Int J Multimed Info Retr manuscript No.
(will be inserted by the editor)

Cross-Specificity: Modelling Data Semantics for Cross-Modal Matching and Retrieval

Yashaswi Verma · Abhishek Jha · C. V. Jawahar

Received: date / Accepted: date

Abstract While dealing with multi-modal data such as pairs of images and text, though individual samples may demonstrate inherent heterogeneity in their content, they are usually coupled with each other based on some higher-level concepts such as their categories. This shared information can be useful in measuring semantics of samples across modalities in a relative manner. In this paper, we investigate the problem of analyzing the degree of specificity in the semantic content of a sample in one modality with respect to semantically similar samples in another modality. Samples that have high similarity with semantically similar samples from another modality are considered to be specific, while others are considered to be relatively ambiguous. To model this property, we propose a novel notion of "cross-specificity". We present two mechanisms to measure cross-specificity: one based on human judgment and other based on an automated approach. We analyze different aspects of cross-specificity, and demonstrate its utility in cross-modal retrieval task. Experiments show that though conceptually simple, it can benefit several existing cross-modal retrieval techniques, and provides significant boost in their performance.

Keywords Cross-media analysis · Semantic matching · Cross-modal retrieval

1 Introduction

As a result of an ever growing multimedia content, an increasing number of real-world applications now deal with multiple modalities. These include multi-modal classification [7,13], multi-modal retrieval [4], multi-modal clustering [5], cross-modal retrieval [12,8,17,6, 16,23], etc. One of the challenges while dealing with multiple modalities is that of the inherent heterogeneity among samples within a modality as well as across different modalities. This can be partly addressed by grouping samples based on some higher-level concepts such as their categories [6,16]. The category of a sample plays a central role in expressing its underlying semantics. Moreover, if two modalities are known to share a common set of categories, this can be useful in modelling their mutual semantics with respect to each other.

In this paper, we make an attempt towards leveraging this shared information to model cross-modal semantics of a sample. For this, we introduce the notion of *cross-specificity*. Given collections of samples from two different modalities that share a common set of semantic categories, cross-specificity measures how well a sample in one modality portrays its (categorical) semantics relative to samples that belong to the same category in another modality. A sample with high cross-specificity score is considered to be specific with respect to its semantically similar samples in another modality, while that with low score is considered to be comparatively ambiguous. Modelling this association can benefit a variety of applications that involve multiple modalities. As we will show later, one such example is the well-known cross-modal retrieval task, where given a query in one modality, the goal is to retrieve semantically relevant samples from another modality. In this task, if a sample is found to be ambiguous in depicting its underlying semantic category with respect to samples in another modality (i.e., has low cross-specificity), an explicit boosting mechanism can be used to enhance

* Corresponding author (E-mail: yashaswiv@iisc.ac.in)
YV is with Indian Institute of Science (IISc), and AJ and CVJ are with International Institute of Information Technology Hyderabad (IIIT-H), India ·

its predictability. However, if a sample is specific, we may not require such boosting.

Given a sample from a particular category (or class) in one modality, we measure its cross-specificity score using two mechanisms: one based on human judgement of its similarity with samples that belong to the same class in another modality, and the other using an automatic similarity measure. We then demonstrate how cross-specificity can benefit cross-modal retrieval task. Experiments show that though simple, it provides consistent improvements in several cross-modal retrieval techniques. Additionally, we also analyze different aspects of cross-specificity such as correlation between human and automated cross-specificity measurements, influence of the size of training data, etc.

2 Related work

Given the increasing amount of data in the form of multiple modalities, there has been considerable interest to leverage it to learn richer models for various applications, rather than relying upon individual modalities independently. This has been found to be particularly useful in modelling visual data, where additional cues in the form of textual tags, captions, GPS coordinates, etc. can provide great amount of semantically valuable information. Examples are work on multi-modal and cross-modal modelling, and understanding image properties. Here we briefly discuss some of the research in these areas in the given context.

Multi-modal and Cross-modal modelling: Several papers study multiple modalities to address a particular task. Guillaumin *et al.* [7] and Li *et al.* [13] use tags as additional features for learning image classification models. McAuley and Leskovec [14] model pairwise relations between images using relational metadata in the form of social connections. In cross-modal analysis, one needs to learn a function that can measure the degree of similarity between a pair of samples from diverse modalities. Several of these approaches try to learn a common subspace for matchings samples from two modalities [8,6,16,18]. In [8], canonical correlation analysis (CCA) is used for learning a common embedding space using paired samples from two modalities. In [17], samples are represented using logistic regressors learned from data semantics, thus allowing direct matching between cross-modal samples. In [23], the problem of cross-modal retrieval is posed in a Structural SVM framework, and is shown to be applicable for both homogeneous as well as heterogeneous representations of cross-modal samples. Recent methods such as [12,6,16] additionally make use of categorical/semantic information, which helps in learning discriminative cross-modal matching functions. Lately, some approaches based on deep neural networks [20,1, 22] have also been proposed for learning associations between images and text.

It is worth noting that while most of these papers focus on *matching* a pair of samples from two modalities based on some cross-modal matching function, our main contribution is to propose a novel *property* of samples in cross-modal data – cross-specificity – that captures the degree of specificity in the content of a sample in one modality relative to its semantically similar samples in another modality. As our second contribution, we also demonstrate how it can benefit existing cross-modal retrieval approaches.

Image properties: There have also been several attempts to study image properties such as object saliency [9,11], likelihood of a certain object being mentioned in its description [2,21], and variance in human perception in perceiving and describing an image [10]. Unlike these, our focus is on estimating properties of a sample based on its semantically similar samples from another modality.

The idea of cross-specificity is closely related to the work of Jas and Parikh [10]. They introduced the notion of "image-specificity" that estimates the degree of specificity in the visual content of an image based on pair-wise similarity among all the captions describing that image. We also aim at estimating the specificity in the content of a sample, however we do this in a relative manner based on similarity among samples across different modalities that share common semantics. Moreover, unlike [10], we do not assume any particular modality, that makes cross-specificity a more general concept. We will conceptually compare cross-specificity and image-specificity in more detail in Sec. 3.2.

3 Cross-Specificity

Here, first we describe two ways to measure cross-specificity, and then present a conceptual contrasting with image-specificity [10].

3.1 Measuring Cross-Specificity

Let S_X and S_Y be sets of samples from two modalities (*e.g.*, images and text). We are also given a set of classes $\mathcal{C}$ such that each sample in both the modalities is associated with a class $c \in \mathcal{C}$. We define the cross-specificity of a sample from a particular class as its average pairwise similarity with all the samples from the same class in the other modality. E.g., let $x \in S_X$ be a sample from

class c, and $Y_c \subset S_Y$ be the subset of all the samples that belong to the same class in the other modality. To compute cross-specificity of x, we compute its pairwise similarity with all the samples $y \in Y_c$ and average the scores. The similarity between x and y can either be graded by humans or computed automatically.

3.1.1 Human Cross-Specificity Measurement

For this, N different human subjects are asked to rate the similarity between a pair of samples x and $y \in Y_c$, on a likert scale of 1 (very dissimilar) to 10 (very similar). In our study, subjects were not informed that the two samples belonged to the same semantic category, which ensured that they rated the similarity solely based on their perceived content. These scores are then normalized to lie in $[0, 1]$.

Let $sim^n_{\text{hum}}(x, y)$ denote similarity between x and $y \in Y_c$ as perceived by the n-th subject, and $spec_{\text{hum}}(x)$ be the cross-specificity score for x. Then we get

$$spec_{\text{hum}}(x) = \frac{1}{N|Y_c|} \sum_{y \in Y_c} \sum_{n=1}^{N} sim^n_{hum}(x, y) \tag{1}$$

3.1.2 Automated Cross-Specificity Measurement

To measure cross-specificity automatically, we would require an automatic criterion to compute similarity between a pair of samples from distinct modalities. We pose this as a cross-modal matching task [8,17,6,16, 23]. Given sets of samples from two different modalities, cross-modal matching approaches model a function $\mathcal{F}$ that can measure the degree of similarity between a pair of samples from diverse modalities.

Let us assume we are given a function $\mathcal{F}$ defined and learned using some cross-modal matching technique such as [8,17,6,16,23]. Based on this, let $sim_{\text{auto}}(x, y)$ denote the similarity score between x and $y \in Y_c$ computed using $\mathcal{F}$, normalized to lie between 0 and 1. The automated cross-specificity score $spec_{\text{auto}}(x)$ for x is then obtained by averaging these similarity scores across all (x, y) pairs. That is,

$$spec_{\text{auto}}(x) = \frac{1}{|Y_c|} \sum_{y \in Y_c} sim_{auto}(x, y) \tag{2}$$

Analogous to Eq. 1 and 2, we can compute $spec_{\text{human}}(y)$ and $spec_{\text{auto}}(y)$ for some $y \in S_Y$. Note that both $spec_{\text{hum}}(\cdot)$ as well as $spec_{\text{auto}}(\cdot)$ lie in $[0, 1]$.

Figure 1 shows example images from the PASCAL-50S dataset [10] along with their human-annotated and automatically computed cross-specificity scores. We can observe that though each of these images contains the corresponding ground-truth category object, the cross-specificity scores using both the measures reduce as the perceptibility of the ground-truth category becomes ambiguous. E.g., in the third and fourth images, their categories "bicycle" (present in between the two persons) and "potted-plant" (placed on a platform in the background) respectively are perceptually small and ambiguous[1].

3.2 Comparison with Image-Specificity

As discussed in Sec. 2, the concept of image-specificity [10] allows us to measure the degree of specificity in the content of an *image*, that relies on linguistic pair-wise similarity *among* multiple captions of that image (i.e., text-to-text matching). The aim of cross-specificity is also to estimate the degree of specificity in the content of a *sample* (that can be from any modality). However, here this is measured based on the semantic category of that sample, and is defined in terms of *its* similarity with samples from *another* modality (i.e., cross-modal matching) that belong to the same category. More importantly, unlike in image-specificity, cross-specificity does not require explicit coupling (or one-to-one correspondence) among cross-modal samples. Below we list some of the critical distinctions between these two concepts[2]:

- To compute image-specificity, we require each image to be associated with multiple (at least two) captions. However, this can be too much to expect from the real-world data. In contrary, cross-specificity requires just a single category label associated with a given sample. This requirement is independent of the particular modality under consideration, and is much more relaxed and practically feasible than the former. E.g., photos on websites such as Flickr, Picasa and Instagram that are tagged with a category are much more than those with multiple captions.
- The way image-specificity is defined makes it restricted to a single data point (that consists of an image and all its captions), and thus disconnected from the rest of the data points. On the other hand, cross-specificity of a sample relies on several other cross-modal samples that belong to the same cateogry. Since a category is not bound to a single

[1] During our analysis, we found that the similarity scores assigned by humans were quite sensitive to subjective aspects such as perceptibility and presence of multiple objects, which led to relatively less scores for human-computed cross-specificity compared to automated cross-specificity.

[2] The reader is suggested to refer Sec. 3.1 of [10] to get further details on the notion of image-specificity, and to better appreciate its distinctions with the proposed cross-specificity.

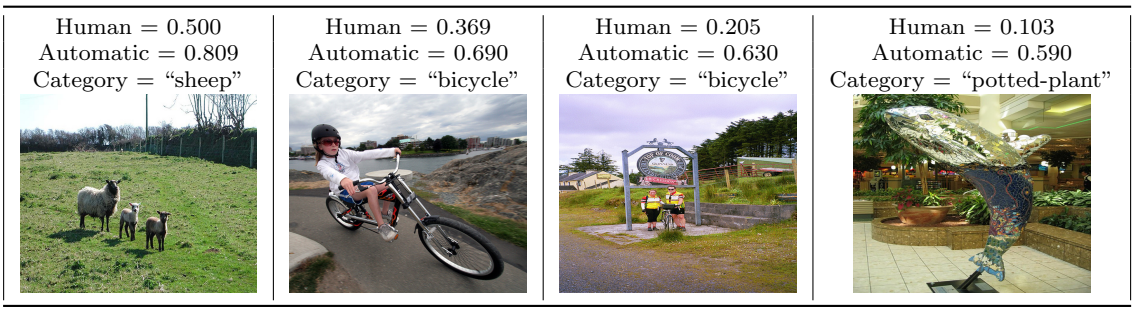

Fig. 1 Example images with high to low human-annotated and automatically computed cross-specificity scores, along with their ground-truth category. Note that from left to right, the category becomes more and more ambiguous, and hence cross-specificity scores (both human and automatic) also reduce.

sample, cross-specificity is expected to encode data abstraction much better than image-specificity.

– As shown in [10], the application of image-specificity is limited to the task of textual (caption) query based image retrieval. Whereas, the notion of cross-specificity is generic, i.e., it does not make any assumption on particular modality, and can benefit cross-modal retrieval between any two modalities, as discussed next.

4 Application of Cross-Specificity: Cross-Modal Retrieval

Let $\mathcal{T}_X^r \subset S_X$ and $\mathcal{T}_Y^r \subset S_Y$ denote the collections of training samples, and $\mathcal{T}_X^e \subset S_X$ and $\mathcal{T}_Y^e \subset S_Y$ be the collections samples in the retrieval set, such that $\mathcal{T}_X^r \cap \mathcal{T}_X^e = \mathcal{T}_Y^r \cap \mathcal{T}_Y^e = \varnothing$. The training samples are used to learn a cross-modal matching function using some cross-modal learning technique such as [8,17,6, 16,23]. During evaluation, the samples in the retrieval sets are matched using the learned function, considering one modality as a query set. In cross-modal retrieval, given a query q from the retrieval set of one modality (say $\mathcal{T}_X^e$), the goal is to rank the samples from another modality ($\mathcal{T}_Y^e$) based on their relevance with the query, such that the samples with more relevance should be ranked higher and vice-versa. Without loss of generality, from now onwards we assume the query $q \in \mathcal{T}_X^e$ (the query set), and retrieval is performed over samples in $\mathcal{T}_Y^e$ during the testing phase.

4.1 Baseline Approach

Here, we learn a cross-modal matching function using a baseline cross-modal learning approach [8,17,6,16, 23] (will be discussed in detail in Sec. 5.4.1). Using the learned function, we compute similarity $sim_{\text{auto}}(q, y)$ between q and $y \in \mathcal{T}_Y^e$ that denotes the baseline relevance rel^y_{baseline} between q and y:

$$rel^y_{\text{baseline}} = sim_{\text{auto}}(q, y) \tag{3}$$

Finally, all the samples in $\mathcal{T}_Y^e$ are ranked (sorted) in descending order of this relevance score.

4.2 Proposed Approach

In the proposed approach, we additionally take into consideration the cross-specificity of a sample rather than ranking based on just the similarity with the query. The rationale is if the underlying (unknown) semantic category of the query is the same as that of the (known) semantic category of a sample in the retrieval set, then that sample is semantically relevant to the query and hence should be ranked higher, and vice-versa. To do so, rather than sorting just based on $sim_{\text{auto}}(q, y)$ as done in the baseline approach, we model $P(\text{match}|sim_{\text{auto}}(q, y))$ which captures the probability that the query matches a sample. To this end, we model this probability using Logistic Regression (LR):

$$\begin{aligned} rel^y_{\text{cross-spec}} &= P(\text{match}|sim_{\text{auto}}(q, y)) \\ &= (1 + exp(-\beta_0^y - \beta_1^y sim_{\text{auto}}(q, y)))^{-1} \end{aligned} \tag{4}$$

This model is trained for each sample in the retrieval set $\mathcal{T}_Y^e$. Assume a sample $y \in \mathcal{T}_Y^e$ that belongs to class c, and let $X_c^r \subset \mathcal{T}_X^r$ be the subset of samples in the training set that are from the same class. For y, positive examples are the similarity scores between y and all $x \in X_c^r$, and negative examples are the similarity scores between y and some random $|X_c^r|$ number of samples from $\mathcal{T}_X^r \setminus X_c^r$ (i.e., the training samples that do not belong to class c). Using these positive and negative examples, the LR model y is trained. In real-world applications, since the retrieval is usually performed on a fixed and known set, this makes learning LRs a one-time process and thus can be done off-line.

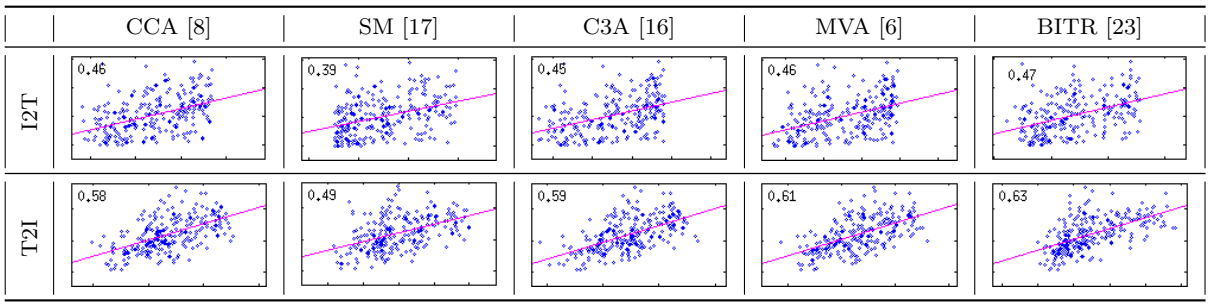

Fig. 2 Correlation between human cross-specificity and automatic cross-specificity computed using different methods for I2T and T2I on the PASCAL-50S dataset. On top of each distribution we show the Spearman's rank correlation score.

The parameters of the LR model (β_0^y and β_1^y) inherently capture the cross-specificity of each sample y. This is because for a sample, LR tries to boost its relevance with samples from the same class and suppress it for others. Once we train a separate LR model for each $y \in \mathcal{T}_Y^e$, it is used to compute $P(\text{match}|sim_{\text{auto}}(q, y))$ for that sample. Finally, all the samples are sorted based on the probability outputs of the LR models.

5 Experiments

Now we analyze different aspects of cross-specificity, and evaluate the proposed cross-specificity based cross-modal retrieval approach.

5.1 Datasets and Features

We experiment on two publicly available datasets, viz. Wikipedia and PASCAL-50S. The Wikipedia dataset was compiled by Rasiwasia et al. [17] from the Wikipedia articles, and is widely used as a de facto benchmark in cross-modal retrieval. It consists of 2173 train and 693 test pairs of images and text articles from 10 different classes. The PASCAL-50S dataset was introduced by Jas and Parikh [10], and is an extended version of the UIUC PASCAL Sentence dataset [15]. It contains 1000 images from 20 different classes, each of which is captioned with 50 unique captions. In our experiments, we consider only the last caption (the "query" caption from [10]) for each image in order to maintain balance between the two modalities, and a random split of 750/250 train/test samples. For both the datasets, we use 4096-dimensional image features computed using the penultimate model [19] pre-trained on the ImageNet dataset. For text, we use a 10-topic representation learned using latent Dirichlet allocation model (LDA) [3].

5.2 Performing Cross-modal Matching

To compute automatic cross-specificity (Sec. 3.1.2), we consider five cross-modal matching techniques: Canonical Correlation Analysis (CCA) [8], Semantic Matching (SM) [17], Cluster Canonical Correlation Analysis (C3A) [16], Multiview approach (MVA) [6] and Bilateral Image-Text Retrieval (BITR) [23]. CCA learns a common embedding space between cross-modality samples by maximizing their correlation. Because of its effectiveness, CCA is considered as a de facto benchmark in cross-modal matching tasks. In SM, a sample is represented using a $\mathcal{C}$-dimensional feature vector. Each dimension of this vector denotes its relevance with a particular category, that is computed using a classifier. C3A and MVA additionally make use of semantic information during the training phase, which helps in learning discriminative cross-modal matching functions. In BITR, a Structural SVM based framework is used for performing cross-modal retrieval. We refer the reader to the respective papers for further details.

While computing cross-specificity of an image, we compute its similarity with text samples from the same class. We will refer to it as "image-to-text" matching or "I2T". Similarly, while computing cross-specificity of a text sample, we compute its similarity with images from the same class. We will refer to it as "text-to-image" matching or "T2I".

5.3 Consistency Analysis

As described in Sec. 3.1, cross-specificity of a sample can be measured using two mechanisms. In the first, humans rate the similarity between pairs of cross-modality samples and in the second, an automatic method is used. We collect human measurements on the PASCAL-50S dataset. For I2T, we collect similarity ratings of each image in the retrieval set with all the captions in the training set that belong to the same class. Simi-

larly, for T2I, we collect similarity ratings between each caption in the retrieval set with all the images in the training set from the same class. This gives 9170 pairs for each I2T and T2I. For each pair, we collected judgments from 2 human subjects.

Figure 2 shows the Spearman's rank correlation between human-annotated and automatically measured cross-specificity using the five cross-modal matching methods. Overall, the correlation scores lie between 0.39 to 0.47 for I2T and 0.49 to 0.63 for T2I[3]. Note that even though the matching is performed between two diverse sources of information (image and text) and is inherently quite subjective, we achieve statistically significant positive correlations between the two mechanisms. These results confirm that cross-specificity is a well-defined phenomenon.

5.4 Cross-modal Retrieval

5.4.1 Baselines

Since a cross-modal matching method is evaluated on cross-modal retrieval task, we consider the five methods discussed in Sec. 5.2 as our baselines for cross-modal retrieval. Following these approaches, we consider two cross-modal retrieval tasks: retrieving text samples given a query image, and retrieving images given a query text. For convenience, we do a slight abuse of notation, and will refer to these tasks as "I2T" and "T2I" respectively. To denote the proposed approach that integrates cross-specificity with baseline cross-modal techniques (Sec. 4.2), we use "CS" as a suffix. To measure cross-modal retrieval performance, we use the standard mean average precision (mAP) [17].

5.4.2 Results and Discussion

Table 1 shows the results for cross-modal retrieval obtained using the baseline approach (Sec. 4.1) and those using the proposed approach based on cross-specificity (Sec. 4.2). From these results, we observe that the proposed approach consistently performs significantly better than the baselines techniques. It achieves up to $10-12\%$ of absolute improvements in some cases, and does better than the baseline for around $50-85\%$ of the queries. By comparing the results corresponding to "%>BL" (the percentage of queries where cross-specicity based retrieval does better than the corresponding baseline methods), we observe that they are comparable for both Wikipedia and Pascal datasets for I2T. However for T2I, we can see that the improvements are much more in the Wikipedia dataset compared to the Pascal dataset. Recall that the textual descriptions associated with images in the Wikipedia dataset are quite long and ambiguous with a lot of irrelevant information, whereas those in the Pascal dataset are short and precise. These results suggest that the promise of the proposed approach gets more pronounced as the degree of ambiguity in the query modality increases. This can be particularly useful in the case of on-line search where users are sometimes not able to precisely describe what they are looking for, thus resulting into ambiguous queries. These results confirm the utility of modelling cross-specificity in cross-modal retrieval tasks.

	Method →	CCA	SM	C3A	MVA	BITR
Wiki (I2T)	Baseline	41.72	41.47	41.71	31.49	43.03
	Cross-Spec	46.57	44.81	46.12	35.92	44.55
	%>BL	61.47	66.09	63.20	58.87	53.82
Wiki (T2I)	Baseline	40.04	28.34	39.85	28.34	40.53
	Cross-Spec	50.29	41.93	50.67	41.93	45.52
	%>BL	84.56	80.66	85.43	84.42	70.71
Pas (I2T)	Baseline	25.40	22.14	26.52	24.02	26.17
	Cross-Spec	35.40	32.36	37.98	34.60	31.60
	%>BL	62.00	62.40	70.80	65.60	53.60
Pas (T2I)	Baseline	29.53	26.12	32.59	29.03	30.49
	Cross-Spec	35.53	34.91	37.41	34.36	32.71
	%>BL	59.20	64.80	59.20	59.20	60.80

Table 1 Cross-modal retrieval results (%mAP) using different methods (CCA [8], SM [17], C3A [16], MVA [6] and BITR [23]). The first row ("Baseline") denotes the performance obtained using the baseline methods. The second row ("Cross-Spec") denotes the performance obtained by integrating cross-specificity with these techniques (i.e., the proposed cross-specificity based approach as discussed in Sec. 4.2). The last row ("% >BL") indicates the percentage of queries where cross-specificity based retrieval does better than the corresponding baseline methods.

In Figure 3, we analyze what percentage of queries benefit the most from the proposed approach. The horizontal axis denotes the percentage of queries, and the vertical axis denotes the margin range by which the baseline is beaten: the first bin denotes the percentage of queries where baseline is beaten by ≤ 0.05 mAP, the second bin denotes the percentage of queries where baseline is beaten by > 0.05 and ≤ 0.10 mAP, and so on till the last bin that denotes the percentage of queries where baseline is beaten by > 0.35 mAP. From these results, we observe that using the proposed approach, around $12-30\%$ of queries achieve up to 5%

[3] All the correlations were found to be statistically significant at $p < 0.0001$.

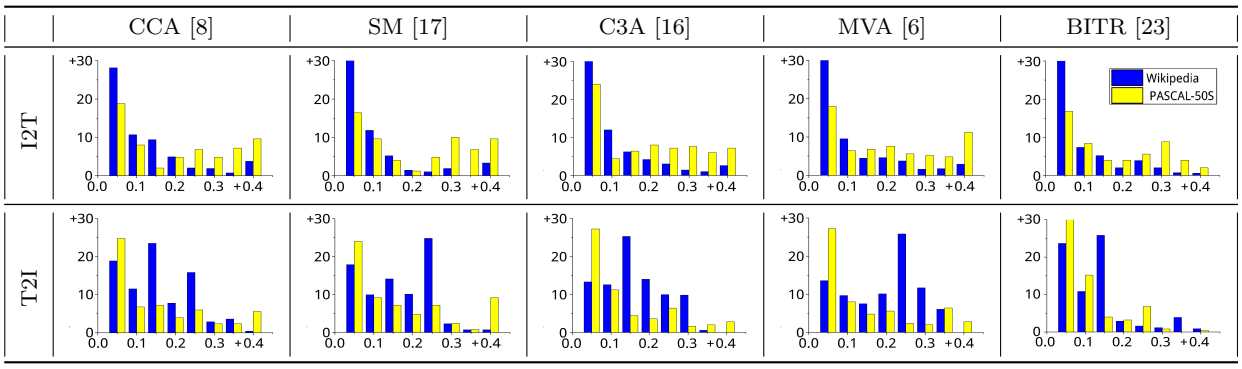

Fig. 3 Cross-modal retrieval results for I2T and T2I using different methods. The horizontal-axis denotes the margin (in terms of mAP) by which baseline is beaten, and the vertical-axis denotes the percentage of queries where baseline is beaten.

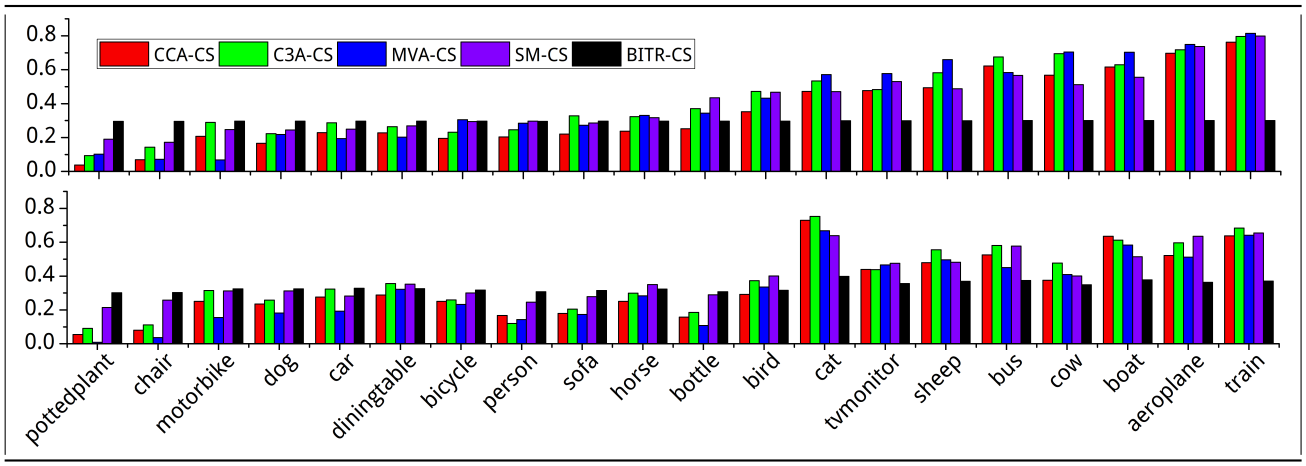

Fig. 4 Average cross-specificity per category for I2T (top) and T2I (bottom) for the PASCAL-50S dataset.

of improvement, and around $2-10\%$ of queries achieve more than 35% of improvement in mAP. Also, for T2I, the improvements are more spread-out over the different margin ranges compared to I2T for the Wikipedia dataset. Recall that in the Wikipedia dataset, each text sample is a long article and contains a lot of ambiguous information. By modelling this ambiguity using cross-specificity, we are able to retrieve semantically more relevant images compared to the baselines. This is also validated from the results in Table 1 where cross-specificity achieves improvements for $70-85\%$ of queries.

In Figure 4, we show the average cross-specificity values for individual categories in the PASCAL-50S dataset using automatic measurements. Here we observe that the scores for categories such as "pottedplant", "chair" and "bottle" are generally low, whereas those for categories such as "cow", "train", "boat", "aeroplane", "bus" and "cat" are generally high. This is justifiable because in this dataset, the former objects are usually not the central component in their images/captions, and occupy only a small portion. This leads to reduced specificity, and hence low cross-specificity scores. However, the latter ones are usually quite prominent and unambiguous, thus achieving higher cross-specificity scores.

Figure 5 shows some qualitative examples from the PASCAL-50S dataset where cross-specificity helps (left) and does not help (right) in improving the rank of ground-truth for a given query. In practice, we observe that the retrieval rank improves in most of the cases, sometimes by a big margin (as is also evident from the "%>BL" results in Table 1). In cases where it degrades, it is usually by a small margin. These results revalidate the practical advantages of cross-specificity in improving cross-modal retrieval performance.

5.5 Empirical Comparison with Image-Specificity

While there are fundamental distinctions between the definition and applicability of image-specificity [10] and the proposed notion of cross-specificity (Sec. 3.2), both aim at measuring degree of specificity in the content of

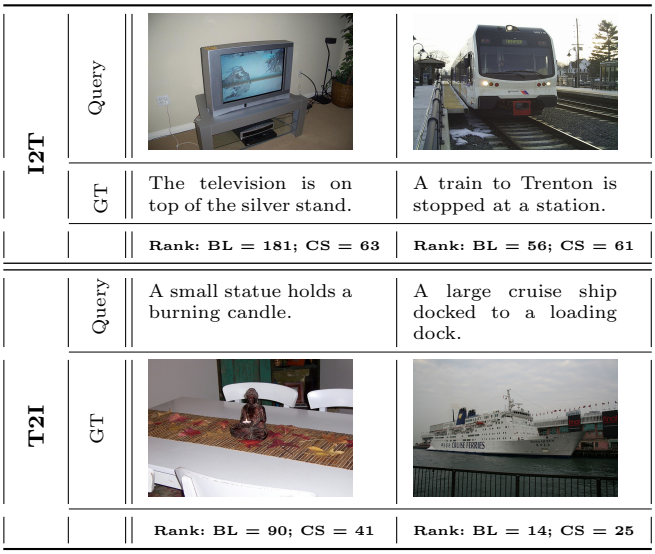

Fig. 5 Some examples where the proposed cross-specificity (CS) based cross-modal retrieval approach improves (left) and does not improve (right) the retrieval rank of the ground-truth (GT) in comparison to the baseline (BL) approach. (Here, CCA is considered as the baseline approach.)

sample in a broad sense. Here we present a brief comparison between these two on the PASCAL-50S dataset, which was also used in [10].

For this, we quantitatively compare them in terms of text-to-image retrieval performance. On this task, image-specificity based retrieval scheme (*c.f.* Sec. 3.2 of [10]) achieves 32.50% mAP, while the proposed cross-specificity based approach (Sec. 4.2) achieves better mAP for all the five baselines (*c.f.* Table 1), and the best performance of 37.41% using C3A [16] as the baseline. Recall that image-specificity computation makes use of all the 50 captions corresponding to an image, while cross-specificity uses just one caption per image. Also, image-specificity relies on text-to-text matching while cross-specificity relies on cross-modal matching. Hence, it may not be fair to do a direct comparison between the two. Nevertheless, these results validate the practical advantages of cross-specificity over image-specificity.

6 Conclusions

We have introduced the notion of cross-specificity, and showed that it is a well-defined phenomenon. We studied various aspects of cross-specificity, and demonstrated its applicability on cross-modal retrieval task. Experiments showed that the proposed approach can provide significant boost in the performance of several existing cross-modal retrieval techniques.

Acknowledgement YV would like to thank the Department of Science and Technology (India) for the INSPIRE Faculty Award.

References

1. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: ICML (2013)
2. Berg, A.C., Berg, T.L., Daumé, H., Dodge, J., Goyal, A., Han, X., Mensch, A., Mitchell, M., Sood, A., Stratos, K., Yamaguchi, K.: Understanding and predicting importance in images. In: CVPR (2012)
3. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. JMLR **12**(1), 234–278 (2003)
4. Chen, X., Hero, A., Savarese, S.: Multimodal video indexing and retrieval using directed information. IEEE Transactions on Multimedia **14**(1), 3–16 (2012)
5. Duan, K., Crandall, D., Batra, D.: Multimodal learning in loosely-organized web images. In: CVPR (2014)
6. Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. IJCV **106**(2), 210–233 (2013)
7. Guillaumin, M., Verbeek, J., Schmid, C.: Multimodal semi-supervised learning for image classification. In: CVPR (2010)
8. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE TPAMI **20**(11), 1254–1259 (1998)
10. Jas, M., Parikh, D.: Image specificity. CVPR (2015)
11. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV (2009)
12. Kang, C., Xiang, S., Liao, S., Xu, C., Pan, C.: Learning consistent feature representation for cross-modal multimedia retrieval. IEEE Transactions on Multimedia **17**(3), 370–381 (2015)
13. Li, Y., Crandall, D., Huttenlocher, D.: Landmark classication in large-scale image collections. In: ICCV (2009)
14. McAuley, J.J., Leskovec, J.: Image labeling on a network: Using social-network metadata for image classification. In: ECCV (2012)
15. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotation using amazon's mechanical turk. In: NAACLHLT Workshop (2010). URL `http://vision.cs.uiuc.edu/pascal-sentences/`
16. Rasiwasia, N., Mahajan, D., Mahadevan, V., Aggarwal, G.: Cluster canonical correlation analysis. In: AISTATS (2014)
17. Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: ACM MM (2010)
18. Sharma, A., Kumar, A., III, H.D., Jacobs, D.W.: Generalized multiview analysis: A discriminative latent space. In: CVPR (2012)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
20. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng., A.Y.: Grounded compositional semantics for finding and describing images with sentences. TACL (2013)
21. Spain, M., Perona, P.: Measuring and predicting object importance. IJCV **91**(1), 59–76 (2011)

22. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. In: NIPS (2012)
23. Verma, Y., Jawahar, C.V.: A support vector approach for cross-modal search of images and texts. Computer Vision and Image Understanding **154**, 48–63 (2017)