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

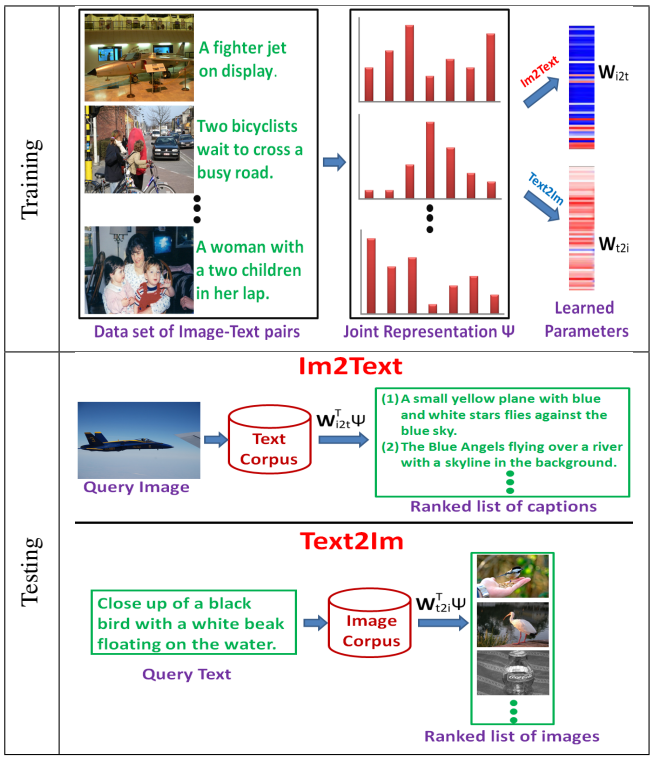

Figure 1: While training, given a dataset consisting of pairs of images and corresponding texts (here captions), we learn models for the two tasks (Im2Text and Text2Im) using a joint image-text representation. While testing for Im2Text, given a query image, we perform retrieval on a collection of only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

To automatically describe image content using text is one of the challenging and interesting research problems in computer vision. A complementary problem to this is to automatically associate semantically relevant image(s) given a piece of text, and is commonly referred as the image retrieval task. In this work, we address the 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"). While several existing methods (*e.g.*, [1]) assume presence of data from both the modalities during the testing phase, the motivation of this work is similar to the few known works (*e.g.*, [2]) that do not make such assumption. This means that for Im2Text, given a query image, our method retrieves 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 retrieves a ranked list of images from an independent image collection without any associated textual meta-data. The major contributions of this work are:

(1) We propose a novel Structural SVM based unified framework for both these tasks. We use vector representations for both visual (image) and textual data that are based on probability distributions over latent topics. From these, we form a joint feature vector using tensor product of input and output representations. Because the output data is represented in the form of a vector, we use Manhattan (M) and Euclidean (E) distances as our loss functions. As the proposed approach performs the two complementary tasks (Im2Text and Text2Im) under a single unified framework, we refer to it as Bilateral Image-Text Retrieval (or BITR). Figure 1 explains the gist of our framework.

(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 best knowledge, ours is the first such study in this domain.

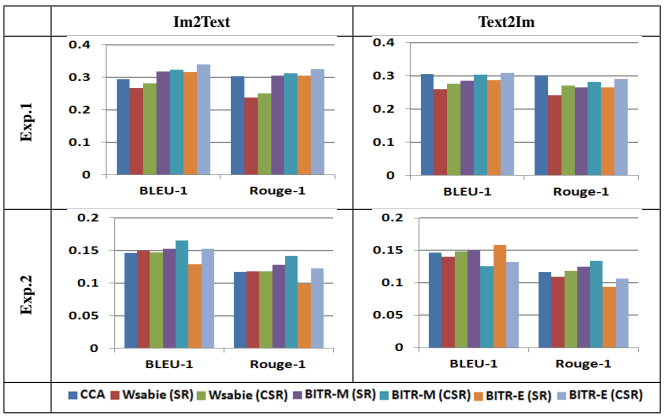

Figure 2: Results on IAPR TC-12 dataset for within-dataset (top) and cross-dataset (bottom) image-caption retrieval.

We conduct experiments on three datasets (UIUC Pascal Sentence dataset, IAPR TC-12 benchmark, and SBU-Captioned Photo dataset), and compare our approach with WSABIE [3] and CCA. These are two well-known methods that can scale to large datasets and have been shown to work well for learning cross-modal associations. While CCA based methods have been used previously under such settings [2], WSABIE was originally proposed for the task of label-ranking and hence can not be directly applied for captions. We do this by adapting it for captions, the details of which are provided in the supplementary file. 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 learned using latent Dirichlet allocation. We refer to this as semantic representation (or SR). The second representation combines SR with cross-modal correlations learned between input and output space. We refer to this as correlated semantic representation (or CSR).

We perform experiments under different settings when textual data is in the form of either captions, or phrases, or labels. Here we discuss the two experiments when textual data is in the form of captions. In the first experiment (Exp.1), we learn dataset-specific models separately for both the tasks (Im2Text and Text2Im). And in the second experiment (Exp.2), we analyze the generalization ability of different methods across datasets. For this, instead of learning models for each dataset individually, we use the models learned using SBU dataset in Exp.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. In both Exp.1 and Exp.2, we use BLEU and Rouge metrics for evaluation.

Figure 2 compares the performances of different methods on IAPR TC-12 dataset (please refer the paper for more results). Here, we can observe that: (a) For all the three methods, the performance usually improves by using CSR as compared to SR. This indicates the advantage of explicitly infusing cross-correlations into data representation. (b) In cross-dataset experiment (Exp.2), the performance of all the methods degrades significantly compared to that in Exp.1. This reflects the impact of dataset specific biases, and thus emphasizes the necessity of performing cross-dataset evaluations. (c) For most of the cases, the proposed method achieves promising results and mostly outperforms existing techniques.

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

[2] 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.

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