Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections

Yunchao Gong[1], Liwei Wang[2], Micah Hodosh[2], Julia Hockenmaier[2] and Svetlana Lazebnik[2]

[1]University of North Carolina at Chapel Hill
yunchao@cs.unc.edu
[2]University of Illinois at Urbana-Champaign
{lwang97,mhodosh2,juliahmr,slazebni}@illinois.edu

Abstract. This paper studies the problem of associating images with descriptive sentences by embedding them in a common latent space. We are interested in learning such embeddings from hundreds of thousands or millions of examples. Unfortunately, it is prohibitively expensive to fully annotate this many training images with ground-truth sentences. Instead, we ask whether we can learn better image-sentence embeddings by augmenting small fully annotated training sets with millions of images that have weak and noisy annotations (titles, tags, or descriptions). After investigating several state-of-the-art scalable embedding methods, we introduce a new algorithm called Stacked Auxiliary Embedding that can successfully transfer knowledge from millions of weakly annotated images to improve the accuracy of retrieval-based image description.

1 Introduction

Describing images with natural language sentences is an ambitious goal at the intersection of computer vision and natural language processing. Previous approaches to this problem can be roughly categorized into two groups: novel sentence generation and retrieval-based description. Approaches in the former group, e.g., [1–6], use natural language models or templates for generating sentences, and learn predictors to "fill in" or compose parts of these models. However, image descriptions automatically composed in this way can often be unnatural. More importantly, as argued by Hodosh et al. [7], it is difficult to objectively compare the quality of novel sentences produced by different generation methods for an image – not least because the sentences can vary in specificity, or exhibit different types of quirks or artifacts. Retrieval-based systems, e.g., [7–9], describe images by retrieving pre-existing sentences from a dataset. One representative method, that of Ordonez et al. [8], uses millions of images from Flickr and their corresponding descriptions as a source of image captions. For each query image, it finds similar images in the Flickr database and transfers the descriptions of these retrieved images to the query. However, since this method relies on image-to-image matching to transfer sentences, it cannot return any sentences

that have no images associated with them. Hybrid retrieval- and generation-based methods are also possible: in follow-up to [8], Kuznetsova et al. [10] adopt a template-based approach of composing parts of retrieved sentences to create more query-specific and relevant descriptions.

To automatically evaluate the quality of image captioning systems, many previous works have relied on the BLEU score [11], which is based on the n-gram precision of the caption returned by a system against a human-produced reference caption (or set of captions). However, BLEU was originally developed for machine translation, and it has widely recognized shortcomings for the much more open-ended image description task [7, 2, 8]: BLEU penalizes captions that are relevant to the image but do not happen to overlap with the reference set; it does not measure vision output quality directly; and it has poor correlation with human judgment. As an automatic alternative to BLEU, Hodosh et al. [7] propose a retrieval-based protocol: given a query image, use the model being evaluated to retrieve sentences from a pool that also contains some reference sentences associated with that image, and see how highly the model ranks them. This protocol can be used with any systems that can score image-sentence pairs. It can still underestimate performance by not reflecting when the system returns a valid caption that was not originally associated with the image, but Hodosh et al. [7] show that recall of the original caption has better correlation with human judgment than BLEU.

In this paper, we adopt the retrieval-based protocol of [7], as well as their idea of image-to-sentence retrieval in a joint image-sentence embedding space. To establish a baseline, they use Kernel Canonical Correlation Analysis (KCCA) [12] with multiple visual and linguistic kernels to map images and sentences into a space where similarity between them can be computed directly. They train this embedding on 6,000 images associated with five ground-truth captions each. However, to enable substantial further progress in techniques for mapping between images and sentences, we believe that a much larger amount of training data is required. This leads to two fundamental challenges:

1. Nonlinear image-sentence embedding methods, such as KCCA, tend not to scale to large training sets.
2. Obtaining high-quality sentence descriptions for millions of images is a prohibitively expensive task.

To address the first challenge, we conduct a comparative evaluation of scalable image-sentence embedding methods and show that linear Canonical Correlation Analysis (CCA) with proper normalization [13] outperforms several state-of-the-art alternatives in terms of both accuracy and efficiency, and is therefore a promising framework on top of which to build a large-scale image-to-sentence retrieval approach. To address the second challenge, we ask: *Can the addition of a large amount of Internet images with noisy textual annotations to a smaller set of images with clean sentence annotations help us induce a better latent space*? Figure 1 shows an illustration of this scenario. It is a multi-view transfer learning setting that, to our knowledge, has not been studied before. It has connections

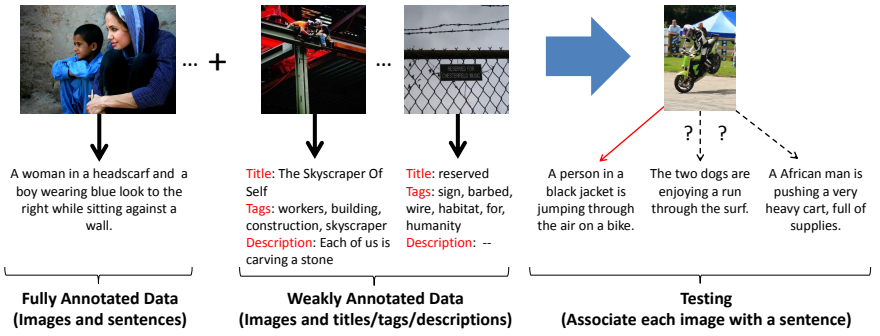

Fig. 1. The problem setting of our paper. We want to use large amounts of Flickr images annotated with noisy tags, titles, and descriptions to help with learning of an image-sentence embedding on a small dataset of images and clean ground truth sentences. At test time, we embed images and sentences in the learned latent space and perform image-to-sentence retrieval.

to multi-view learning [13, 7], transfer learning [14–16], and methods that use Internet data to help recognition [17–21]. Starting with the normalized CCA formulation, we propose a novel transfer learning approach we call Stacked Auxiliary Embedding (SAE) and show its effectiveness in transferring knowledge from two large-scale Flickr datasets of a million images each.

The rest of our presentation is organized as follows. Section 2 will introduce our datasets, evaluation protocols, and feature representations for images and text. In Section 3, we begin by conducting a comparative evaluation of several scalable image-sentence embedding models in the fully supervised scenario – i.e., trained on tens of thousands of images annotated with ground-truth sentences. Next, in Section 4, we take the winning embedding, CCA with normalization [13] and consider how to improve it by adding millions of images weakly annotated with noisy tags, titles, and descriptions. We introduce our Stacked Auxiliary Embedding model and demonstrate that it outperforms a number of alternative baselines in terms of image-sentence retrieval accuracy.

2 Datasets, Protocols, and Features

2.1 Datasets

We begin by describing our datasets for learning image-sentence embeddings. Our fully annotated data comes from the dataset of Young et al. [22], which is an expanded version of the one from [7]. This dataset, referred to as **Flickr30K**, contains 31,783 images collected from different Flickr groups and focusing on events involving people and animals. Each image is associated with five sentences independently written by native English speakers from Mechanical Turk. Sample data from Flickr30K is shown in Figure 2(a).

For the weakly annotated data for the transfer task, we experiment with two datasets of about a million images each that do not overlap with Flickr30K or each other. For the first one, referred to as **Flickr1M**, we used queries based on

Fig. 2. (a) Sample images and sentences from the Flickr30K dataset [22]. (b) Sample images from the Flickr1M dataset. These images come with titles, tags, and descriptions, some of which may be missing.

the most frequent 350 keywords in Flickr30K to download more than two million additional images from Flickr. After removing duplicates and images lacking tags, we were left with one million images. We use these images and their tags, titles and descriptions as weak supervision. Sample data from Flickr1M is shown in Figure 2(b). As our second weakly annotated dataset, we use the **SBU1M** dataset of [8], which also comes from Flickr, but has very different statistics from Flickr30K because it was collected totally independently. We took the Flickr IDs of the SBU1M images and downloaded all their titles, tags and descriptions. We are interested in experimenting on both datasets because we would like to investigate to what an extent the success of transfer embedding methods depends on the similarity between the fully and the weakly supervised domains.

2.2 Evaluation Protocol

As stated in the Introduction, we follow the retrieval-based protocol of [7, 22]. For the Flickr30K dataset, we use the 3,000 test images from the split of [22] and for each test image, we keep only the first sentence out of five. Each method being evaluated is used to separately map the 3,000 images and 3,000 sentences to the learned latent space, and then each of these images is used as a query to retrieve the sentences based on some similarity measure in the latent space. If the ground-truth sentence is within the top k retrieved sentences, we mark this query as successful, otherwise, it is a failure. We report Recall@10, which is the percentage of query images that have successfully found their ground truth sentence within $k = 10$ nearest neighbors (numbers for other k exhibit exactly the same trends). To learn the latent spaces, we use fixed training subsets ranging in size from 5,000 to 25,000, together with all five sentences per each image. That is, if we report results for a given training set size, we are in fact using five times as many image/sentence pairs. We also use a disjoint set of 3,000 validation images (also from the split of [22]) to tune parameters.

Table 1. Recall@10 for CNN activations versus a combination of standard visual features. The standard features consist of a 960-dimensional GIST [23], a 512-dimensional RGB histogram, and three local descriptors densely sampled on a regular grid: CSIFT [24], RGBSIFT [24], and HOG [25]. Each local descriptor type is quantized using a codebook of size 500 and VLAD-pooled [26] to obtain $6,400$-dimensional image descriptors. GIST and RGB histograms are PCA-reduced to 500 dimensions and VLAD to 1,000 dimensions, and concatenated to get 4,000-dimensional combined descriptors. The sentence features are 3,000-dimensional BoW vectors and the embedding model is normalized CCA (Section 3).

method / training set size	5,000	15,000	25,000
Standard combined features (4,000 dim.)	11.07	18.40	22.13
CNN activations (4,096 dim.)	**19.77**	**27.03**	**31.13**

2.3 Visual and Textual Features

We represent the visual content of images using activations from a deep convolutional neural network (CNN) [27]. CNNs have recently produced state-of-the-art results on a wide range of recognition tasks. Specifically, we use the 4,096-dimensional activations from the sixth hidden layer of the Decaf system [28] pre-trained on ImageNet [29]. Table 1 confirms that CNN activations give significantly higher accuracy on our problem than a combination of standard visual descriptors like GIST and bags of visual words.

For the textual features, we use a standard bag-of-words (BoW) representation. In the following, we will refer as a "document" to each separate piece of text associated with an image: a sentence, a title, a set of tags, or a description. We first pre-process all the documents with WordNet's lemmatizer [30] and remove stop words. After that, we construct a dictionary by taking a few thousand most common words, and represent documents as tf-idf-weighted BoW vectors. Table 2 compares different dictionary sizes for sentence features. We have found that using 3,000 words is sufficient for good performance. For sentences, we have also experimented with a bigram feature, but did not observe any improvement.

For our weakly labeled datasets, Flickr1M and SBU1M, each image is associated with up to three document types: tags, titles, and descriptions (Figure 2(b)). Among other things, we are interested in investigating which of these types (or their combination) gives the best cues for transfer learning. Table 3 compares the BoW features constructed from each document type separately, as well as a single BoW feature computed from a concatenation of all of them. Surprisingly, titles achieve the highest performance despite having the shortest average length. Thus, while tags are more commonly used, titles might actually be the most informative source of annotations from Flickr. On the other hand, descriptions of the Flickr images are by far the longest, but their predictive power is the worst. In the end, combining all three document types achieves the best performance, so in the following experiments, we will use the combined text feature for Flickr1M and SBU1M.

Table 2. Recall@10 for sentence features with different dictionary sizes and different training set sizes. The embedding technique is normalized CCA (Section 3).

dictionary / training set size	5,000	15,000	25,000
1,000	19.03	23.30	26.53
3,000	19.77	27.03	31.13
5,000	20.20	27.40	31.07

Table 3. Recall@10 for different text cues on the weakly annotated datasets, together with the average number of words for each type of cue. We train normalized CCA (Section 3) on Flickr1M or SBU1M and directly apply it to the Flickr30K test set (the Flickr30K training set is not used). All text is represented using 3,000-dimensional tf-idf-weighted BoW vectors.

	Average length	Flickr1M→ Flickr30K	SBU1M → Flickr30K
Title	2.93	17.83	14.57
Tag	5.09	15.97	12.90
Description	23.41	16.67	14.57
Combined	31.03	**18.33**	**15.50**

3 Fully Supervised Image-Sentence Embedding

To provide a foundation for developing transfer embeddings, we first conduct a comparative evaluation of scalable methods for joint embedding of images and sentences in the fully supervised scenario, i.e., training on images paired with clean ground-truth sentences and no auxiliary data of any kind. The methods we compare include textbook baselines of ridge regression and canonical correlation analysis (CCA), as well as several state-of-the-art methods: CCA with normalization [13], Wsabie with stochastic gradient descent [31], and Wsabie with an adaptive learning rate [32, 33].

Assuming images and sentences are represented by vectors of dimension d and D, respectively, our training data consists of a set of images $X \in \mathbb{R}^{n \times D}$ and associated sentences $Y \in \mathbb{R}^{n \times d}$, for n image/sentence pairs. Each image $\boldsymbol{x}$ corresponds to a row in X, and each sentence $\boldsymbol{y}$ corresponds to a row in Y. The goal of all the embedding methods is to find matrices $W \in \mathbb{R}^{D \times c}$ and $U \in \mathbb{R}^{d \times c}$ to map images and sentences respectively as XW and YU to a common c-dimensional latent space in which image-to-sentence retrieval can be done by directly computing a distance or similarity function between pairs of projected image and sentence features.

Ridge Regression: Socher et al. [34] suggest mapping images to a sentence space for zero-shot learning by minimizing the sum of squared distances between the two views. This formulation is close to ridge regression, which we take as our first baseline. The projection matrix U for sentences is given by the top c PCA directions of Y. Then the mapping W from the image features X to the PCA-projected sentence features $\hat{Y} = YU$ is found by minimizing $\|\hat{Y} - XW\|_F^2 + \lambda\|W\|_F^2$. The optimal W is found in closed form as $(X^T X + \lambda I)^{-1} X^T \hat{Y}$. The regularization parameter λ is found on the validation set. Given a query image feature $\boldsymbol{x}$, image-to-sentence retrieval is performed by projecting this feature as

$\boldsymbol{x}W$ and finding the closest k sentences $\hat{\boldsymbol{y}} = \boldsymbol{y}U$ according to their Euclidean distance $\|\boldsymbol{x}W - \boldsymbol{y}U\|_2$.

Canonical Correlation Analysis (CCA) [35] aims to find projections W and U for the two views X and Y such that the normalized correlation between the projected data is maximized:

$$\max_{W,U} \ \text{trace}(W^T X^T Y U) \quad \text{s.t.} \quad W^T X^T X W = I,\ U^T Y^T Y U = I\,. \quad (1)$$

The CCA objective function can be solved as a generalized eigenvalue problem, and entries of the top c leading eigenvectors are concatenated to form W and U. As with ridge regression, the distance function for image-to-sentence retrieval in the projected space is Euclidean.

Normalized Canonical Correlation Analysis: Recently, Gong et al. [13] reported significantly improved results for cross-modal retrieval by scaling the columns of the CCA projection matrices by a power of the corresponding eigenvalues, and using cosine similarity instead of Euclidean distance. Specifically, given the projection matrices W and U obtained by solving the CCA objective (eq. 1) with columns corresponding to c eigenvectors, and their eigenvalues $\lambda_1, \ldots, \lambda_c$, the similarity between image $\boldsymbol{x}$ and sentence $\boldsymbol{y}$ is measured as:

$$\frac{\Big(\boldsymbol{x}W\,\text{diag}(\lambda_1^t, \ldots, \lambda_c^t)\Big)\Big(\boldsymbol{y}U\,\text{diag}(\lambda_1^t, \ldots, \lambda_c^t)\Big)^T}{\|\boldsymbol{x}W\,\text{diag}(\lambda_1^t, \ldots, \lambda_c^t)\|_2 \|\boldsymbol{y}U\,\text{diag}(\lambda_1^t, \ldots, \lambda_c^t)\|_2}\,, \quad (2)$$

where t is the power to which the eigenvalues are taken (we use $t = 4$, the same value as in [13]). The cosine similarity is a natural choice for test data as it is exactly the quantity that the CCA objective function is maximizing for the training data. In this work, we would like to see whether this similarity also improves image-to-sentence retrieval, a task that was not considered in [13].

Wsabie with SGD: Weston et al. [31] have proposed the Wsabie approach for mapping images and tags to the same space using stochastic gradient descent. Several other works, e.g., [36], have also reported good results for this model. We adapt Wsabie to our problem as follows. Given the training set of n image/sentence pairs, we iterate through them in random order. For each pair of image feature $\boldsymbol{x}_i$ and positive sentence $\boldsymbol{y}_i$, we keep sampling negative sentences (i.e., sentences not originally associated with this image) until we find a negative sentence $\boldsymbol{y}_j$ that violates the margin constraint:

$$\boldsymbol{x}_i W U^T \boldsymbol{y}_j^T > \boldsymbol{x}_i W U^T \boldsymbol{y}_i^T - 1$$

(here, we use correlation as the similarity function between images and sentences in the latent space). Assuming we have sampled s sentences until we find a violation, we estimate the rank of the positive sentence given the current model by $r_i = \lfloor \frac{n-1}{s} \rfloor$. Then we weight the amount of margin violation by the ranking loss $L(r) = \sum_l^r 1/l$ as in [31]. For a small rank (corresponding to a good model), the value of the loss is small, and for a large one, it is large. This leads to the

Table 4. Recall@10 for different image-sentence embedding methods.

method / training set size	5,000	15,000	25,000
Ridge regression	10.63	11.40	12.77
CCA	8.76	12.37	15.43
CCA+Normalization	**19.77**	**27.03**	**31.13**
Wsabie with SGD	15.43	17.86	18.10
Wsabie with AdaGrad	18.20	24.33	26.60

following stochastic objective function:

$$\sum_{i=1}^{n} L(r_i) \max(0, 1 - \boldsymbol{x}_i W U^T \boldsymbol{y}_i^T + \boldsymbol{x}_i W U^T \boldsymbol{y}_j^T) \quad (3)$$

$$\text{s.t.} \quad \|\boldsymbol{w}_k\|_2^2 \leq \alpha, \ \|\boldsymbol{u}_k\|_2^2 \leq \alpha, \quad k = 1, \ldots, c, \quad (4)$$

where $\boldsymbol{w}_k$ and $\boldsymbol{u}_k$ denote the columns of W and U. To minimize the objective function, whenever we find a violation, we take a gradient step to adjust the weights (entries of U and W) and project them to enforce the constraints (eq. 4). We initialize the weights using a random Gaussian with zero mean and unit variance, tune the learning rate by searching a grid of values [0.01, 0.05, 0.1, 0.2, 0.5, 1] on the validation set, and run the algorithm for 300 epochs. The parameter α is also tuned on the validation set using a grid of $[50, 100, 150, 200]$. At retrieval time, we use normalized correlation or cosine similarity between projected images and sentences: $(\boldsymbol{x}WU^T\boldsymbol{y}^T)/(\|\boldsymbol{x}W\|_2\|\boldsymbol{y}U\|_2)$ (we have found it to work better than unnormalized correlation or Euclidean distance).

Wsabie with AdaGrad: We also minimize the loss of eq. (3) with AdaGrad [32, 33], a per-dimensional learning rate adjustment method that has been shown to improve performance of SGD. We tune the global learning rate over a grid of [0.2, 0.4, 0.6, 0.8, 1] on the validation set. Once again, we initialize the weights using a random Gaussian and train for 300 epochs. As with the regular Wsabie, we use cosine similarity for image-to-sentence retrieval.

Comparative evaluation. Table 4 compares the performance of the above image-sentence embedding methods. For all methods, we set the dimension of the latent space to $c = 96$, which we have found to work the best in all cases. We can see that neither ridge regression nor vanilla CCA are competitive with the rest of the approaches. However, when combined with the normalized similarity function (eq. 2), CCA yields dramatically better performance, which is consistent with the findings of [13] on other cross-modal search tasks. As for Wsabie, the SGD version is better than CCA but much worse than normalized CCA, while Wsabie with AdaGrad is only 2-5% below normalized CCA. The advantage of normalized CCA over Wsabie with AdaGrad is probably due to two reasons. First, our experiments seem to indicate that cosine similarity (i.e., normalized correlation) works the best for image-to-sentence retrieval in the latent space, and the CCA objective function, unlike the Wsabie one, directly optimizes this measure. Furthermore, CCA finds the globally optimal solution in closed form. By contrast, our current Wsabie objective (eq. 3) is already non-convex and SGD might not be able to obtain its global optimum (and reformulating the objective in terms of normalized correlation would only make matters worse).

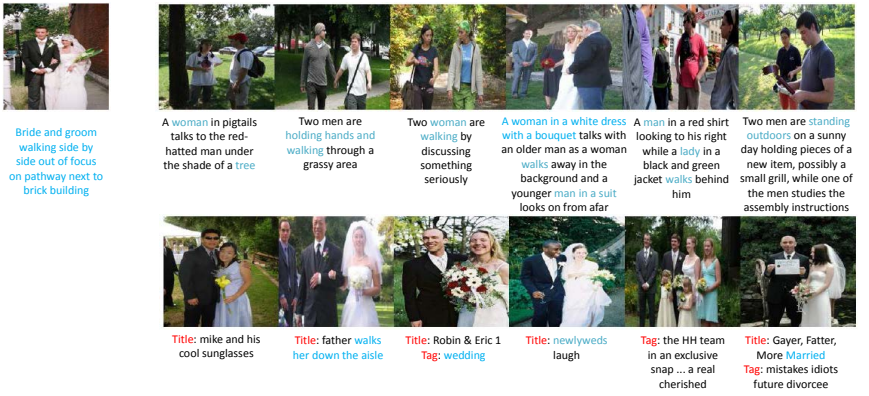

Fig. 3. A Flickr30K query image (left) with its nearest neighbors (according to CNN visual features) from Flickr30K (top) and Flickr1M (bottom). Associated sentences (resp. Flickr text) are shown beneath the retrieved images. Words relevant to the content of the query are manually highlighted in blue for visualization purposes.

In terms of computational requirements, normalized CCA is faster and easier to tune than Wsabie. CCA only requires solving a generalized eigenvalue problem involving the cross-covariance matrix. The complexity of this step scales roughly quadratically in the combined input feature dimension and is insensitive to training set size. In practice, it is very fast: on our four-core Xeon 3.33GHz machine with 64GB RAM, it takes 5 minutes for 5,000 training examples or 15 minutes for one million. On the other hand, training for Wsabie involves multiple passes over the data and validation for parameter tuning. For 5,000 examples, just one epoch of Wsabie already takes around 15 minutes on the same machine, and the time scales linearly with the training set size. Thus, we will use the normalized CCA approach as the basis for our transfer embedding model.

4 Transfer Embedding

In this section, we get to the main focus of our work: adding a large amount of weakly annotated images to a smaller amount of fully annotated ones to learn a better image-sentence embedding. In this setting, the weakly annotated data comes from the Flickr1M or SBU1M datasets (described in Section 2.1), and the fully annotated data comes from Flickr30K. Training is done on one of Flickr1M or SBU1M, plus the training subset of Flickr30K. Testing is done on the same test subset of Flickr30K as all the preceding experiments.

4.1 Stacked Auxiliary Embedding

Our basic assumption is that images and annotations in Flickr1M share some similarity with the images and sentences in Flickr30K. To illustrate this, Figure 3 shows a sample image from Flickr30K together with its nearest neighbors in Flickr1M and Flickr30K. We can see that the Flickr1M neighbors have much

more relevant content to the query than the Flickr30K ones. This suggests that Flickr1M can provide additional useful information for learning the embedding (although, as will be shown in Section 4.3, a naive attempt to transfer text from nearest neighbors via the method of [21] does not succeed).

We follow related work where embedded features learned from auxiliary sources are concatenated with the original features to form a stacked representation [37, 38]. As the first step, we use CCA to learn a joint c_1-dimensional embedding from our weakly annotated dataset, say Flickr1M. Let $A \in \mathbb{R}^{d \times c_1}$ and $B \in \mathbb{R}^{D \times c_1}$ denote the resulting projection matrices for visual and textual features, respectively, with each column already scaled by the t-th power of its eigenvalue. We then apply these projections to X and Y, the visual and textual feature vectors from the Flickr30K training set. Next, we nonlinearly transform the embedded features XA and YB using a mapping $\phi(\cdot)$ and concatenate the result with the original features to form the stacked representation:

$$\hat{X} = [X, \phi(XA)], \qquad \hat{Y} = [Y, \phi(YB)]. \tag{5}$$

The goal of $\phi(\cdot)$ is to raise the dimensionality of its input and help avoid degradation of the stacked model. We use the random Fourier feature (RFF) mapping [39]: $\phi(\boldsymbol{x}) = \sqrt{2}\cos(\boldsymbol{x}R + \boldsymbol{b})$, where R is drawn from Normal$(0, \sigma^2)$ (σ is set to the average distance to the 50th nearest neighbor) and $\boldsymbol{b}$ is drawn from Unif$[0, 1]$. For the CCA embedding, we set the output dimensionality to $c_1 = 128$, and then use RFF to raise the dimensionality to 3,000 (note that we have found the results to be insensitive to the exact choice of these values). We have also tested other nonlinear functions such as sigmoid or tanh, but found they do not work well for our case.

Given the augmented Flickr30K features $\hat{X}$ and $\hat{Y}$ as defined by eq. (5), we again learn a CCA model on top of them to obtain the projections $\hat{W}$ and $\hat{U}$ for images and sentences. The dimensionality of the final output space is 96 as in Section 3 (this value is much more sensitive than the $c_1 = 128$ of the first round of CCA and needs to be tuned on the validation set). At test time, we apply the entire sequence of learned transformations to the test images and sentences and use the cosine similarity of eq. (2) to perform image-to-sentence retrieval.

We dub our method **Stacked Auxiliary Embedding (SAE)**. It is inspired by stacked denoising autoencoders [40, 41] and the recent work on using stacked corrupted features for domain adaptation [38]. Like these approaches, we also use an embedding learned from noisy data to augment the feature representation. Unlike them, we are trying to use a large amount of noisily annotated images as auxiliary sources, instead of randomly added corruptions. Also, to our knowledge, we are the first to apply such techniques to a multi-view setting.

4.2 Baseline Models

We compare our proposed SAE model to a number of baselines.

Fully Supervised Only: We only use the clean annotated images and sentences from Flickr30K to learn the normalized CCA model. This corresponds to the setting of Section 3.

Weakly Supervised Only: We only use the images and noisy textual information (titles, tags, descriptions) from Flickr1M or SBU1M to learn the normalized CCA model, and no clean data from Flickr30K.

Joint Training: We treat the fully and weakly annotated training samples as being the same, merge them together into a single dataset, and train a normalized CCA embedding. That is, if X and Y denote the image and sentence features of the Flickr30K training set, and F and T denote the image and noisy text features of Flickr1M or SBU1M, we concatenate them vertically as $[X; \beta F]$ and $[Y; \beta T]$. The weight β controls the contribution of the weakly annotated data.

Text Feature: This method was proposed by Wang et al. [21] for using large noisily annotated image collections to improve image classification. To obtain the text feature for each image in the Flickr30K dataset, we find its k nearest neighbors in the weakly annotated dataset based on visual similarity of CNN features. Then we construct a single text feature for each Flickr30K image by averaging the BoW vectors (formed from combined titles, tags, and descriptions) of the retrieved images. We denote the new text feature as as $\hat{T}$. Next, we concatenate the original visual features and text features as $\hat{X} = [X, \hat{T}]$, and perform CCA on $\hat{X}$ and the clean sentences Y to obtain the image-sentence embedding. We have experimented with different values of k and did not find much variation in performance, so we report results for $k = 50$ in the following.

Stacked Training: We first learn a c_1-dimensional CCA embedding of images and text from Flickr1M or SBU1M, and embed the images and sentences from Flickr30K in that latent space. Then we learn another CCA embedding on top of these features. This corresponds to setting $\hat{X} = XA$ and $\hat{Y} = YB$ in eq. (5).

SAE (linear): We apply our SAE framework, only without the nonlinear mapping. That is, we set $\hat{X} = [X, XA]$ and $\hat{Y} = [Y, YB]$ in eq. (5). Together with stacked training, this baseline examines whether every component of SAE is indeed necessary in order to obtain good performance.

4.3 Empirical Results

Table 5 compares SAE to all the baselines. We separately report results for using Flickr1M and SBU1M as the weakly annotated domains. The most important observation is that none of the methods except SAE can consistently exceed the fully supervised baseline – i.e., they are unable to benefit from the million weakly annotated images. For joint training, we have varied the weight β of the weakly annotated dataset (two of the values tried are shown in the table), but could only obtain an improvement for the smallest amount of fully annotated data (5,000 examples). For stacked training, we could not obtain any improvement by varying the dimensionality c_1 of the intermediate embedding learned from the weakly annotated dataset, or by nonlinearly transforming the output of the intermediate embedding. Text features also fail to make a difference over the fully supervised baseline.

By contrast, both the linear and the nonlinear versions of our proposed SAE method achieve a substantial improvement over the fully supervised model, with

Table 5. Recall@10 for methods that train both on the weakly annotated images and Flickr30K. See Section 4 for description of methods and parameters.

	Flickr1M			SBU1M		
method / training set size	5,000	15,000	25,000	5,000	15,000	25,000
Fully Supervised Only	19.77	27.03	31.13	19.77	27.03	31.13
Weakly Supervised Only	18.33	18.33	18.33	15.10	15.10	15.10
Joint Training ($\beta = 0.01$)	20.80	25.90	28.47	20.87	25.87	28.47
Joint Training ($\beta = 1$)	20.63	23.50	25.37	20.07	24.10	25.63
Text Feature ($k = 50$)	19.67	27.00	30.97	19.63	27.03	30.93
Stacked Training (c_1 =256)	19.30	22.93	24.30	19.13	21.97	22.73
Stacked Training (c_1 =1024)	15.10	22.83	26.17	15.17	22.63	25.87
SAE (linear)	23.53	28.57	30.73	22.67	28.43	30.97
SAE (nonlinear)	**23.60**	**29.80**	**32.83**	**23.17**	**29.50**	**32.40**

Table 6. Recall@10 for training the SAE model on different numbers of weakly annotated images. The number of Flickr30K training images is 5,000.

Internet dataset size	Flickr1M	SBU1M
0 (fully annotated only)	19.77	19.77
1,000	20.93	20.20
10,000	20.23	20.53
100,000	21.90	22.60
1,000,000	**23.60**	**23.17**

the nonlinear consistently being the best. Figure 4 shows the top-ranked sentences for a few sample images for the fully supervised baseline vs. SAE. Note that even the incorrect sentences retrieved by SAE tend to contain many keywords closely related to the content of the image. Interestingly, we get very similar results with SAE by using either Flickr1M or SBU1M. This is unexpected, as we have specifically downloaded Flickr1M to match the statistics of Flickr30K – indeed, by looking at the results of the weakly supervised baseline (second line of Table 5), we can see that directly training on Flickr1M does produce a better embedding for Flickr30K than training on SBU1M (18.33% vs. 15.10%). However, after applying SAE, the advantage of Flickr1M disappears, which suggests that a sufficiently complex statistical model is somehow able to extract roughly the same information from any sufficiently large-scale weakly annotated dataset.

Next, Tables 5 and 6 allow us to examine how the performance of SAE changes when we vary the amounts of fully and weakly supervised training data. By comparing the first and last lines of Table 5, it is easy to ascertain that as we increase the number of Flickr30K training examples, the benefit afforded by the Flickr1M or SBU1M examples diminishes. Nevertheless, even when we use the largest number of fully supervised training examples available (25,000), SAE still gives us around 1.3-1.5% improvement. It is important to note that Flickr30K is already the largest dataset of images and sentences available to date; increasing the amount of available fully annotated data by orders of magnitude is likely to be prohibitively expensive, whereas weakly annotated data can be downloaded in unlimited quantities essentially for free. To our knowledge, SAE

is the first attempt at combining the two sources of annotation to improve image description. Our main contribution is to confirm that weakly labeled data can improve image-sentence embeddings *in principle* – and, as our extensive baseline comparisons show, getting any kind of improvement is not trivial. Future research should result in methods that can give bigger improvements.

Finally, it is interesting to compare the absolute accuracy of our image-to-sentence retrieval to other results reported in the literature. In fact, Hodosh et al. [7] have intended their dataset and protocol to constitute a standard benchmark that can be used to automatically compare different methods as "black boxes" to gauge the absolute state of the art. For their own KCCA approach, they report a Recall@10 of 30.3% on a 6K/1K training/test split of their original Flickr8K dataset. For the visual features, they use spatial pyramid kernels on color, texture, and SIFT features. While this representation is not exactly equivalent to our "standard" visual features (Table 1, top line), we expect it to have a similar expressive power. For the text features, they use a sophisticated trigram kernel with distributional and alignment-based similarities – a representation we could not easily accommodate in our linear CCA framework. For comparison, our fully supervised normalized CCA model trained and tested on the same 6K/1K split with the "standard" visual features has a Recall@10 of 30.1% – a remarkably similar number despite our system being totally unrelated to that of [7]. For the SAE approach with additional Flickr1M training data we get 38.2% – a significant improvement. With the CNN visual features, the numbers for our CCA and SAE models go up to 43.8% and 48.8%, respectively. In the future, we would like to experiment with encoding more complex linguistic features in our linear CCA framework to see what additional benefit we can obtain from improving that part of the representation (Hodosh et al. [7] have observed a big advantage for their trigram feature over a simple BoW).

5 Discussion

Our paper is the first to show that Internet images annotated with noisy titles, tags, and descriptions can provide useful information for improving joint embeddings of images and sentences for the application of retrieval-based image description, despite the fact that these sources of textual information have very different distributions and are collected in completely different ways. We have introduced a novel method named Stacked Auxiliary Embedding that convincingly outperforms a number of alternatives, and is, in fact, the only method we have considered that is able to obtain a non-trivial improvement over the baseline that uses fully supervised image-sentence data only.

Apart from this main contribution, we have obtained several other interesting findings along the way. In particular, we have shown that CNN features work much better than traditional visual features for our task, with very affordable dimensionality. This adds to the growing list of recent results in the vision community showing the effectiveness of pre-trained CNN activations as a generic representation for recognition. We have also found that Flickr image titles seem

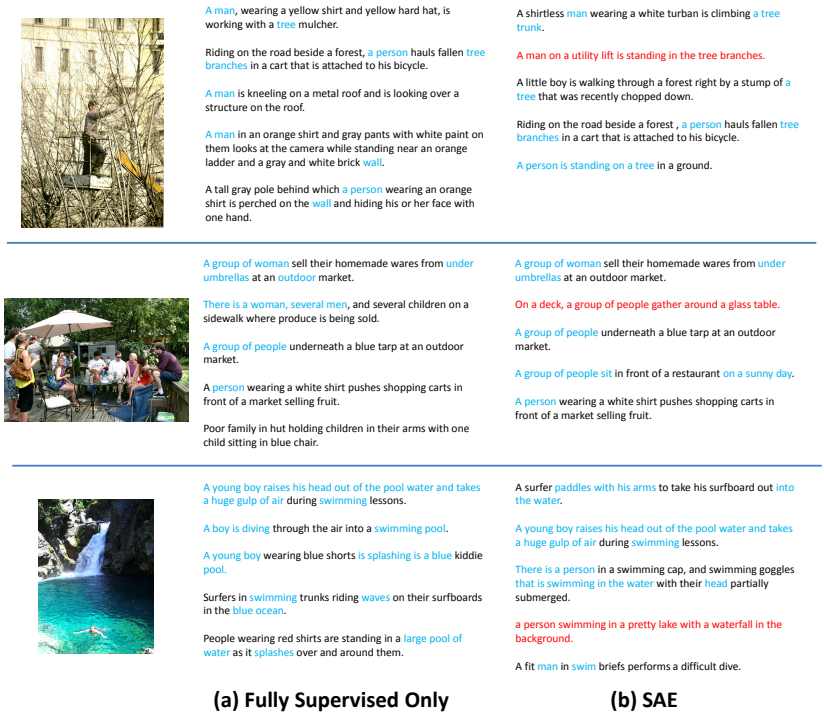

Fig. 4. Image-to-sentence retrieval examples for our fully supervised model vs. SAE. Sentences in red are the ground truth. In the other sentences, words relevant to the query image are manually highlighted in blue for visualization purposes.

to be more discriminative than the more commonly used tags, despite being much shorter. Next, we have confirmed the somewhat surprising findings of [13] that a simple modification of the similarity function used for retrieval with CCA dramatically improves its accuracy, to the point of outperforming sophisticated state-of-the-art ranking models such as Wsabie. While we were able to improve Wsabie in turn with the addition of AdaGrad, normalized CCA still emerged as the more accurate and scalable method.

In the future, we would like to gain more insight into what makes SAE effective. While our baseline comparisons have empirically confirmed the necessity of every implementation choice (i.e., stacking, nonlinearly transforming the intermediate embedded features, and concatenating them with the original features), the resulting technique is frustratingly opaque.

Acknowledgments. Lazebnik's research was partially supported by NSF grants 1228082 and 1302438, the DARPA Computer Science Study Group, Xerox UAC, Microsoft Research, and the Sloan Foundation. Hockenmaier's research was partially supported by NSF grants 1053856 and 1205627. Gong was supported by the 2013 Google Ph.D. Fellowship in Machine Perception.

References

1. Farhadi, A., Hejrati, S., Sadeghi, A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.A.: Every picture tells a story: Generating sentences from images. In: ECCV. (2010)
2. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: Understanding and generating image descriptions. In: CVPR. (2011)
3. Li, S., Kulkarni, G., Berg, T.L., Berg, A.C., Choi, Y.: Composing simple image descriptions using web-scale n-grams. In: CoNLL. (2011)
4. Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., Daumé, III, H.: Midge: Generating image descriptions from computer vision detections. In: EACL. (2012)
5. Fidler, S., Sharma, A., Urtasun, R.: A sentence is worth a thousand pixels. In: CVPR. (2013)
6. Yao, B.Z., Yang, X., Lin, L., Lee, M.W., Zhu, S.C.: I2T: Image parsing to text description. Proceedings of the IEEE **98** (2010)
7. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research (2013)
8. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2Text: Describing images using 1 million captioned photographs. NIPS (2011)
9. Socher, R., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. In: ACL. (2013)
10. Kuznetsova, P., Ordonez, V., Berg, A.C., Berg, T.L., Choi, Y.: Collective generation of natural image descriptions. In: ACL. (2012)
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL. (2002) 311–318
12. Hardoon, D., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis; an overview with application to learning methods. Neural Computation **16** (2004)
13. Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. IJCV (2013)
14. Gong, B., Grauman, K., Sha, F.: Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In: ICML. (2013) 222–230
15. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: ECCV. (2010) 213–226
16. Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A.: Data-driven visual similarity for cross-domain image matching. ACM SIGGRAPH ASIA **30**(6) (2011)
17. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Transactions on Graphics (SIGGRAPH) **26**(3) (2007)
18. Guillaumin, M., Ferrari, V.: Large-scale knowledge transfer for object localization in imageNet. In: CVPR. (2012) 3202–3209
19. Guillaumin, M., Verbeek, J., Schmid, C.: Multimodal semi-supervised learning for image classification. In: CVPR. (2010) 902–909
20. Quattoni, A., Collins, M., Darrell, T.: Learning visual representations using images with captions. In: CVPR. (2007)
21. Wang, G., Hoiem, D., Forsyth, D.: Building text features for object image classification. In: CVPR. (2009)
22. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In: TACL. (2014)

23. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV (2001)
24. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. PAMI **32**(9) (2010) 1582–1596
25. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
26. Jégou, H., Douze, M., Schmid, C., Perez, P.: Aggregating local descriptors into a compact image representation. In: CVPR. (2010)
27. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS. (2012)
28. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: A deep convolutional activation feature for generic visual recognition. CoRR **abs/1310.1531** (2013)
29. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. (2009)
30. Loper, E., Bird, S.: Nltk: The natural language toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1. (2002)
31. Weston, J., Bengio, S., Usunier, N.: Wsabie: Scaling up to large vocabulary image annotation. In: IJCAI. (2011)
32. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. JMLR (2011)
33. Zeiler, M.D.: ADADELTA: An adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)
34. Socher, R., Ganjoo, M., Sridhar, H., Bastani, O., Manning, C.D., Ng, A.Y.: Zero-shot learning through cross-modal transfer. In: NIPS. (2013)
35. Hotelling, H.: Relations between two sets of variables. Biometrika **28** (1936) 312–377
36. Gordo, A., Rodrıguez-Serrano, J.A., Perronnin, F., Valveny, E.: Leveraging category-level labels for instance-level image retrieval. In: CVPR. (2012)
37. Gopalan, R., Li, R., Chellappa, R.: Domain adaptation for object recognition: An unsupervised approach. In: ICCV. (2011)
38. Xu, Z., Chen, M., Weinberger, K.Q., Sha, F.: From sBoW to dCoT: Marginalized encoders for text representation. In: CIKM. (2011)
39. Rahimi, A., Recht, B.: Random features for large-scale kernel machines. NIPS (2007)
40. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: ICML. (2008) 1096–1103
41. Bengio, Y.: Learning deep architectures for AI. Foundations and Trends in Machine Learning **2**(1) (January 2009) 1–127