Webly Supervised Joint Embedding for Cross-Modal Image-Text Retrieval

Niluthpol Chowdhury Mithun
University of California, Riverside, CA
nmith001@ucr.edu

Rameswar Panda
University of California, Riverside, CA
rpand002@ucr.edu

Evangelos E. Papalexakis
University of California, Riverside, CA
epapalex@cs.ucr.edu

Amit K. Roy-Chowdhury
University of California, Riverside, CA
amitrc@ece.ucr.edu

ABSTRACT

Cross-modal retrieval between visual data and natural language description remains a long-standing challenge in multimedia. While recent image-text retrieval methods offer great promise by learning deep representations aligned across modalities, most of these methods are plagued by the issue of training with small-scale datasets covering a limited number of images with ground-truth sentences. Moreover, it is extremely expensive to create a larger dataset by annotating millions of images with sentences and may lead to a biased model. Inspired by the recent success of webly supervised learning in deep neural networks, we capitalize on readily-available web images with noisy annotations to learn robust image-text joint representation. Specifically, our main idea is to leverage web images and corresponding tags, along with fully annotated datasets, in training for learning the visual-semantic joint embedding. We propose a two-stage approach for the task that can augment a typical supervised pair-wise ranking loss based formulation with weakly-annotated web images to learn a more robust visual-semantic embedding. Experiments on two standard benchmark datasets demonstrate that our method achieves a significant performance gain in image-text retrieval compared to state-of-the-art approaches.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**;

KEYWORDS

Image-Text Retrieval, Joint Embedding, Webly Supervised Learning

ACM Reference Format:
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 (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240712

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '18, October 22–26, 2018, Seoul, Republic of Korea
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240712

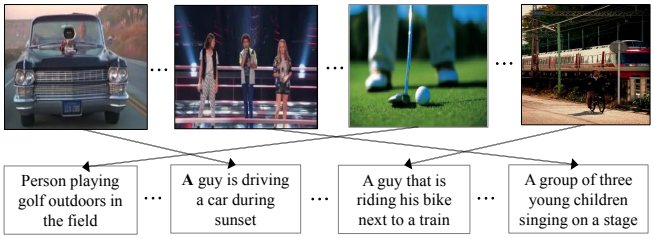

Figure 1: Illustration of image-text retrieval task: Given a text query, retrieve and rank images from the database based on how well they depict the text or vice versa.

1 INTRODUCTION

Joint embeddings have been widely used in multimedia data mining as they enable us to integrate the understanding of different modalities together. These embeddings are usually learned by mapping inputs from two or more distinct domains (e.g., images and text) into a common latent space, where the transformed vectors of semantically associated inputs should be close. Learning an appropriate embedding is crucial for achieving high-performance in many multimedia applications involving multiple modalities. In this work, we focus on the task of cross-modal retrieval between images and language (See Fig. 1), i.e., the retrieval of images given sentence query, and retrieval of text from a query image.

The majority of the success in image-text retrieval task has been achieved by the joint embedding models trained in a supervised way using image-text pairs from hand-labeled image datasets (e.g., MSCOCO [3], Flickr30k[42]). Although, these datasets cover a significant number of images (e.g., about 80k in MSCOCO and 30K in Flickr30K), creating a larger dataset with image-sentence pairs is extremely difficult and labor-intensive [32]. Moreover, it is generally feasible to have only a limited number of users to annotate training images, which may lead to a biased model [17, 50, 59]. Hence, while these datasets provide a convenient modeling assumption, they are very restrictive considering the enormous amount of rich descriptions that a human can compose [23]. Accordingly, although trained models show good performance on benchmark datasets for image-text retrieval task, applying such models in the open-world setting is unlikely to show satisfactory cross-dataset generalization (training on a dataset, testing on a different dataset) performance.

On the other hand, streams of images with noisy tags are readily available in datasets, such as Flickr-1M [21], as well as in nearly

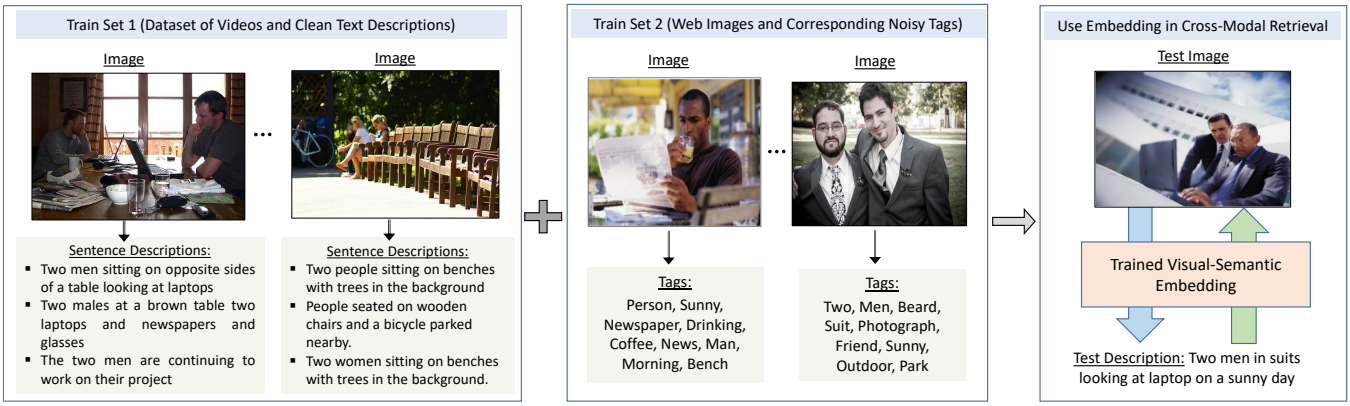

Figure 2: The problem setting of our paper. Our goal is to utilize web images associated with noisy tags to learn a robust visual-semantic embedding from a dataset of clean images with ground truth sentences. We test the learned latent space by projecting images and text descriptions from the test set in the embedding and perform cross-modal retrieval.

infinite numbers on the web. Developing a practical system for image-text retrieval considering a large number of web images is more likely to be robust. However, inefficient utilization of weakly-annotated images may increase ambiguity and degrade performance. Motivated by this observation, we pose an important question in this paper: *Can a large number of web images with noisy annotations be leveraged upon with a fully annotated dataset of images with textual descriptions to learn better joint embeddings?* Fig. 2 shows an illustration of this scenario. This is an extremely relevant problem to address due to the difficulty and non-scalability of obtaining a large amount of human-annotated training set of image-text pairs.

In this work, we study how to judiciously utilize web images to develop a successful image-text retrieval system. We propose a novel framework that can augment any ranking loss based supervised formulation with weakly-supervised web data for learning robust joint embeddings. Our approach consistently outperforms previous approaches significantly in cross-modal image-text retrieval tasks. We believe our efforts will provide insights to the researchers working in this area to focus on the importance of large scale web data for efficiently learning a more comprehensive representation from multimodal data.

1.1 Overview of the Proposed Approach

In the cross-modal image-text retrieval task, an embedding network is learned to project image features and text features into the same joint space, and then the retrieval is performed by searching the nearest neighbor in the latent space. In this work, we attempt to utilize web images annotated with noisy tags for improving joint embeddings trained using a dataset of images and ground-truth sentence descriptions. However, combining web image-tag pairs with image-text pairs in training the embedding is non-trivial. The greatest obstacle arises from noisy tags and the intrinsic difference between the representation of sentence description and tags. A typical representation of text is similar to, and yet very different from the representation of tags. Sentences are usually represented using RNN-based encoder with word-to-vec (Word2Vec) model, providing sequential input vectors to the encoder. In contrast, tags do not have sequential information and a useful representation of tags can be tf-idf weighted BOW vectors or the average of all Word2Vec vectors corresponding to the tags.

To bridge this gap, we propose a two-stage approach that learns the joint image-text representation. Firstly, we use a supervised formulation that leverages the available clean image-text pairs from a dataset to learn an aligned representation that can be shared across three modalities (e.g., image, tag, text). As tags are not available directly in the datasets, we consider nouns and verbs from a sentence as dummy tags (Fig. 3). We leverage ranking loss based formulation with image-text and image-tags pairs to learn a shared representation across modalities. Secondly, we utilize weakly-annotated image-tags pairs from the web (e.g., Flickr) to update the previously learned shared representation, which allows us to transfer knowledge from thousands of freely available weakly annotated images to develop a better cross-modal retrieval system. Our proposed approach is also motivated by learning using privileged information (LUPI) paradigm [45, 51] and multi-task learning strategies in deep neural networks [2, 43] that share representations between closely related tasks for enhanced learning performance.

1.2 Contributions

We address a novel and practical problem in this paper—how to exploit large-scale web data for learning an effective multi-modal embedding without requiring a large amount of human-crafted training data. Towards solving this problem, we make the following main contributions.

- We propose a webly-supervised approach utilizing web image collection with associated noisy tags, and a clean dataset containing images and ground truth sentence descriptions for learning robust joint representations.
- We develop a novel framework with pair-wise ranking loss for augmenting a typical supervised method with weakly-supervised web data to learn a more robust joint embedding.
- We demonstrate clear performance improvement in image-text retrieval task using proposed web-supervised approach on Flickr30K [42] and MSCOCO datasets [35].

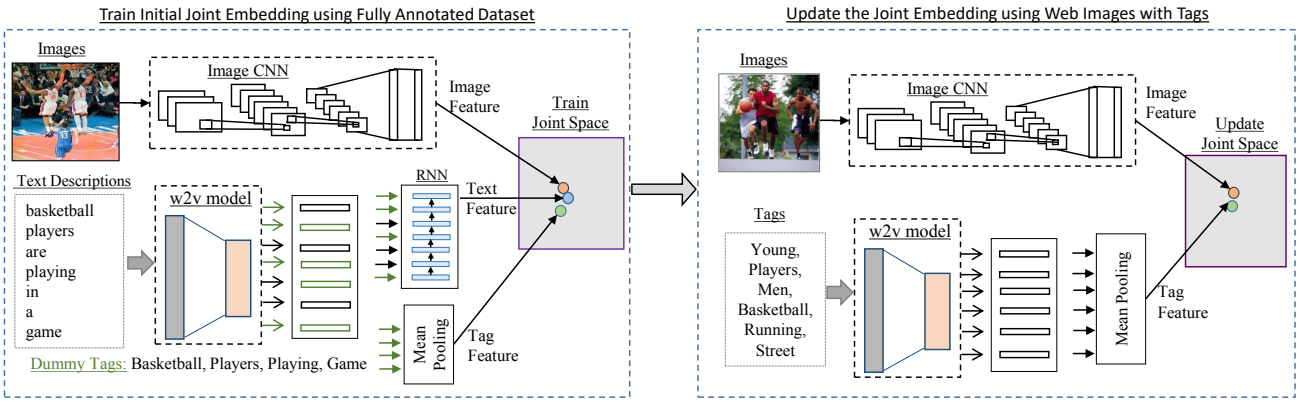

Figure 3: A brief illustration of our proposed framework for learning visual-semantic embedding model utilizing image-text pairs from a dataset and image-tag pairs from the web. First, a dataset of images and their sentence descriptions are used to learn an aligned image-text representation. Then, we update the joint representation using web images and corresponding tags. The trained embedding is used in image-text retrieval task. Please see Section 3 for details.

2 RELATED WORK

Visual-Semantic Embedding: Joint visual-semantic models have shown excellent performance on several multimedia tasks, e.g., cross-modal retrieval [18, 31, 38, 55], image captioning [23, 36], image classification [10, 12, 20] video summarization [4, 41]. Cross-modal retrieval methods require computing semantic similarity between two different modalities, i.e., vision and language. Learning joint visual-semantic representation naturally fits to our task of image-text retrieval since it is possible to directly compare visual data and sentence descriptions in such a joint space [8, 40].

Image-Text Retrieval: Recently, there has been significant interest in developing powerful image-text retrieval methods in multimedia, computer vision and machine learning communities [15, 24]. In [9], a method for mapping visual and textual data to a common space based on extracting a triplet of object, action, and scene is presented. A number of image-text embedding approaches has been developed based on Canonical Correlation Analysis (CCA) [12, 16, 47, 57]. Ranking loss has been used for training the embedding in most recent works relating image and language modality for image-text retrieval [8, 10, 30, 40, 53]. In [10], words and images are projected to a common space utilizing a ranking loss that applies a penalty when an incorrect label is ranked higher than the correct one. A bi-directional ranking loss based formulation is used to project image features and sentence features to a joint space for cross-modal image-text retrieval in [30].

Several image-text retrieval methods extended this work [30] with slight modifications in the loss function [8], similarity calculation [52, 53] or input features [40]. In [8], the authors modified the ranking loss based on violations incurred by relatively hard negatives and is the current state-of-the art in image-text retrieval task. An embedding network is proposed in [53] that uses the bi-directional ranking loss along with neighborhood constraints. Multi-modal attention mechanism is proposed in [40] to selectively attend to specific image regions and sentence fragments and calculate similarity. A multi-modal LSTM network is proposed in [19] that recurrently select salient pairwise instances from image and text, and aggregate local similarity measurement for image-sentence matching. Our method complements the works that project words and images to a common space utilizing a bi-directional ranking loss. The proposed formulation could be extended and applied to most of these approaches with little modifications.

Webly Supervised Learning: The method of manually annotating images for training does not scale well to the open-world setting as it is impracticable to collect and annotate images for all relevant concepts [33, 39]. Moreover, there exists different types of bias in the existing datasets [28, 49, 50]. In order to circumvent these issues, several recent studies focused on using web images and associated metadata as auxiliary source of information to train their models [11, 34, 48]. Although web images are noisy, utilizing such weakly-labeled images has been shown to be very effective in many multimedia tasks [13, 22, 34]

Our work is motivated by these works on learning more powerful models by realizing the potential of web data. As the largest MSCOCO dataset for image-sentence retrieval has only 80K training images, we believe it is extremely crucial and practical to complement scarcer clean image-sentence data with web images to improve the generalization ability of image-text embedding models. Most relevant to our work is [13], where authors constructed a dictionary by taking a few thousand most common words and represent text as tf-idf weighted bag of words (BoW) vectors that ignore word order and represents each caption as a vector of word frequencies. Although, such a textual feature representation allows them to utilize the same feature extractor for sentences and set of tags, it fails to consider the inherent sequential nature present in sentences in training image-sentence embedding models.

3 APPROACH

In this section, we first describe the network structure (Section 3.1). Then, we revisit the basic framework for learning image text mapping using pair-wise ranking loss (Section 3.2). Finally, we present

our proposed strategy to incorporate the tags in the framework to learn an improved embedding (Section 3.3).

3.1 Network Structure and Input Feature

Network Structure: We learn our joint embedding model using a deep neural network framework. As shown in Fig. 3, our model has three different branches for utilizing image, sentence, and tags. Each branch has different expert network for a specific modality followed by two fully connected embedding layers. The idea is that the expert networks will focus on identifying modality-specific features at first and the embedding layers will convert the modality-specific features to modality-robust features. The parameters of these expert networks can be fine-tuned together with training the embedding layers. For simplicity, we keep image encoder (e.g., pre-trained CNN) and tag encoder (e.g., pre-trained Word2Vec model) fixed in this work. The word embedding and the GRU for sentence representation are trained end-to-end.

Text Representation: For encoding sentences, we use Gated Recurrent Units (GRU) [5], which has been used for representing sentences in many recent works [8, 30]. We set the dimensionality of the joint embedding space, D, to 1024. The dimensionality of the word embeddings that are input to the GRU is 300.

Image Representation: For encoding image, we adopt a deep CNN model trained on ImageNet dataset as the encoder. Specifically, we experiment with state-of-the-art 152 layer ResNet model [14] and 19 layer VGG model [46] in this work. We extract image features directly from the penultimate fully connected layer. The dimension of the image embedding is 2048 for ResNet152 and 4096 for VGG19.

Tag Representation: We generate the feature representation of tags by summing over the Word2Vec [37] embeddings of all tags associated with an image and then normalizing it by the number of tags. Averaged word vectors has been shown to be a strong feature for text in several tasks [26, 27, 58].

3.2 Train Joint Embedding with Ranking Loss

We now describe the basic framework for learning joint image-sentence embedding based on bi-directional ranking loss. Many prior approaches have utilized pairwise ranking loss as the objective for learning joint embedding between visual input and textual input [24, 30, 55, 60]. Specifically, these approaches minimize a hinge-based triplet ranking loss in order to maximize the similarity between an image embedding and corresponding text embedding and minimize similarity to all other non-matching ones.

Given an image feature representation $\bar{i}$ ($\bar{i} \in \mathbb{R}^V$), the projection on the joint space can be derived as $i = W^{(i)}\bar{i}$ ($i \in \mathbb{R}^D$). Similarly, the projection of input text embedding $\bar{s}$ ($\bar{s} \in \mathbb{R}^T$) to joint space can be derived by $s = W^{(s)}\bar{s}$ ($s \in \mathbb{R}^D$). Here, $W^{(i)} \in \mathbb{R}^{D \times V}$ is the transformation matrix that maps the visual content into the joint space and D is the dimensionality of the space. In the same way, $W^{(s)} \in \mathbb{R}^{D \times T}$ maps input sentence embedding to the joint space. Given feature representation for words in a sentence, the sentence embedding $\bar{s}$ is found from the hidden state of the GRU. Here, given the feature representation of both images and corresponding text, the goal is to learn a joint embedding characterized by θ (i.e., $W^{(i)}$, $W^{(s)}$ and GRU weights) such that the image content and semantic content are projected into the joint space. Now, the image-sentence loss function $\mathcal{L}_{IS}$ can be written as,

$$\mathcal{L}_{IS} = \sum_{(i,s)} \Big\{ \sum_{s^-} max\big[0, \Delta - f(i,s) + f(i,s^-)\big] + \sum_{i^-} max\big[0, \Delta - f(s,i) + f(s,i^-)\big] \Big\} \tag{1}$$

where s^- is a non-matching text embedding for image embedding i, and s is the matching text embedding. This is similar for image embedding i and non-matching image embedding i^-. Δ is the margin value for the ranking loss. The scoring function $f(i,s)$ measures the similarity between the images and text in the joint embedded space. In this work, we use cosine similarity in the representation space to calculate similarity, which is widely used in learning image-text embedding and shown to be very effective in many prior works [8, 30, 60]. However, note that our approach does not depend on any particular choice of similarity function.

The first term in Eq. (1) represents the sum over all non-matching text embedding s^- which attempts to ensure that for each visual feature, corresponding/matching text features should be closer than non-matching ones in the joint space. Similarly, the second term attempts to ensure that text embedding that corresponds to the image embedding should be closer in the joint space to each other than non-matching image embeddings.

Recently, focusing on hard-negatives has been shown to be effective in learning joint embeddings [8, 44, 56, 60]. Subsequently, the loss in Eq. 1 is modified to focus on hard negatives (i.e., the negative closest to each positive (i,s) pair) instead of sum over all negatives in the formulation. For a positive pair (i,s), the hardest negative sample can be identified using $\hat{i} = \arg\max_{i^-} f(s,i^-)$ and $\hat{s} = \arg\max_{s^-} f(i,s^-)$. The loss function can be written as follows,

$$\mathcal{L}_{IS} = \sum_{(i,s)} \Big\{ max\big[0,\ \Delta - f(i,s) + f(i,\hat{s})\big] + max\big[0,\ \Delta - f(s,i) + f(s,\hat{i})\big] \Big\} \tag{2}$$

We name Eq. 1 as VSE loss and Eq. 2 as VSEPP loss. We utilize both of these loss functions in evaluating our proposed approach.

3.3 Training Joint Embedding with Web Data

In this work, we try to utilize image-tag pairs from the web for improving joint embeddings trained using a clean dataset with images-sentence pairs. Our aim is to learn a good representation for image-text embedding that ideally ignores the data-dependent noise and generalizes well. Utilization of web data effectively increases the sample size used for training our model and can be considered as implicit data augmentation. However, it is not possible to directly update the embedding (Sec. 3.2) using image-tag pairs. GRU based approach is not suitable for representing tags since tags do not have any semantic context as in the sentences.

Our task can also be considered from the perspective of learning with side or privileged information strategies [45, 51], as in our case an additional tag modality is available at training time and we would like to utilize this extra information to train a stronger model. However, directly employing LUPI strategies are also not possible in

our case as the training data do not provide information for all three modalities at the same time. The training datasets (e.g., MSCOCO, Flickr30K) provide only image-sentence pairs and do not provide tags. On the other hand, a web source usually provides images with tags, but no sentence descriptions. To bridge this gap, we propose a two-stage approach to train the joint image-text representation. In the first stage, we leverage the available clean image-text pairs from a dataset to learn an aligned representation that can be shared across three modalities (e.g., image, tag, text). In the second stage, we adapt the model trained in the first stage with web data.

Stage I: Training Initial Embedding. We leverage image-text pairs from an annotated dataset to learn a joint embedding for image, tags, and text. As tags are not available directly in the datasets, we consider nouns and verbs from the relevant sentence as dummy tags for an image (Fig. 3). For learning the shared representation, we combine the image-text ranking loss objective (Sec. 3.2), with image-tag ranking loss objective. We believe combining image-tag ranking loss objective provides a regularization effect in training that leads to more generalized image-text embedding.

Now the goal is to learn a joint embedding characterized by θ (i.e., $W^{(i)}$, $W^{(t)}$, $W^{(s)}$ and GRU weights) such that the image, sentence, and tags are projected into the joint space. Here, $W^{(t)}$ projects the representation of tags $\bar{t}$ on the joint space as, $t = W^{(t)}\bar{t}$. The resulting loss function can be written as follows,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{IS} + \lambda_2 \mathcal{L}_{IT} \tag{3}$$

where, $\mathcal{L}_{IT}$ represent image-tag ranking loss objective, which is similar to image-sentence ranking loss objective $\mathcal{L}_{IS}$ in Sec. 3.2. Similar to VSEPP loss in Eq. 2, $\mathcal{L}_{IT}$ can be written as,

$$\begin{aligned}\mathcal{L}_{IT} = \sum_{(i,t)} \Big\{ & max\big[0,\ \Delta - f(i,t) + f(i,\hat{t})\big] \\ & +\ max\big[0,\ \Delta - f(t,i) + f(t,\hat{i})\big]\Big\}\end{aligned} \tag{4}$$

where for a positive image-tag pair (i, t), the hardest negative sample tag representation can be identified as $\hat{t}$. Note that, all tags associated with an image are considered for generating tag representation in creating an image-tag pair rather than considering a single tag related to that image. In Eq. 3, λ_1 and λ_2 are predefined weights for different losses. In the first training stage, both losses are used ($\lambda_1 = 1$ and $\lambda_2 = 1$) while in the second stage, the image-text loss is not used ($\lambda_1 = 0$ and $\lambda_2 = 1$).

Stage II: Model Adaptation with Web Data. After Stage I converges, we have a shared representation of image, sentence description and tags with a learned image-tag embedding model. In Stage II, we utilize weakly-annotated image-tags pairs from Flickr to update the previously learned embedding network using $\mathcal{L}_{IT}$ loss. This enables us to transfer knowledge from thousands of freely available weakly annotated images in learning the embedding. We utilize a smaller learning rate in Stage II, as the network achieves competitive performance after Stage I and tuning the embedding network with a high learning rate from weakly-annotated data may lead to catastrophic forgetting [25].

As web data is very prone to label noise, we found it is hard to learn good representation for our task in many cases. Hence, in Stage II, we adopt a curriculum learning-based strategy in training. Curriculum learning allows the model to learn from easier instances first so they can be used as building blocks to learn more complex ones, which leads to a better performance in the final task. It has been shown in many previous works that appropriate curriculum strategies guide the learner towards better local minima [1]. Our idea is to gradually inject difficult information to the learner such that in the early stages of training, the network is presented with images related to frequently occurring concepts/keywords in the clean training set. Images related to rarely occurring concepts are presented at a later stage. Since the network trained in Stage I is more likely to have learned well about frequently occurring concepts, label noise is less likely to affect the network.

4 EXPERIMENTS

We perform experiments on two standard benchmark datasets with the main goal of analyzing the performance of different supervised methods by utilizing large scale web data using our curriculum guided webly supervised approach. Ideally, we would expect an improvement in performance irrespective of the loss function and features used to learn the embedding in Sec. 3.

We first describe the details on the datasets and evaluation metric in Sec. 4.1 and training details in Sec. 4.2. We report the results of different methods on MSCOCO dataset in Sec. 4.3 and results on Flickr30K dataset in Sec. 4.4.

4.1 Datasets and Evaluation Metric

We present experiments on standard benchmark datasets for sentence-based image description: MSCOCO Dataset [3] and Flickr30K dataset [42] to evaluate the performance of our proposed framework.

MSCOCO. The MSCOCO is a large-scale sentence-based image description dataset. This is the largest image captioning dataset in terms of the number of sentences and the size of the vocabulary. This dataset contains around 123K images. Each image comes with 5 captions. Following [23], we use the training, testing and validation split. In this split, the set contains 82,783 training images, 5000 validation images and 5000 test images. About 30K images were left out in this split. Some previous works utilize this images with for training to improve accuracy. We also report results using this images in training. In most of the previous works, the results are reported by averaging over 5 folds of 1K test images [7, 30, 54].

Flickr30K. Flickr30K is another standard benchmark dataset for sentence-based image description. Flickr30K dataset has a standard 31,783 images and 158,915 English captions. Each image comes with 5 captions, annotated by AMT workers. In this work, we follow the dataset split provided in [23]. In this dataset split, the training set contains 29,000 images, validation set contains 1000 and test set contains 1000 images.

Web Image Collection. We use photo-sharing website Flickr to retrieve web images with tags and use those images without any additional manual labeling. To collect images, we create a list of 1000 most occurring keywords in MSCOCO and Flickr30K dataset text descriptions and sort them in descending order based on frequency. We remove stop-words and group similar words together after performing lemmatization. We then use this list of keywords to query Flickr and retrieve around 200 images per query, together

Table 1: Image-to-Text Retrieval Results on MSCOCO Dataset.

#	Method	Image-to-Text Retrieval			Text-to-Image Retrieval		
		R@1	R@10	Med R	R@1	R@10	Med R
1.1	Embedding-Net	54.9	92.2	-	43.3	87.5	-
	2Way-Net	55.8	-	-	39.7	-	-
	Sm-LSTM	53.2	91.5	1.0	40.7	87.4	2.0
	Order-Embedding	46.7	88.9	2.0	37.9	85.9	2.0
	SAE–VGG19	46.8	87.7	2.0	35.8	82.9	2.4
	SAE–ResNet152	59.2	95.2	1.0	44.7	88.4	2.0
1.2	VSE–VGG19	46.8	89.0	1.8	34.2	83.6	2.6
	VSEPP–VGG19	51.9	90.4	1.0	39.5	85.6	2.0
	VSE–ResNet152	52.7	91.8	1.0	36.0	85.5	2.2
	VSEPP–ResNet152	58.3	93.3	1.0	43.6	87.8	2.0
1.3	Ours (VSE–VGG19)	47.2	90.9	1.6	35.1	85.3	2.0
	Ours (VSEPP–VGG19)	53.7	92.5	1.0	41.2	89.7	2.0
	Ours (VSE–ResNet152)	52.9	94.3	1.0	42.2	89.1	2.0
	Ours (VSEPP–ResNet152)	61.5	96.1	1.0	46.3	89.4	2.0

with their tags. In this way, we collect about 210,000 images with tags. We only collect images having at least two English tags and we don't collect more than 5 images from a single owner. We also utilize first 5 tags to remove duplicate images.

Evaluation Metric. We use the standard evaluation criteria used in most prior work on image-text retrieval task [6, 8, 30]. We measure rank-based performance by Recall at K ($R@K$) and Median Rank($MedR$). $R@K$ calculates the percentage of test samples for which the correct result is ranked within the top-K retrieved results to the query sample. We project sentences, tags, and images into the embedded space and calculate similarity scores. We report results of $R@1$ and $R@10$. Median Rank calculates the median of the ground-truth matches in the ranking results.

4.2 Training Details

We start training with a learning rate of 0.0002 and keep the learning rate fixed for 10 epochs. We then lower the learning rate by a factor of 10 every 10 epochs. We continue training Stage I for initial 20 epochs. Then we start updating the learned model in Stage I with web images in Stage II for another 20 epochs. The embedding networks are trained using ADAM optimizer [29]. Gradients are clipped when the $L2$ norm of the gradients (for the entire layer) exceeds 2. We tried different values for margin Δ in training and empirically choose Δ as 0.2, which we found performed well consistently on the datasets. We evaluate the model on the validation set after every epoch. The best model is chosen based on the sum of recalls in the validation set to deal with the over-fitting issue. We use a batch-size of 128 in the experiment. We also tried with other mini-batch sizes of 32 and 64 but didn't notice significant impact on the performance. We used two Telsa K80 GPUs and implemented the network using PyTorch toolkit.

4.3 Results on MSCOCO Dataset

We report the result of testing on MSCOCO dataset [35] in Table 1. To understand the effect of the proposed webly supervised approach, we divide the table in 3 rows (1.1-1.3). We compare our results with several representative image-text retrieval approaches, i.e., Embedding-Net [53], 2Way-Net [7], Sm-LSTM [19], Order-Embedding [52], SAE [13], VSE [30] and VSEPP [8]. For these approaches, we directly cite scores from respective papers when available and select the score of the best performing method if scores for multiple models are reported.

In row-1.2, we report the results on applying two different variants of pair-wise ranking loss based baseline VSE and VSEPP with two different feature representation from [8]. VSE[30] is based on the basic triplet ranking loss similar to Eq. 1 and VSEPP[8] is based on the loss function that emphasizes hard-negatives as shown in Eq. 2. We consider VSE and VSEPP loss based formulation as the main baseline for this work. Finally, in row-1.3, results using the proposed approach are reported. To enable a fair comparison, we apply our webly supervised method using the same VSE and VSEPP loss used by methods in row-1.2.

Effect of Proposed Webly Supervised Training. For evaluating the impact of our approach, we compare results reported in row-1.2 and row-1.3. Our method utilizes the same loss functions and features used in row-1.2 for a fair comparison. From Table 1, We observe that the proposed approach improves performance consistently in all the cases. For the retrieval task, the average performance increase in text-to-image retrieval is 7.5% in R@1 and 3.2% in R@10.

We also compare the proposed approach with web supervised approach SAE[13] (reported in row-1.1). In this regard, we implement SAE based webly supervised approach following [13]. We use the same feature and VSEPP ranking loss for a fair comparison and follow the exact same settings for experiments. We observe that our approach consistently performs better.

Effect of Loss Function. While evaluating the performance of different ranking loss, we observe that our webly supervised approach shows performance improvement for both VSE and VSEPP based formulation, and the performance improvement rate is similar for both VSE and VSEPP (See row-1.2 and row-1.3). Similar to the previous works [8, 60], we also find that methods using VSEPP loss perform better than VSE loss. We observe that in the image-to-text retrieval task, the performance improvement using VSEPP based formulation is higher and in the text-to-image retrieval task, the performance improvement for VSE based formulation is higher.

Table 2: Image-to-Text Retrieval Results on Flickr30K Dataset.

#	Method	Image-to-Text Retrieval			Text-to-Image Retrieval		
		R@1	R@10	Med R	R@1	R@10	Med R
2.1	Embedding-Net	43.2	79.8	-	31.7	72.4	-
	2Way-Net	49.8	-	-	36.0	-	-
	Sm-LSTM	42.5	81.5	2.0	30.2	72.3	3.0
	Order-Embedding	43.8	83.0	2.0	32.7	73.9	4.0
	SAE – VGG19	32.8	70.3	3.0	25.2	63.5	5.0
	SAE – ResNet152	43.4	80.7	2.0	31.0	71.3	3.0
2.2	VSE –VGG19	29.8	71.9	3.0	23.0	61.0	6.0
	VSEPP–VGG19	31.9	68.0	4.0	26.8	66.8	4.0
	VSE–ResNet152	38.2	80.8	2.0	26.6	67.0	4.0
	VSEPP–ResNet152	43.7	82.1	2.0	32.3	72.1	3.0
2.3	Ours (VSE–VGG19)	32.4	74.1	3.0	24.9	64.3	5.0
	Ours(VSEPP–VGG19)	37.8	77.1	3.0	27.9	68.9	4.0
	Ours(VSE–ResNet152)	41.4	84.5	2.0	29.7	71.9	4.0
	Ours (VSEPP–ResNet152)	47.4	85.9	2.0	35.2	74.8	3.0

Figure 4: Examples of 4 test images from Flickr30K dataset and the top 1 retrieved captions for our web supervised VSEPP-ResNet152 and standard VSEPP-ResNet as shown in Table. 2. The value in brackets is the rank of the highest ranked ground-truth caption in retrieval. Ground Truth (GT) is a sample from the ground-truth captions. Image 1,2 and 4 show a few examples where utilizing our approach helps to match the correct caption, compared to using the typical approach.

Effect of Feature. For evaluating the impact of different image feature in our web-supervised learning, we compare VGG19 feature based results with ResNet152 feature based results. We find consistent performance improvement using both VGG19 and ResNet152 feature. However, the performance improvement is slightly higher when ResNet152 feature is used. In image-to-text retrieval, the average performance improvement in R@1 using ResNet152 feature is 4%, compared to 2.3% using VGG19 feature. In the text-to-image retrieval task, the average performance improvement in R@1 using ResNet152 feature is 11.18%, compared to 3.5% using VGG19 feature.

Our webly supervised learning approach is agnostic to the choice loss function used for cross-modal feature fusion and we believe more sophisticated ones will only benefit our approach. We use two different variants of pairwise ranking loss (VSE and VSEPP) in the evaluation and observe that our approach improves the performance in both cases irrespective of the feature used to represent the images.

4.4 Results on Flickr30K Dataset

Table 2 summarizes the results on Flickr30K dataset [42]. Similar to Table 1, we divide the table into 3 rows (2.1-2.3) to understand the effect of the proposed approach compared to other approaches. From Table 2, we have the following key observations: (1) Similar to the results on MSCOCO dataset, our proposed approach consistently improves the performance of different supervised method (row-2.2 and row-2.3) in image-to-text retrieval by a margin of about 3%-6% in R@1 and 3%-9% in R@10. The maximum improvement of 6%-9% is observed in the VSEPP-VGG19 case while the least

mean improvement of 4.8% is observed in VSE-VGG19 case. (2) In text-to-image retrieval task, the average performance improvement using our webly-supervised approach are 2.25% and 3.25% in R@1 and R@10 respectively. These improvements once again show that learning by utilizing large scale web data covering a wide variety of concepts lead to a robust embedding for cross-modal retrieval tasks. In Fig. 4, we show examples of few test images from Flickr30K dataset and the top 1 retrieved captions for the VSEPP-ResNet152 based formulations.

5 CONCLUSIONS

In this work, we showed how to leverage web images with tags to assist training robust image-text embedding models for the target task of image-text retrieval that has limited labeled data. We attempt to address the challenge by proposing a two-stage approach that can augment a typical supervised pair-wise ranking loss based formulation with weakly-annotated web images to learn better image-text embedding. Our approach has benefits in both performance and scalability. Extensive experiments demonstrate that our approach significantly improves the performance in the image-text retrieval task in two benchmark datasets. Moving forward, we would like to improve our method by utilizing other types of metadata (e.g., social media groups, comments) while learning the multi-modal embedding. Furthermore, the objective of webly supervised learning may suffer when the amount of noisy tags associated with web images is unexpectedly high compared to clean relevant tags. In such cases, we plan to improve our method by designing loss functions or layers specific to noise reduction, providing a more principled way for learning the multi-modal embedding in presence of significant noise.

Acknowledgement. This work was partially supported by NSF grants IIS-1746031 and CNS-1544969. We thank Sujoy Paul for helpful suggestions and Victor Hill for setting up the computing infrastructure used in this work.

REFERENCES

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*. ACM, 41–48.
[2] Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 2. 164–169.
[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).
[4] Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2017. Textually Customized Video Summaries. *arXiv preprint arXiv:1702.01528* (2017).
[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
[6] Jianfeng Dong, Xirong Li, and Cees GM Snoek. 2016. Word2VisualVec: Image and video to sentence matching by visual feature prediction. *CoRR, abs/1604.06838* (2016).
[7] Aviv Eisenschtat and Lior Wolf. 2017. Linking Image and Text With 2-Way Nets. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4601–4611.
[8] Fartash Faghri, David J. Fleet, Ryan Kiros, and Sanja Fidler. 2017. VSE++: Improved Visual-Semantic Embeddings. *CoRR* abs/1707.05612 (2017). arXiv:1707.05612 http://arxiv.org/abs/1707.05612
[9] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*. Springer, 15–29.
[10] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129.
[11] Dihong Gong, Daisy Zhe Wang, and Yang Peng. 2017. Multimodal Learning for Web Information Extraction. In *Proceedings of the 2017 ACM Multimedia Conference*. ACM, 288–296.
[12] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision* 106, 2 (2014), 210–233.
[13] 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.
[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[15] Christian Andreas Henning and Ralph Ewerth. 2017. Estimating the information gap between textual and visual representations. In *International Conference on Multimedia Retrieval*. ACM, 14–22.
[16] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.
[17] Zeyuan Hu and Julia Strout. 2018. Exploring Stereotypes and Biased Data with the Crowd. *arXiv preprint arXiv:1801.03261* (2018).
[18] Feiran Huang, Xiaoming Zhang, Zhoujun Li, Tao Mei, Yueying He, and Zhonghua Zhao. 2017. Learning Social Image Embedding with Deep Multimodal Attention Networks. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. ACM, 460–468.
[19] Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-aware image and sentence matching with selective multimodal lstm. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2310–2318.
[20] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. 2017. Learning Robust Visual-Semantic Embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3571–3580.
[21] Mark J Huiskes and Michael S Lew. 2008. The MIR flickr retrieval evaluation. In *International Conference on Multimedia Information Retrieval*. ACM, 39–43.
[22] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. 2016. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*. Springer, 67–84.
[23] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3128–3137.
[24] Andrej Karpathy, Armand Joulin, and Fei Fei F Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*. 1889–1897.
[25] Ronald Kemker, Angelina Abitino, Marc McClure, and Christopher Kanan. 2017. Measuring Catastrophic Forgetting in Neural Networks. *arXiv preprint arXiv:1708.02072* (2017).
[26] Tom Kenter, Alexey Borisov, and Maarten de Rijke. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640* (2016).
[27] Tom Kenter and Maarten de Rijke. 2015. Short text similarity with word embeddings. In *ACM Int. Conf. Information and Knowledge Management*. 1411–1420.
[28] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. 2012. Undoing the damage of dataset bias. In *European Conference on Computer Vision*. Springer, 158–171.
[29] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[30] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).
[31] 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*. IEEE, 4437–4446.
[32] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. 2016. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*. Springer, 301–320.
[33] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2017. Learning visual n-grams from web data. In *International Conference on Computer Vision*.
[34] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2017. Attention Transfer from Web Images for Video Recognition. In *Proceedings of the 2017 ACM Multimedia Conference*. ACM, 1–9.
[35] 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.
[36] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632* (2014).

[37] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[38] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. 2018. Learning Joint Embedding with Multimodal Cues for Cross-Modal Video-Text Retrieval. In *International Conference on Multimedia Retrieval*. ACM, 19–27.

[39] Niluthpol Chowdhury Mithun, Rameswar Panda, and Amit K Roy-Chowdhury. 2016. Generating diverse image datasets with limited labeling. In *Proceedings of the 2016 ACM Multimedia Conference*. ACM, 566–570.

[40] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*. 299–307.

[41] Bryan Plummer, Matthew Brown, and Svetlana Lazebnik. 2017. Enhancing Video Summarization via Vision-Language Embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

[42] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision*. IEEE, 2641–2649.

[43] Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* (2017).

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.

[45] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. 2013. Learning to rank using privileged information. In *International Conference on Computer Vision*. IEEE, 825–832.

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

[47] Richard Socher and Li Fei-Fei. 2010. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 966–973.

[48] 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 2015 ACM Multimedia Conference*. ACM, 371–380.

[49] Antonio Torralba, Alexei Efros, et al. 2011. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1521–1528.

[50] Emiel van Miltenburg. 2016. Stereotyping and bias in the flickr30k dataset. *arXiv preprint arXiv:1605.06083* (2016).

[51] Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural networks* 22, 5-6 (2009), 544–557.

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

[53] 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).

[54] 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).

[55] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5005–5013.

[56] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. 2017. Sampling Matters in Deep Embedding Learning. In *IEEE International Conference on Computer Vision*. IEEE, 2859–2867.

[57] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3441–3450.

[58] Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632* (2014).

[59] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).

[60] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2017. Dual-Path Convolutional Image-Text Embedding. *arXiv preprint arXiv:1711.05535* (2017).