$

where the combination $\{\hat{\beta}, \hat{\gamma}\}$ is chosen on a validation set, with $\{\beta, \gamma\} \in \{\{1,1\}, \{1,2\}, \{2,1\}, \{1,3\}, \{3,1\}, \{1,4\}, \{4,1\}, \{1,5\}, \{5,1\}\}$.

The second combination is more interesting as it explores statistics about the distribution of $\alpha^{DIV}$ and $\alpha^{DIS}$ scores per dataset. To combine the scores for a particular sample, we use the mean and standard deviation of the distribution of the corresponding score over the whole dataset:

$$\alpha_i^{COMB-STAT} = \lambda \times \text{softmax}\Big(\mu_{DIV} * \sigma_{DIV} * \Upsilon_i^{DIV} + \mu_{DIS} * \sigma_{DIS} * \Upsilon_i^{DIS}\Big) \tag{8}$$

where $\mu = mean(\alpha_i, \ldots, \alpha_n)$ and $\sigma = stdev(\alpha_i, \ldots, \alpha_n)$.

The intuition for the second approach is that the contribution of each measure should be proportional to the mean diversity/discrepancy in the dataset. The more challenging the dataset according to a metric, the larger the contribution of this metric, for an individual sample's relative importance in the cross-modal retrieval loss. We also weigh samples proportional to standard deviation according to the given metric (diversity or discrepancy); larger dataset fluctuations (stdev) according to the metric indicates this metric is more discriminative of the challenge of cross-modal retrieval for this particular pair. This approach is less computationally intensive as it does not require a sweep for $\{\hat{\beta}, \hat{\gamma}\}$. We show the mean/stdev statistics per dataset in Table 1, where we observe the means for GoodNews and Politics are larger than for COCO.

Note that while we present combination methods as an intuitive extension for our two single-metric methods, arriving at a good combination entails significant engineering (e.g. $\beta, \gamma$ tuning) and is *not our focus*.

### 3.5. Implementation details

All methods use ResNet-50 [14] initialized with ImageNet features for images, and Gated Recurrent Units [6] for text, with hidden state size 512. All layers train in these networks. All methods and baselines are built on top of PVSE [27]; we select $K$ per dataset on the validation set, with optimal values shown in Tab. 1. The margin $m$ is set to 0.1. Images are scaled to 224x224 and augmented with random horizontal flipping. We use Xavier initialization [12] on all non-pretrained learnable weights. GRUs are initialized with 200D word embeddings learned on the dataset on which they are applied. We perform $L_2$ normalization on embeddings produced by each model. We use Adam [16] with minibatch size of 32, learning rate 1.0e-4 (decayed by a factor of 10 after every 5 epochs of no decrease in val loss), and weight decay 1e-5. We use a train-val-test split of 80-10-10 for all datasets. We use [29]'s implementation for computing semantic neighborhoods on text embeddings, which uses [19] to efficiently compute approximate nearest neighbors for $\Psi$; we use $N = 200$ neighbors. We probabilistically sample at most 1000 neighbors at a time from $s_i''$ in Eq. 2. We cache embeddings from the prior epoch for efficient computation of Eqs. 2 and 4. We use the validation set to pick $\lambda$ per dataset; because of the summation in Eq. 1, $\lambda$ should be at least equal to the batch size, but larger values could bring improvements by increasing the impact of the triplet loss in PVSE's [27] multitask loss. We show the impact of $\Gamma = \pm 1$ for all methods in Tab. 3.

| | COCO | GoodNews | Politics | ConcCap |
|---|---|---|---|---|
| $K$ (PVSE) | 3 | 5 | 9 | 7 |
| $\lambda$ | 196 | 160 | 96 | 196 |
| $\hat{\beta}$ | 1 | 4 | 3 | 3 |
| $\hat{\gamma}$ | 4 | 1 | 1 | 1 |
| $\mu_{DIV} * \sigma_{DIV}$ | 0.04 | 0.01 | 0.01 | 0.01 |
| $\mu_{DIS} * \sigma_{DIS}$ | 0.03 | 0.03 | 0.06 | 0.01 |
| $\mu_{DIV}$ | 0.3173 | 0.3979 | 0.6279 | 0.1569 |
| $\sigma_{DIV}$ | 0.1156 | 0.0136 | 0.0123 | 0.0753 |
| $\mu_{DISC}$ | 0.2705 | 0.3986 | 0.6309 | 0.1182 |
| $\sigma_{DISC}$ | 0.1068 | 0.0676 | 0.0885 | 0.0742 |
| $\beta = 5, \gamma = 1$ | 0.4922 | 0.7842 | 0.4734 | 0.6649 |
| $\beta = 1, \gamma = 5$ | 0.4850 | 0.7803 | 0.4722 | 0.6406 |

Table 1. Hyperparameters chosen or computed, and (last two rows) extreme settings evaluated for OURS-COMBINED-VAL on a small subset of the training set, avg over I→T, T→I.

## 4. Experimental Validation

**Baselines.** We compare our weighting strategies, OURS-DIVERSITY, OURS-DISCREPANCY, and their combinations, to five very recent methods. We chose these methods because they are either appropriate for more abstract image-text matching, or because they also compute *weights for samples*. Thus, we explicitly test the quality of the sample weights that our method computes.

- PVSE [27] is a recent cross-modal retrieval method with multi-head self-attention, which computes multiple embeddings to account for polysemy.
- THOMAS [29] uses semantic neighborhoods to compute within-modality losses.
- HAL [18] is a sample weighting cross-modal retrieval method which up-weighs samples likely to be the closest sample to multiple queries.
- MITHUN [21] weighs samples based on hardness (using ranks of matching images/text, larger values denoting worse match hence more challenging sample).

| | **COCO** | | **GoodNews** | | **Politics** | | **ConcCap** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | **I→T** | **T→I** | **I→T** | **T→I** | **I→T** | **T→I** | **I→T** | **T→I** | **Avg** |
| PVSE [27] | 0.6541 | 0.6561 | 0.8516 | 0.8526 | 0.5919 | 0.6057 | 0.7138 | 0.7168 | 0.7053 |
| Thomas [29] | 0.6552 | 0.6494 | 0.8637 | 0.8667 | 0.6184 | 0.6187 | 0.7379 | 0.7473 | 0.7197 |
| HAL [18] | 0.6665 | 0.6845 | 0.8623 | 0.8579 | 0.5919 | 0.5903 | 0.7638 | 0.7685 | 0.7232 |
| Mithun [21] | 0.6967 | 0.6950 | 0.8439 | 0.8463 | 0.5792 | 0.5839 | 0.7523 | 0.7497 | 0.7184 |
| Amrani [3] | 0.6746 | 0.6756 | 0.8629 | 0.8678 | 0.6117 | 0.6117 | 0.7376 | 0.7356 | 0.7222 |
| Ours-Discrepancy | 0.6851 | 0.6844 | 0.8751 | 0.8766 | 0.6211 | 0.6228 | 0.7548 | 0.7588 | 0.7348 |
| Ours-Diversity | 0.6729 | 0.6718 | **0.8774** | **0.8803** | 0.6268 | **0.6366** | **0.7767** | **0.7719** | 0.7393 |
| Ours-Combined-Val | **0.7056** | **0.7013** | 0.8750 | 0.8780 | **0.6285** | **0.6291** | **0.7715** | **0.7723** | **0.7452** |
| Ours-Combined-Stats | **0.7062** | **0.7007** | **0.8791** | **0.8782** | **0.6274** | 0.6239 | 0.7714 | 0.7713 | **0.7448** |

Table 2. Retrieval results (top-1 accuracy) for image to text (**I→T**) and text to image (**T→I**). The best two methods per task are shown in **bold**. Our method's gains over the baselines are large: the stddev across baselines' Avg is 0.7%, while our methods' gains over the strongest baseline's Avg are 1.6%, 1.2%, 2.2% and 2.2%. The better of our single-metric methods is underlined.

- Amrani [3] up-weighs samples where both image/text in a sample belong to tight clusters.

**Datasets.** We demonstrate our approach on four large-scale recent datasets. *ConcCap* [25] contains ∼3.3M images with alt-text descriptions for the web. *GoodNews* [4] contains ∼466k images and captions from the New York Times. *Politics* [28] contains ∼246k pairs of images with sentences from news articles. *COCO* [17] contains ∼120k images with captions. While COCO and Flickr30K are among the most popular retrieval datasets, both contain descriptive captions which heavily overlap with the image. We use ConcCap, GoodNews and Politics in place of Flickr30K, to demonstrate the challenge of retrieval when there is weak alignment of the image and text. We forego other datasets, e.g. Wikipedia and XMediaNet, since they are very small (3k/36k samples).

**Metrics and training losses.** We evaluate Top-1 accuracy on image to text, and text to image matching. Following [29], because image-text alignment in GoodNews and Politics are very challenging, we use a 5-way multiple-choice task (1 correct and 4 incorrect options), rather than rank or recall, to ensure differences between methods are more visible; with other metrics, all results are low due to the challenge of the task. A further challenge with traditional Recall@k over the entire test set is that abstract retrieval is subjective, i.e. should a method be penalized since it ranks an image with a scale of justice higher than Lady Justice for text about justice? The smaller retrieval tasks reduce the likelihood subjective samples are in each task. For consistency, we also use a $c$-way task for ConcCap and COCO, but $c = 20$ for ConcCap and $c = 100$ for COCO, commensurate with the challenge of retrieval in these datasets. *We did verify that methods' relative performance is the same for Top-1, Recall@3, and Rank (please see supplementary file).* We use a triplet loss as the main loss for cross-modal retrieval, in all methods. This is the loss these methods originally used, except for [29] which showed results with both triplet and angular loss (but the contribution was *not* angular loss). For fair comparison, since most baselines use triplet loss, we use triplet for this method as well, and include angular results in our supplementary file. For most methods [18, 21, 27, 29] we used the original authors' code to compute performance.

## 4.1. Main result

**Our methods outperform prior art:** At the top of Table 2 are the five state-of-the-art methods. At the bottom are our two single-metric weighting techniques and their combinations. The best performers per dataset/task are always among our methods. Recall HAL, Mithun and Amrani are sample weighting methods, and **each of *our* sample weighting methods outperform them** on average, and for most datasets individually. The biggest gains Ours-Combined-Val achieves are: 6% on ConcCap and 5% on COCO over PVSE, 5% on COCO over Thomas, and 5% on Politics over Mithun. The largest gains of Ours-Diversity are: 6% on ConcCap over PVSE, 5% on Politics over HAL and Mithun, and 4% on ConcCap over Thomas and Amrani.

We also trained models using [33]'s distance-weighted sampling method with triplet loss. [33] achieves 0.7972 (I→T) and 0.7964 (T→I) on GoodNews, but performed significantly worse than PVSE on other datasets (e.g. ∼0.65 on ConcCap). One possible reason is only one of the $k$ embeddings (the closest to the negative) receives training signal in [27]'s model. An alternative approach would be to dynamically select the negative for each of the $k$ embeddings, but this is a non-trivial extension of [33] and we leave it as future work.

**Discrepancy for literal, diversity for abstract data:** Among single-strategy methods, Ours-Diversity is stronger on average. As we alluded to in the Approach section, different methods are appropriate for the more literal dataset (COCO) vs the more abstract datasets. In particular, Ours-Discrepancy is the stronger method for COCO, but Ours-Diversity is stronger for GoodNews, Politics, and ConcCap. For COCO, the notion of neighborhood is very precise and dominated by object presence. Thus, what we wish to capture is how much the captions for similar images differ; it is sufficient to capture the relation between a sample and its $N$ neighbors or neighbors-of-neighbors (as Ours-Discrepancy does). In contrast, for all three other datasets which are the focus of our study, images and captions are much more diverse, thus examining the relations between all $N^2$ pairs of neighbors (as Ours-Diversity does) is useful. We include a detailed discussion of the motivation of our measures in our supplementary material, as well as the weight distribution produced by each measure per dataset. We observe performance correlations with the weight distributions: small variance for Diversity correlates with strong performance, while the opposite holds for Discrepancy.

**Combining discrepancy and diversity boosts results:** Our strongest method on average is Ours-Combined-Val closely followed by Ours-Combined-Stats. Both combinations outperform both single-metric methods on COCO, which is the literal dataset where Ours-Discrepancy outperforms Ours-Diversity. The combinations outperform the stronger single-metric method (Ours-Diversity) in several settings on the other datasets: Ours-Combined-Val outperforms Ours-Diversity on 2 of the remaining 6 tasks (Politics I→T, ConcCap T→I), and Ours-Combined-Stats outperforms Ours-Diversity on GoodNews I→T, Politics I→T. The weakest performance is on Conceptual Captions, because Ours-Discrepancy performs much worse on that dataset than Ours-Diversity. This can also be observed from Tab. 1 where we show the difference between the $\beta = 5, \gamma = 1$ and $\beta = 1, \gamma = 5$ settings for the combination weights. This difference is largest for ConcCap, consistent with the relative performance of our two methods in Tab. 2. Thus, while for COCO, GoodNews and Politics combinations are useful, for ConcCap, using Ours-Diversity is optimal. We found discrepancy and diversity were very slightly correlated ($\rho = 0.0520$ Spearman's rank), thus the two metrics are complementary.

**Difference between combination methods:** We present the scalars $\hat{\beta}$ and $\hat{\gamma}$ chosen to combine the diversity and discrepancy metrics (Eq. 7) in Table 1. These are generally similar for GoodNews, Politics and ConcCap, but different for COCO (larger contribution of discrepancy). Note that standard deviation across the possible $\hat{\beta}, \hat{\gamma}$ settings was 0.7-1.3%, i.e. smaller than the gain between most of our methods and the strongest baseline. We also show the means and standard deviations of the cosine similarity scores used to compute diversity and discrepancy scores, in Tab. 1. We observe that although the relative magnitude of the calculated $\mu * \sigma$ statistics are different than the swept $\beta/\gamma$ hyperparameters, the weighting using the statistics works comparably well. One possible cause is that we swept for $\beta/\gamma$ by training on a subset of the train set for computational reasons. However, $\mu * \sigma$ are easily calculated on the *full* train set. We suspect $\beta/\gamma$ may perform better if the full train set were used to sweep, at the expense of computational overhead.

| | Diversity | | Discrepancy | |
|---|---|---|---|---|
| **Method** | **I→T** | **T→I** | **I→T** | **T→I** |
| Random $\alpha$ | 0.5777 | 0.5780 | 0.5777 | 0.5780 |
| Ablation for $\Gamma$ | | | | |
| $\Gamma = -1$ | **0.6268** | **0.6366** | **0.6211** | **0.6228** |
| $\Gamma = 0$ | 0.6046 | 0.6020 | 0.6046 | 0.6020 |
| $\Gamma = +1$ | 0.6206 | 0.6226 | 0.6158 | 0.6187 |
| Ablation for $\alpha$ combinations from image and text | | | | |
| $\alpha_{\mathbf{X}} + \alpha_{\mathbf{Y}}$ | 0.6188 | 0.6251 | 0.6130 | 0.6184 |
| $\lvert\alpha_{\mathbf{X}} - \alpha_{\mathbf{Y}}\rvert$ | **0.6268** | **0.6366** | **0.6211** | **0.6228** |
| $\alpha_{\mathbf{X}}$ | 0.6155 | 0.6192 | 0.6058 | 0.6040 |
| $\alpha_{\mathbf{Y}}$ | 0.6059 | 0.6113 | 0.6032 | 0.6006 |

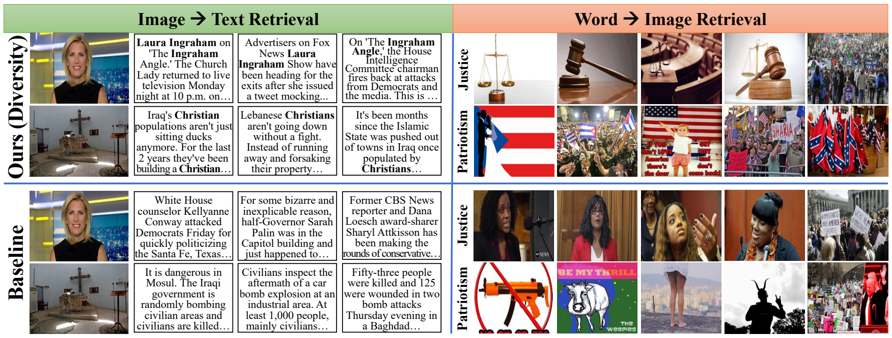
Table 3. Ablation on Politics [28]. The first group show results for $\Gamma = -1/0/+1$. The second group shows strategies for combining the image/text weight vectors (summing vs. absolute difference). Random and $\Gamma = 0$ are method-agnostic so are the same for each.

## 4.2. Ablation results

Next, we verify the contribution of our methods' components, using the **Politics** dataset as it contains some of the most abstract image-text relations. We show results in Table 3.

**Gamma:** We first motivate the choice of directionality for our weighting mechanisms. For Ours-Diversity and Ours-Discrepancy, the weighting could be implemented with the opposite sign (via $\Gamma$), e.g. we could prioritize samples that come from homogeneous rather than diverse regions. From Table 3 (top block), we see that emphasizing samples with low homogeneity (high "diversity"), and asymmetric neighborhoods (high "discrepancy") perform better. The differences between $\Gamma = \pm 1$ on Diversity are similar to the standard deviation over the baselines. We also trained a model which used all equal weights ($\Gamma = 0$, still scaled by $\lambda$) and found it performed even worse than the suboptimal $\Gamma$. Purely random weights performed worst. Using $\Gamma$ opposite to the optimal setting still has benefit over uniform weights since it allows the model to focus; even focusing on easy samples is better than no focus [35].

**Combining $\alpha$ scores:** In the second block, we explore how to combine the $\boldsymbol{\alpha_X}$ and $\boldsymbol{\alpha_Y}$ scores. We observe taking a difference between the two modalities is better, so weights are larger for samples whose measures differ more across modalities. This underscores the emphasis on prioritizing



**Laura Ingraham** on 'The **Ingraham** Angle.' The Church Lady returned to live television Monday night at 10 p.m. on…

Advertisers on Fox News **Laura Ingraham** Show have been heading for the exits after she issued a tweet mocking...

On 'The **Ingraham Angle**,' the House Intelligence Committee chairman fires back at attacks from Democrats and the media. This is …

Iraq's **Christian** populations aren't just sitting ducks anymore. For the last 2 years they've been building a **Christian**…

Lebanese **Christians** aren't going down without a fight. Instead of running away and forsaking their property…

It's been months since the Islamic State was pushed out of towns in Iraq once populated by **Christians**…

White House counselor Kellyanne Conway attacked Democrats Friday for quickly politicizing the Santa Fe, Texas…

For some bizarre and inexplicable reason, half-Governor Sarah Palin was in the Capitol building and just happened to…

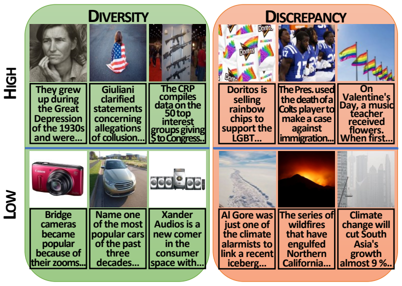
Former CBS News reporter and Dana Loesch award-sharer Sharyl Attkisson has been making the rounds of conservative…

It is dangerous in Mosul. The Iraqi government is randomly bombing civilian areas and civilians are killed…

Civilians inspect the aftermath of a car bomb explosion at an industrial area. At least 1,000 people, mainly civilians…

Fifty-three people were killed and 125 were wounded in two bomb attacks Thursday evening in a Baghdad…

Figure 3. Retrieval results for Ours-Diversity and Amrani on Politics.



They grew up during the Great Depression of the 1930s and were…

Giuliani clarified statements concerning allegations of collusion…

The CRP compiles data on the 50 top interest groups giving $ to Congress…

Doritos is selling rainbow chips to support the LGBT…

The Pres. used the death of a Colts player to make a case against immigration…

On Valentine's Day, a music teacher received flowers. When first…

Bridge cameras became popular because of their zooms…

Name one of the most popular cars of the past three decades…

Xander Audios is a new comer in the consumer space with…

Al Gore was just one of the climate alarmists to link a recent iceberg…

The series of wildfires that have engulfed Northern California…

Climate change will cut South Asia's growth almost 9 %..

Figure 4. Example image-text pairs from Politics receiving highest/lowest weights by our measures.

image-text pairs where image and text are different: samples differing across modalities suggest a complementary, rather than overlapping cross-modal alignment, while emphasizing uniformity or overlap (via sum) performs worse. Note that using $\alpha_{\mathbf{X}}$ and $\alpha_{\mathbf{Y}}$ individually performed much worse.

$\lambda$**:** The stdev over settings of scaling parameter $\lambda$ was 0.5%-1.5%, much smaller than our gains over baselines.

### 4.3. Qualitative results

**Weighted samples:** In Fig. 4 we show samples receiving the highest or lowest weights. For diversity, high-scoring samples concern abstract subjects in which image and text are complementary (sad woman-"Great Depression", American flag-"collusion"), while low-scoring ones are more concrete. For discrepancy, we observe cases where the image-text pairing is more atypical (e.g. football players-"immigration", pride flags-"Valentine's day" and "flowers"), while low-scoring ones are more literal (iceberg-"iceberg", fire-"wildfires").

**Retrievals:** In Fig. 3 we show retrievals using Ours-Diversity vs. Amrani on Politics [28]. We bold words in the text that highly align with the image. For image to text, our method correctly retrieves texts mentioning "Laura Ingraham" for the first image, while the baseline retrieves text mentioning women which aren't shown. For the second image, both methods retrieve text about the Middle East, but ours retrieved text mentioning Christians (which aligns with the cross in the image). For word to image, our method performs much better for abstract concepts like "justice" (ours retrieves gavels, balances, and protests, while the baseline retrieves people related to specific court cases). For "patriotism", ours retrieves flags and protests, while the baseline retrieves largely irrelevant images.

## 5. Broader Impacts

Sophisticated cross-modal retrieval techniques have a variety of applications, including news curation and image captioning (beyond the literal, descriptive level) for visually impaired readers. More broadly, understanding the intricate relationship between images and text has ramifications for understanding persuasion, as well as bias, in multimodal news media sources. In particular, if a system can understand that the image included with a particular text actually contradicts the surface meaning of the text, it may detect cases of irony and mockery, and thus, detect hateful use of conventional or social media. How such detections are used is a matter of policy and not the subject of this paper. Instead, the goal of our work is to enable better, more nuanced, modeling of image-text relationships. Eventually, we hope methods like ours will help build more socially aware systems. We believe that in order to ensure AI systems do social good rather than harm, they need to understand subtleties, and our method is a step in this direction.

## 6. Conclusion

Each modality in real-world data often exhibits complementarity (the degree to which the image and text complement one another), yet most methods assume a parallel alignment. We thus proposed a method for emphasizing cross-modal samples containing abstract, non-literal relationships which relies on two measures of cross-modal complementarity. Diversity emphasizes samples whose neighbors, in image or text space, are diverse in their semantics. Discrepancy computes the distance of a sample to the semantic neighbors of its semantic neighbors (emphasizing samples that could have multiple senses). We perform experiments on three large-scale datasets containing challenging image-text relations, as well as a standard cross-modal retrieval benchmark. Our experiments demonstrate that our method yields substantial performance gains compared to numerous baselines.

**Acknowledgements:** This work was supported by gifts from Google, Amazon and Adobe.

## References

[1] Juan Leon Alcazar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbelaez, and Bernard Ghanem. Apes: Audiovisual person search in untrimmed video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1720–1729, 2021. 2

[2] Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, July 2020. 2

[3] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6644–6652, 2021. 3, 6

[4] Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12466–12475, 2019. 2, 6

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2017. 2

[6] Kyunghyun Cho, B van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*, 2014. 5

[7] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021. 2

[8] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903, 2018. 2

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improved visual-semantic embeddings. In *British Machine Vision Conference (BMVC)*, 2018. 2, 3

[10] 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 *Advances in neural information processing systems*, pages 2121–2129, 2013. 2

[11] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–285, 2018. 2

[12] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010. 5

[13] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7181–7189, 2018. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[15] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1945–1954, 2018. 2

[16] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 5

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6

[18] Fangyu Liu, Rongtian Ye, Xun Wang, and Shuaipeng Li. Hal: Improved text-image matching by mitigating visual semantic hubs. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020. 3, 5, 6

[19] Yury A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2016. 5

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 2

[21] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018. 3, 4, 5, 6

[22] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017. 2

[23] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096, 2018. 2

[24] 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. 2

[25] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2, 6

[26] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1857–1865, 2016. 2, 3

[27] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019. 3, 5, 6

[28] Christopher Thomas and Adriana Kovashka. Predicting the politics of an image using webly supervised data. *Advances in Neural Information Processing Systems*, 2019. 2, 6, 7, 8

[29] Christopher Thomas and Adriana Kovashka. Preserving semantic neighborhoods for robust cross-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 4, 5, 6

[30] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 154–162, 2017. 2

[31] Kai Wang, Luis Herranz, and Joost van de Weijer. Continual learning in cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3628–3638, 2021. 2

[32] 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*, pages 5005–5013, 2016. 2, 3

[33] 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. 2, 4, 6

[34] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6609–6618, 2019. 2

[35] Hong Xuan, Abby Stylianou, and Robert Pless. Improved embeddings with easy positive triplet mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2, 7

[36] Baosheng Yu, Tongliang Liu, Mingming Gong, Changxing Ding, and Dacheng Tao. Correcting the triplet selection bias for triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–87, 2018. 2

[37] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4709–4717, 2017. 2

[38] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang. Hard-aware deeply cascaded embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 814–823, 2017. 2

[39] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3536–3545, 2020. 2

[40] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019. 2