Preserving Semantic Neighborhoods for Robust Cross-modal Retrieval

Christopher Thomas[0000−0002−3226−396X] and Adriana Kovashka[0000−0003−1901−9660]

University of Pittsburgh, Pittsburgh PA 15260, USA
{chris, kovashka}@cs.pitt.edu

Abstract. The abundance of multimodal data (e.g. social media posts) has inspired interest in cross-modal retrieval methods. Popular approaches rely on a variety of metric learning losses, which prescribe what the proximity of image and text should be, in the learned space. However, most prior methods have focused on the case where image and text convey redundant information; in contrast, real-world image-text pairs convey complementary information with little overlap. Further, images in news articles and media portray topics in a visually diverse fashion; thus, we need to take special care to ensure a meaningful image representation. We propose novel within-modality losses which encourage semantic coherency in both the text and image subspaces, which does not necessarily align with visual coherency. Our method ensures that not only are paired images and texts close, but the expected image-image and text-text relationships are also observed. Our approach improves the results of cross-modal retrieval on four datasets compared to five baselines.

1 Introduction

Vision-language tasks such as image captioning [2, 27, 58] and cross-modal generation and retrieval [40, 60, 63] have seen increased interest in recent years. At the core of methods in this space are techniques to bring together images and their corresponding pieces of text. However, most existing cross-modal retrieval methods only work on data where the two modalities (images and text) are well aligned, and provide fairly redundant information. As shown in Fig. 1, captioning datasets such as COCO contain samples where the overlap between images and text is significant (both image and text mention or show the same objects). In this setting, cross-modal retrieval means finding the manifestation of a single concept in two modalities (e.g. learning embeddings such that the word "banana" and the pixels for "banana" project close by in a learned space).

In contrast, real-world news articles contain image and text pairs that cover the same topic, but show complementary information (protest signs vs information about the specific event; guns vs discussion of rights; rainbow flag vs LGBT rights). While a human viewer can still guess which images go with which text, the alignment between image and text is abstract and symbolic. Further, images in news articles are ambiguous *in isolation*. We show in Fig. 2 that an image

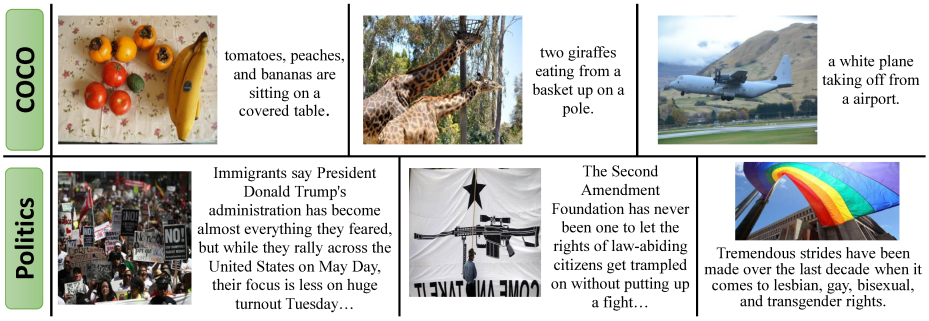

Fig. 1. Image-text pairs from COCO [25] and Politics [49]. Traditional image captions (top) are descriptive of the image, while we focus on the more challenging problem of aligning images and text with a non-literal complementary relationship (bottom).

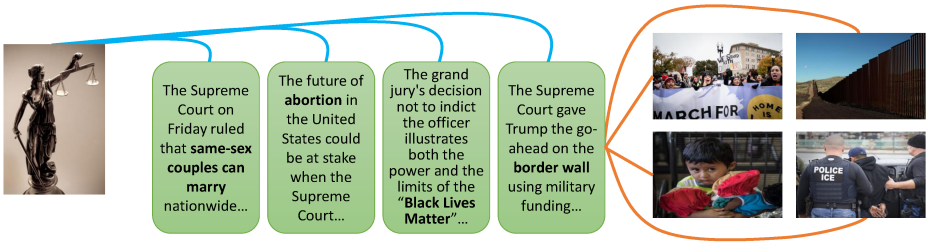

Fig. 2. The image on the left symbolizes justice and may be paired with text about a variety of subjects (e.g. abortion, same sex marriage). Similarly, text regarding immigration (right) may be paired with visually dissimilar images. Our approach enforces that *semantically* similar content (images on the right) is close in the learned space. To discover such content, we use semantic neighbors of the text and their paired images.

might illustrate multiple related texts (shown in green), and each text in turn could be illustrated with multiple visually distant images (e.g. the four images on the right-hand side could appear with the border wall text). Thus, we must first resolve any ambiguities in the image, and figure out "what it means".

We propose a metric learning approach where we use the semantic relationships between text segments, to guide the embedding learned for corresponding images. In other words, to understand what an image "means", we look at what articles it appeared with. Unlike prior approaches, we capture this information not only across modalities, but within the image modality itself. If texts y_i and y_j are semantically similar, we learn an embedding where we explicitly encourage their paired images x_i and x_j to be similar, using a new unimodal loss. Note that in general x_i and x_j need not be similar in the original visual space (Fig. 2). In addition, we encourage texts y_i and y_j, who were close in the unimodal space, to remain close.

Our novel loss formulation explicitly encourages *within-modality semantic coherence*. Fig. 3 shows the effect. On the left, we show the proximity of sam-

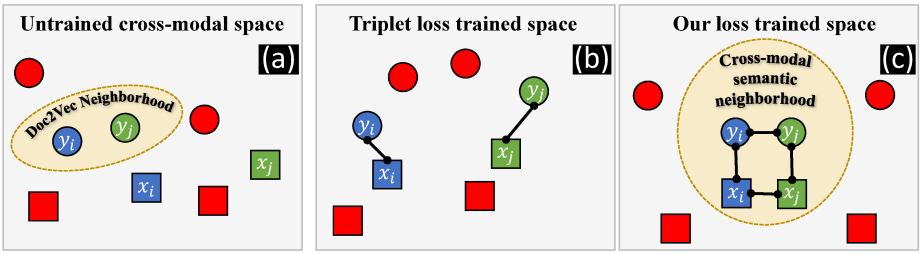

Fig. 3. We show how our method enforces cross-modal semantic coherence. Circles represent text and squares images. In (a), we show the untrained cross-modal space. Note y_i and y_j are neighbors in Doc2Vec space and thus semantically similar. (b) shows the space after triplet loss training. y_i and x_i, and y_j and x_j, are now close as desired, but y_i and y_j have moved apart, and x_i and x_j remain distant. (c) shows our loss's effect. Now, all semantic neighbors (both images and text) are pulled closer.

ples before cross-modal learning; specifically, while two texts are close in the document space, their paired articles may be far from the texts. In the middle, we show the effect of using a standard triplet loss, which pulls image-text pairs close, but does not necessarily preserve the similarity of related articles; they are now further than they used to be in the original space. In contrast, on the right, we show how our method brings paired images and text closer, while also preserving a semantically coherent region, i.e. the texts remained close.

In our approach, we use neighborhoods in the original text document space, to compute semantic proximity. We also experiment with an alternative approach where we compute neighborhoods using the visual space, then guide the corresponding texts to be close. This approach is a variant of ours, and is novel in the sense that it uses proximity in one unimodal space, to guide the other space/modality. While unimodal losses based on visual similarity are helpful over a standard cross-modal loss (e.g. triplet loss), our main approach is superior.

Next, we compare to a method [52] which utilizes the *set* of text annotations available for an image in COCO, to guide the structure of the learned space. We show that when these ground-truth annotations are available, using them to compute neighborhoods in the textual space is the most reliable. However, on many datasets, such sets of annotations (more than one for the same image) are not available. We show that our approach offers a comparable alternative.

Finally, we test the contribution of our additional losses using PVSE [48], a state-of-the-art visual semantic embedding model, as a backbone. We show that our proposed loss further improves the performance of this model.

To summarize, our contributions are as follows.

- We preserve relationships in the original *semantic* space. Because images do not clearly capture semantics, we use the semantic space (from text) to guide the image representation, through a unimodal (within-modality) loss.
- We perform detailed experimental analysis of our proposed loss function, including ablations, on four recent large-scale image-text datasets. One [3]

contains multimodal articles from New York Times, and another contains articles from far-left/right media [49]. We also conduct experiments on [25, 43]. Our approach significantly improves the state-of-the-art in most cases. The more abstract the dataset/alignment, the more beneficial our approach.
– We tackle a new cross-modal retrieval problem where the visual space is much less concrete. This scenario is quite practical, and has applications ranging from automatic caption generation for news images, to detection of fake multimodal articles (i.e. detecting whether an image supports the text).

2 Related Work

Cross-modal learning. A fundamental problem in cross-modal inference is the creation of a shared semantic manifold on which multiple modalities may be represented. The goal is to learn a space where content about related semantics (e.g. images of "border wall" and text about "border wall") projects close by, regardless of which modality it comes from. Many image-text embedding methods rely on a two-stream architecture, with one stream handling visual content (e.g. captured by a CNN) and the other stream handling textual content (e.g. through an RNN). Both streams are trained with paired data, e.g. an image and its captions, and a variety of loss functions are used to encourage both streams to produce similar embeddings for paired data. Recently, purely attention-based approaches have been proposed [6, 26]. One common loss used to train retrieval models is triplet loss, which originates in the (single-modality) metric learning literature, e.g. for learning face representations [42]. In cross-modal retrieval, the triplet loss has been used broadly [9,32,34,38,57,66]. Alternative choices include angular loss [51], N-pairs loss [47], hierarchical loss [11], and clustering loss [36].

While single-modality losses like triplet, angular and N-pairs have been used across and within modalities, they are not sufficient for cross-modal retrieval. These losses do not ensure that the general semantics of the text are preserved in the new cross-modal space; thus, the cross-modal matching task might distort them too much. This phenomenon resembles forgetting [14, 24] but in the cross-modal domain. Our method preserves within-modal structure, and a similar effect can be achieved by leveraging category labels as in [5, 31, 50, 64]; however, such labels are not available in the datasets we consider, nor is it clear how to define them, since matches lie beyond the presence of objects. Importantly, classic retrieval losses losses do not tackle the complementary relationship between images and text, which makes the space of topically related images more visually diffuse. In other words, two images might depict substantially *different visual* content but nonetheless be *semantically related*.

Note that we do not propose a new *model* for image-text alignment, but instead propose cross-modal embedding *constraints* which can be used to train any such model. For example, we compare to Song et al. [48]'s recent polysemous visual semantic embedding (PVSE) model, which uses global and local features to compute self-attention residuals. Our loss improves [48]'s performance.

Our work is also related to cross-modal distillation [10, 12, 15, 46], which transfers supervision across modalities, but none of these approaches exploit the semantic signal that text neighborhoods carry to constrain the visual representations. Finally, [1, 22, 61] detect different types of image-text relationships (e.g. parallel, complementary) but do not retrieve across modalities.

Metric learning approaches learn distance metrics which meaningfully measure the similarity of objects. These can be broadly categorized into: 1) sampling-based methods [17, 18, 28, 29, 36, 44, 45, 53, 54, 56, 59], which intelligently choose easy/hard samples or weight samples; or 2) loss functions [8, 11, 16, 42, 47, 51, 55] which impose intuitions regarding neighborhood structure, data separation, etc. Our method relates to the second category. Triplet loss [20, 42] takes into account the *relative* similarity of positives and negatives, such that positive pairs are closer to each other than positives are to negatives. [62] generalize triplet loss by fusing it with classification loss. [37] integrate all positive and negative pairs within a minibatch, such that all pair combinations are updated jointly. Similarly, [47]'s N-pair loss pushes multiple negatives away in each triplet. [52] propose a structural loss, which pulls multiple text paired with the same image together, but requires more than one ground truth caption per image (which most datasets lack). In contrast, our approach pulls semantically similar images *and* text together and only requires a single caption per image. More recently, [51] propose an angular loss which leverages the triangle inequality to constrain the angle between points within triplets. We show how cross-modal complementary information (semantics paired with diverse visuals) can be leveraged to improve the learned embedding space, regardless of the specific loss used.

3 Method

Consider two image-text pairs, $\{x_i, y_i\}$ and $\{x_j, y_j\}$. To ground the "meaning" of the images, we use proximity in a generic, pre-trained textual space between the texts y_i and y_j. If y_i and y_j are semantically close, we expect that they will also be relatively close in the learned space, and further, that x_i and x_j will be close also. We observed that, while intuitive, this expectation does not actually hold in the learned cross-modal space. The problem becomes more severe when image and paired text do not exhibit literal alignment, as shown in Fig. 1, because images paired via text neighbors could be visually different.

We describe how several common existing loss functions tackle cross-modal retrieval, and discuss their limitations. We then propose two constraints which pull within-modality semantic neighbors close to each other. Fig. 4 illustrates how our approach differs from standard metric learning losses.

3.1 Problem formulation and existing approaches

We assume a dataset $\boldsymbol{\mathcal{D}} = \{\mathbf{I}, \mathbf{T}\}$ of n image-text pairs, where $\mathbf{I} = \{x_1, x_2, \ldots, x_n\}$ and $\mathbf{T} = \{y_1, y_2, \ldots, y_n\}$ denote the set of paired images and text, respectively. By pairs, we mean y_i is text related to or co-occurring with image x_i. Let f_I

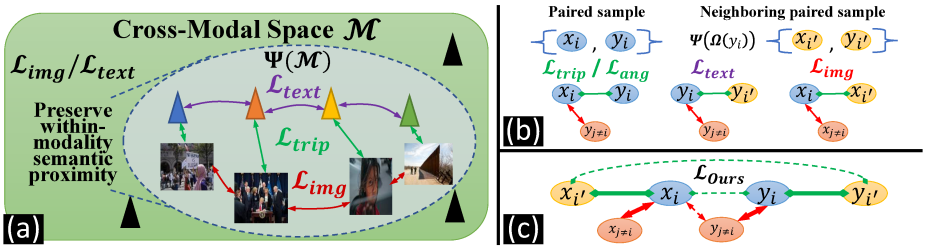

Fig. 4. (a): $\mathcal{L}_{text}$ and $\mathcal{L}_{img}$ pull semantic neighbors of the same modality closer. The images are visually distinct, but semantically similar. (b): Pull connections are shown in green, and push in red. $\mathcal{L}_{trip}$ and $\mathcal{L}_{ang}$ operate cross-modally, but impose no within-modality constraints. (c): $\mathcal{L}_{ours}$ (which combines all three losses above) exploits the paired nature of the data to enforce the expected inter/intra-modal relationships. Solid lines indicate connections that our loss enforces but triplet/angular do not.

denote a convolutional neural network which projects images into the joint space and f_T a recurrent network which projects text. We use the notational shorthand $f_T(y) = y$ and $f_I(x) = x$. The goal of training f_I and f_T is to learn a cross-modal manifold $\mathcal{M}$ where semantically similar samples are close. At inference time, we wish to retrieve a ground-truth paired text given an input image, or vice versa. One common technique is triplet loss [42] which posits that paired samples should be closer to one another than they are to non-paired samples. Let $\mathcal{T} = \left(x_i^a, y_i^p, y_j^n\right)$ denote a triplet of samples consisting of an anchor (a), positive or paired sample (p), and negative or non-paired sample (n) chosen randomly such that $i \neq j$. Let m denote a margin. The triplet loss $\mathcal{L}_{trip}$ is then:

$$\mathcal{L}_{trip}(\mathcal{T}) = \left[\|x_i^a - y_i^p\|_2^2 - \|x_i^a - y_j^n\|_2^2 + m\right]_+ \tag{1}$$

This loss is perhaps the most common one used in cross-modal retrieval tasks, but it has some deficiencies. For example, the gradient of the triplet loss wrt. each point only considers two points, but ignores their relationship with the third one; for example, $\frac{\partial \mathcal{L}_{trip}}{\partial x_i^a} = 2\left(y_j^n - y_i^p\right)$. This allows for degenerate cases, so angular loss $\mathcal{L}_{ang}$ [51] accounts for the angular relationship of all three points:

$$\mathcal{L}_{ang}(\mathcal{T}) = \left[\|x_i^a - y_i^p\|_2^2 - 4\tan^2\alpha\|y_j^n - \mathcal{C}_i\|_2^2\right]_+ \tag{2}$$

where $\mathcal{C}_i = \left(x_i^a + y_i^p\right)/2$ is the center of a circle through anchor and positive.

One challenging aspect of these losses is choosing a good negative term in the triplet. If the negative is too far from the anchor, the loss becomes 0 and no learning occurs. In contrast, if negatives are chosen too close, the model may have difficulty converging to a reasonable solution as it continuously tries to move samples to avoid overlap with the negatives. How to best sample triplets to avoid these issues is an active area of research [8]. One recent technique, the N-pairs loss [47], proposes that instead of a single negative sample being used, all negatives within the minibatch should be used. The N-pairs loss $\mathcal{L}_{ang}^{NP}$ pushes the

anchor and positive embedding away from *multiple* negatives simultaneously:

$$\mathcal{L}_{ang}^{NP}(\mathcal{T}) = \sum_{y_j \in \text{minibatch},\ j \neq i} \mathcal{L}_{ang}\left(x_i^a, y_i^p, y_j^n\right) \tag{3}$$

The symmetric constraint [65] can also be added to explicitly account for bidirectional retrieval, i.e. text-to-image, by swapping the role of images and text to form symmetric triplets $\mathcal{T}_{sym} = (y_i^a, x_i^p, x_i^n)$:

$$\mathcal{L}_{ang}^{NP+SYM}(\mathcal{T}, \mathcal{T}_{sym}) = \mathcal{L}_{ang}^{NP}(\mathcal{T}) + \mathcal{L}_{ang}^{NP}(\mathcal{T}_{sym}) \tag{4}$$

Limitations. While these loss functions have been used for cross-modal retrieval, they do not take advantage of several unique aspects of the multi-modal setting. Only the dashed pull/push connections in Fig. 4 (c) are part of triplet/angular loss. The solid connections are intuitive, but only enforced in our novel formulation. We argue the lack of explicit *within-modality* constraints allows discontinuities within the space for semantically related content from the same modality.

3.2 Our proposed loss

The text domain provides a semantic fingerprint for the image-text pair, since vastly dissimilar visual content may still be semantically related (e.g. image of White house, image of protest), while similar visual content (e.g. crowd in church, crowd at mall) could be semantically unrelated. We thus use the text domain to constrain within-modality semantic locality for both images and text.

To measure ground-truth semantic similarity, we pretrain a Doc2Vec [23] model Ω on the train set of text. Specifically, let d be the document embedding of article y_i, T denote the number of words in y_i, w_t represent the embedding learned for word t, $p(\cdot)$ be the probability of the given word, and k denote the look-around window. Ω learns word embeddings and document embeddings which maximize the average log probability: $\frac{1}{T}\sum_{t=1}^{T} \log p\left(w_t | d, w_{t-k}, \ldots, w_{t+k}\right)$. After training Ω, we use iterative backpropagation to compute the document embedding which maximizes the log probability for every article in the dataset: $\Omega(\mathbf{T}) = \{\Omega(y_1), \ldots, \Omega(y_n)\}$.

Because Doc2Vec has been shown to capture latent topics within text documents well [35], we seek to enforce that locality originally captured in $\Omega(\mathbf{T})$'s space also be preserved in the cross-modal space $\mathcal{M}$. Let

$$\Psi(\Omega(y_i)) = \langle x_{i'}, y_{i'} \rangle \tag{5}$$

denote a nearest neighbor function over $\Omega(\mathbf{T})$, where $\langle \cdot, \cdot \rangle$ is an image-text pair in the train set randomly sampled from the $k = 200$ nearest neighbors to y_i, and $i \neq i'$. $\Psi(\Omega(y_i))$ thus returns an image-text pair semantically related to y_i.

We formulate two loss functions to enforce within-modality semantic locality in $\mathcal{M}$. The first, $\mathcal{L}_{text}$, enforces locality of the text's projections:

$$\begin{aligned} \mathcal{T}'_{text} &= \left(y_i^a, y_{i'}^p, y_j^n\right) \\ \mathcal{L}_{text}(\mathcal{T}'_{text}) &= \mathcal{L}_{ang}(\mathcal{T}'_{text}) \\ \mathcal{L}_{ang}(\mathcal{T}'_{text}) &= \left[\|y_i^a - y_{i'}^p\|_2^2 - 4\tan^2\alpha \|y_j^n - \mathcal{C}_i\|_2^2\right]_+ \end{aligned} \tag{6}$$

where y_j^n is the negative sample chosen randomly such that $i \neq j$ and $\mathcal{C}_i = (y_i^a + y_i^p)/2$. $\mathcal{L}_{text}$ is the most straightforward transfer of semantics from $\Omega(\mathbf{T})$'s space to the joint space: nearest neighbors in Ω should remain close in $\mathcal{M}$.

As Fig. 4 (c) shows, $\mathcal{L}_{text}$ also indirectly causes semantically related images to move closer in $\mathcal{M}$: there is now a weak connection between x_i and $x_{i'}$ through the now-connected y_i and $y_{i'}$. To directly ensure smoothness and semantic coherence between x_i and $x_{i'}$, we propose a second constraint, $\mathcal{L}_{img}$:

$$\begin{aligned} \mathcal{T}'_{img} &= \left(x_i^a, x_{i'}^p, x_j^n\right) \\ \mathcal{L}_{img}\left(\mathcal{T}'_{img}\right) &= \mathcal{L}_{ang}\left(\mathcal{T}'_{img}\right) \\ \mathcal{L}_{ang}\left(\mathcal{T}'_{img}\right) &= \left[\|x_i^a - x_{i'}^p\|_2^2 - 4\tan^2\alpha\|x_j^n - \mathcal{C}_i\|_2^2\right]_+ \end{aligned} \tag{7}$$

where x_j^n is the randomly chosen negative sample such that $i \neq j$ and $\mathcal{C}_i = (x_i^a + x_i^p)/2$. Note that x_i and $x_{i'}$ are often not going to be neighbors in the original visual space. We use N-pairs over all terms to maximize discriminativity, and symmetric loss to ensure robust bidirectional retrieval:

$$\begin{aligned} &\mathcal{L}_{ang}^{OURS}\left(\mathcal{T}, \mathcal{T}_{sym}, \mathcal{T}'_{text}, \mathcal{T}'_{img}\right) = \\ &\mathcal{L}_{ang}^{NP+SYM}\left(\mathcal{T}, \mathcal{T}_{sym}\right) + \alpha\mathcal{L}_{text}^{NP}\left(\mathcal{T}'_{text}\right) + \beta\mathcal{L}_{img}^{NP}\left(\mathcal{T}'_{img}\right) \end{aligned} \tag{8}$$

where α, β are hyperparameters controlling the importance of each constraint.

Second variant. We also experiment with a variant of our method where the nearest neighbor function in Eq. 5 (computed in Doc2Vec space) is replaced with one that computes nearest neighbors in the space of visual (e.g. ResNet) features. Now $x_i, x_{i'}$ are neighbors in the original visual space before cross-modal training, and $y_i, y_{i'}$ are their paired articles (which may not be neighbors in the original Doc2Vec space). We denote this method as Ours (Img NNs) in Table 1, and show that while it helps over a simple triplet- or angular-based baseline, it is inferior to our main method variant described above.

Discussion. At a low level, our method combines three angular losses. However, note that our losses in Eq. 6 and Eq. 7 do not exist in prior literature. While [52] leverages ground-truth neighbors (sets of neighbors provided together for the same image sample in a dataset), we are not aware of prior work that estimates neighbors. Importantly, we are not aware of prior work that uses the text space to construct a loss over the image space, as Eq. 7 does. We show that the choice of space in which semantic coherency is computed is important; doing this in the original textual space is superior than using the original image space. We show the contribution of both of these losses in our experiments.

3.3 Implementation details

All methods use a two-stream architecture, with the image stream using a ResNet-50 [19] architecture initialized with ImageNet features, and the text stream using Gated Recurrent Units [7] with hidden state size 512. We use image size 224x224 and random horizontal flipping, and initialize all non-pretrained

learnable weights via Xavier init. [13]. Text models are initialized with word embeddings of size 200 learned on the target dataset. We apply a linear transformation to each model's output features ($\mathbb{R}^{2048\times256}$ for image, $\mathbb{R}^{512\times256}$ for text) to get the final embedding, and perform L_2 normalization. We use Adam [21] with minibatch size 64, learning rate 1.0e-4, and weight decay 1e-5. We decay the learning rate by a factor of 0.1 after every 5 epochs of no decrease in val. loss. We use a train-val-test split of 80-10-10. For Doc2Vec, we use [41] with $d \in \mathbb{R}^{200}$ and train using distributed memory [23] for 20 epochs with window $k = 20$, ignoring words that appear less than 20 times. We use hierarchical softmax [33] to compute $p(\cdot)$. To efficiently compute approximate nearest neighbors for Ψ, we use [30]; our method adds negligible computational overhead as neighbors are computed prior to training. We choose $\alpha = 0.3, \beta = 0.1$ for $\mathcal{L}_{trip}^{OURS}$, and $\alpha = 0.2, \beta = 0.3$ for $\mathcal{L}_{ang}^{OURS}$, on a held-out val. set.

4 Experiments

We compare our method to five baselines on four recent large-scale datasets. Our results consistently demonstrate the superiority of our approach at bidirectional retrieval. We also show our method better preserves within-modality semantic locality by keeping neighboring images and text closer in the joint space.

4.1 Datasets

Two datasets feature challenging indirect relations between image and text, compared to standard captioning data. These also exhibit longer text paired with images: 59 and 18 words on average, compared to 11 in COCO.

Politics [49] consists of images paired with news articles. In some cases, multiple images were paired with boilerplate text (website headliner, privacy policy) due to failed data scraping. We removed duplicates using MinHash [4]. We were left with 246,131 unique image-text pairs. Because the articles are lengthy, we only use the first two sentences of each. [49] do not perform retrieval.

GoodNews [3] consists of ∼466k images paired with their captions. All data was harvested from the New York Times. Captions often feature abstract or indirect text in order to relate the image to the article it appeared with. The method in [3] takes image and text as input, hence cannot serve as a baseline.

We also test on two large-scale standard image captioning datasets, where the relationship between image and text is typically more direct:

COCO [25] is a large dataset containing numerous annotations, such as objects, segmentations, and captions. The dataset contains ∼120k images with captions. Unlike our other datasets, COCO contains more than one caption per image, with each image paired with four to seven captions.

Conceptual Captions [43] is composed of ∼3.3M image-text pairs. The text comes from automatically cleaned alt-text descriptions paired with images harvested from the internet and has been found to represent a much wider variety of style and content compared to COCO.

	Img-Text Non-Literal				Img-Text Literal			
	Politics [49]		**GoodNews** [3]		**ConcCap** [43]		**COCO** [25]	
Method	**I→T**	**T→I**	**I→T**	**T→I**	**I→T**	**T→I**	**I→T**	**T→I**
Ang+NP+Sym	0.6270	0.6216	0.8704	0.8728	0.7687	0.7695	**0.6976**	**0.6964**
Ours (Img NNs)	0.6370	0.6378	0.8840	0.8852	0.7636	0.7666	0.6819	0.6876
Ours	**0.6467**	**0.6492**	**0.8849**	**0.8865**	**0.7760**	**0.7835**	0.6900	0.6885
PVSE	0.6246	0.6199	0.8724	0.8709	0.7746	0.7809	0.6878	0.6892
PVSE+Ours	**0.6264**	**0.6314**	**0.8867**	**0.8864**	**0.7865**	**0.7924**	**0.6932**	**0.6925**
Trip+NP+Sym	0.4742	0.4801	0.7203	0.7216	**0.5413**	0.5332	**0.4957**	**0.4746**
Ours (Trip)	**0.4940**	**0.4877**	**0.7390**	**0.7378**	0.5386	**0.5394**	0.4790	0.4611

Table 1. We show retrieval results for image to text (**I→T**) and text to image (**T→I**) on all datasets. The best method per group is shown in bold.

4.2 Baselines

We compare to N-Pairs Symmetric Angular Loss (Ang+NP+Sym, a combination of [47, 51, 65], trained with $\mathcal{L}_{ang}^{NP+SYM}$). For a subset of results, we also replace the angular loss with the weaker but more common triplet loss (Trip+NP+Sym). We show the result of choosing to enforce coherency within the image and text modalities by using images rather than text; this is the second variant of our method, denoted Ours (Img NNs).

We also compare our approach against the deep structure preserving loss [52] (Struc), which enforces that captions paired with the same image are closer to each other than to non-paired captions.

Finally, we show how our approach can improve the performance of a state-of-the-art cross-modal retrieval model. PVSE [48] uses both images and text to compute a self-attention residual before producing embeddings.

4.3 Quantitative results

We formulate a cross-modal retrieval task such that given a query image or text, the embedding of the paired text/image must be closer to the query embedding than non-paired samples also of the target modality. We sample random (non-paired) samples from the test set, along with the ground-truth paired sample. We then compute Recall@1 within each task: that is, whether the ground truth paired sample is closer to its cross-modal embedding than the non-paired embeddings. For our most challenging datasets (GoodNews and Politics), we use a 5-way task. For COCO and Conceptual Captions, we found this task to be too simple and that all methods easily achieved very high performance due to the literal image-text relationship. Because we wish to distinguish meaningful performance differences between methods, we used a 20-way task for Conceptual Captions and a 100-way task for COCO. Task complexities were chosen based on the baseline's performance, before our method's results were computed.

We report the results in Table 1. The first and second group of results all use angular loss, while the third set use triplet loss. We observe that our method

significantly outperforms all baselines tested for both directions of cross-modal retrieval for three of the four datasets. Our method achieves a 2% relative boost in accuracy (on average across both retrieval tasks) vs. the strongest baseline on GoodNews, and a 4% boost on Politics. We also observe recall is much worse for all tasks on the Politics dataset compared to GoodNews, likely because the images and article text are much less well-aligned. The performance gap seems small but note that given the figurative use of images in these datasets, often there may not be a clear ground-truth answer. In Fig. 2, Themis may be constrained to be close to protestors or border wall. At test time, the ground-truth text paired with Themis may be about the Supreme Court, but one of the "incorrect" answers could be about immigration or freedom, which still make sense. Our method keeps more neighbors closer to the query point as shown next, thus may retrieve plausible, but technically "incorrect" neighbors for a query.

Importantly, we see that while the variant of our method using neighborhoods computed in image space (Ours Img NNs) does outperform Ang+NP+Sym, it is weaker than our main method variant (Ours). We also observe that when adding our loss on top of the PVSE model [48], accuracy of retrieval improves. In other words, our loss is complementary to advancements accomplished by network model-based techniques such as attention.

Our method outperforms the baselines on ConcCap also, but not on COCO, since COCO is the easiest, least abstract of all datasets, with the most literal image-text alignment. Our approach constrains neighboring texts and their images to be close, and for datasets where matching is on a more abstract, challenging level, the benefit of neighbor information outweighs the disadvantage of this inexact similarity. However, for more straightforward tasks (e.g. in COCO), it may introduce noise. For example, for caption "a man on a bicycle with a banana", the model may pull that image and text closer to images with a banana in a bowl of fruit. Overall, our approach of enforcing within-modality semantic neighborhoods substantially improves cross-view retrieval, particularly when the relationship between image and text is complementary, rather than redundant.

To better ground our method's performance in datasets typically used for retrieval, we also conducted an experimented on Flickr30K [39]. Since that dataset does not exhibit image-text complementarity, we do not expect our method to improve performance, but it should not significantly reduce it. We compared the original PVSE against PVSE with our novel loss. We observed that our method slightly outperformed the original PVSE, on both text-to-image and image-to-text retrieval (0.5419 and 0.5559 for ours, vs 0.5405 and 0.5539 for PVSE).

In Table 2, we show a result comparing our method to Deep Structure Preserving Loss [52]. Since this method requires a *set* of annotations (captions) for an image, i.e. it requires *ground-truth neighbor relations* for texts, we can only apply it on COCO. In the first column, we show our method. In the second, we show [52] using ground-truth neighbors. Next, we show using [52] with *estimated neighbors*, as in our method. We see that as expected, using estimated rather than ground-truth text neighbors reduces performance (third vs. second columns). When estimated neighbors are used in [52]'s structural constraint, our

		Ours (Trip)	Struc (GT, Text)	Struc (NN_Ω, Text)	Struc (NN_Ω, Img)
COCO	**I→T**	<u>0.4790</u>	**0.4817**	0.4635	0.4752
	T→I	<u>0.4611</u>	**0.4867**	0.4594	0.4604

Table 2. We show retrieval results for image to text (**I→T**) and text to image (**T→I**) on COCO using [52]'s loss vs. ours. GT requires multiple **G**round **T**ruth captions per image, while NN uses **N**earest **N**eighbors. The best method per row is shown in bold, while the best method which does not require a set of neighboring text is underlined.

	GoodNews [3]		**Politics** [49]	
Method	**I**	**T**	**I**	**T**
Trip+NP+Sym	0.1183	0.1294	0.1135	0.1311
Ours (Trip)	**0.1327**	**0.1426**	**0.1319**	**0.1483**
Ang+NP+Sym	0.1032	0.1131	0.1199	0.1544
Ours (Ang)	**0.1270**	**0.1376**	**0.1386**	**0.1703**

Table 3. We test how well each method preserves the semantic neighborhood (see text) of Ω in $\mathcal{M}$. Higher values are better. Best method is shown in bold.

method performs better (third vs. first columns). Interestingly, we observe that defining [52]'s structural constraint in image rather than text space is better (fourth vs. third columns). In both cases, neighborhoods are computed in *text* space (Eq. 5). This may be because the structural constraint, which requires the group of neighbors to be closer together than to others, is too strict for estimated text neighbors. That is, the constraint may require the text embeddings to lose useful discriminativity to be closer to neighboring text. Neighboring images are likely to be much more visually similar in COCO than in GoodNews or Politics as they will contain the same objects.

We next test how well each method preserves the *semantic neighborhood* given by Ω, i.e. Doc2Vec space. We begin by computing the embeddings in $\mathcal{M}$ (cross-modal space) for all test samples. For each such sample s_i (either image or text), we compute $\Psi_{\mathcal{M}}(s_i)$, that is, we retrieve the neighbors (of the same modality as s_i) in $\mathcal{M}$. We next retrieve the neighbors of s_i in Ω, $\Psi_\Omega(s_i)$, described in Sec. 3.2. For each sample, we compute $|\Psi_{\mathcal{M}}(s_i) \cap \Psi_\Omega(s_i)| \,/\, |\Psi_\Omega(s_i)|$, i.e. the percentage of the nearest neighbors of the sample in Ω which are also its neighbors in $\mathcal{M}$. That is, we measure how well each method preserves within-modality semantic locality through the number of neighbors in Doc2Vec space which remain neighbors in the learned space. We consider the 200 nearest neighbors. We report the result for competitive baselines in Table 3. We find that our constraints are, indeed, preserving within-modality semantic locality, as sample proximity in Ω is more preserved in $\mathcal{M}$ with our approach than without it, i.e. we better reconstruct the semantic neighborhood of Ω in $\mathcal{M}$. We believe this allows our model to ultimately perform better at cross-modal retrieval.

We finally test the contribution of each component of our proposed loss. We test two variants of our method, where we remove either $\mathcal{L}_{text}$ or $\mathcal{L}_{img}$. We present our results in Table 4. In every case, *combining* our losses for our full

Method	GoodNews [3]		Politics [49]	
	I→T	T→I	I→T	T→I
Ours (Ang)	**0.8849**	**0.8865**	**0.6467**	**0.6492**
Ours (Ang)-$\mathcal{L}_{text}$	<u>0.8786</u>	0.8813	0.6387	<u>0.6467</u>
Ours (Ang)-$\mathcal{L}_{img}$	0.8782	<u>0.8817</u>	<u>0.6390</u>	0.6413

Table 4. We show an ablation of our method where we remove either component of our loss. The best method is shown in **bold** and the best ablation is underlined.

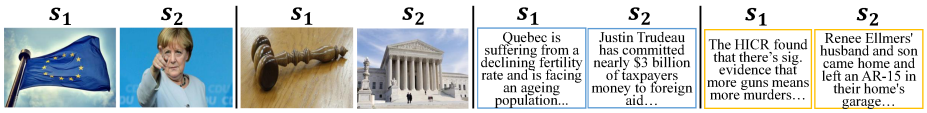

Fig. 5. Uncurated results showing image/text samples that our method keeps closest in $\mathcal{M}$ compared to the baseline, i.e. pairs where $\frac{d_{ours}(s_1,s_2)}{d_{baseline}(s_1,s_2)}$ is smallest. Our method keeps semantically related images and text closer in the space, relative to the baseline. While the images are not visually similar, they are semantically similar (EU and Merkel; judge's gavel and Supreme Court).

method performs the best, suggesting that each loss plays a complementary role in enforcing semantic locality for its target modality.

4.4 Qualitative results

In this section, we present qualitative results illustrating how our constraints both improve semantic proximity and demonstrate superior retrieval results.

Semantic proximity: In Fig. 5, we perform an experiment to discover what samples our constraints affect the most. We randomly sampled 10k image-image and text-text neighbor pairs (in Ω) and computed their distance in $\mathcal{M}$ using features from our method vs. the baseline Ang+NP+Sym. Small ratios indicate the samples were closer in $\mathcal{M}$ using our method, relative to the baseline, while larger indicate the opposite. We show the samples with the *two smallest* ratios for images and text. We observe that visually dissimilar, but semantically similar images have the smallest ratio (e.g. E.U. flag and Merkel, Judge's gavel and Supreme Court), which suggests our $\mathcal{L}_{img}$ constraint has moved the samples closer than the baseline places them. For text, we observe articles about the same issue are brought closer even though specifics differ.

Cross-modal retrieval results: In Fig. 6 we show the top-3 results for a set of queries, retrieved by our method vs. Ang+NP+Sym. We observe increased semantic homogeneity in the returned samples compared with the baseline. For example, images retrieved for "drugs" using our method consistently feature marijuana, while the baseline returns images of pills, smoke, and incorrect retrievals; "wall" results in consistent images of the border wall; "immigration" features arrests. For text retrieval, we find that our method consistently performs better at recognizing public figures and returning related articles.

Fig. 6. We show cross-modal retrieval results on Politics [49] using our method and the strongest baseline. We bold text aligning with the image. For text retrieval, ours returns more relevant (and semantically consistent) results. For image retrieval, our method exhibits more consistency (e.g. drug images are marijuana, immigration images show arrests), while the baseline returns more inconsistent and irrelevant images.

5 Conclusions

We proposed a novel loss function which improves semantic coherence for cross-modal retrieval. Our approach leverages a latent space learned on text alone, in order to enforce proximity within samples of the same modality, in the learned cross-modal space. We constrain text and image embeddings to be close in joint space if they or their partners were close in the unimodal text space. We experimentally demonstrate that our approach significantly improves upon several state-of-the-art loss functions on multiple challenging datasets. We presented qualitative results demonstrating increased semantic homogeneity of retrieval results. Applications of our method include improving retrieval of abstract, non-literal text, visual question answering over news and multimodal media, news curation, and learning general-purpose robust visual-semantic embeddings.

Acknowledgements: This material is based upon work supported by the National Science Foundation under Grant No. 1718262. It was also supported by Adobe and Amazon gifts, and an NVIDIA hardware grant. We thank the reviewers and AC for their valuable suggestions.

References

1. Alikhani, M., Sharma, P., Li, S., Soricut, R., Stone, M.: Clue: Cross-modal coherence modeling for caption generation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2020)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
3. Biten, A.F., Gomez, L., Rusinol, M., Karatzas, D.: Good news, everyone! context driven entity-aware captioning for news images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
4. Broder, A.: On the resemblance and containment of documents. In: Proceedings of the Compression and Complexity of Sequences (1997)
5. Carvalho, M., Cadène, R., Picard, D., Soulier, L., Thome, N., Cord, M.: Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In: ACM SIGIR Conference on Research & Development in Information Retrieval (2018)
6. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
7. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. In: Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8) (2014)
8. Duan, Y., Zheng, W., Lin, X., Lu, J., Zhou, J.: Deep adversarial metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
9. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improved visual-semantic embeddings. In: British Machine Vision Conference (BMVC) (2018)
10. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems (NIPS) (2013)
11. Ge, W.: Deep metric learning with hierarchical triplet loss. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
12. Girdhar, R., Tran, D., Torresani, L., Ramanan, D.: Distinit: Learning video representations without a single labeled video. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
13. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS) (2010)
14. Goodfellow, I.J., Mirza, M., Da Xiao, A.C., Bengio, Y.: An empirical investigation of catastrophic forgeting in gradient-based neural networks. In: Proceedings of International Conference on Learning Representations (ICLR) (2014)
15. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
16. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2006)

17. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C.: Matchnet: Unifying feature and metric learning for patch-based matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
18. Harwood, B., Kumar, B., Carneiro, G., Reid, I., Drummond, T., et al.: Smart mining for deep metric learning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
20. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International Workshop on Similarity-Based Pattern Recognition (2015)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. Proceedings of the International Conference on Learning Representations (ICLR) (2015)
22. Kruk, J., Lubin, J., Sikka, K., Lin, X., Jurafsky, D., Divakaran, A.: Integrating text and image: Determining multimodal document intent in instagram posts. In: Empirical Methods in Natural Language Processing (EMNLP) (2019)
23. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning (ICML) (2014)
24. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2017)
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV) (2014)
26. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
27. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
28. Lu, J., Xu, C., Zhang, W., Duan, L.Y., Mei, T.: Sampling wisely: Deep image embedding by top-k precision optimization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
29. Lu, J., Hu, J., Tan, Y.P.: Discriminative deep metric learning for face and kinship verification. IEEE Transactions on Image Processing **26**(9), 4269–4282 (2017)
30. Malkov, Y.A., Yashunin, D.A.: Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence (2016)
31. Marin, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., Torralba, A.: Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
32. Mithun, N.C., Paul, S., Roy-Chowdhury, A.K.: Weakly supervised video moment retrieval from text queries. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
33. Morin, F., Bengio, Y.: Hierarchical probabilistic neural network language model. In: International Workshop on Artificial Intelligence and Statistics (AISTATS) (2005)
34. Murrugarra-Llerena, N., Kovashka, A.: Cross-modality personalization for retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

35. Niu, L., Dai, X., Zhang, J., Chen, J.: Topic2vec: learning distributed representations of topics. In: International Conference on Asian Language Processing (IALP) (2015)
36. Oh Song, H., Jegelka, S., Rathod, V., Murphy, K.: Deep metric learning via facility location. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
37. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
38. Pang, K., Li, K., Yang, Y., Zhang, H., Hospedales, T.M., Xiang, T., Song, Y.Z.: Generalising fine-grained sketch-based image retrieval. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
39. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
40. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning (ICML) (2016)
41. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (2010)
42. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
43. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2018)
44. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
45. Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., Moreno-Noguer, F.: Discriminative learning of deep convolutional feature point descriptors. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
46. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Zero-shot learning through cross-modal transfer. In: Advances in Neural Information Processing Systems (NIPS) (2013)
47. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: Advances in Neural Information Processing Systems (NIPS) (2016)
48. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
49. Thomas, C., Kovashka, A.: Predicting the politics of an image using webly supervised data. Advances in Neural Information Processing Systems (NeurIPS) (2019)
50. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: ACM International Conference on Multimedia (2017)
51. Wang, J., Zhou, F., Wen, S., Liu, X., Lin, Y.: Deep metric learning with angular loss. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)

52. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
53. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
54. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
55. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research **10**(Feb), 207–244 (2009)
56. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
57. Ye, K., Kovashka, A.: Advise: Symbolism and external knowledge for decoding advertisements. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
58. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
59. Yuan, Y., Yang, K., Zhang, C.: Hard-aware deeply cascaded embedding. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
60. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
61. Zhang, M., Hwa, R., Kovashka, A.: Equal but not the same: Understanding the implicit relationship between persuasive images and text. In: British Machine Vision Conference (BMVC) (2018)
62. Zhang, X., Zhou, F., Lin, Y., Zhang, S.: Embedding label structures for fine-grained feature representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
63. Zhang, Z., Xie, Y., Yang, L.: Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
64. Zhen, L., Hu, P., Wang, X., Peng, D.: Deep supervised cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
65. Zhou, S., Wang, J., Wang, J., Gong, Y., Zheng, N.: Point to set similarity based deep feature learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
66. Zhu, B., Ngo, C.W., Chen, J., Hao, Y.: R2gan: Cross-modal recipe retrieval with generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)