Annotation Efficient Cross-Modal Retrieval with Adversarial Attentive Alignment

Po-Yao Huang[1], Guoliang Kang[1], Wenhe Liu[1], Xiaojun Chang[2*], and Alexander G. Hauptmann[1]
poyaoh,gkang,wenhel@cs.cmu.edu,cxj273@gmail.com,alex@cs.cmu.edu
[1]Language Technologies Institute, Carnegie Mellon University
[2]Faculty of Information Technology, Monash University

ABSTRACT

Visual-semantic embeddings are central to many multimedia applications such as cross-modal retrieval between visual data and natural language descriptions. Conventionally, learning a joint embedding space relies on large parallel multimodal corpora. Since massive human annotation is expensive to obtain, there is a strong motivation in developing versatile algorithms to learn from large corpora with fewer annotations. In this paper, we propose a novel framework to leverage automatically extracted regional semantics from un-annotated images as additional weak supervision to learn visual-semantic embeddings. The proposed model employs adversarial attentive alignments to close the inherent heterogeneous gaps between annotated and un-annotated portions of visual and textual domains. To demonstrate its superiority, we conduct extensive experiments on sparsely annotated multimodal corpora. The experimental results show that the proposed model outperforms state-of-the-art visual-semantic embedding models by a significant margin for cross-modal retrieval tasks on the sparse Flickr30k and MS-COCO datasets. It is also worth noting that, despite using only 20% of the annotations, the proposed model can achieve competitive performance (Recall at 10 > 80.0% for 1K and > 70.0% for 5K text-to-image retrieval) compared to the benchmarks trained with the complete annotations.

KEYWORDS

Cross-modal Retrieval, Joint Embedding, Adversarial Learning, Annotation Efficiency

ACM Reference Format:
Po-Yao Huang[1], Guoliang Kang[1], Wenhe Liu[1], Xiaojun Chang[2*], and Alexander G. Hauptmann[1]. 2019. Annotation Efficient Cross-Modal Retrieval with Adversarial Attentive Alignment . In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3343031.3350894

1 INTRODUCTION

Learning robust visual-semantic embeddings is central to the success of many multimedia applications involving multiple modalities such as cross-modal search and data mining [46]. The embedding model aims at encoding and mapping knowledge of multimodal entities into a joint embedding space. The transformation function is typically learned by aligning paired-inputs from two or more distinct domains (*e.g.*, images and natural language descriptions) into the common latent space where the embeddings are close if they are semantically associated or distant if uncorrelated.

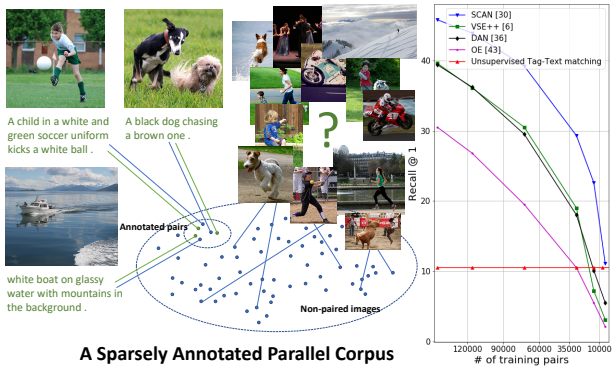


Figure 1: (Left) We consider learning under a sparsely annotated parallel corpus with abundant un-annotated images and limited (image, natural language sentence) pairs. (Right) Performance degeneration of state-of-the-art cross-modal retrieval models in the text-to-image retrieval task on Flick30K. (5 sentences/image.)

Recently, deep neural networks have made significant advancement for learning joint embeddings [6, 21, 30, 36, 52]. Such success is largely attributed to the availability of large-scale human-annotated parallel corpora such as the MS-COCO [32] and Flickr30K [51] datasets. Essentially, there are more than 610,000 and 150,000 annotated image-text pairs in MS-COCO and Flickr30K, respectively. As pointed out in [4], on par to quantity, the annotation diversity is also crucial for downstream tasks. Although models trained with affluent amount of well-annotated image-text pairs can achieve reasonable performance, we observe that the trend does not generalize to more common cases where only a limited amount of parallel annotations are available. As shown in Figure 1, recent VSE models [6, 30, 36, 43] all suffer greater degeneration as annotations become more sparsely available. (See Sec. 4 for experimental details.) Since collecting massive and high-quality human annotations for multimedia corpus is often prohibitively expensive and impractical, there is a strong incentive to designing annotation efficient algorithms to reduce the cost.

In this paper, we deal with the **sparse parallel corpus** scenario (Figure 1) where for cross-modal search and retrieval, a large collection of visual data is available but only a small amount of them are annotated with corresponding text descriptions. We pose an

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350894

challenging yet rewarding question: *Can we learn satisfactory visual-semantic embedding with a sparse parallel corpus?* Despite some recent progress [25, 31, 41], learning with small amount of parallel data is still challenging and to be developed in urgent need.

A straightforward way to deal with a sparse parallel corpus is to directly utilize the machine generated semantics of the images. In [35], Mithun *et al.* proposed a webly approach to utilize the global tags of the images. However, without handling the inevitable domain gap between the natural language description and the machine generated tags properly, the visual-semantic embedding learning could be negatively affected, which largely limits the the performance.

To circumvent these issues, inspired by the observation in [2] where bottom-up attention over regional objects aligns well with human's visual system, we propose to utilize "*regional semantics*" which correspond to the regions-of-interest in the un-annotated images and leverage the textual sequences of them to form "pseudo" image-text pairs as the additional weak supervision to conquer the sparsity of image-text annotation. Each regional semantic consists of the category of visual object and its attributes (*e.g. white cat*) which can be automatically extracted with object detection modules [1, 39]. With the inferred regional semantics, we develop a novel method to learn the joint visual-semantic embedding space from both the annotated pairs and the inferred pairs efficiently. To minimize the inherent domain gaps between annotated and un-annotated portion of visual and textual domains, we further impose an attentive alignment with adversarial learning objectives to selectively improve the correlation of semantically close components.

We conduct extensive experiments to quantify the degeneration of current state-of-the-art cross-modal retrieval models in the practical sparse parallel corpus scenario and to show the superiority of the proposed adversarial attentive alignment model for learning visual-semantic embeddings (A3VSE). In terms of reducing annotation effort, in comparison to various recently benchmarks trained with the complete annotations, the proposed model achieves a competitive performance with only 20% of annotations (Recall at 10 > 80.0% for 1K text-to-image and 70.0% for 5K text-to-image retrieval on Flickr30K and MS-COCO, respectively).

In a nutshell, our contributions can be summarized as

- We quantify the impact of learning with common sparse parallel corpora for the state-of-the-art cross-modal retrieval models and shed new insight for annotation efficiency.
- We propose to extract and leverage regional semantics to weakly supervise visual-semantic representation learning.
- We introduce adversarial attentive alignment to deal with multiple heterogeneous domain gaps. The attention mechanism emphasizes the visual or textural informative part to enable effective alignment.
- Experimental results of cross-modal retrieval on the Flickr30k and MS-COCO datasets demonstrate the superiority of our method to the state-of-the-art methods, under the same sparse parallel corpus setting. It is worth noting that, even trained with only 20% of the annotations, our model achieves competitive performance to recent models trained with the complete annotations.

2 RELATED WORKS

Visual-Semantic Embeddings for Cross-Modal Retrieval: Joint visual-semantic embeddings (VSE) have shown great potential in many multimedia tasks, including cross-modal retrieval [8, 21, 26], visual question answering [3, 12], image captioning [2, 49], muiltimodal classification [15], etc. Recently, there are increasing interest in developing system to match natural language descriptions to visual data with VSE [6, 21, 44, 48] for cross-modal retrieval.

In former works, the improvements in VSE are mainly processed on two perspectives: feature learning model and loss function. Various feature learning models have been extensively studied. For the textual feature, the conventional models introduce Fisher vectors [38] for word embeddings [34, 37] as in [8, 27, 47, 48]. Alternatively, recurrent neural networks (RNNs) [14] have been applied in many latest models [6, 17, 18, 20–22, 30, 36] and Zheng *et al.* suggest a convolutional structure in [52]. For the visual feature, VGG [40] and ResNet [13] models are widely implemented in previous works. Recently, Lee *et al.* [30] proposed to extract regional features from Faster-RCNN model [39]. Attention mechanisms also have been studied in the area [17, 22, 30, 36]. These works learn to select input fragments based on the context from either the same modality [17, 21, 36] or from another modality [30] or both [16]. In [18, 45], additional semantic features has been utilized in a multi-task schema. In contrast, in this work, we use image-semantic pair as the weak supervision for learning VSE with sparse corpora.

Most recent works in VSE leverage triplet loss [6, 8, 21, 27, 30, 36, 47, 48]. In [26], Kiros *et al.* proposed to use a triplet ranking loss to penalize the model with individual violations across the negatives. In [47, 48], Wang *et al.* add a within-view neighborhood structure-preserving constraints to further preserve the intra-modal structure. In VSE++ [6], Faghri *et al.* empirically show that emphasizing hard negative examples results in robust joint embeddings. Adversarial objective for cross-modal retrieval is firstly introduced in [45, 50] which narrow down the gap between different modalities by regularization via a domain discriminator. Our work generalize the idea about domain alignment and target on a more common but challenging sparse corpora scenario, where all the above models struggle without a plethora of parallel annotations.

Learning with Limited Supervision: Training models with sufficient amount of annotated data could achieve considerable performance for cross-model retrieval. However, in practice it is difficult to obtain a large amount of well-annotated data [25]. To address this problem, several previous works proposed to utilize web images and their meta data as an auxiliary source of training data [31, 41]. Meanwhile, there are studies focusing on learning with limited supervision. Jiang *et al.* proposed a coupled dictionary learning method to learn the class prototypes that utilize the discriminative information of visual space to improve the less discriminative semantic space in [19]. Tsai *et al.* augmented a typical supervised formulation with unsupervised techniques for learning joint embeddings of visual and textual data in [42]. Although promising performance has been obtained, none of these works consider the sparse parallel corpus setting.

To the best of our knowledge, the most relevant work to ours are [11, 35], where the authors resource meta data and image tags (*i.e.* global semantics) to improve learning of joint embedding space.

Our work complements their effort in two perspectives: First, we explore the feasibility of automatic regional semantics as they are more similar to natural language descriptions and leverage them for training improved sequential text encoder. Furthermore, we consider to close the inherent heterogeneous domain gaps with adversarial attentive alignment.

3 METHODOLOGY

We consider a common scenario where annotated image-text pairs are sparsely available and un-annotated images are abundant. While manually annotating images with natural language descriptions is expensive, automatically indexing them with semantic tags is relatively efficient [5]. Inspired by the bottom-up approach by [2], instead of resourcing global semantic tags as in [35], we seek to leverage semantics of salient regional objects which aligns well with the natural attention in human's cognition system to form additional image-semantic pairs for training. However, the inferred regional semantics exhibit clear difference to the natural language descriptions as in the annotated image-text pairs. A judicious way incorporating in these "pseudo" image-semantic pairs across heterogeneous domains for learning visual-semantic embeddings is therefore important.

Figure 2 illustrates the proposed adversarial attentive alignment model for learning visual-semantic embeddings (**A3VSE**). The proposed model jointly leverages the strong supervision from the annotated image-text pairs and the weak supervision from the inferred image-semantics pairs. Furthermore, A3VSE employs attentive adversarial objectives to selectively align entities from the annotated and un-annotated portion of visual and textual inputs and narrow the domain gaps in between.

3.1 Problem Formulation

Let $\mathcal{D}^l = \{I_1, \cdots, I_{N_l}\}$ be an annotated collection of instances where each instance $I_i = (v, t)$ consists of the image v and the corresponding natural language description t. Let $\mathcal{D}^u = \{v_1^u, \cdots, v_{N_u}^u\}$ denotes the collected but un-annotated images. We name $\mathcal{D} = \mathcal{D}^l \cup \mathcal{D}^u$ where $N_l << N_u$, as a **sparse parallel corpus**. We aim to utilize the un-annotated data $\mathcal{D}^u$, together with the annotated data $\mathcal{D}^l$, to learn better visual-semantic embeddings.

3.2 Feature Extractors

Let F^v and F^t denote the visual feature extractor and the textual feature extractor, respectively. We model F^v as a fixed object detection model (*e.g.* Faster RCNN), followed by a trainable fully-connected layer for mapping raw visual features in Faster RCNN into a H-dimension joint embedding space. On the other hand, F^t encodes the word tokens in a sentence with a word embedding matrix, followed by a trainable long short-term memory (LSTM) network to model the sequential text inputs. Note that the encoders F^v and F^t are shared among $\mathcal{D}^l$ and $\mathcal{D}^u$.

The visual feature of an image v is encoded as $\boldsymbol{V} = F^v(v) = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N] \in \mathbb{R}^{H \times N}$, where N is the maximum number of region-of-interest. Similarly, a sentence $t = [t_1, \cdots, t_M]$ is encoded as $\boldsymbol{T} = F^t(t) = [\boldsymbol{t}_1, \cdots, \boldsymbol{t}_M] \in \mathbb{R}^{H \times M}$, where M is the maximum sentence length. $(\boldsymbol{V}_i, \boldsymbol{T}_i)$ represents an annotated feature pair.

For $v^u \in \mathcal{D}^u$, we utilize an object detector (Faster RCNN [39]) to extract sequences of regional semantics (as text tokens, $s = [s_1, \cdots, s_M]$) and generate image-semantic pairs $(\boldsymbol{V}_i^u, \boldsymbol{S}_i)$. The regional semantics are the word tokens of attribute and the class name of the objects detected from an image v^u (*e.g.* "blue car"). The detected textual tokens are sorted by their object-wise confidence scores. We concatenate the regional semantics into one sentence, and then encode it as $\boldsymbol{S} = [\boldsymbol{s}_1, \cdots, \boldsymbol{s}_M] \in \mathbb{R}^{H \times M}$ via the shared F^t.

3.3 Adversarial Attentive Alignment

For learning and aligning instance-wise representation in individual modalities, we apply an attention network which focuses on certain encoded region/ tokens of inputs with respect to the global context from the same modality. We leverage a K-head context-aware attention network to capture the interactions between encoded entities and select informative ones for cross-modal alignment.

Given the feature representations (*i.e.* the visual features $\boldsymbol{V}$ or the texture features $\boldsymbol{T}$), the attentive encoder can be written as (we take visual features as an example):

$$E^v(\boldsymbol{V}) = [\boldsymbol{W}_0^v \boldsymbol{V}^\top, \boldsymbol{W}_1^v \boldsymbol{V}^\top, \cdots, \boldsymbol{W}_{K-1}^v \boldsymbol{V}^\top] \tag{1}$$

where

$$W_{ik}^v = \frac{\exp(\lambda_v \alpha_{ik}^v)}{\sum_{i'} \exp(\lambda_v \alpha_{i'k}^v)},$$

$$\alpha_{ik}^v = \tanh(\boldsymbol{P}_k^v \frac{1}{M^v} \sum_{i'} \boldsymbol{v}_{i'})^\top \tanh(\boldsymbol{Q}_k^v \boldsymbol{v}_i).$$

The $\boldsymbol{W}_k^v \in \mathbb{R}^{1 \times M^v}$, and $\boldsymbol{P}_k^v, \boldsymbol{Q}_k^v \in \mathbb{R}^{K' \times H}$, $k \in \{0, 1, \cdots, K-1\}$ are the parameters of the attentive encoder E^v, *i.e.* $\theta_{v-attn} = \{(\boldsymbol{W}_k^v, \boldsymbol{P}_k^v, \boldsymbol{Q}_k^v) | k \in \{0, 1, \cdots, K-1\}\}$. The λ_v is a constant temperature for the softmax function. The attentive encoder for the textural features (denoted by $E^t(\boldsymbol{T})$) works the same way but with independent parameters $\theta_{t-attn} = \{(\boldsymbol{W}_k^t, \boldsymbol{P}_k^t, \boldsymbol{Q}_k^t) | k \in \{0, 1, \cdots, K-1\}\}$. Note that E^t and E^v are shared among $\mathcal{D}_l$ and $\mathcal{D}_u$.

Thus, for an image v or v^u, the instance-level feature representation can be extracted and selectively encoded through $G^v = E^v \circ F^v$. Correspondingly, for the text description t or s, the instance-level feature can be achieved by $G^t = E^t \circ F^t$. We use $\theta_v = \{\theta_{v-attn}, \theta_{v-enc}\}$ and $\theta_t = \{\theta_{t-attn}, \theta_{t-enc}\}$ to denote the trainable parameters of G^v and G^t, respectively.

Triplet Alignment. For learning the joint embedding, we apply a hinge-based triplet ranking loss with hard negative mining as in [6] to align instance-wise paired visual-textual representations. Let $(\boldsymbol{a}, \boldsymbol{b})$ denotes a sampled image-text or image-semantic pair and $S(\boldsymbol{a}, \boldsymbol{b})$ is the cosine similarity. Let $\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}^-}{\operatorname{argmax}}\, S(\boldsymbol{a}, \boldsymbol{b}^-)$ and $\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}^-}{\operatorname{argmax}}\, S(\boldsymbol{a}^-, \boldsymbol{b})$ denote the hard negatives in the sampled batch. The triplet objective can be written as:

$$\ell^p(\mathcal{A}, \mathcal{B}; \alpha) = \frac{1}{L} \sum_{i=1}^{L} \{ \big[\alpha - S(\boldsymbol{a}_i, \boldsymbol{b}_i) + S(\boldsymbol{a}_i, \hat{\boldsymbol{b}})\big]_+ + \big[\alpha - S(\boldsymbol{a}_i, \boldsymbol{b}_i) + S(\hat{\boldsymbol{a}}, \boldsymbol{b}_i)\big]_+ \}, \tag{2}$$

where $|\mathcal{A}| = |\mathcal{B}| = L$, $[.]_+ = max(0, .)$, and α is the margin between the similarity of positive pair and that of hard-negative pair. Since annotated image-text pairs sampled from $\mathcal{D}^l$ are more reliable than

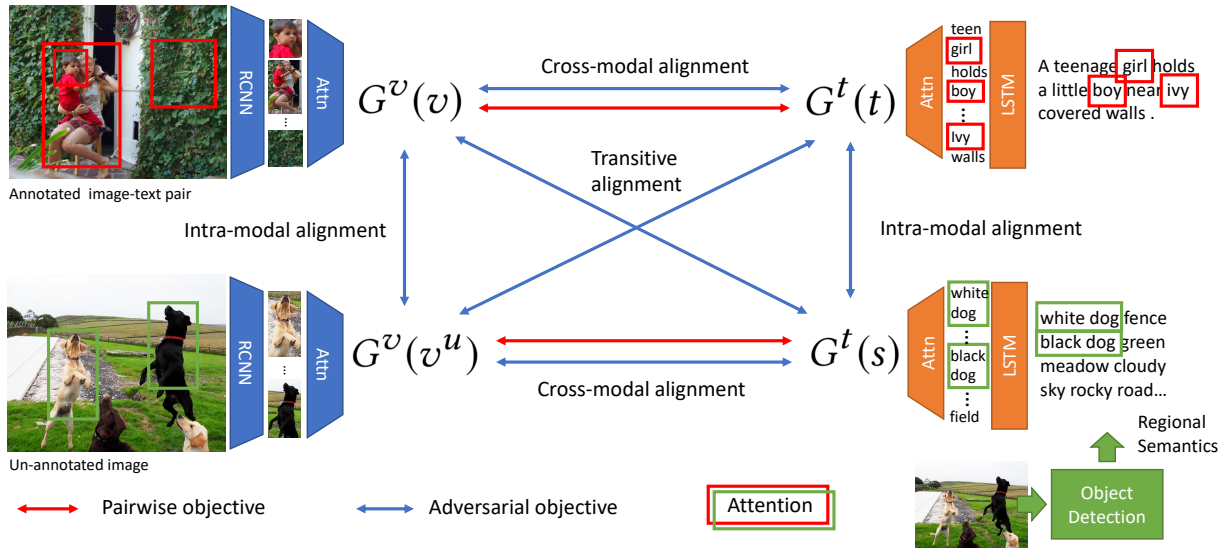


Figure 2: The proposed adversarial attentive alignment model for learning visual-semantic embeddings (A3VSE) for sparsely annotated multimodal corpora. Our model incorporates pseudo "image-text" pairs (illustrated as the bottom image-semantic pair) from the sequence of regional semantics of salient visual objects in un-annotated images. The triplet objectives (colored in red) and adversarial objectives (colored in blue) attend and align semantically correlated instances in the joint embedding space while closing the heterogeneous domain gaps between the annotated/un-annotated portion of visual and textual inputs.

image-semantic pairs sampled from $\mathcal{D}^u$, we differentiate the strong supervision by the former from the later with a hyper-parameter β. We model the triplet alignment objective as:

$$\ell^{tri} = \beta\ell^p(G^v(v), G^v(t); \alpha_{vt}) + (1-\beta)\ell^p(G^v(v^u), G^t(s); \alpha_{vs}) \tag{3}$$

A3VSE takes four different types of data, *i.e.* $\boldsymbol{V}$, $\boldsymbol{T}$, $\boldsymbol{V}^u$, $\boldsymbol{S}$ which are regarded as samples from four different domains. As shown in Figure 2, we propose using adversarial training to minimize the domain gaps among them. Specifically, we introduce six domain discriminators which are parameterized by θ_{vv^u}, θ_{ts}, θ_{vt}, $\theta_{v^u s}$, θ_{vs}, and $\theta_{v^u t}$. On one hand, they are trained to classify samples into correct domains. On the other hand, we employ the gradient reversal layer (GRL) [9] to the reverse the gradients propagated from these discriminators to update G^v and G^t to minimize the domain discrepancy. Such adversarial process can effectively diminish the discrepancy across different domains.

Generally, the adversarial loss for aligning two domains is

$$\ell^d(\mathcal{A}, \mathcal{B}; \theta) = \frac{1}{|\mathcal{A}|}\sum_{i=1}^{|\mathcal{A}|}\log D_\theta(\boldsymbol{a}_i) + \frac{1}{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\log(1 - D_\theta(\boldsymbol{b}_j)) \tag{4}$$

where D_θ is the domain discriminator parameterized by θ. The $\mathcal{A} = \{\boldsymbol{a}_i\}$ and $\mathcal{B} = \{\boldsymbol{b}_j\}$ are the mini-batch data sampled from two domains. The instantiations of $\boldsymbol{a}$ and $\boldsymbol{b}$ can be either two of $\{G^v(v), G^v(v^u), G^t(t), G^t(s)\}$. As shown in Figure. 2, we perform three types of alignments, *i.e. intra-modal alignment*, *Cross-modal alignment*, and *Transitive alignment*, which are described as follows.

Intra-modal Alignment handles the domain gaps between the annotated and un-annotated images, and annotated text descriptions and sequences of regional semantics. Specifically,

$$\begin{aligned}\ell^{intra} &= \lambda_{vv^u}\ell^d(G^v(v), G^v(v^u); \theta_{vv^u}) \\ &+ \lambda_{ts}\ell^d(G^t(t), G^t(s); \theta_{ts})\end{aligned} \tag{5}$$

Cross-modal Alignment aims at aligning the distribution of attended visual and textual features for annotated image-text pairs and inferred image-semantic pairs. That is,

$$\begin{aligned}\ell^{cross} &= \lambda_{vt}\ell^d(G^v(v), G^t(t); \theta_{vt}) \\ &+ \lambda_{v^u s}\ell^d(G^v(v^u), G^t(s); \theta_{v^u s})\end{aligned} \tag{6}$$

Transitive Alignment minimizes the domain gap between annotated images and sequences of regional semantics, and the domain gap between un-annotated images and annotated text descriptions:

$$\begin{aligned}\ell^{trans} &= \lambda_{vs}\ell^d(G^v(v), G^t(s); \theta_{vs}) \\ &+ \lambda_{v^u t}\ell^d(G^v(v^u), G^t(t); \theta_{v^u t})\end{aligned} \tag{7}$$

The overall adversarial objective for the attentive alignment is:

$$\ell^{adv} = \ell^{intra} + \ell^{cross} + \ell^{trans} \tag{8}$$

And the final objective can be formalized as

$$\ell^{A3VSE} = \ell^{adv} + \ell^{tri} \tag{9}$$

3.4 Optimization

Training and Inference. A min-max optimization is performed between the domain discriminators and attentive encoders:

$$(\theta_v, \theta_t) = \text{argmin}_{\theta_v, \theta_t} \ell^{A3VSE}(\boldsymbol{\theta})$$
$$(\theta_{adv}) = \text{argmax}_{\theta_{adv}} \ell^{A3VSE}(\boldsymbol{\theta}), \quad (10)$$

where $\theta_{adv} \triangleq (\theta_{vt}, \theta_{v^u s}, \theta_{ts}, \theta_{vv^u}, \theta_{vs}, \theta_{v^u t})$. In each iteration, we sample a mini-batch of (v, t) from $\mathcal{D}^l$ and (v^u, s) from $\mathcal{D}^u$ then follow the common practice in [9] of adversarial training with GRL to optimize Eq. 9. At the inference stage, we extract the visual embedding for image v and textual embedding for sentence t through G^v and G^t.

Discussion. In A3VSE, attentive encoders and adversarial alignment cooperate to learn satisfactory visual-semantic embeddings. On one hand, attentive encoders emphasize the informative part of the visual regions or textual entities, which helps adversarial training avoid misalignment and learn more discriminative features; on the other hand, adversarial alignment contributes to the improvement of attention mechanism of the attentive encoders in individual modalities which otherwise may be biased by the less amount of parallel image-text data.

4 EXPERIMENT

We perform extensive experiments to confirm the superiority of the proposed A3VSE model over competitive baselines with sparsely annotated multimodal corpora. We evaluate the learned visual-semantic embeddings in cross-modal retrieval tasks on two standard benchmark datasets (Flickr30K [51] and MS-COCO [32]) with the main goal of building an annotation efficient cross-modal retrieval model.

4.1 Dataset and Metric

We consider two commonly used benchmark datasets with natural language image descriptions: Flickr30K [51] and MS-COCO [32]. We constrain the amount of image-text annotations available in the training phase as an analogy to real-world scenarios where annotations are typically sparsely available.

Flickr30K [51]: There are 31,783 images and 158,915 image-text pairs in the Flick30K dataset. Five English descriptions are annotated for each image. We start with the standard split defined in [21] with 29,000 training, 1,000 validation, and 1,000 testing images. For learning with limited parallel pairs, we randomly shuffle once and trim the training set into 14,500 (50%), 5,800 (10%), and 2,900 (10%) subset of images. We sample 1, 2, and 5 text descriptions corresponding to those images. The resulting sparse training set is with size 2,900 (2%) to 72,500 (50%) out of 145,000 (100%) training image-text pairs in the original training split. The statistics of the new training splits of sparse Flickr30K can be found in Table 1. The standard validation and the testing are used for model selection and testing.

MS-COCO [32]: The MS-COCO dataset contains 123,287 images where each image is annotated with five English descriptions. In total, 616,435 image-text pairs are available. We follow the widely used split in [21] to move originally left 30,504 validation images to the training set, resulting a training set of 113,287 training images and 566,435 image-text pairs. We follow the same procedure as performed in Flickr30K and sample 5,664 (5%), 11,382 (10%), and 22,657 (20%) images along with 1, 2, 5 corresponding text descriptions. The statistics and the amount of training pairs can be found in Table 3. We report the testing performance on the whole 5,000 testing set.

Metric: As in most prior work on cross-modal retrieval tasks [6, 30, 36, 52], we measure rank-based performance by recall at K (R@k). Given a query, recall at k (R@k) calculates the percentage of test instances for which the correct one can be found in the top-K retrieved instances. We report R@1, R@5, and R@10.

4.2 Experimental Setup and Baselines

We focus on the text-to-image retrieval task (searching images with a natural language description as the query) and the image-to-text retrieval task (searching sentences with a query image) with the learned visual-semantic embeddings. We train models under different levels of training sparsity. Model selection and testing are with the full validation and the full testing set, respectively.

For all the baselines, we use their best single model settings and the code from their publicly available Github repositories. Since there are much less paired training instances in sparsely annotated dataset, for fair comparison and in prevention of under-fitting, we either keep the number of (mini-batch) training iterations as 50% iterations of the full dataset or extend the training epoch by 1.2x (for 50% annotations), 2.0x (20% annotations) and 2.5x (10% annotations). Early stopping and learning rate adjustment in the baselines follow the same adjustment if feasible.

Unsupervised baseline with image-level semantics We build an unsupervised cross-modal retrieval baseline using *NO* parallel annotations. Image-level semantics (*i.e.*, global semantics) of each image are extracted using pre-trained models from the following datasets: (1) Open Image [29]: 5,000 semantics trained on 9 million images. (2) ImageNet Shuffle [33], 12,073 classes defined in ImageNet. (3) Place365 [53]: 365 visual scene types. (4) Google Sports [23]: 478 sport-related semantics. We remove duplicated semantic concepts, normalize the scores, and then merge them into a 16500-dimension global semantic vector $\boldsymbol{s}_g$ for each image. Each dimension can be referred to a semantic concept in the original dataset. For example, an "aquarium" in Place365.

For retrieval, we directly match image-level semantics (tags) to text. Specifically, we expand the tokens in a sentence with the synsets defined in WordNet[7] and construct a 16500-dimension k-hot query vector $\boldsymbol{q}$, where k is the number of matched concepts. The matching score is calculated as $r = \boldsymbol{s}_g^T \boldsymbol{q}$.

4.3 Implementation Details

We now detail the pre-processing and implementation of the proposed model. To identify and vectorize salient visual objects in images, we use the Faster RCNN model [39] in [2] to detect objects and extract their corresponding visual features $\boldsymbol{V} \in \mathbb{R}^{36\times 2048}$. 36 is the maximum number of ROI in an image and 2,048 is the dimension of the flattened 5-th pooling layer of Faster RCNN [39]. We use raw features without l2 normalization.

For regional semantics in un-annotated images, we use the Faster RCNN model in [2] fine-tuned on Visual Genome [28] to extract English attribute names and class names of the objects detected from an image. Specifically, for every un-annotated image $v_j^u \in \mathcal{D}_u$, we generate $s_j = [s_{j1} \| s_{j2} \cdots \| s_{j|ROI|}]$ where "$\|$" is concatenation and

Sparse Flickr30K				Ours (A3VSE)						SCAN [30] (SOTA)					
%	#	%	# Ann	Text-to-Image			Image-to-Text			Text-to-Image			Image-to-Text		
Img	Sent	Ann	Pairs	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
10%	1/5	2%	2,900	20.7	46.0	58.5	27.6	56.2	68.1	2.0	7.2	11.7	5.1	16.0	22.9
10%	2/5	4%	5,800	28.1	55.6	66.9	42.0	69.7	79.0	16.1	35.7	46.5	18.9	39.8	53.9
10%	5/5	10%	14,500	32.0	60.1	71.0	46.8	72.8	80.7	24.6	48.1	59.3	25.9	56.3	70.9
20%	1/5	4%	5,800	29.1	56.4	68.1	43.3	71.0	81.8	17.2	37.5	47.5	21.2	44.4	55.0
20%	2/5	8%	11,600	32.6	61.6	72.3	44.8	72.7	82.8	28.4	54.0	64.6	39.0	68.0	78.6
20%	5/5	20%	29,000	34.9	64.4	73.6	48.4	77.0	85.1	29.3	56.9	68.3	42.1	71.8	81.3
50%	1/5	10%	14,500	36.7	65.1	75.9	51.6	78.7	85.7	29.5	56.3	67.3	40.2	72.2	81.4
50%	2/5	20%	23,200	42.9	70.5	80.3	61.4	83.7	89.4	33.9	61.3	71.4	46.8	75.2	84.5
50%	5/5	50%	72,500	44.5	73.8	83.3	60.9	85.7	91.6	39.2	67.5	77.2	52.6	80.3	87.5

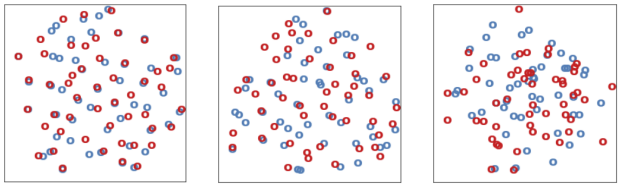
Table 1: Performance comparison on the 1K testing set of Flickr30K. The models are trained with the sparsely annotated training data as specified in the left column. *% Img* stands for the percentage of training images available compared to original training images in Flickr30K. *# Sent* stands for the number of paired text descriptions available for each image. *%/# Ann* is the percentage/number of annotations used for training compared to the complete training annotations in Flickr30K.

	Text-to-Image			Image-to-Text		
Model	R@1	R@5	R@10	R@1	R@5	R@10
Flickr30K **0%** Ann, 0 pairs						
s_g baseline	10.5	21.5	29.2	12.1	24.0	31.1
Flickr30K 10% Img, 5/5 Sent, **10%** Ann, 14,500 pairs						
DPC [52]	8.5	26.0	40.9	11.8	45.5	66.0
DAN [36]	10.1	25.3	42.8	12.2	41.7	64.5
VSE++ [6]	7.2	27.5	40.5	10.5	40.2	62.8
SCAN [30]	24.6	48.1	59.3	25.9	56.3	70.1
Ours (A3VSE)	**32.0**	**60.1**	**71.0**	**46.8**	**73.2**	**80.7**
Flickr30K 50% Img, 2/5 Sent, **20%** Ann, 29,000 pairs						
DPC [52]	26.4	53.0	63.9	35.8	68.5	79.7
DAN [36]	26.9	52.3	64.8	37.2	69.9	78.2
VSE++ [6]	27.3	54.5	66.0	33.5	65.2	78.2
SCAN [30]	33.9	61.3	71.4	46.8	75.2	84.5
Ours (A3VSE)	**42.9**	**70.5**	**80.3**	**61.4**	**83.7**	**89.4**
Flickr30K **100%** Ann, 145,000 pairs						
DPC [52]	39.1	69.2	80.9	55.6	81.9	89.0
DAN [36]	39.4	69.2	79.1	55.0	81.8	89.5
VSE++ [6]	39.6	70.1	79.8	53.1	82.1	87.5
SCAN [30]	45.8	74.4	83.0	61.8	87.5	93.7
Ours (A3VSE)	**49.5**	**79.5**	**86.6**	**65.0**	**89.2**	**94.5**

Table 2: Performance comparison with baselines on two sparse settings in Flickr30K.

s_k = [Attribute$_k$ Class$_k$] (*e.g.* "blue car"). There are 2,000 detectable objects and attributes. These regional semantics are then sorted by the confidence scores and concatenated as a text sequence. We group the image and the sequence and encode them as an image-semantic pair $(\boldsymbol{V}^u, \boldsymbol{S})$.

In our model, we set the embedding dimension H to 512. The same dimension is shared by all the context vectors in the attention modules. For text pre-processing, we tokenize, lower-case, truncate maximum sentence length to 57 on MS-COCO and 82 on Flickr30K, and then remove word tokens which appear less than 4 times. Similar to [52], we initialize word embeddings with pre-trained Glove embeddings [37]. All the weights within the network are initialized with Xavier initialization [10]. Other hyper-parameters



(a) 20% Img, 5/5 Sent (b) 10% Img, 5/5 Sent (c) No s in (b)

Figure 3: t-SNE visualization of the embedded testing images (blue) and sentences (red) under sparse Flick30K. Paired ones are expected to be close to each other.

are set as follows: $K = 3, \alpha_{vt} = 0.2$, $\alpha_{vs} = 0.3$, $\beta = 0.8$, and $\gamma = 2/(1 + exp(-\eta p)) - 1$ as in [9] where $\eta = 10$ and p is linearly increased from 0 to 1 in proportional to the training epoch. The hyper-parameters for the adversarial object is set as: Intra-modal alignments:$\lambda_{vv^u} = 0.2, \lambda_{ts} = 0.1$; Cross-modal alignments:$\lambda_{vt} = 0.5, \lambda_{v^u s} = 0.5$; Transitive alignments:$\lambda_{v^u t} = \lambda_{vs} = 0.3$.

For training, we train 24 epochs with Adam [24] optimizer. Learning rate is first 0.0005 then 0.00005 after 16th epoch. Models with the greatest summation of recall at 1, 5, 10 in the validation set are selected for testing. Weight decay is set to 0.000001 and gradients larger than 2.0 are clipped. The batch size is 128.

4.4 Results on Sparse Flickr30K

Table 1 shows the testing results with various levels of training sparsity on Flickr30K. Comparing the performance under the same percentage of annotations, the first interesting observation is that generally speaking it is preferred to have diverse images annotated than annotating a small amount of images with more text descriptions. With the same 10% annotations, it is better to annotate 50% of images with one sentence each than 10% of images with five sentences. These results suggest that regarding data collection and annotation, visual diversity is likely to be more important than textual diversity. Two cases of t-SNE visualization of the learned embedding are shown in Figure 3a and Figure 3b.

Under all sparse training set settings, the proposed model outperforms current state-of-the-art cross-modal retrieval model [30] by a significant margin. Namely, 4.2 to 18.7 in R@1, 6.3 to 38.8 in R@5, and 5.3 to 46.8 in R@10 text-to-image retrieval tasks. Notably,

Sparse MS-COCO				Ours (A3VSE)						SCAN [30] (SOTA)					
%	#	%	# Ann	Text-to-Image			Image-to-Text			Text-to-Image			Image-to-Text		
Img	Sent	Ann	Pairs	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
5%	1/5	1%	5,664	14.2	35.8	48.9	19.2	44.2	57.4	9.0	24.4	35.0	9.5	27.2	38.9
5%	2/5	2%	11,328	16.1	39.5	52.8	22.2	47.8	61.8	12.7	31.6	42.9	11.9	33.1	46.1
5%	5/5	5%	28,320	19.7	44.4	57.7	27.8	55.9	68.8	16.8	40.0	52.6	21.0	47.3	61.2
10%	1/5	2%	11,328	17.7	41.9	54.8	24.6	51.5	63.7	12.7	31.8	43.2	12.8	34.1	48.2
10%	2/5	4%	22,656	20.3	45.5	58.8	26.5	55.6	68.8	17.3	41.5	54.4	22.4	49.7	62.5
10%	5/5	10%	56,640	23.2	50.5	64.1	30.5	60.4	73.1	19.4	44.3	57.3	25.5	53.8	67.6
20%	1/5	4%	22,657	20.0	45.9	59.5	26.9	54.4	67.9	16.3	37.9	50.3	17.8	43.4	57.0
20%	2/5	8%	45,314	24.5	51.8	64.8	32.4	63.0	75.1	20.3	44.5	57.3	24.2	53.7	67.5
20%	5/5	20%	113,287	27.4	56.0	68.9	38.3	68.1	79.3	21.1	45.2	57.8	24.2	54.8	68.6

Table 3: Performance comparison on the 5K testing set of MS-COCO.

	Text-to-Image			Image-to-Text		
Model	R@1	R@5	R@10	R@1	R@5	R@10
MS-COCO **0%** Ann, 0 pairs						
s_g baseline	7.5	16.8	23.2	8.8	15.0	22.8
MS-COCO 10% Img, 1/5 Sent, **2%** Ann, 11,328 pairs						
DPC [52]	8.1	28.3	38.0	10.5	30.8	41.0
DAN [36]	8.8	28.3	37.1	11.1	30.1	42.5
VSE++ [6]	8.5	27.6	36.5	10.7	30.2	44.5
SCAN [30]	12.7	31.8	43.2	12.8	34.1	48.2
Ours (A3VSE)	**17.7**	**41.9**	**54.8**	**24.6**	**51.5**	**63.7**
MS-COCO 50% Img, 2/5 Sent, **20%** Ann, 113,287 pairs						
DPC [52]	19.1	41.0	55.5	20.5	45.1	60.2
DAN [36]	19.5	40.8	54.0	20.7	47.7	61.7
VSE++ [6]	19.5	41.2	56.5	21.5	48.5	63.5
SCAN [30]	22.3	47.5	60.2	25.5	56.1	70.5
Ours (A3VSE)	**28.2**	**57.9**	**70.6**	**38.4**	**69.5**	**81.1**
MS-COCO **100%** Ann, 566,435 pairs						
DPC [52]	25.3	53.4	66.4	41.2	70.5	81.1
DAN [36]	29.8	58.8	70.0	40.8	70.0	79.8
VSE++ [6]	30.3	56.0	72.4	41.3	69.5	81.2
SCAN [30]	34.4	63.7	75.7	46.4	77.4	87.2
Ours (A3VSE)	**39.0**	**68.0**	**80.1**	**49.3**	**81.1**	**90.2**

Table 4: Performance comparison with baselines on two sparse settings in MS-COCO.

greater improvement over current best model is achieved when less pairwise annotations are available. The improvements converges (but still outperforms) with more annotations available. A similar trend can be observed for the image-to-text retrieval task. These results demonstrate that the proposed A3VSE model can judiciously use regional semantics from un-annotated images for training its encoders and effectively learn the visual-semantic embeddings.

As shown in Table 2, in comparison to other recent models DAN [36], DPC [52], and VSE++ [6], the proposed model significantly outperforms them in all scenarios. In terms of reducing annotation effort, the proposed A3VSE model achieves competitive performance (with the criteria defined as R@10 > 80.0%) trained on only 20% annotations (23,200 pairs).

It is noteworthy that the unsupervised approach with global semantics which use *NO* image-text pairs cannot deliver satisfactory retrieval performance when query with natural language, indicating that there is a clear domain shift between the semantic pool of current image classification/ tagging models and the natural language queries. A similar phenomena is observed in our ablation study. Moreover, from the crossover of 10.5 R@1 in Figure 1 (right), the unsupervised global semantics from external classification datasets is worth as many as 14,000 image-text annotation pairs for the recent cross-modal retrieval models. Notably, A3VSE achieves 29.1 R@1 even trained with only 5,800 pairs.

4.5 Results on Sparse MS-COCO

Table 3 shows the results on the harder 5K testing set of MS-COCO. We sample 5%, 10%, 20% of images in MS-COCO to keep the number training of pairs more comparable to Flickr30K. The proposed model delivers the best performance on most metrics under all sparsity settings. For text-to-image retrieval, it outperforms SCAN [30] by 2.9 to 6.3 in R@1, 4.0 to 11.4 in R@5, and 4.4 to 13.9 in R@10. Similar trend can be observed in image-to-text retrieval task. The comparison with other recently published models is shown in Table 4 where the proposed model achieves the best performance in all sparse corpus scenarios.

Despite using only 20% of image-text annotations, the proposed model still achieves competitive performance (with the criteria defined as R@10 > 70.0%) in the more challenging 5K testing set in MS-COCO. More than 80% of annotation effort for the image-text pairs could potentially be relieved. Based on the quantitative results on multiple datasets, we validate the superiority and the annotation efficiency of the proposed A3VSE model.

4.6 Ablation Study

To quantify the contribution from individual components, we conduct ablation studies evaluating the cross-modal retrieval performance with models trained with 10% of images and 5/5 corresponding text descriptions (10% annotations) in Flickr30K. In each experiment, we remove one or change component of concern to quantify its relative importance. The larger the drop implies that the component is more important. For the experiment without semantics ($\boldsymbol{s}$), we remove all the regional semantics from the input and show the performance of the vanilla model. Then we swap the sequence of regional semantics with global semantics $\boldsymbol{s}_g$ and encode global semantics (can be viewed as image-level tags after applying a 0.3 threshold) with the shared word embedding matrix. For the internal modules and adversarial objectives, we either remove the attention layer with mean pooling over encoded visual/textual entities as the final instance-level representation, or we purge an adversarial objective from Eq. 9 during the training phase.

(a) 50% Img, 2/5 Sent, 23200 pairs **(b) 10% Img, 2/5 Sent, 5,800 pairs** **(c) Failures of (b)**

Figure 4: Qualitative examples of the proposed A3VSE model in text-to-image retrieval task (the upper two rows) and image-to-text retrieval task (the bottom row) on Flickr30K.

Flickr30K 10% Img 5/5 Sent, **10**% Ann, 14,500 pairs						
	Text-to-Image			Image-to-Text		
Model	R@1	R@5	R@10	R@1	R@5	R@10
No s	23.4	47.9	58.2	26.5	58.1	71.5
Swap s with s_g	29.0	56.3	67.2	40.5	67.4	77.6
s, without attention	23.8	50.1	62.7	35.8	64.3	75.1
s, without L_{adv}	30.9	58.5	69.0	43.8	70.9	79.5
Without ℓ^{intra}	31.8	59.6	**71.0**	44.8	72.5	**80.8**
Without ℓ^{cross}	31.3	59.2	70.5	45.2	71.8	80.1
Without ℓ^{trans}	31.5	59.7	70.9	46.1	71.8	80.3
Full model	**32.0**	**60.1**	**71.0**	**46.8**	**72.8**	80.7

Table 5: Ablation study of the proposed model

Table 5 shows the results of the ablation study. We observe that while global semantics boost model performance from the vanilla model, the regional semantics is the better choice even if they have a relatively small vocabulary size (1,104 versus 1,576) for the un-annotated images in sparse Flickr30K. The visualization of learned embeddings in Figure 3b and Figure 3c double confirms the difference. One possible explanation for this phenomena is that regional semantics are more similar to natural language descriptions. We observe that the distribution of vocabulary is closer (13.1% Intersection over Union (IoU)) between the natural language queries and the regional semantics than the global semantics (9.8% IoU). For instance, in a natural language description, people tend to describe an image with "frog" or "dog" rather than the detected global semantics "Amphibian" and "havanese".

Additionally, the attentive adversarial learning with domain discriminators plays an important role for closing the domain gaps between annotated and un-annotated inputs, delivers improved performance over models without adversarial objectives. However, we observe small variants among the best metrics over various configurations, suggesting that a careful hyper-parameter tuning may be required to achieve the optimal performance. We leave the robust automatic tuning for aligning multiple heterogeneous domains as our future work.

4.7 Qualitative Results

Figure 4 illustrates sampled qualitative testing results in the image-to-text and text-to-image retrieval tasks on sparse Flickr30K. The top two rows show the top four retrieved images given the natural language query above. The one and only one correct image is marked in green or red if rank > 10. The image-to-text retrieval results are depicted in the bottom row. We list the top five retrieved sentences and the corresponding query image. The correct sentences (up to five) are colored in green otherwise red.

In most cases the proposed model generates satisfactory results. As less parallel image-text pairs are available for training, we observe performance degeneration. For the failure cases, as expected, we observe that many failures result from out-of-vocabulary words (*e.g.* "amplifier" and "harp") in the sentences.

5 CONCLUSION

To reduce expensive human annotation cost, we have presented a novel annotation efficient A3VSE model for learning improved visual-semantic embeddings (VSEs) with sparsely annotated multimodal corpora. The proposed model jointly leverages strong supervision from image-text pairs and weak supervision from image-semantic pairs where the regional semantics are extracted from the un-annotated image collection. To further unify the heterogeneous inputs in the joint embedding space, our model employs attention-enhanced adversarial objectives to model intra-modal, cross-modal, and transitive alignment to selectively align annotated and un-annotated portion of visual and textual inputs.

In sparse Flickr30K and MS-COCO, the proposed model consistently and significantly outperforms recent competitive baselines. In comparison to global semantic tags, we have shown that regional semantics are more feasible for learning VSEs under sparsity. With regard to reducing annotation effort, we have presents insights towards efficient annotation collection and utilization. We have demonstrated that nearly 80% of the annotations can be reduced with the proposed model while achieving competitive results to recent models trained with the complete annotations.

ACKNOWLEDGEMENT

This research is supported by DARPA grant FA8750-18-2-0018 and FA8750-19-2-0501 funded under the AIDA program and the LwLL program. It is also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00340.

REFERENCES

[1] Waleed Abdulla. 2017. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. https://github.com/matterport/Mask_RCNN.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

[4] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards Diverse and Natural Image Descriptions via a Conditional GAN. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2989–2998.

[5] Jia Deng, Alexander C Berg, and Li Fei-Fei. 2011. Hierarchical semantic indexing for large scale image retrieval. In *CVPR 2011*. IEEE, 785–792.

[6] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. (2018). https://github.com/fartashf/vsepp

[7] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

[8] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. 2121–2129.

[9] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *International Conference on Machine Learning*. 1180–1189.

[10] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 249–256.

[11] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *European conference on computer vision*. Springer, 529–545.

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[15] Po-Yao Huang, Junwei Liang, Jean-Baptiste Lamare, and Alexander G. Hauptmann. 2018. Multimodal Filtering of Social Media for Temporal Monitoring and Event Analysis. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (ICMR '18)*. ACM, New York, NY, USA, 450–457. https://doi.org/10.1145/3206025.3206079

[16] Po-Yao Huang, Vaibhav, Xiaojun Chang, and Alexander G. Hauptmann. 2019. Improving What Cross-Modal Retrieval Models Learn Through Object-Oriented Inter- and Intra-Modal Attention Networks. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval (ICMR '19)*. ACM, New York, NY, USA, 244–252. https://doi.org/10.1145/3323873.3325043

[17] Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-Aware Image and Sentence Matching with Selective Multimodal LSTM. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 7254–7262.

[18] Yan Huang, Qi Wu, and Liang Wang. 2017. Learning semantic concepts and order for image and sentence matching. *arXiv preprint arXiv:1712.02036* (2017).

[19] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2018. Learning class prototypes via structure alignment for zero-shot recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 118–134.

[20] Xinyang Jiang, Fei Wu, Xi Li, Zhou Zhao, Weiming Lu, Siliang Tang, and Yueting Zhuang. 2015. Deep compositional cross-modal learning to rank via local-global alignment. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 69–78.

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

[22] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*. 1889–1897.

[23] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale Video Classification with Convolutional Neural Networks. In *CVPR*.

[24] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

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

[26] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *NIPS Workshop* (2014).

[27] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating neural word embeddings with deep image representations using Fisher Vectors. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 4437–4446.

[28] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. https://arxiv.org/abs/1602.07332

[29] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982* (2018).

[30] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. *arXiv preprint arXiv:1803.08024* (2018).

[31] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2017. Attention transfer from web images for video recognition. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 1–9.

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

[33] Pascal Mettes, Dennis C. Koelma, and Cees G.M. Snoek. 2016. The ImageNet Shuffle: Reorganized Pre-training for Video Event Detection. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR '16)*. ACM, New York, NY, USA, 175–182. https://doi.org/10.1145/2911996.2912036

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

[35] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos E Papalexakis, and Amit K Roy-Chowdhury. 2018. Webly Supervised Joint Embedding for Cross-Modal Image-Text Retrieval. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 1856–1864.

[36] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2156–2164.

[37] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[38] Florent Perronnin and Christopher Dance. 2007. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[40] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[41] Chen Sun, Sanketh Shetty, Rahul Sukthankar, and Ram Nevatia. 2015. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 371–380.

[42] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. 2017. Learning robust visual-semantic embeddings. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 3591–3600.

[43] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361* (2015).

[44] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-Embeddings of Images and Language. *CoRR* abs/1511.06361 (2015).

[45] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial Cross-Modal Retrieval. In *ACM MM*.

[46] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215* (2016).

[47] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[48] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 5005–5013.

[49] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on*

machine learning. 2048–2057.

[50] Xing Xu, Li He, Huimin Lu, Lianli Gao, and Yanli Ji. [n. d.]. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web* ([n. d.]), 1–16.

[51] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[52] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2017. Dual-Path Convolutional Image-Text Embedding. *CoRR* abs/1711.05535 (2017). arXiv:1711.05535 http://arxiv.org/abs/1711.05535

[53] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 487–495. http://papers.nips.cc/paper/5349-learning-deep-features-for-scene-recognition-using-places-database.pdf