Im2Text and Text2Im: Associating Images and Texts for Cross-Modal Retrieval

Yashaswi Verma
http://researchweb.iiit.ac.in/~yashaswi.verma/
C. V. Jawahar
http://www.iiit.ac.in/~jawahar/

CVIT
IIIT-Hyderabad, India
http://cvit.iiit.ac.in

Abstract

Building bilateral semantic associations between images and texts is among the fundamental problems in computer vision. In this paper, we study two complementary cross-modal prediction tasks: (i) predicting text(s) given an image ("Im2Text"), and (ii) predicting image(s) given a piece of text ("Text2Im"). We make no assumption on the specific form of text; i.e., it could be either a set of labels, phrases, or even captions. We pose both these tasks in a retrieval framework. For Im2Text, given a query image, our goal is to retrieve a ranked list of semantically relevant texts from an *independent* text-corpus (i.e., texts with no corresponding images). Similarly, for Text2Im, given a query text, we aim to retrieve a ranked list of semantically relevant images from a collection of *unannotated* images (i.e., images without any associated textual meta-data).

We propose a novel Structural SVM based unified formulation for these two tasks. For both visual and textual data, two types of representations are investigated. These are based on: (1) unimodal probability distributions over topics learned using latent Dirichlet allocation, and (2) explicitly learned multi-modal correlations using canonical correlation analysis. Extensive experiments on three popular datasets (two medium and one web-scale) demonstrate that our framework gives promising results compared to existing models under various settings, thus confirming its efficacy for both the tasks.

1 Introduction

During the past decade, there has been a massive explosion of multimedia content on the Internet. As a result, several interesting as well as challenging research problems have emerged, one of them being automatically describing image content using text. While most of the earlier as well as recent works have focused on automatically annotating images using semantic labels [4, 6, 10, 19, 35, 38], in the past few years, describing images using phrases [11, 15, 29, 37], or one or more simple captions [5, 11, 14, 15, 16, 21, 22, 26, 37, 39, 40] have attained significant attention. A complementary problem to these is to automatically associate one or more semantically relevant images given a piece of text (i.e., a label, phrase or caption), and is commonly referred as the image retrieval task [2, 5, 6, 10, 24].

Although huge amount of *independent* visual and textual data are available today, only a small portion of them is linked with semantic associations. Hence, it comes as a natural choice to develop new models that can learn complex associations between the two modalities using this small portion, and then to apply them to automatically build associations

between the two in the larger independent space. In this work, we address this problem of learning bilateral associations between visual and textual data. We study two complementary tasks: (i) predicting text(s) given an image (*Im2Text*), and (ii) predicting image(s) given a piece of text (*Text2Im*). In contrast to several popular methods such as [5, 6, 10, 11, 15, 19, 22, 35, 37] that assume presence of data from both the modalities (visual and textual) during the testing phase, our approach has a motivation similar to the few known works (e.g. [12, 26]) that *do not* make such assumption. This means that for Im2Text, given a query image, our method can retrieve a ranked list of semantically relevant texts from a plain text-corpus that has *no* associated images. Similarly, for Text2Im, given a query text, it can retrieve a ranked list of images from an independent collection of images *without* any associated textual meta-data. The major contributions of this work are: **(1)** We propose a novel Structural SVM [32] based unified framework for both these tasks, which provides at least two advantages. First, Structural SVM provides a natural framework to work with complex and structured input/output spaces, and a unified framework helps in better understanding and appreciating the complementary nature of the two problems. And second, in practice this allows us to implement a general method, and adapt it for different forms of data with minimal modifications. **(2)** We examine generalization of different methods *across* datasets when textual data is in the form of captions. For this, we learn models from one dataset, and perform retrieval on other. To our knowledge, ours is the first such study in this domain.

To validate the applicability of our method, we conduct experiments on three popular datasets under different settings. We investigate two types of data representations: the first representation is based on probability distributions over high-level topics, and the second is based on explicitly learned cross-correlations based on the first representation. Rather than using raw features, such representations provide a semantically more meaningful and coherent platform for matching visual and textual data. Extensive evaluations demonstrate the superiority of the proposed framework compared to existing techniques.

2 Related Works

The problems of image and text retrieval are well-studied research topics [2, 20, 24, 30, 31]. Most of the existing approaches are based on retrieval of unimodal data; i.e., both query as well as retrieved data belong to the same modality (e.g., either image [31] or text [20]). Another approach that is popular among web-based search engines is to use textual meta-data associated with images during retrieval. Given a textual query, it is directly matched with this meta-data instead of looking at corresponding image. However, such images constitute only a small portion of the enormous amount of images available on the Internet, most of which are without such meta-data. This limitation has led to a growing interest in the problem of automatic image annotation [4, 8, 9, 10, 19, 35, 36, 38]. Such models can support label-based queries during image retrieval without assuming availability of any associated textual meta-data. Among these methods, perhaps WSABIE [38] is the only method that has been applied for large-scale image annotation task. However, such models fail to capture the relationships among different objects present in an image (e.g., "dog *in* car"). A recent work [29] relaxes this constraint, and learns models for *visual phrases* (e.g. "person riding bicycle"). Lately, there have been several attempts that use short captions to describe images [5, 11, 14, 15, 16, 21, 22, 33, 37, 39, 40] (and a few recent efforts such as [9, 28] to describe videos). Most of these works first try to predict the visual content of an image using some off-the-shelf computer vision technique (such as pre-trained object detectors and/or scene classifiers [14, 16, 21, 39], feature-based similarity with database images [11, 22, 37], or both [15, 22]).

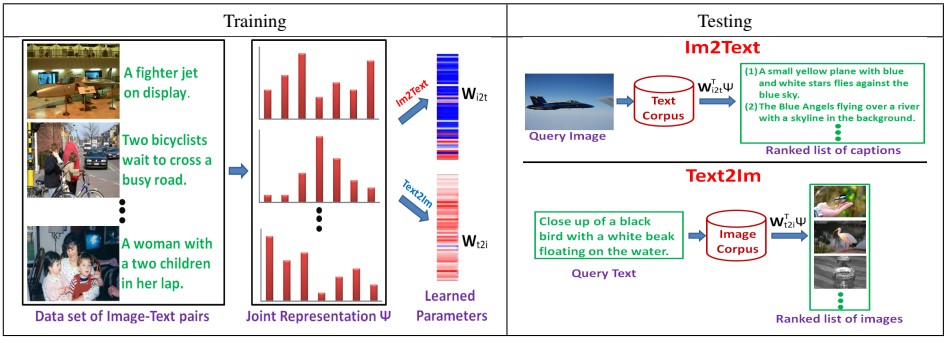

Figure 1: While training, given a dataset consisting of pairs of images and corresponding texts (here captions), we learn models for the two tasks (Im2Text and Text2Im) using a joint image-text representation. While testing for Im2Text, given a query image, we perform retrieval on a collection of only textual samples using the learned model. Similarly, for Text2Im, given a query text, retrieval is performed on a database consisting only of images.

This information is then fused using some Natural Language Generation (NLG) technique to construct image descriptions. All these works have shown that though the idea of *generating* captions provides a much larger space of possible descriptions that one could come-up with, most of these usually fail to match human-standards. One primary reason for this is the limitations of NLG which is still an emerging field. Few other works [5, 12, 22, 26] try to partly address this by directly transferring existing (human-written) captions to new images using visual clues. While [5, 22] do this by matching query image with database images, [12, 26] perform this by matching images and captions in a projected space learned using canonical correlation analysis (CCA). Though these demonstrate applicability of CCA on small captioned data, a recent work [7] demonstrates its applicability to large scale datasets for image annotation task when data has multiple views.

Our work relates with [3] which deals with multimodal clustering of web images that could be associated with noisy and/or sparse metadata (e.g., text, GPS coordinates, etc.). However, our primary aim is cross-modal retrieval by learning image-text associations without using similarities within a modality. Our work also closely relates with [27], in which images of text and text strings are first embedded into a vector space, and then a compatibility function is learned using Structural SVM that allows to perform both image retrieval as well as recognition. However, in our case, textual data is also complex and structured (e.g., captions), unlike [27] where each text string is considered as an individual category.

3 Bilateral Image-Text Retrieval

Now we present our framework for cross-modal retrieval. First we consider the task of retrieving semantically relevant text(s) given a query image (i.e., Im2Text). In Sec. 3.5, we will discuss how the same framework can be adopted for performing Text2Im as well.

Let $\mathcal{D} = \{(I_1, T_1), \ldots, (I_N, T_N)\}$ be a collection of images and corresponding texts. Each image I_i is represented using a p-dimensional feature vector $\mathbf{x}_i$ in space $\mathcal{X} = \mathbb{R}^p$. Similarly, each text T_i is represented using a q-dimensional feature vector $\mathbf{y}_i$ in space $\mathcal{Y} = \mathbb{R}^q$. We consider the problem of learning functions $f : \mathcal{X} \rightarrow \mathcal{Y}$ using the input-output pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\} \in \mathcal{X} \times \mathcal{Y}$. Similar to the Structural SVM framework [32], our objective is to learn a

discriminant function $F : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ that can be used to predict the optimal output $\mathbf{y}^*$ given an input $\mathbf{x}$ by maximizing F over the space $\mathcal{Y}$. That is, $\mathbf{y}^* = f(\mathbf{x};\mathbf{w}) = \underset{\mathbf{y}\in\mathcal{Y}}{\text{argmax}}\ F(\mathbf{x},\mathbf{y};\mathbf{w})$, where $\mathbf{w}$ is the parameter vector that needs to be learned. We make the assumption of $F = \mathbf{w}\cdot\Psi(\mathbf{x},\mathbf{y})$; i.e., F is a linear function of the joint feature representation $\Psi(\cdot)$ of input-output pair. The task of learning $\mathbf{w}$ is then formulated as the following optimization problem:

$$\min_{\mathbf{w},\xi\geq 0}\ \frac{1}{2}||\mathbf{w}||_2^2 + \frac{C}{N}\sum_{i=1}^{N}\xi_i \quad s.t.\ \mathbf{w}\cdot\Psi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}\cdot\Psi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y}_i,\mathbf{y}) - \xi_i\ \ \forall i, \mathbf{y}\in\mathcal{Y}\setminus\{\mathbf{y}_i\} \quad (1)$$

where $||\cdot||_2^2$ denotes squared L_2-norm, $C > 0$ is a constant that controls the trade-off between regularization term and loss term, ξ_i denotes slack variable, and $\Delta(\mathbf{y}_i,\mathbf{y})$ denotes loss function that acts as a margin for penalizing any prediction other than the true output. The set of constraints in the above optimization problem signify that for every sample $\mathbf{x}_i$, the parameter vector $\mathbf{w}$ should be learned such that the prediction score for the true output (i.e. $F(\mathbf{x}_i,\mathbf{y}_i;\mathbf{w})$) remains higher than the prediction score for any other output by a margin.

3.1 Joint Image-Text Representation

The purpose of $\Psi(\cdot)$ is to provide a joint representation for input and output data depending on their individual representations. As discussed in Sec 4, we use identical representations for both visual (image) and textual data based on probability distributions over latent topics learned from corresponding modalities. Thus, for a given sample (image or text), each dimension of its feature vector corresponds to the probability of that particular topic. Since these topics are learned independently for images and text (as the two modalities are fundamentally different in their original forms), it is not straightforward to obtain direct correspondence among the topics of the two modalities. However, given an image-text pair (I,T), we know that essentially both I and T represent similar semantic content, though in different forms. This means that there should exist some (indirect) correspondence across their topics as well. To learn this correspondence, one feasible choice is to consider all possible pairs of topics across the two modalities (by taking a cross-product between them), and then learn weights over these pairs. These weights would signify the relative correspondence of every topic-pair; i.e., if a topic-pair has high score, then it is quite likely that the individual topics in that pair represent similar semantic concept in the two modalities. With this motivation, we propose to use the joint representation constructed from the input-output representations $\mathbf{x}$ and $\mathbf{y}$ using their tensor product. That is, each dimension of $\mathbf{x}$ is multiplicatively combined with every dimension of $\mathbf{y}$ to get $\Psi(\mathbf{x},\mathbf{y}) = \mathbf{x}\otimes\mathbf{y} \in \mathbb{R}^r$, where $r = p\times q$. This representation has apparent advantage not only in efficiently capturing linear correlations between input-output modalities, but also provides computational benefits.

3.2 Loss Function

The function $\Delta(\cdot)$ in Eq. 1 is a problem specific loss function. It acts as a margin in the Structural SVM framework, and is used to penalize incorrect predictions against the true output. It is defined such that given an input-output pair $(\mathbf{x}_i,\mathbf{y}_i)$ and any incorrect output $\mathbf{y}$, the value of $\Delta(\mathbf{y}_i,\mathbf{y})$ should depend on *similarity* between $\mathbf{y}_i$ and $\mathbf{y}$. If $\mathbf{y}_i$ and $\mathbf{y}$ are similar, the loss value should be small and vice-versa.

Projecting output (textual) data into a semantic space defined in the form of a vector space $\mathcal{Y}$ allows us to adopt any suitable distance metric defined in vector space as our choice of loss function. Though this projection can be highly non-linear in nature, the assumption

here is that the projected space keeps the semantic proximity of the data intact; i.e., data points that are semantically similar are closer to each other in the projected vector space, than the data points that are semantically dissimilar to each other [1]. Here we consider two popular distance metrics as loss functions: Manhattan distance $\Delta_M(\cdot)$ and squared Euclidean distance $\Delta_E(\cdot)$. Thus, the two loss function are defined as:

$$\Delta_M(\mathbf{y}_i,\mathbf{y}) = ||\mathbf{y}_i-\mathbf{y}||_1, \;\; and \;\; \Delta_E(\mathbf{y}_i,\mathbf{y}) = ||\mathbf{y}_i-\mathbf{y}||_2^2\,, \tag{2}$$

where $||\cdot||_1$ denotes L_1-norm. Since both $\Delta_M(\cdot)$ and $\Delta_E(\cdot)$ are distance metrics, they satisfy the properties of a valid loss function; i.e., $\Delta_Z(\mathbf{y}_i,\mathbf{y}_i)=0$, $\Delta_Z(\mathbf{y}_i,\mathbf{y}_j)\geq 0$ for $i\neq j$, and $\Delta_Z(\mathbf{y}_i,\mathbf{y}_j)\geq\Delta_Z(\mathbf{y}_i,\mathbf{y}_i)$ for $i\neq j$ (where $Z\in\{M,E\}$). These loss functions can be evaluated efficiently, which also helps in efficiently computing the most violated constraint [32] while solving Eq. 1, as discussed next.

3.3 Solving the Optimization Problem

We solve Eq. 1 using a cutting-plane algorithm [32]. It requires efficient computation of the most violated constraint during each iteration. Given an input-output pair $(\mathbf{x}_i,\mathbf{y}_i)$, the most violated constraint is the constraint corresponding to the incorrect prediction $\hat{\mathbf{y}}$ with maximum score predicted using the current learned parameter vector $\mathbf{w}$. It is given by:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}\in\mathcal{Y}\setminus\{\mathbf{y}_i\}}{\operatorname{argmax}} \;\; \Delta(\mathbf{y}_i,\mathbf{y}) + \mathbf{w}\cdot\Psi(\mathbf{x}_i,\mathbf{y}) - \mathbf{w}\cdot\Psi(\mathbf{x}_i,\mathbf{y}_i) \tag{3}$$

Since the last term is constant with respect to $\mathbf{y}$, it can be dropped. For the two loss functions in Eq. 2, this maps to the following problems respectively:

$$\hat{\mathbf{y}}_M = \underset{\mathbf{y}\in\mathcal{Y}\setminus\{\mathbf{y}_i\}}{\operatorname{argmax}} \;\; ||\mathbf{y}_i-\mathbf{y}||_1 + \mathbf{w}\cdot\Psi(\mathbf{x}_i,\mathbf{y}), \;\; and \;\; \hat{\mathbf{y}}_E = \underset{\mathbf{y}\in\mathcal{Y}\setminus\{\mathbf{y}_i\}}{\operatorname{argmax}} \;\; ||\mathbf{y}_i-\mathbf{y}||_2^2 + \mathbf{w}\cdot\Psi(\mathbf{x}_i,\mathbf{y}) \tag{4}$$

It can be easily verified that both the equations correspond to maximizing a convex function. In practice, since we consider normalized feature representations (Sec. 4), every element of a feature vector remains bounded within a range. Precisely, we assume that every $\mathbf{y}\in[a,b]^q$, and is either L_1- or L_2-normalized. This allows us to solve the problems in Eq. 4 efficiently using an iterative gradient-ascent and projection method. To solve the optimization problem in Eq. 1, we use [34].

3.4 Retrieving a Ranked List of Output

Consider an independent database $\mathcal{T}'=\{T'_1,\ldots,T'_{|\mathcal{T}'|}\}$ consisting of only textual samples, where each T'_k is represented using a feature vector $\mathbf{y}'_k\in\mathcal{Y}$. Once the parameter vector $\mathbf{w}$ is learned, given a query image J represented by $\mathbf{x}_J\in\mathcal{X}$, Im2Text requires ranking the elements of $\mathcal{T}'$ according to their semantic relevance with J. This can be performed by sorting the elements of $\mathcal{T}'$ based on the score $F(\mathbf{x}_J,\mathbf{y}_k;\mathbf{w})=\mathbf{w}\cdot\Psi(\mathbf{x}_J,\mathbf{y}_k)$, $\forall k\in\{1,\ldots,|\mathcal{T}'|\}$ (with higher score corresponding to greater relevance and vice-versa), thus allowing to retrieve a ranked list of texts given an image.

3.5 Performing Text2Im

It is simple to verify that in order to perform Text2Im under the same framework, all we require is to swap the input and output spaces; i.e. now $\mathbf{y}\in\mathcal{Y}$ would represent the *input*

[1] Since an analogous projection is also applied on the input (image) data, this allows us to learn a mapping (i.e., $\mathbf{w}$) between input-output space in a discriminative manner.

space defined over textual features, and $\mathbf{x} \in \mathcal{X}$ would represent the *output* space defined over image features. Because we use identical representations for both visual and textual data (discussed in Sec. 4), the loss functions defined in Eq. 2 over textual feature will remain valid for visual features as well. However, since this is an inverse problem of Im2Text, we learn separate model in this case. As the proposed approach performs the two complementary tasks (Im2Text and Text2Im) under a single unified framework, we shall refer to it as *Bilateral Image-Text Retrieval* (or *BITR*). Figure 1 explains the gist of our framework.

4 Representing Visual and Textual Data

We consider two types of representations for visual and textual data. The first representation captures high-level semantics of data in the form of unimodal topic distributions. We refer to this as *semantic representation* (or *SR*). The second representation combines SR with cross-modal correlations learned using CCA between input and output space. We refer to this as *correlated semantic representation* (or *CSR*). The two representations are described below.

4.1 Semantic Representation (SR)

This representation is based on probability distribution over topics learned using latent Dirichlet allocation model [1]. It is one of the most popular probabilistic generative topic models, and has been known to efficiently capture complex semantics of data. It first discovers topics from documents (collections of discrete units) based on multinomial distribution, and then provides a representation for each document as a probability distribution over these topics.

4.1.1 Representing Images

Each image is first represented as a histogram of Bag-of-Words (BoW) of dense SIFT features [18]. From training images of SBU dataset [22], $0.5M$ SIFT features are randomly sampled, and a vocabulary of 1000 visual words is learned using k-means algorithm. Using this vocabulary, visual topics are learned using 5000 random images from SBU dataset. Finally, each image is represented as a probability distribution over the learned topics.

4.1.2 Representing Text

Representation of textual data varies under different settings depending upon its given form:
Representing Captions: To learn textual topics, we use the captions of SBU dataset [22]. First, we build a vocabulary of around $0.18M$ (textual) keywords after removing stop-words, and then use captions of training data to learn topics. Finally, each caption is represented as a probability distribution over these topics.
Representing Phrases: Here we assume an annotated dataset where each image is tagged with a set of phrases. We learn textual topics by considering each phrase as a discrete unit, and then represent each phrase as a probability distribution over them.
Representing Labels: Similar to the previous case, we assume an annotated dataset of images tagged with a set of labels. While learning topics, each label is considered as a discrete unit. After that, each label is represented as a probability distribution over the learned topics.

4.2 Correlated Semantic Representation (CSR)

This is based on the assumption that an image and its corresponding text are two heterogeneous representations of similar information. Under this assumption, given the independently obtained semantic representations for visual and textual data (Sec. 4.1), they are

mapped into maximally correlated vector subspaces using CCA [13]. In practice, the new representations are L_2-normalized before forming the joint representation.

5 Experiments

Now we demonstrate the applicability of our method for different forms of textual data. We learn 100 topics individually for each modality; i.e. each visual and textual sample is represented using a 100-dimensional feature vector ($p = q = 100$). In all the experiments, we report results using the two loss functions given in Eq. 2, and will refer to them as BITR-M and BITR-E respectively. Also, we will denote the particular representation being employed using (SR) or (CSR).

As discussed in Sec. 2, WSABIE [38] and CCA [13] are two well-known methods that can scale to large datasets and have been shown to work well for learning cross-modal associations. Hence, we show comparisons with these in all our experiments. While CCA based methods have been used previously under such settings [12, 26], WSABIE [38] was originally proposed for the task of label-ranking and hence can not be directly applied for captions. We do this by adapting the WSABIE algorithm such that instead of learning a separate parameter vector for each label, we learn a single parameter matrix for all the captions. This is analogous to the parameter matrix being learned for visual features in the WSABIE algorithm [2]. Both CCA and WSABIE learn separate projection matrices for input and output data. In practice, they both may converge to a lower dimensional space compared to the dimensionality of given data. However, in all our experiments, we project data into the same space for both these methods. This not only minimizes information loss, but also allows fair comparisons and avoids the need of tuning on optimal number of projections required by each. Also, for CCA, we use normalized correlation in order to compute similarity between two projected features, since it was found to achieve best results in [26]. Along with these two methods, we also consider weighted k-nearest neighbours (wKNN) algorithm (similar to [10]) and one-vs.-rest SVM for additional comparisons in Experiment-3 while considering phrases/labels as textual data, as these methods are easily applicable in that setting.

5.1 Experiment-1 (Image-Caption Retrieval)

Here we consider textual data to be in the form of captions. We conduct experiments on three datasets: UIUC Pascal Sentence dataset [25], IAPR TC-12 benchmark [8], and SBU-Captioned Photo dataset [22]. While Pascal Sentence and IAPR TC-12 benchmark are medium-scale datasets that contain 1000 and 19627 captioned images respectively, SBU Photo is a web-scale dataset with $1M$ captioned images. In both IAPR TC-12 and SBU datasets, there is one caption per image, while in Pascal Sentence dataset, each image is captioned with five independent captions.

Experimental Setting: For SBU dataset, we follow the train/test split of [22], which includes 500 test images and rest ($0.9995M$) as training images. To learn parameters for different models, we use a subset of $0.1M$ samples randomly sampled from training data. During testing phase, we perform retrieval over all training samples (captions for Im2Text, and images for Text2Im). For the other two datasets, we compute performance over all the samples similar to [11, 39]. This is done by creating ten partitions, and each time considering one for testing and rest for training. For evaluation, we use BLEU [23] and Rouge [17] metrics [3].

[2] More details on our extension of WSABIE for captions are provided in the supplementary file.

[3] To compute BLEU scores, we use the code provided by NIST (version-13a). And to compute Rouge scores, we use Release-1.5.5 provided by http://www.berouge.com/Pages/default.aspx.

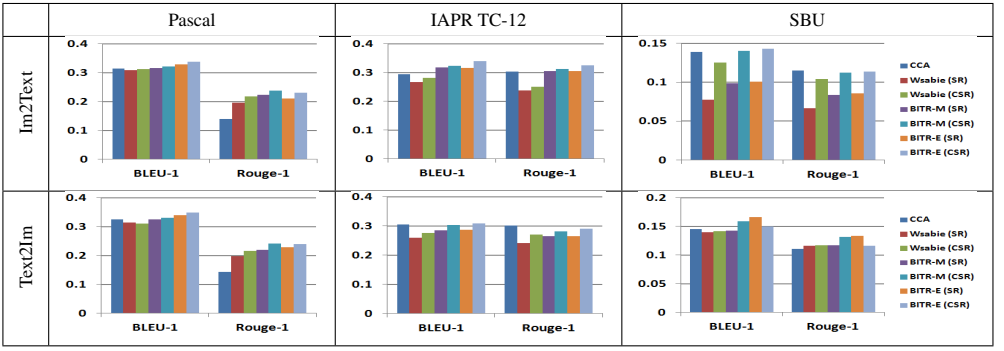

Figure 2: Results for image-caption retrieval (Sec. 5.1).

Data Set →	Pascal				IAPR TC-12				SBU			
Method ↓	B-1	B-2	B-3	R-1	B-1	B-2	B-3	R-1	B-1	B-2	B-3	R-1
Ordonez *et al.* [22]	–	–	–	–	–	–	–	–	0.13	–	–	–
Gupta *et al.* [11]	0.36	0.09	0.01	0.21	0.15	0.06	0.01	0.14	–	–	–	–
Ours	0.33	0.10	0.06	0.23	0.32	0.13	0.07	0.33	0.14	0.05	0.03	0.10

Table 1: Comparison between previously reported results and our best results for Im2Text under Experiment-1 (B-n means n-gram BLEU score, and R-1 means 1-gram Rouge score).

These have also been used by previous methods [11, 14, 16, 22, 37] that describe images (i.e., Im2Text). For both these metrics, higher score means better performance. For both Im2Text and Text2Im, we report mean one-gram BLEU and Rouge scores. For Im2Text, these scores are averaged over the top five retrieved captions, by matching them with the ground-truth caption of query image. For Text2Im, we compute these scores by matching the query caption with ground-truth captions of top five retrieved images.

Results: Figure 2 shows performance of different methods on the three datasets for the two tasks. Following observations can be made from these results: (i) For most of the cases, BITR-E (CSR) outperforms all other methods. This implies that Euclidean distance based loss better models dataset specific patterns. (ii) For all the three methods (i.e., WSABIE, BITR-E and BITR-M), the performance usually improves by using CSR as compared to SR. This reflects the advantage of explicitly infusing cross-correlations into data representation. (iii) For Pascal dataset, relative performances of different methods follow almost similar trends for Im2Text and Text2Im. However, there is comparatively more diversity on the other two datasets. This could be because Pascal dataset is relatively much smaller than the other two datasets, and the number of semantic concepts it covers is quite less. This may result into dataset specific biases into the learned models, and thus reflects the necessity of evaluations on big and diverse datasets such as IAPR TC-12 and SBU.

In Table 1, we compare our best results on Im2Text with [22] and [11] [4]. Since both [22] and [11] use a dataset consisting of both the modalities during testing phase, and [11] generates captions rather than retrieving them, these results are not directly comparable. How-

[4] While [22] and [11] are the only works that have previously reported results on SBU and IAPR TC-12 datasets respectively, it was shown in [11] that their method outperformed other methods such as [14, 39] on the Pascal dataset. Hence we compare only with these two methods.

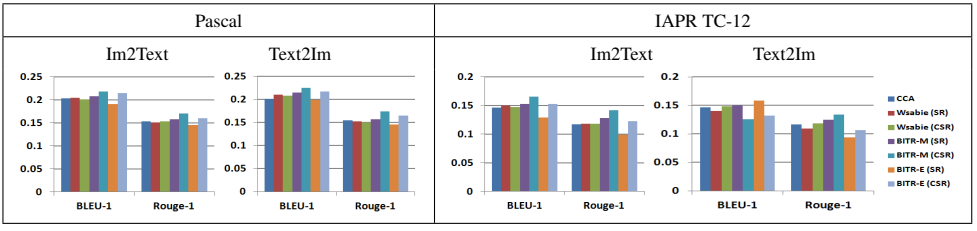

Figure 3: Results for cross-dataset image-caption retrieval (Sec. 5.2).

ever, it is worth noticing that even by matching images directly with captions, our method performs comparable or superior than the other two. This reflects the effectiveness of our framework in learning semantic associations between the two modalities.

5.2 Experiment-2 (Cross-dataset Image-Caption Retrieval)

In this experiment, we analyze the generalization ability of different methods across datasets. For this, we consider textual data to be in the form of captions as in Experiment-1, and follow the same evaluation protocol. Here, instead of learning models for each dataset individually, we use the models learned using SBU dataset in Experiment-1 and evaluate the performance on the other two datasets, i.e. Pascal and IAPR TC-12. Precisely, for Im2Text, we consider query images from Pascal or IAPR TC-12 dataset, and perform retrieval on the captions of SBU dataset. Similarly, for Text2Im, we consider query caption from Pascal or IAPR TC-12 dataset, and perform retrieval on the image collection of SBU dataset. The goal of this experiment is to analyze the effect of dataset specific biases, and to the best of our knowledge, ours is the first such study in this domain.

Results: Figure 3 shows the results for this experiment. Following observations can be made from these results: (i) For all the methods, the performance degrades significantly compared to that in Experiment-1. This reflects the impact of dataset specific biases, and thus emphasizes the necessity of performing cross-dataset evaluations. (ii) Unlike Experiment-1, BITR-M usually performs better than BITR-E. This is because Manhattan distance is known to be more robust than Euclidean distance against noise/outliers in data. Moreover, it (BITR-M (CSR)) also mostly outperforms other methods. This suggests that $\Delta_M(\cdot)$ could practically be a better choice than $\Delta_E(\cdot)$ for real-world applications.

5.3 Experiment-3 (Image-Phrase and Image-Label Retrieval)

Here we consider textual data to be in the form of either phrases or labels, and demonstrate results on IAPR TC-12 dataset [8]. In case of phrases, we extract them from available captions using the Stanford CoreNLP toolkit [5]. In practice, we extract three types of phrases: (*noun*, *verb*), (*noun*, *preposition*, *noun*) and (*verb*, *preposition*, *noun*). In case of labels, we use the same set of annotations as in [10, 19, 35].

Experimental Setting: We create ten partitions of the dataset and report averaged results over ten trials, each time considering one partition for testing and rest for training. For Im2Text, given a query image, we rank the phrases (labels) as discussed in Sec. 3.4. For Text2Im, given a query phrase (label), we rank images in an analogous manner. Here, since

[5] http://nlp.stanford.edu/software/corenlp.shtml

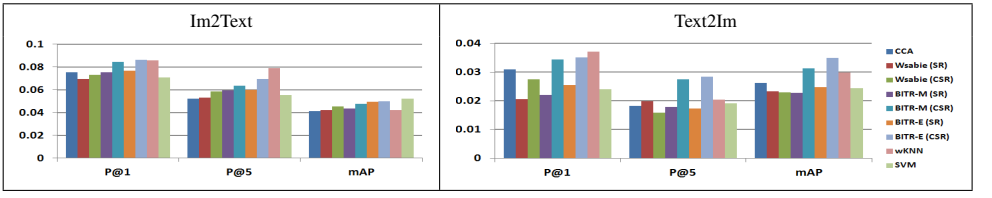

Figure 4: Results for image-phrase retrieval on IAPR TC-12 dataset.

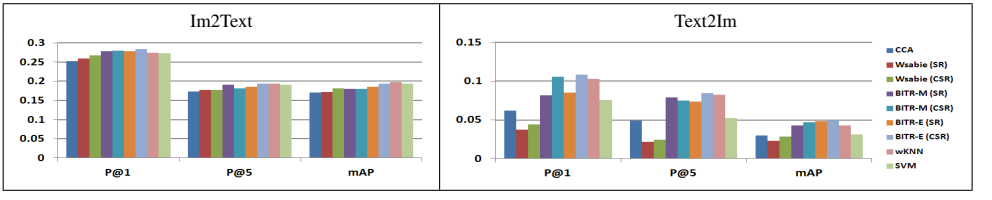

Figure 5: Results for image-label retrieval on IAPR TC-12 dataset.

we are dealing with individual phrases (labels), we use the original WSABIE algorithm [38], and not the modified one as in Experiment-1. We follow the popular metrics of Precision@1 (P@1), Precision@5 (P@5) and mean Average Precision (mAP) for performance evaluation.
Results: Figure 4 and Figure 5 compare different methods when textual data is in the form of phrases and labels respectively. Note that in contrast to all other methods, wKNN makes use of data from both the modalities during testing phase. Due to this, despite its simplicity, it mostly achieves very encouraging results compared to other methods. Under this setting also, our methods (particularly BITR-E (CSR)) demonstrate competitive performance, and perform comparable/superior to all other methods. We can also observe that the results for phrases and labels follow quite similar trends. This is expected since in both the experiments, we consider each phrase/label as a discrete unit, thus focusing only on the co-occurrence of phrases/labels. An interesting direction for future work would be to build better representations for phrases that could capture hierarchical semantic correlations (among words co-occurring in a phrase, and among phrases co-occurring in an annotation).

6 Conclusion

We have presented a novel Structural SVM based framework to perform cross-modal multimedia retrieval. Under this framework, we have investigated two types of data representations based on high-level semantic topics and cross-correlations. We have demonstrated the applicability of our method for different forms of textual data using two medium and one web-scale dataset. For both Im2Text and Text2Im, our method achieved promising results and mostly outperformed existing techniques. In this work, we have considered visual (image) and textual data as the two modalities, nevertheless the fundamental ideas discussed are straightaway applicable to cross-modal retrieval tasks in other domains as well.

Acknowledgment

Yashaswi Verma is supported by Microsoft Research India PhD fellowship 2013.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 12(1):234–278, 2003.

[2] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences and trends of new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

[3] Kun Duan, David J. Crandall, and Dhruv Batra. Multimodal learning in loosely-organized web images. In *CVPR*, 2014.

[4] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.

[5] Ali Farhadi, Mohsen Hejrati, Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences for images. In *ECCV*, 2010.

[6] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[7] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2): 210–233, 2013.

[8] M. Grubinger. *Analysis and Evaluation of Visual Information Systems Performance.* PhD thesis, Victoria University, Melbourne, Australia, 2007.

[9] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[10] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbour models for image auto-annotation. In *ICCV*, 2009.

[11] Ankush Gupta, Yashaswi Verma, and C. V. Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, 2012.

[12] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 47:853–899, 2013.

[13] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[14] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Baby Talk: Understanding and generating simple image descriptions. In *CVPR*, 2011.

[15] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. In *ACL*, 2012.

[16] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In *CoNLL*, 2011.

[17] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACLHLT*, 2003.

[18] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60 (2):91–110, 2004.

[19] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. A new baseline for image annotation. In *ECCV*, 2008.

[20] C. Meadow, B. Boyce, D. Kraft, and C. Barry. Text information retrieval systems. *Emerald Group Pub Ltd*, 2007.

[21] Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Sratos, Xufeng Han, Alysssa Mensch, Alexander C. Berg, Tamara L. Berg, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections . In *EACL*, 2012.

[22] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[23] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, 2002.

[24] M. Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: overview of the imageclefphoto task 2009. *CLEF working notes*, 2009.

[25] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotation using amazon's mechanical turk. In *NAACLHLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[26] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 2010.

[27] Jose Rodriguez and Florent Perronnin. Label embedding for text recognition. In *BMVC*, 2013.

[28] Marcus Rohrbach, Wei Qiu, and Ivan Titov. Translating video content to natural language descriptions. In *ICCV*, 2013.

[29] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011.

[30] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, 2006.

[31] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *PAMI*, 22(12):1349–1380, 2000.

[32] Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004.

[33] Yoshitaka Ushiku, Tatsuya Harada, and Yasuo Kuniyoshi. Understanding images with natural sentences. In *ACM MM*, 2011.

[34] A. Vedaldi. A MATLAB wrapper of SVM^{struct}. http://www.vlfeat.org/~vedaldi/code/svm-struct-matlab.html, 2011.

[35] Yashaswi Verma and C. V. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *ECCV*, 2012.

[36] Yashaswi Verma and C. V. Jawahar. Exploring SVM for image annotation in presence of confusing labels. In *BMVC*, 2013.

[37] Yashaswi Verma, Ankush Gupta, Prashanth Mannem, and C. V. Jawahar. Generating image descriptions using semantic similarities in the output space. In *V&L Net Workshop on Language for Vision, in conjunction with CVPR*, 2013.

[38] Jason Weston, Samy Bengio, and Nicolas Usunier. WSABIE: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011.

[39] Y. Yang, C. L. Teo, Hal Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011.

[40] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2T: Image parsing to text description. In *Proceedings of the IEEE*, 2008.