2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Webly Supervised Image-Text Embedding with Noisy Tag Refinement

Niluthpol C Mithun‡, Ravdeep Pasricha†, Evangelos Papalexakis†, and Amit K. Roy-Chowdhury†
†University of California, Riverside, CA, ‡Center for Vision Technologies, SRI International, Princeton, NJ
Email: niluthpol.mithun@sri.com, rpasr001@ucr.edu, epapalex@cs.ucr.edu, amitrc@ece.ucr.edu

***Abstract*—In this paper, we address the problem of utilizing web images in training robust joint embedding models for the image-text retrieval task. Prior webly supervised approaches directly leverage weakly annotated web images in the joint embedding learning framework. The objective of these approaches would suffer significantly when the ratio of noisy and missing tags associated with the web images is very high. In this regard, we propose a CP decomposition based tensor completion framework to refine the tags of web images by modeling observed ternary inter-relations between the sets of labeled images, tags, and web images as a tensor. To effectively deal with the high ratio of missing entries likely in our case, we incorporate intra-modal correlation as side information in the proposed framework. Our tag refinement approach combined with existing web supervised image-text embedding approaches provide a more principled way for learning the joint embedding models in the presence of significant noise from web data and limited clean labeled data. Experiments on benchmark datasets demonstrate that the proposed approach helps to achieve a significant performance gain in image-text retrieval.**

I. Introduction

Cross-modal retrieval between visual data and natural language description has gained considerable momentum in recent years due to the the success of deep neural networks in learning aligned representations across modalities [1], [2], [3], [4]. The deep network based approaches often suffer from the requirement of huge amount of labeled cross-modal samples for training robust models. However, constructing datasets containing large number of annotated image-text pairs is extremely labor-intensive and prone to errors. In practical scenarios, given labeling budget, it is generally feasible to annotate only a few training images with text. Hence, although existing image-text retrieval models show promising results on datasets, they are unlikely to generalize well to uncontrolled scenarios typical in both social media and sensor networks as they are reliant on large volumes of training data that closely mimic what is expected in the test cases.

To deal with the issue of limited training data and inspired by the availability of streams of images with associated text in the web, a few methods [5], [6] have explored the use of web images with fully annotated datasets in training improved joint image-text embedding models (See Fig. 1 for a brief illustration of webly supervised image-text embedding task). Although these approaches [5], [6] show performance improvement in retrieval task, the improvement is quite limited considering the utilization of large amount of additional web data. The raw tags associated with web images are often incomplete and error-prone. Hence, directly utilizing such data without any refinement (as in the prior works [5], [6]) in the objective of learning may lead to an increased ambiguity and degraded performance, when the amount of noisy tags associated with web images is unexpectedly high compared to clean relevant tags [5]. The learning approach should be able to deal with huge amount missing information as encountered frequently in our setting (i.e., most social media images may contain a few relevant tags). These challenges make the problem of learning robust joint embedding models using web images extremely difficult.

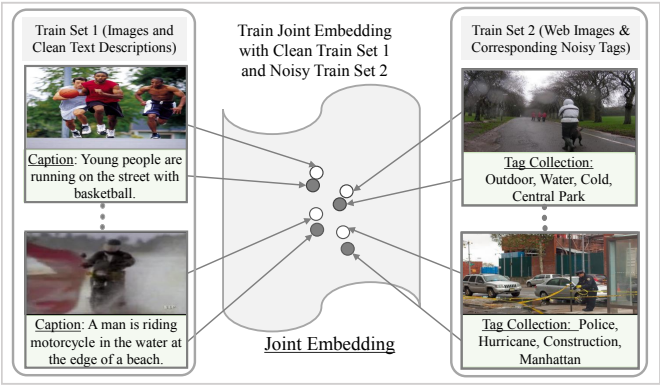

Fig. 1. Brief illustration of webly supervised learning of image-text embedding: the goal is to utilize freely available web images with noisy tags along with a small dataset of images with manually annotated text in learning.

Motivated by above, we explore the research question - *Based on a limited fully annotated set of images with textual descriptions, is it possible to refine the tags of web image and utilize them in boosting the performance of joint image-text embedding models?* For example, can we build a reasonable joint image-text embedding model when we have access to only 5% of labeled data from image-text datasets (e.g., MSCOCO) and the remaining 95% data are weakly annotated? Although, existing largest image-text datasets cover a limited number of image-text (e.g., about 500k in MSCOCO and 150K pairs Flickr30K), it is critical to consider availability of a significantly smaller number of cross-modal pairs (e.g., 2K pairs) focusing on specific practical applications, such as cross-modal retrieval focusing on a sudden emergency scenario. In such a case, it is extremely crucial to complement scarcer clean set of pairs with freely available web images to improve the performance of image-text embedding models. However,

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412106

availability of a small clean set makes it extremely difficult to train a reliable model, considering significantly high amount of noisy and missing entries typical in web image tagging.

Towards tackling the challenges, our idea is to first refine the tags of weakly annotated web image collection utilizing their latent relationships with the small clean set of images. The two set of image collections can be inter-related easily based on associated tags, however, we can only have partial observations of the relationships due to the noisy nature of web image tags. We propose to utilize the observed incomplete relationships in a tensor completion framework to predict the missing tags and remove the noisy ones. The proposed image tag refinement approach is motivated by the success of tensor completion approaches in multi-way data analysis [7], [8], [9]. In this work, we formulate the web image-tag refinement as a CP decomposition based tensor completion approach that leverages ternary interactions among dataset images, tags and web images in refining web image tags. To efficiently recover missing dynamics, we also incorporate intra-modal similarity as auxiliary information to regularize the tensor completion problem. Refined web images are then used with webly supervised learning frameworks for training joint image-text embeddings. Experiments demonstrate that our approach is successful in cleaning tags and improving the performance of joint embedding models significantly.

Contributions. The main contributions of this work can be summarized as follows: **First**, we present an efficient framework for learning of image-text embedding in the presence of limited clean labeled data and web images containing significant noise. In the framework, the web images associated with noisy tags are first refined using the proposed tensor completion approach and then used with a small clean dataset in webly supervised learning frameworks for training joint image-text embedding models. **Second**, we propose to refine tags of web images by modeling the inter-relation between web image collection and clean dataset images (based on associated tags) as a tensor and utilizing intra-modal similarity as side information in a CP decomposition based tensor completion framework. **Third**, experiments demonstrate that the proposed approach is promising in refining tags from web image collection and helps to train improved joint image-text embedding models with limited clean labeled data.

II. Related Works

Cross-modal image-text retrieval. Cross-modal image/video to text retrieval has received intense attention recently [1], [3], [10], [11], [12]. Learning joint embedding models are very popular in this task since it is possible to compare representations from different modalities in such a joint space [3], [13], [4], [14], [15]. Canonical correlation analysis (CCA) based approaches is a popular choice in learning joint image-text embeddings and have also been explored in several prior works [16], [17], [18].

In recent years, triplet ranking loss based approaches have been widely used in learning joint visual-semantic embedding [1], [19], [3], [13], [20], [4], [15] and have achieved state-of-the-art performance in cross-modal retrieval tasks [3], [20], [4], [14], [10]. Although these supervised approaches show strong performance on benchmark datasets, they would suffer in practical scenarios where it is possible to label only a few examples with a given labeling budget and time.

Web supervised image-text retrieval. It is extremely difficult and labour-intensive to create large-scale datasets by manually annotating images [21], [5]. Moreover, it has been reported that models trained on dataset images fails to generalize well across datasets [22], [23], [24]. On the other hand, several recent works have shown that utilizing large-amount of weakly labeled training data can be very effective in training powerful models for many multimedia and computer vision tasks [6], [25], [26], [27], [28].

Prior works [6] and [5] attempted to improve image-text embedding models using weakly annotated image collection. In [6], authors considered a stacked joint embedding approach for incorporating information from weakly annotated Flickr1M dataset in learning image-text embedding. In [5], authors presented an approach that can augment a typical pairwise ranking based supervised approach for joint image-text embedding with web images and tags. Another relevant work is [29], which does not use web data with tags, but proposes an approach to use a un-annotated image collection with an annotated collection in learning joint embedding models. These approaches [6], [5], [29] are unlikely to be successful when either the clean training set is very small, or the web image collection has a high ratio of noisy and missing tags.

Tensor completion for multi-modal data analysis. Tensor completion approaches focus on estimating the missing elements of partially observed tensors [30]. CP decomposition [31], [32] and Tucker decomposition [33], [34] are most widely used approaches for low-rank decomposition of tensors. There are several works on completing tensors to estimate missing data based on tensor decomposition [9], [35], [36], [37]. In this work, we follow [37], [38] and develop a tensor decomposition based tensor completion approach. We use CP decomposition as it has been found that Tucker decomposition based approaches are computationally less flexible than CP approaches in handling large datasets in a distributed manner as it needs to deal with complex core tensor [30].

There have been a few works on exploiting tensor decomposition in tag refinement [36], [8], [39], [40], [41]. Most of these works [36], [8], [39], [40] assume the availability of additional user information along with images and tags from social media sources and utilize Tucker decomposition based approach for refinement. Although user information may provide important cues in refining tags, it is unlikely to be available in most cases. In this work, we explore the use of a small clean dataset containing images and tags in refining web image tags so that we can limit the propagation of noisy tags in recovering missing tags. Several previous works have shown that utilizing relationships among data sources as auxiliary information helps to improve the quality of tensor

decomposition significantly when limited entries are observed [9], [42], [38], [43]. Inspired by these works, we use intra-modal similarity matrices as side information in the proposed approach to deal with a high ratio of missing entries.

III. Approach

Our approach consists of two major steps. First, we present the proposed tensor completion based approach for refining web image tags. Then, we present the approach for learning joint image-text embedding using pair-wise ranking loss.

A. Image Tag Refinement

The intuition is multi-dimensional relation that exists between web image with noisy tags and images with clean tags can be modeled as a tensor. Analyzing the tensor can be beneficial in refining web image tags. We consider having three types of entities (i.e., web images, dataset images, and selected tags) and the ternary relationship (based on tag association) among the entities is modeled as a tensor. We follow AirCP framework for our image-tag refinement approach [37]. A brief illustration of the approach is shown in Fig. 2. We start by giving notations and then present the approach.

Preliminaries: Throughout this paper, we use calligraphic bold uppercase letters to denote tensors, uppercase letters to denote matrices and lowercase letters to denote vectors. For a third order tensor $\boldsymbol{\mathcal{Y}}$, its entries are denoted by $\mathcal{Y}_{ijk}$. The Frobenius norm of $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{|D|\times|W|\times|T|}$ is defined as $||\boldsymbol{\mathcal{Y}}||_F = \sqrt{\sum_{i=1}^{|D|}\sum_{j=1}^{|W|}\sum_{k=1}^{|T|}\mathcal{Y}_{ijk}^2}$.

The CP tensor decomposition aims to approximate an order-N tensor with R latent factors as a sum of R rank-one tensors [44], [45]. For a third order tensor $\boldsymbol{\mathcal{Y}}$, it can be written as :

$$\boldsymbol{\mathcal{Y}} \approx \tilde{\boldsymbol{\mathcal{Y}}} = [[Z^{(1)}, Z^{(2)}, Z^{(3)}]]$$

Here, $[[Z^{(1)}, Z^{(2)}, Z^{(3)}]]$ represents a weighted sum of rank-1 tensors where the vectors that specify the rank-1s are columns of the factor matrices $Z^{(1)}$, $Z^{(2)}$, and $Z^{(3)}$.

*1) **Tag Refinement using CP tensor completion model**:* We consider that we have access to three types of data, i.e., the images from dataset $D = \{d_i\}_{i=1}^{|D|}$ (for which we know the associated tags correctly), the images from the web $W = \{n_i\}_{j=1}^{|W|}$ (for which we know *some* associated *noisy* tags), and the selected tag set $T = \{t_i\}_{k=1}^{|T|}$. $\boldsymbol{\mathcal{Y}}$ denotes the tensor with complete tri-modal dynamics. Since very few tags are found in most images, $\boldsymbol{\mathcal{Y}}$ is likely to be sparse and low-rank. If the i-th image from the dataset and the j-th image from web image collection are both annotated with the kth tag from the selected tag set, $\mathcal{Y}_{ijk} = 1$. Otherwise, $\mathcal{Y}_{ijk} = 0$. However, as web images mostly have a few associated noisy tags, we only have a partial observation of $\boldsymbol{\mathcal{Y}}$ at the start.

Our goal is to refine tags of web image set W by predicting missing tags. We propose to model the recovery of missing tags based on CP tensor completion model (1) as follows:

$$\min_{Z^{(n)},\boldsymbol{\mathcal{Y}}} \frac{1}{2}||\boldsymbol{\mathcal{Y}} - [[Z^{(1)}, Z^{(2)}, Z^{(3)}]]\ ||_F^2 + \frac{\lambda}{2}\sum_{n=1}^{3}||Z^{(n)}||_F^2; \quad \text{s.t. } \boldsymbol{\Upsilon} * \boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{O}}, \tag{1}$$

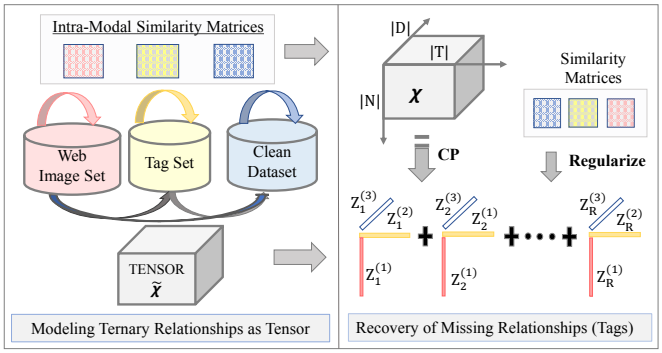

Fig. 2. Brief Illustration of our CP decomposition based Tensor Completion approach for Image-Tag Refinement.

$\boldsymbol{\mathcal{O}}$ is the observations for $\boldsymbol{\mathcal{Y}}$. The latent factor matrices for the clean dataset images, web image set, and tag set are denoted respectively by $Z^{(1)} \in \mathbb{R}^{|D|\times R}, Z^{(2)} \in \mathbb{R}^{|W|\times R}$ and $Z^{(3)} \in \mathbb{R}^{|T|\times R}$. Here, R is the number of latent factors. $\boldsymbol{\Upsilon}$ is a non-negative weight tensor with the same size as $\boldsymbol{\mathcal{Y}}$. If $\mathcal{Y}_{ijk}$ is observed, $\boldsymbol{\Upsilon}_{\mathbf{ijk}} = \mathbf{1}$. Otherwise $\boldsymbol{\Upsilon}_{\mathbf{ijk}} = \mathbf{0}$.

Our goal is to estimate $\boldsymbol{\mathcal{Y}}$ for recovering the missing dynamics of tags based on the partially observed data. To effectively recover $\boldsymbol{\mathcal{Y}}$, we also consider intra-relationships in the three types of data as side information as described below.

*2) **Regularize model with auxiliary information**:* In our case, we not only have ternary relational information among inter-modal objects, but also information on the objects (images and tags) themselves. We consider using intra-modal relations between entities as additional side information in our tensor factorization framework. We can calculate the intra-modal relationship between images based on image similarity measures. Similarly, we can model the relationship between tags by calculating the similarity between tags. In this work, we use the cosine similarity. Given, feature representation of images, the similarity between the images of the dataset can be calculated as $\Gamma_{Dataset}(i,j) = d_i^T d_j/(||d_i||_2\ ||d_j||_2)$. Γ_{Web} and Γ_{Tag} similarity matrices are also calculated in a similar way using cosine similarity measure. We utilize these similarity matrices as auxiliary information. The idea is that if two images are similar, the latent representations of these two images should be similar. Therefore, we want to make the latent representations of two similar entities to be close. This can be obtained by minimizing the following loss:

$$\begin{aligned} L_{AUX} &= \sum_{i,j}\Gamma(i,j)\ ||Z_{i,:}^{(n)} - Z_{j,:}^{(n)}||^2 \\ &= \sum_{i,j} {Z_{i,:}^{(n)}}^T \Gamma(i,j) Z_{i,:}^{(n)} - \sum_{i,j} {Z_{i,:}^{(n)}}^T \Gamma(i,j) Z_{j,:}^{(n)} \\ &= tr({Z^{(n)}}^T \mathcal{L}\ Z^{(n)}) \end{aligned}$$

where $Z_{i,:}^{(n)}$ is the ith row of the factor matrix $Z^{(n)}$ for the nth mode of tensor $\boldsymbol{\mathcal{Y}}$ ($n \in \{1,2,3\}$). D is a diagonal matrix with $D_{ij} = \sum_j \Gamma_{ij}$ and $\mathcal{L} = D - \Gamma$ is the Laplacian of similarity

matrix Γ. Adding auxiliary information, the Eq. 1 becomes:

$$\min_{Z^{(n)},\boldsymbol{\mathcal{Y}}} \frac{1}{2}||\boldsymbol{\mathcal{Y}} - [[Z^{(1)}, Z^{(2)}, Z^{(3)}]]\ ||_F^2 + \frac{\lambda}{2}\sum_{n=1}^{3}||Z^{(n)}||_F^2 + \sum_{n=1}^{3}\alpha_n tr(Z^{(n)^T}\mathcal{L}_n\ Z^{(n)}); \quad \text{s.t. } \boldsymbol{\Upsilon} * \boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{O}}, \tag{2}$$

α_n is a hyper-parameter to control the weight of auxiliary information from different factors.

*3) **ADMM Optimization**:* We use alternating direction method of multipliers (ADMM) [46], [47], [37] to solve our optimization problem in Eq.2. The ADMM algorithm consists of three main steps. First, an auxiliary variable is introduced to separate the objective function into two different objectives. Second, an augmented Lagrangian is formed combining both linear and quadratic terms through a scaled dual variable. Third, the augmented Lagrangian is minimized iteratively with respect to the primal variables and the dual variable until convergence. To facilitate optimization, we consider an equivalent form of Eq. 2 introducing an auxiliary variable U:

$$\min_{Z^{(n)},\boldsymbol{\mathcal{Y}}} \frac{1}{2}||\boldsymbol{\mathcal{Y}} - [[Z^{(1)}, Z^{(2)}, Z^{(3)}]]\ ||_F^2 + \frac{\lambda}{2}\sum_{n=1}^{3}||Z^{(n)}||_F^2 + \sum_{n=1}^{3}\alpha_n tr(U^{(n)^T}\mathcal{L}_n\ U^{(n)}); \quad \text{s.t. } \boldsymbol{\Upsilon} * \boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{O}}, \mathbf{Z^{(n)}} = \mathbf{U^{(n)}} \tag{3}$$

We can form augmented Lagrangian $L_\mu(U^{(n)}, Z^{(n)}, \Lambda^{(n)})$ with both linear and quadratic terms as follows:

$$L_\mu(U^{(n)}, Z^{(n)}, \Lambda^{(n)}, \boldsymbol{\mathcal{Y}}) = \frac{1}{2}||\boldsymbol{\mathcal{Y}} - [[Z^{(1)}, Z^{(2)}, Z^{(3)}]]\ ||_F^2 + \frac{\lambda}{2}\sum_{n=1}^{3}||Z^{(n)}||_F^2 + \sum_{n=1}^{3}\alpha_n tr(U^{(n)^T}\mathcal{L}_n\ U^{(n)}) + \frac{1}{2}||\boldsymbol{\Upsilon} * \boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{O}}||_{\mathbf{F}}^2 + \sum_{n=1}^{3} < \Lambda^{(n)}, Z^{(n)} - U^{(n)} > + \sum_{n=1}^{3}\frac{\mu}{2}||Z^{(n)} - U^{(n)}||_F^2 \tag{4}$$

where Λ is a dual variable, $\langle ., . \rangle$ denote the inner product, $\|.\|_F$ is the Frobenius norm and $\mu > 0$ is a penalty parameter.

To solve the problem in Eq. 4 at each iteration t, ADMM updates the variables in alternating fashion as:

$$U_{t+1}^{(n)} = \arg\min_{U^{(n)}} L_\mu(U_t, Z_t, \Lambda_t, \boldsymbol{\mathcal{Y}}_t) \tag{5}$$

$$Z_{t+1}^{(n)} = \arg\min_{Z^{(n)}} L_\mu(U_{t+1}, Z_t, \Lambda_t, \boldsymbol{\mathcal{Y}}_t) \tag{6}$$

$$\boldsymbol{\mathcal{Y}}_{t+1} = \arg\min_{\boldsymbol{\mathcal{Y}}} L_\mu(U_{t+1}, Z_{t+1}, \Lambda_{t+1}, \boldsymbol{\mathcal{Y}}_t) \tag{7}$$

$$\Lambda_{t+1}^{(n)} = \arg\min_{\Lambda} L_\mu(U_{t+1}, Z_{t+1}, \Lambda_t, \boldsymbol{\mathcal{Y}}_t) \tag{8}$$

In the following, we present the derivation of specific update rules for Eq. 6, Eq. 5, Eq. 8 and Eq.13.

Update $U^{(n)}$ when fixing others: To update $U^{(n)}$ (e.g., $U^{(1)}$ or $U^{(2)}$ or $U^{(3)}$) after ignoring the variables that are irrelevant to $U^{(n)}$, the problem (6) becomes:

$$\min_{U^{(n)}} \alpha_n tr(U^{(n)^T}\mathcal{L}\ U^{(n)}) + < \Lambda^{(n)}, Z^{(n)} - U^{(n)} > + \frac{\mu}{2}||Z^{(n)} - U^{(n)}||_F^2$$

On combining linear and quadratic error terms into a single term by scaling the dual variable Λ, we get the following:

$$\min_{U^{(n)}} \alpha_n tr(U^{(n)^T}\mathcal{L}_n\ U^{(n)}) + \frac{\mu}{2}\|Z^{(n)} - U^{(n)} + \Lambda^{(n)}/\mu\|_F^2 \tag{9}$$

It is a convex quadratic problem. Solving for $U^{(n)}$ yields:

$$U_{t+1}^{(n)} = (\mu I + \alpha_n\mathcal{L}_n)^{-1}(\mu Z_t^{(n)} - \Lambda_t^{(n)}) \tag{10}$$

Update $Z^{(n)}$ when fixing others: To update $Z^{(n)}$ ($n \in 1, 2, 3$), the method alternates among the modes, fixing every factor matrix but $Z^{(n)}$ and solving for it. The objective function can be written as follows:

$$\min_{Z^{(n)}} \frac{1}{2}||\mathcal{Y}_{(n)} - Z^{(n)}A^{(n)^T}||_F^2 + \frac{\lambda}{2}||Z^{(n)}||_F^2 + \frac{\mu}{2}\|Z^{(n)} - U^{(n)} + \Lambda^{(n)}/\mu\|_F^2 \tag{11}$$

Here, $\mathcal{Y}_{(n)}$ represents the mode-n matrix unfolding of tensor $\boldsymbol{\mathcal{Y}}$. the mode-n matricization of $\tilde{\boldsymbol{\mathcal{Y}}}$ can be written in terms the factor matrices as $\tilde{\mathcal{Y}_{(n)}} = Z^{(n)}A^{(n)^T}$ where $A^{(n)} = (Z^{(M)} \odot ..Z^{(n+1)} \odot Z^{(n-1)} \odot .. \odot Z^{(1)})|_{M=3}$. Here, $\odot$ denotes Khatri-Rao product. Now, solving Eq. 11 for $Z^{(n)}$ yields:

$$Z_{t+1}^{(n)} = (A^{(n)}\ A^{(n)^T} + \lambda I + \mu I)^{-1}\ (\mathcal{Y}_{(n)}^t\ A^{(n)^T} + \mu U_{t+1}^{(n)} + \Lambda_t^{(n)}) \tag{12}$$

Update $\mathcal{Y}$: To solve for $\boldsymbol{\mathcal{Y}}$, we can write the objective in Eq. 4 as follows:

$$\min_{\boldsymbol{\mathcal{Y}}} \frac{1}{2}||\boldsymbol{\mathcal{Y}} - [[Z^{(1)}, Z^{(2)}, Z^{(3)}]]\ ||_F^2 + \frac{1}{2}||\boldsymbol{\Upsilon} * \boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{O}}||_{\mathbf{F}}^2 \tag{13}$$

Now solving for $\boldsymbol{\mathcal{Y}}$ yields:

$$\boldsymbol{\mathcal{Y}}_{t+1} = \boldsymbol{\mathcal{O}} + (1 - \boldsymbol{\Upsilon}) * [[\mathbf{Z^{(1)}}, \mathbf{Z^{(2)}}, \mathbf{Z^{(3)}}]] \tag{14}$$

Update $\Lambda^{(n)}$: Having (U, Z) fixed, perform a gradient ascent update with step size μ on Lagrange multipliers as

$$\Lambda_{t+1}^{(n)} = \Lambda_t^{(n)} + \mu(U_{t+1}^{(n)} - Z_{t+1}^{(n)}) \tag{15}$$

The overall ADMM procedure is shown in Algo. 1. After convergence, we have the final completed tensor $\boldsymbol{\mathcal{Y}}$. From $\boldsymbol{\mathcal{Y}}$, we can recover the tags for our web image collection. Summing $\boldsymbol{\mathcal{Y}}$ over dataset image dimensions, we can have a matrix whose values indicate the strength of association between web images and tags.

Algorithm 1 An ADMM solver for (Eq. 4)

1: **Input:** $\boldsymbol{\mathcal{O}}$, $\boldsymbol{\Upsilon}$, $\Gamma^{(n)}$ and λ, $N = 3$, $n = 1 : N$, $\mu > 0$, $Th = 10^{-5}$, $nIter = 500$
2: **Initialization:** Initialize $U^{(n)}, Z^{(n)}, \Lambda^{(n)}, iter$ to zero, $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{O}}$.
3: **while** $(\max\{||Z^{(n)} - U^{(n)}||_F; n = 1, ..., N\} < Th)$ **or** $(iter \leq nIter)$ **do**
4: $\quad U_{t+1}^{(n)} \leftarrow (\mu I + \alpha_n\mathcal{L}_n)^{-1}(\mu Z_t^{(n)} - \Lambda_t^{(n)})$;
5: $\quad Z_{t+1}^{(n)} \leftarrow (A^{(n)}\ A^{(n)^T} + \lambda I + \mu I)^{-1}\ (\mathcal{Y}_{(n)}^t\ A^{(n)^T} + \mu U_{t+1}^{(n)} + \Lambda_t^{(n)})$;
6: $\quad \boldsymbol{\mathcal{Y}}_{t+1} \leftarrow \boldsymbol{\mathcal{O}} + (1 - \boldsymbol{\Upsilon}) * [[\mathbf{Z^{(1)}}, \mathbf{Z^{(2)}}, \mathbf{Z^{(3)}}]]$;
7: $\quad \Lambda_{t+1}^{(n)} \leftarrow \Lambda_t^{(n)} + \mu(U_{t+1}^{(n)} - Z_{t+1}^{(n)})$;
8: $\quad iter \leftarrow iter + 1$;
9: **end while**
10: **Output:** Tensor $\boldsymbol{\mathcal{Y}}$, Factor Matrices $Z^{(1)}$, $Z^{(2)}$ and $Z^{(3)}$.

B. Joint Embedding

Our framework includes a web image tag refinement approach in the webly supervised joint embedding framework. The framework attempts at attaining better performance compared to the image-text embedding baselines that directly uses raw image and tags in training without any refinement. Pairwise ranking loss has been shown to be successful in many prior works to learn robust joint embedding models for image-text retrieval [1], [3], [19], [13], and also in web supervised learning of joint embeddings [5]. We also optimize ranking loss function in learning webly supervised joint embedding models to show the benefits of the proposed tag refinement step in the overall image-text retrieval performance.

We now briefly describe the method for learning joint visual-semantic embedding utilizing pair-wise ranking loss with a two-branch neural network framework [1], [3], [5], [20]. Optimizing the pair-wise ranking loss attempts to ensure that for each visual input, the matching text inputs should be closer than the non-matching text inputs in the joint image-text embedding space or vice versa. In the two-branch framework, one of the branches consists of a neural network to extract features from text and the other consists of a network to extract features from visual inputs. These branches are followed by fully connected layers focusing on projecting the text and visual features to a joint space where it is possible to directly compare the cross-modal inputs.

We denote the image feature vector as $\boldsymbol{v}_p \quad (\boldsymbol{v}_p \in \mathbb{R}^V)$ and the paired text feature vector as $\boldsymbol{w}_p$ $(\boldsymbol{w}_p \in \mathbb{R}^T)$. Given the representation of training pairs, the goal is to learn a joint embedding characterized by W^V and W^T such that the matching image-text pairs are closer in the joint space than the non-matching pairs. W^V and W^T denotes the weights of fully connected embedding layers. D is the dimension of the joint space. The visual features $\boldsymbol{v}_p$ is projected into the embedding as $\boldsymbol{i}_p = W^V \boldsymbol{v}_p$ $(\boldsymbol{i}_p \in \mathbb{R}^D)$. Similiarly, and W^T that projects input text contents $\boldsymbol{w}_p$ to the joint space as $\boldsymbol{t}_p = W^T \boldsymbol{w}_p$ $(\boldsymbol{t}_p \in \mathbb{R}^D)$. For a matching pair $(\boldsymbol{i}_p, \boldsymbol{t}_p)$, $\boldsymbol{t}_p^-$ is a non-matching text embedding and $\boldsymbol{i}_p^-$ is a non-matching image embedding. The ranking loss function $\mathcal{L}_p$ for the matching image-text pair $(\boldsymbol{i}_p, \boldsymbol{t}_p)$ can be expressed as follows,

$$\begin{aligned}\mathcal{L}_p = &\sum_{\boldsymbol{t}_p^-} max\big[0, \alpha - S(\boldsymbol{i}_p, \boldsymbol{t}_p) + S(\boldsymbol{i}_p, \boldsymbol{t}_p^-)\big] \\ &+ \sum_{\boldsymbol{i}_p^-} max\big[0, \alpha - S(\boldsymbol{t}_p, \boldsymbol{i}_p) + S(\boldsymbol{t}_p, \boldsymbol{i}_p^-)\big]\end{aligned} \tag{16}$$

where $S(\boldsymbol{i}_p, \boldsymbol{t}_p)$ is the function to measure the similarity between the image embedding and text embedding in the joint space. α represents the margin value for the loss. We use cosine similarity which has been widely used in prior works on learning joint image-text embedding models [1], [3].

IV. Experiments

We perform experiments on two benchmark datasets with the goal of evaluating the performance of including our tag refinement approach with joint embedding baselines. As existing works on webly supervised learning of image-text embedding directly use web images and associated tags without any refinement in training, our work is the first to show results with using an initial refinement step. Our framework would attempt at attaining significantly better performance than the baselines that web image and tags without refinement.

A. Datasets and Evaluation Metrics

In this section, we present a brief description of the datasets used in this work. We also discuss the evaluation metrics.

Datasets. We experiment with Flickr30K dataset ([48], [49]) and MSCOCO dataset ([50]). Flickr30K includes about 31K images harvested from Flickr and each image has five captions. We follow the widely used split provided by [51], where, train, validation and test set contain 29K, 1K, and 1K images, respectively. MSCOCO dataset ([50]) contains about 120K images and each image is described by 5 captions. We use the split following [51]. In this split, the dataset is divided into about 82K training, 5K validation, and 5K test images. In most of the previous works, the results are reported by averaging over 5 folds of 1K test images [1], [52].

Evaluation Metrics. For the evaluation of tensor completion performance, we use relative error, which is one of the most commonly used metrics in evaluating the performance of tensor completion algorithms. The relative error for predicted tensor is calculated by the standard error in tensor prediction (Frobenius norm difference between ground-truth tensor and the predicted tensor), divided by the Frobenius norm of the ground-truth tensor. The relative error for observed tensor is calculated in a similar way.

For image-text retrieval, we use a rank based metric which is commonly used in prior works on different cross-modal retrieval tasks [1], [3], [5], [20], [53]. We measure performance by $R@K$ (Recall at K) and $MedR$ (Median Rank). $R@K$ calculates the percentage of samples for which the ground-truth (GT) annotation is ranked within the top-K retrieved results to the query. We also report the median rank which calculates the median of the GT results in the retrieval ranking.

B. Experimental Setup

Data Preparation. We are interested in estimating the influence of noisy or missing tags on the performance of our approach. However, it is very difficult to collect a large number of web images with tags and label them. Hence, we create synthetic data based on image-text pairs from datasets (e.g., Flickr30K) to evaluate the effect of our image-tag refinement approach. First, we create a synthetic clean image-tag dataset from the training sets of the datasets. For each image, we collect the unique nouns and verbs as image tags from the associated 5 sentences. We retain only the top 1000 occurring words in the training set.

We then create noisy image-tag datasets from the synthetic clean set based on the missing ratio of tags (e.g., 30%, 50%, 70%) we would like to consider in evaluating the approach. In this regard, given a missing ratio (%), we randomly select the overall number of tags to be replaced. We remove most

TABLE I

THIS TABLE PRESENTS THE RETRIEVAL RESULTS ON FLICKR30K AND MSCOCO DATASET. **ACTUAL** INDICATES THE INITIAL SYNTHETIC CLEAN IMAGE-TAG SET (NO MISSING DATA) CREATED BY EXTRACTING UNIQUE NOUN AND VERBS FROM CAPTIONS ASSOCIATED WITH IMAGES AS TAGS. **OBSERVED** INDICATES THE SYNTHETIC NOISY WEB IMAGE-TAG SET CONSTRUCTED BY REMOVING TAGS BASED ON A GIVEN MISSING (%). **PREDICTED** INDICATES THE REFINED IMAGE-TAG SET OBTAINED BY REFINING THE OBSERVED SET APPLYING THE PROPOSED TENSOR COMPLETION APPROACH. FOLLOWING [1], WE USE VGG16 FEATURE AND PAIRWISE RANKING LOSS FOR TRAINING JOINT EMBEDDING MODELS.

		FLICKR30K									**MSCOCO**								
Task	Tag Quality	Missing Data (%) = 30			Missing Data (%) = 50			Missing Data (%) = 70			Missing Data (%) = 30			Missing Data (%) = 50			Missing Data (%) = 70		
		R@1	R@10	MedR	R@1	R@10	MedR	R@1	R@10	MedR	R@1	R@10	MedR	R@1	R@10	MedR	R@1	R@10	MedR
Image to Text	Actual	5.5	24.7	57	5.5	24.7	57	5.5	24.7	57	9.7	40.6	17	9.7	40.6	17	9.7	40.6	17
	Observed	4.7	18.7	97	1.4	9.3	280	1.8	7.7	365	8.8	37.5	20	8.6	33.7	27	3.8	19.3	136
	Predicted	**4.3**	**21.2**	**79**	**2.8**	**14.7**	**186**	**2.5**	**14.7**	**163**	**9.7**	**40.0**	**19**	**9.2**	**35.4**	**25**	**6.8**	**28.9**	**34**
Text to Image	Actual	2.8	14.7	137	2.8	14.7	137	2.8	14.7	137	6.0	30.2	35	6.0	30.2	35	6.0	30.2	35
	Observed	1.6	10.1	200	0.7	4.3	328	0.4	4.4	337	5.2	22.8	70	4.0	20.9	67	2.7	16.6	107
	Predicted	**2.3**	**11.4**	**191**	**1.1**	**7.5**	**230**	**0.5**	**3.6**	**385**	**4.9**	**22.8**	**69**	**3.7**	**18**	**105**	**3.2**	**14.9**	**110**

of the tags and replace a few tags with random English words from the dictionary. In this way, we create several noisy datasets based on different missing ratios. These noisy datasets are considered as our observed set. From the synthetic clean datasets, we utilize the first 1K images from the training set as our small clean image-text set as D in tensor completion ($|D| = 1000$). The noisy image-tag dataset is created from the remaining training images and these images are considered as images from web W. The top 1000 occurring words in the training set are considered as the tags set T ($|T|$=1000).

Implementation Details. In training the joint embedding model, we use VGG16 (pre-trained on ImageNet) as the visual feature encoder. As we consider having tags for the image set, we do not have sequential information from sentences. We utilize the average of the Word2Vec vectors corresponding to the tags as the text representation. The dimension of the joint space is set to 1024. We implemented the network in PyTorch [54]. The embedding network is trained using Stochastic Gradient Descent dividing the dataset into batches of size 128. Hence, the approach considers only the negative samples selected in a stochastic manner from a mini-batch and is more efficient to train. We use a starting learning rate of 0.002, and it is lowered every 15 epoch by a factor of 10. The margin parameter alpha is empirically chosen as 0.2. The best performing model on the validation set is selected.

The tensor completion approach is implemented using Matlab tensor toolbox [55]. In the constructed observed tensor $\mathcal{T}$, we only know the observed non-zero entries. However, we do not have any prior information about zero entries whether they are missing or not relevant. However, for a good reconstruction of the tensor, a certain amount of observed entries is often required [30]. We randomly sample zeros from remaining entries to have an equal observed ratio as non-zeros. We vary tensor rank from 10 to 20, and empirically fix as 20, which we found to be consistently performing well in terms of lower relative standard error in tensor completion. We use 500 as the maximum number of iterations allowed.

C. Results

In this section, we present quantitative results to evaluate the proposed approach. We evaluate the performance of the image tag set predicted by our approach compared to the initial clean set, observed set and baselines in image-text retrieval on the Flickr30K and MSCOCO dataset (Sec. IV-C1). We varied the percentage of missing tags among [30%, 50%, 70%]. The percentages are chosen for experimentation. We also compare the proposed tensor completion approach with baselines in terms of relative error in image tag refinement (Sec. IV-C2).

*1) **Experiments on Cross-Modal Image-Text Retrieval** :* To understand the effect of the proposed tag refinement approach in overall cross-modal image-text retrieval performance, we evaluate on data prepared from Flickr30K and MSCOCO. While utilizing the complete image-sentence retrieval train sets (e.g., MSCOCO) are useful to experiment with web-supervised learning, we consider the more practical case of utilizing web images with a very small subset of the clean datasets. Hence, the goal of the experiments is not to compete with the best reported results in these datasets. The goal is to evaluate whether the proposed approach is capable of boosting the performance of state-of-the-art joint embedding based image-text retrieval approaches trained with a small clean set of image-text pairs utilizing freely available web images. We believe, our setting, therefore, more precisely reflect practical challenges caused by noisy web images. The performance is reported in Table I where Observed(% missing of actual) indicates that the observed image-tag set is created by removing tags from actual/clean image-tag set by chosen missing/unobserved data percentage(%).

Results on Flickr30K. We compare retrieval results based on the joint embedding models trained using the actual set, the observed set, and the predicted set. From Table I, we have several key observations. First, we observe that the predicted set shows better performance compared to the observed in almost all metrics. We find in case of 30% missing data, the observed set performs better than predicted set in R@1 in the image to text retrieval. However, the predicted set shows performance improvement in the other metrics. We observe similar improvements in the case of other missing data ratios. Second, we see that as we increase the percentage of missing data, the model learned using the predicted set in general performs significantly better than the model learned using the observed set. Initially, in the case of 30% missing data, the performance of predicted and observed set is comparable. As

the missing percentage increases, the observed set shows a significant drop in performance. However, the predicted set is able to limit the performance drop by recovering some related tags. We see the predicted set improves the average median rank by about 27% (relative) from the observed set.

Results on MSCOCO. Similar to Flickr30K in Table I, it is evident from the Table that our proposed tag refinement approach helps to improve performance over directly using images with raw tags (observed set). As expected, the performance drops for both observed set and predicted set as the percentage of missing entries from the actual set increase. We again see that in case of a low missing ratio, the observed and predicted set shows comparable performance. However, the performance of the prediction method is promising as it shows significant improvement compared to the observed set when the missing data percentage is high (70%). We see a 3% absolute improvement in R@1 and 102 point decrease in median rank using the proposed approach in the image to text retrieval task with 70% missing data.

*2) **Relative Errors in Tensor Completion**:* In Table II, we compare relative errors of predicted tensor and observed tensor for different percentages of missing entries. We also compare with matrix-based image tag refinement approach Matrix-LRCTE [56], and baseline multi-way tucker and parafac model for handling missing data [57]. From the Table II, we find that the predicted tensor results in consistently decreasing the relative error compared to the observed tensor across missing percentages. The average improvement using the proposed approach in relative error is about 10.21%. The maximum improvement of 13.3% is found in MSCOCO with 30% missing data and the minimum improvement of 8.7% is found in Flickr30K with 30% missing data.

Comparison with Baselines. We first evaluate the contribution of the CP regularization terms in overall improvement. From Table II for Flickr30k with missing ratio[70%, 50%, 30%], the relative error of observed tensor is [0.839, 0.721, 0.563] and the proposed regularized CP achieves significantly less [0.762, 0.649, 0.514]. However, the relative error for the proposed without regularization is [0.826, 0.705, 0.533], which is on par with observed and significantly worse than the proposed approach. In MSCOCO, we again find that the regularization term contributes significantly to reducing relative error. We also noticed that cross-modal retrieval performance degrades when regularization term is not used.

We compare with the image tag refinement approach Matrix-LRCTE [56], which is a matrix refinement approach based on low-rank, content-tag prior and error sparsity [56]. For missing [30%,50%,70%] in Flickr30K, [56] achieves limited improvement of [2.9%, 1.7%, 0.6%) from the observed, whereas the proposed achieves a larger improvement of [8.7%, 10.0%, 9.2%] as reported in Table II. We also find a similar trend in MSCOCO evaluation. We could not compare with prior tensor completion based tag refinement approaches [8], [40], [39] as these works assume the availability of additional user information from social media for tag refinement.

TABLE II
RELATIVE ERRORS (COMPARED TO THE GROUND-TRUTH/ACTUAL TENSOR) FOR RECOVERING MISSING TAGS (BEFORE AND AFTER TENSOR COMPLETION) FOR DIFFERENT PERCENTAGE OF MISSING ENTRIES. WE OBSERVE THAT THE PREDICTED TENSOR GIVES ON AVERAGE 10.2% IMPROVEMENT OVER THE OBSERVED TENSOR.

Method	Flickr30K			MSCOCO		
	30%	50%	70%	30%	50%	70%
Observed	0.563	0.721	0.839	0.534	0.703	0.838
Parafac	0.732	0.830	0.891	0.748	0.848	0.902
Tucker	0.726	0.825	0.880	0.728	0.813	0.883
Matrix-LRCTE	0.546	0.709	0.834	0.521	0.686	0.828
Proposed (No Regularization)	0.533	0.705	0.826	0.516	0.689	0.822
Proposed	0.514	0.649	0.762	0.463	0.635	0.751

We also experimented with baseline multi-way tucker and parafac model for handling missing data following n-way toolbox [57], which uses an expectation-maximization scheme on top of masking away the missing data from the residuals. We observed these approaches fail to improve performance over the observed set and even consistently show higher relative error in all the cases. In Flickr30k, We find the predicted tensor shows a relative error of [0.726, 0.821, 0.880] with the tucker model, and [0.732, 0.815, 0.893] with the parafac model, for missing [30%,50%,70%] of data, which is significantly worse than observed and the proposed approach. We also observe a similar trend in MSCOCO. We believe this is because of our observed tensor is extremely sparse. As only a small set of tags from the large tag list are associated with each image from dataset and web, the tensor capturing tri-modal dynamics has very few non-zero entries. These baseline tucker and parafac completion approaches [57] fail to handle missing data in such an extremely sparse tensor.

V. CONCLUSION

We address the problem of utilizing web images in training robust image-text embedding models. Directly using web images with raw tags in training may significantly affect the performance of the webly supervised approaches when the ratio of missing tags is high and available clean labeled data is very limited. In this regard, we propose a CP decomposition based approach to refine tags of web images by modeling the ternary inter-relation between the web image collection and the clean dataset images (based on associated tags) as a tensor and utilizing intra-modal similarity as side information. Our image tag refinement approach combined with supervised image-text embedding approaches provide a way to learn robust joint embedding models in the presence of significant noise from web data and limited clean labeled data. Experiments on data prepared from benchmark datasets with different percentage of missing data demonstrate that the proposed approach can successfully recover more than 10% missing data on average and consequently helps to achieve a consistent performance gain in cross-modal image-text retrieval task.

Acknowledgements. This work was partially supported by the National Science Foundation Grant no. 1901379.

REFERENCES

[1] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[2] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *CVPR*, 2016.

[3] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improved visual-semantic embeddings," *arXiv preprint arXiv:1707.05612*, 2017.

[4] H. Wu, J. Mao, Y. Zhang, Y. Jiang, L. Li, W. Sun, and W.-Y. Ma, "Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations," in *CVPR*, 2019.

[5] N. C. Mithun, R. Panda, E. Papalexakis, and A. Roy-Chowdhury, "Webly supervised joint embedding for cross-modal image-text retrieval," in *ACM MM*, 2018.

[6] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *ECCV*, 2014.

[7] D. Rafailidis and A. Nanopoulos, "Modeling users preference dynamics and side information in recommender systems," *IEEE Trans. on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 6, pp. 782–792, 2016.

[8] J. Tang, X. Shu, G.-J. Qi, Z. Li, M. Wang, S. Yan, and R. Jain, "Tri-clustered tensor completion for social-aware image tag refinement," *IEEE TPAMI*, vol. 39, no. 8, pp. 1662–1674, 2017.

[9] A. Narita, K. Hayashi, R. Tomioka, and H. Kashima, "Tensor factorization using auxiliary information," *Data Mining and Knowledge Discovery*, vol. 25, no. 2, pp. 298–324, 2012.

[10] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, and J. Song, "Matching images and text with multi-modal tensor fusion and re-ranking," in *ACM MM*, 2019.

[11] C. Liu, Z. Mao, A.-A. Liu, T. Zhang, B. Wang, and Y. Zhang, "Focus your attention: A bidirectional focal attention network for image-text matching," in *ACM MM*, 2019.

[12] F. Liu, R. Ye, X. Wang, and S. Li, "Hal: Improved text-image matching by mitigating visual semantic hubs," in *AAAI*, 2020.

[13] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *CVPR*, 2017.

[14] Y. Wu, S. Wang, G. Song, and Q. Huang, "Learning fragment self-attention embeddings for image-text matching," in *ACM MM*, 2019.

[15] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han, "Cross-modal image-text retrieval with semantic consistency," in *ACM MM*, 2019.

[16] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *CVPR*, 2015.

[17] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[18] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *IJCV*, vol. 106, no. 2, pp. 210–233, 2014.

[19] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE TPAMI*, 2018.

[20] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *ICMR*. ACM, 2018.

[21] A. Li, A. Jabri, A. Joulin, and L. van der Maaten, "Learning visual n-grams from web data," in *ICCV*, 2017.

[22] E. van Miltenburg, "Stereotyping and bias in the flickr30k dataset," *arXiv preprint arXiv:1605.06083*, 2016.

[23] A. Torralba and A. Efros, "Unbiased look at dataset bias," in *CVPR*, 2011.

[24] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, "Undoing the damage of dataset bias," in *ECCV*, 2012.

[25] D. Gong, D. Z. Wang, and Y. Peng, "Multimodal learning for web information extraction," in *ACM MM*, 2017.

[26] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Attention transfer from web images for video recognition," in *ACM MM*, 2017.

[27] A. Joulin, L. van der Maaten, A. Jabri, and N. Vasilache, "Learning visual features from large weakly supervised data," in *ECCV*, 2016.

[28] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *ICCV*, 2017.

[29] P.-Y. Huang, G. Kang, W. Liu, X. Chang, and A. G. Hauptmann, "Annotation efficient cross-modal retrieval with adversarial attentive alignment," in *ACM MM*, 2019.

[30] Q. Song, H. Ge, J. Caverlee, and X. Hu, "Tensor completion algorithms in big data analytics," *ACM TKDD*, vol. 13, no. 1, p. 6, 2019.

[31] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *Journal of Mathematics and Physics*, vol. 6, no. 1-4, pp. 164–189, 1927.

[32] R. A. Harshman, "Foundations of the parafac procedure: Models and conditions for an explanatory multi-modal factor analysis," 1970.

[33] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[34] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[35] Y. Liu, F. Shang, L. Jiao, J. Cheng, and H. Cheng, "Trace norm regularized candecomp/parafac decomposition with missing data," *IEEE Trans. Cybernetics*, vol. 45, no. 11, pp. 2437–2448, 2015.

[36] J. Sang, J. Liu, and C. Xu, "Exploiting user information for image tag refinement," in *ACM MM*, 2011.

[37] H. Ge, J. Caverlee, N. Zhang, and A. Squicciarini, "Uncovering the spatio-temporal dynamics of memes in the presence of incomplete information," in *CIKM*. ACM, 2016.

[38] H. Ge, J. Caverlee, and H. Lu, "Taper: A contextual tensor-based approach for personalized expert recommendation," in *RecSys*, 2016.

[39] J. Sang, C. Xu, and J. Liu, "User-aware image tag refinement via ternary semantic analysis," *IEEE TMM*, vol. 14, no. 3, pp. 883–895, 2012.

[40] J. Tang, X. Shu, Z. Li, Y.-G. Jiang, and Q. Tian, "Social anchor-unit graph regularized tensor completion for large-scale image retagging," *IEEE TPAMI*, vol. 41, no. 8, pp. 2027–2034, 2019.

[41] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. D. Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval," *ACM CSUR*, vol. 49, no. 1, pp. 1–39, 2016.

[42] H. Zhou, D. Zhang, K. Xie, and Y. Chen, "Spatio-temporal tensor completion for imputing missing internet traffic data," in *IPCCC*, 2015.

[43] H. Wang, F. Nie, and H. Huang, "Low-rank tensor completion with spatio-temporal consistency," in *AAAI*, 2014.

[44] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[45] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, 2017.

[46] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[47] A. P. Liavas and N. D. Sidiropoulos, "Parallel algorithms for constrained tensor factorization via alternating direction method of multipliers," *IEEE Trans. Signal Process*, vol. 63, no. 20, pp. 5450–5463, 2015.

[48] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *TACL*, vol. 2, pp. 67–78, 2014.

[49] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *ICCV*, 2015.

[50] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[51] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[52] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE TPAMI*, 2018.

[53] N. C. Mithun, K. Sikka, H.-P. Chiu, S. Samarasekera, and R. Kumar, "Rgb2lidar: Towards solving large-scale cross-modal visual localization," in *ACM MM*, 2020.

[54] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[55] B. W. Bader, T. G. Kolda *et al.*, "Matlab tensor toolbox version 2.6," Available online, February 2015. [Online]. Available: http://www.sandia.gov/ tgkolda/TensorToolbox/

[56] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *ACM MM*, 2010.

[57] C. A. Andersson and R. Bro, "The n-way toolbox for matlab," *Chemometrics and intelligent laboratory systems*, vol. 52, no. 1, pp. 1–4, 2000.