Aligning Salient Objects to Queries: A Multi-modal and Multi-object Image Retrieval Framework

Sounak Dey[1\[0000−0001−6704−6810\]], Anjan Dutta[1\[0000−0002−1667−2245\]], Suman K. Ghosh[1\[0000−0002−5462−8499\]], Ernest Valveny[1\[0000−0002−0368−9697\]], and Josep Lladós[1\[0000−0002−4533−4739\]]

Computer Vision Center, Autonomous University of Barcelona, Barcelona, Spain
{sdey,adutta,sghosh,ernest,josep}@cvc.uab.es

Abstract. In this paper we propose an approach for multi-modal image retrieval in multi-labelled images. A multi-modal deep network architecture is formulated to jointly model sketches and text as input query modalities into a common embedding space, which is then further aligned with the image feature space. Our architecture also relies on a salient object detection through a supervised LSTM-based visual attention model learned from convolutional features. Both the alignment between the queries and the image and the supervision of the attention on the images are obtained by generalizing the Hungarian Algorithm using different loss functions. This permits encoding the object-based features and its alignment with the query irrespective of the availability of the co-occurrence of different objects in the training set. We validate the performance of our approach on standard single/multi-object datasets, showing state-of-the art performance in every dataset.

1 Introduction

Content Based Image Retrieval (CBIR) has been for decades one of the prevalent topics in Computer Vision. Rapidly the field has evolved towards a more human-centered view. Thus, cross-media image retrieval systems emerged, allowing users to express the queries in a more natural way using different input modalities, such as text, speech, or sketches.

Text-based image retrieval (TBIR) is the most widely established modality, due to the simplicity of matching text queries with manually assigned keywords describing the image content. Human annotation of large databases of images is time consuming and may lack completeness. Thus, current trends explore TBIR systems that could bridge the semantic gap with queries based on natural language descriptions of the image content. Despite its wider expressiveness there can still be circumstances where text cannot be adequate to portray the query and it can be difficult to establish the link between text and image contents.

With the advent of touch screen and pen input devices, sketches have emerged as an alternative mode to provide the query that can deal with the limitations
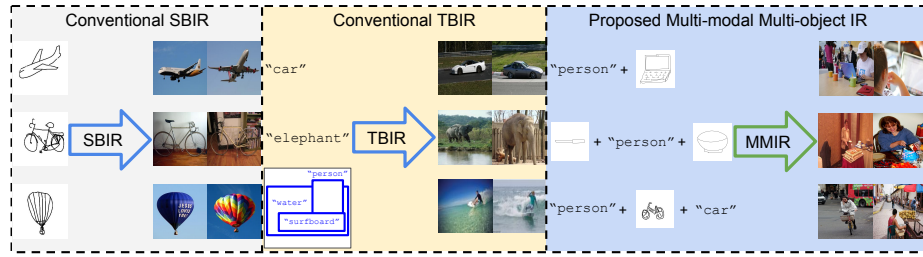
Fig. 1: In conventional SBIR and TBIR, during the training phase, respectively a sketch and a text representation is mapped to the image representation of corresponding class. Querying images with multiple labels has been explored within the TBIR domain [18] (as shown in the last row of Conventional TBIR column). Querying images with multiple objects using multi-modal queries provides convenience in searching, but it is an extremely challenging task and has not been addressed yet.

of text and images. Sketches are a natural way to conceptualize visual objects in terms of simplified shapes and their pose, however they have a few constraints. Sketching needs some idea and ability in drawing shapes that can make some users uncomfortable with this modality. Thus a SBIR (Sketch Based Image Retrieval) framework can not replace conventional text based retrieval which has its own benefits (e.g. utilization of keyboard versus stylus). Thus both modalities can complement each other. Although numerous methodologies for SBIR have been proposed, none of them allow to utilize text as an extra or complementary query modality. Thus, the first motivation for the work presented in this paper is to propose a multi-modal image retrieval approach where the query can be either sketch or text or both. To make the different modalities compatible, a common semantic embedding space is defined.

Another noteworthy constraint of most existing image retrieval pipelines is that they can only manage situations where just a single salient object is significant – see Figure 1. This motivates the second challenge of the present work, i.e. allowing to express queries that can refer to multiple objects. In this way, the proposed model provides more expressiveness to the search language since users can construct queries consisting of different concepts that are aligned to the salient objects of the target images. To the best of our understanding, none of the SBIR techniques can deal with multi-object scenarios. Although some of the strategies [7] based on textual queries can recover relevant images containing multiple objects.

Consequently, we propose a unified multi-modal and multi-object image retrieval (MMIR) framework, which permits to retrieve images containing multiple objects, expressing the query using text, sketches or a combination of both. The framework is based on a deep network architecture in which multi-modality is addressed by projecting word2vec representations of sketches and text into a common semantic space aligned with the image feature space. To deal with

multi-object search we integrate an LSTM-based visual attention model that learns to discover relevant zones of the image. To match the set of attention glimpses with the set of queries we propose a Hungarian loss that finds the best correspondence between both sets. The Hungarian loss is also used to introduce supervision while training the visual attention model by guiding the result of the attention towards object bounding boxes.

The main contributions of this work are: (1) The proposal of a common semantic space among text and sketches, obtained through word2vec representation of both input modalities, and aligned with the image feature space; (2) A visual attention model that automatically detects salient objects from an image, that is trained in a supervised way in order to minimize the assignment cost between attention output and object bounding boxes.

The rest of the paper is organized as follows: in Section 2, we review the relevant state-of-the-art. Section 3 describes in detail our proposed cross-modal/multi-object image retrieval framework. In Section 4, we describe the experimental framework and present the results of the experiments. Finally, Section 5 draws the conclusions and outlines the future directions.

2   Related Work

In this section we review image retrieval using text – text based image retrieval (TBIR) – and sketches – sketch based image retrieval (SBIR). Subsequently, as our method detects object as part of multi-object image retrieval by means of an attention procedure, we also discuss about the state of the art on attention models.

Advances in feature learning have recently provided effective feature representations for different modalities such as text [4, 20], images [6, 22], and hand-drawn sketches [43, 29] which have been shown to greatly improve the retrieval performance. A common approach when dealing with multi-modal data is to learn a joint embedding to map all modalities into a common latent space [25]. However, in multi-modal image retrieval [35], the complementary use of text and sketch to express the queries has not been much explored [2].

TBIR, dating back to the late 1970s, has evolved from just a keyword-based task to a more challenging task based on natural language descriptions (e.g., sentences and paragraphs) [10]. Queries in the form of sentences rather than keywords refer not only to object categorical information but also interactions, such as spatial relationships between objects [12, 39]. In our work we keep text in the simple form of keywords, but we permit to express objects relationships as combination of several text or sketch based queries. Recently, projecting text into the word2vec space has been shown to achieve a high level of accuracy in TBIR [5]. Thus, we also rely on word2vec to represent text queries, but we also extend this idea to obtain a semantic embedding of sketches into the word2vec space.

SBIR is one of the alternative ways of searching and overcoming the limitations imposed by TBIR systems. Apart from images, sketches have been success-

fully used for 3D shape retrieval purpose as well [44, 38]. Since sketch is more close to the semantic representation, it tends to help retrieving target results in the user mind from a semantic perspective [37]. Here, the main challenge is to bridge the domain gap between sketches and natural images. In literature, methods addressing this issue can be grouped into two categories: (1) hand-crafted methods, (2) deep learning-based methods. The hand-crafted features (e.g. SIFT [16]), gradient field HOG (GF-HOG [9]) are extracted from both the sketches and edge maps of natural images and further clustered using 'Bag-of-Words' (BoW), histogram of edge local orientations (S-HELO) [27] or Learned Key Shapes (LKS) [28]). One of the major limitations for such methods is the difficulty to match the edge maps to non-aligned sketches with large variations and ambiguity. To address the domain shift issue, convolutional neural networks (CNNs) methods [11] have recently been used to learn domain-transformable features from sketches and images with an end-to-end framework [29, 43]. In our work, we address these semantic gap by directly projecting sketches to a semantic space using word2vec, that is further aligned with the image space.

The current deep SBIR methods tend to perform well only in a single object scenario with a simple contour shape on a clean background [15, 29, 43]. Recently, there have been attempts to apply deep learning to multi-label image recognition task [41, 36, 34]. Razavian *et al.* [24] applies off-the-shelf features extracted from a deep network pre-trained on ImageNet [26] for multi-label image classification. Wang *et al.* [34] utilize RNNs to learn a joint image-label embedding to characterize the semantic label dependency as well as the image-label relevance. Some works exploit object proposals to only focus on the informative regions, which effectively eliminate the influences of the non-object areas and thus demonstrate significant improvement in multi-label image recognition task [36, 41]. More specifically, Wei *et al.* [36] propose a Hypotheses-CNN-Pooling framework to aggregate the label scores of each specific object hypotheses to achieve the final multi-label predictions. Yang *et al.* [41], formulate the multi-label image recognition problem as a multi-class multi-instance learning problem to incorporate local information and enhance the discriminative ability of the features by encoding the label view information. As an alternative we propose to use an attention model to discover image regions relevant to each of the objects.

Visual attention model has gained a lot of interest recently. In image captioning [40] visual attention assists the generation of descriptive captions. [42] targeted attention on a set of concepts extracted from the image to generate captions. In visual question answering [31, 45] several models have been proposed which attend to image regions or questions when generating an answer. Concurrently, [1] analyzed the consistency between human and deep network attention in visual question answering. Our goal differs in that we are interested in how attention on salient objects can be aligned with the queried object. We use the attention correction proposed in [14] to create a supervised attention for salient object detection. A Hungarian Loss [32] function is proposed to match the salient objects with the queries projected to a semantic embedding space.

3   Multi-modal and Multi-object Image Retrieval

In this section we describe the proposed methodology (see the architecture in Figure 2). The query (text and sketch) is embedded into a common semantic space. For each image in the image database, an LSTM based attention map generator finds several attention maps (one at every time step of LSTM) based on a CNN feature map and the previous attention map. These attention maps are trained in a supervised way and can be thought of as the relative importance of the different areas of the image in order to get the feature representation of the different salient objects in the image. For every attention map, the CNN features are weighted and averaged to get the final feature representation at every step. Thus, we obtain a set of features corresponding to different salient objects in the image. On the other hand, we allow for multiple queries each of them embedded in the semantic space. Consequently we have a set of query features and a set of image features that have to be matched and aligned. To compute a distance between these two sets, we use a Hungarian loss that gives the minimum cost assignment between them. Cosine distance between query features and object features is used to compute the individual cost between each pair query/object. In the following sections we introduce the details of each of the components of this global architecture.
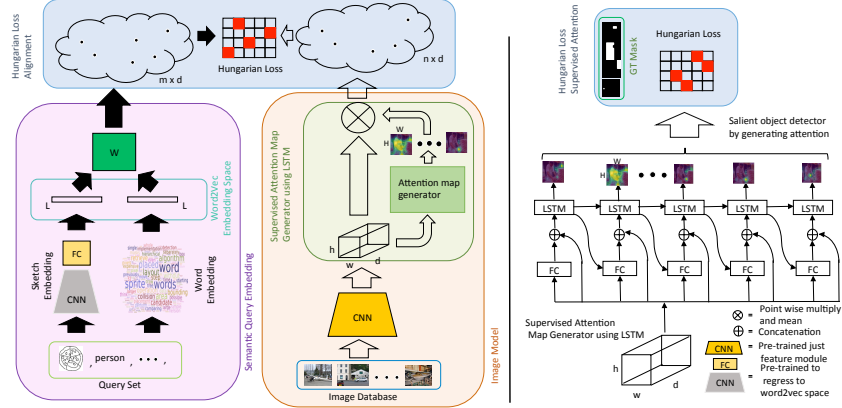

Fig. 2: Architecture of the proposed unified MMIR framework. The semantic query network (in purple) and the image model (in orange) are aligned using the Hungarian loss (in blue). Right side of the figure elaborates supervised attention map generator (in green). The Hungarian loss for attention (in blue) computes the assignments in test time and computes the assignments and loss during training.

3.1   Semantic Query Embedding

For successful retrieval of images given text or sketch as query, a proper embedding space must be defined, where both text and sketch can be directly

compared to the image with a distance measure. In the case of sketches this has been achieved by learning a global feature using triplet loss [29] or similarity loss [23]. For text, either one hot vector encoding based on a fixed vocabulary or some semantic embedding mapping words into a continuous vector space can be used. In our case, we chose to use a semantic embedding as it gives the opportunity to use generic words as query and the model is not restricted to a certain vocabulary of words. For the combination of sketch and text, in [2] two different subspaces were proposed, one between text and image another between sketch and image. However, end to end training of different subspaces is difficult and unstable.

We argue that a better option is to find a subspace where all three modalities can be compared. Therefore, we propose a semantic space to embed the queries, capturing the common properties of text and sketch. In particular, we use word2vec [19] representation to obtain this semantic space. For the text we use directly the word2vec embedding of the words. Sketch images are regressed to the word2vec space by using a CNN based regressor. In the following we provide the details of each embedding.

For word encoding, we have used the standard word2vec [19] representation, which is pre-trained on the set of words from the English Wikipedia[1]. This word representation produces a feature vector of 1000 dimensions.

For the regression of sketch images into semantic space, we adopt a modified version of the VGG-16 network [30]. We modified the top fully connected layer module to accommodate the output vector dimension to that of word2vec. The entire network was then trained as a regression framework with cosine embedding loss to project the sketches in a space parallel to word2vec. For doing so, we used the sketch images from the Sketchy dataset [29] to produce the corresponding word2vec representation of the class name. Once trained, we used the network for obtaining the sketch representation mapped into the word2vec space.

3.2   Supervised Attention for Image Representation

With the goal of retrieving images relevant to a set of query objects, our aim is to extract a set of feature vectors representing salient objects from a particular image. Another possible alternative would be to encode the image into a global signature and aggregate multiple queries into a single representation making retrieval a nearest neighbour problem in this feature space. However we argue that this approach has severe limitations. Firstly, to find a suitable global image representation that can encode the presence of multiple objects we should train with a well curated dataset containing possible combination of different objects. Such a dataset would be huge and training would be complex and take a lot of time. Secondly, although aggregation of queries could be solved in various ways, there is not an easy way to encode their relative position, which can be crucial to match with that of image. In [18] this problem is dealt with by using a spatial query box and then learning the relative position by using a conventional CNN.

---

[1] https://www.wikipedia.org/

However, the position has to be provided by user which can limit the usability of the query interface. Another possibility would be to use a state-of-the-art object detector to detect object and compute the corresponding feature vectors for each object. However, most object detection pipelines assume rectangular objects and do not consider the image context around them. In addition, such an approach would eventually make the pipeline hard to train end-to-end due to different loss functions.

With all this in mind, we propose a different alternative by using a more flexible method based on an LSTM together with an attention model to detect multiple objects (in the form of mask) one at every step. The input to the attention model is a set of features extracted with a conventional CNN corresponding to a spatial grid in which every point represents an area in the image (through its receptive field). In a nutshell, our LSTM based object detector is trained to output one attention map at every step which depends on the previous attention map and the CNN features. The LSTM remembers the attended regions through the hidden vector, which prohibits it to attend the same object multiple times.

The soft attention mechanism was first used in computer vision by [40] where the attention model is not supervised in a sense that there is no loss calculated directly on the attention weights. Thus, the attention mechanism is free to attend anywhere, being only guided by the final output, which is calculated from the result of the attended features. However this model can be extended by applying a direct supervision i.e by directing the model "where to attend" [14]. In this work we used a similar framework but we did the following changes. Firstly, we changed the cross entropy loss between the target and the generated attention map and the softmax over the grid locations by a sigmoid cross entropy loss and a sigmoid activation function to generate a binary map over the grid. This follows from the hypothesis that that every grid location is independent and can be a potential region to attend. Secondly, in our case the order in which the different targets (corresponding to objects) are attended is not relevant. Thus, we need a way to match the unordered set of targets with the unordered set of generated attention maps. We solve this by finding minimum cost between the two sets by using a generalized Hungarian loss. We outline the details of the formulation for this matching later.

More formally, the attention model computes n different attention maps and corresponding image representations for a single image using an LSTM network. The input to the LSTM is a set of features extracted from a pre-trained CNN-based feature extractor resulting in L vectors, $\{\mathbf{a}_1, ..., \mathbf{a}_L\}, \mathbf{a}_i \in \mathbb{R}^d$ that correspond to L spatial locations of an image. The LSTM generates n attention maps, $\{\alpha_t \in [0, 1]^L, t = 1, \ldots, n\}$ and their corresponding image representations: $\{\mathbf{x}_t \in \mathbb{R}^d, t = 1, \ldots, n\}$. Our LSTM model can be visualized in Figure 2. At every step, the input the LSTM takes as intput the hidden representation $\hat{h}$, the local features given by $\{\mathbf{a}_1, ..., \mathbf{a}_L\}$ and the attention map generated at previous time step. To generate the attention map the hidden vector is passed through an MLP layer followed by sigmoid activation, which provides attention maps that can be interpreted as the importance of every spatial location for the de-

tection/retrieval of a certain object. Considering α_{i_l}, for $l = 1, 2, \ldots, L$ to be the weights on the L spatial grids for the ith step, the final image representation for i^{th} attention map is the weighted average of image features over all spatial locations, $\mathbf{x}_i = \sum_{l=1}^{L} \alpha_{i_l} \mathbf{a}_l$

3.3   Hungarian Loss

In our framework we have to match two unordered sets of elements keeping two constraints. On one hand, we have m queries (represented as points in the regressed word2vec space) and n (where $n > m$) different image representations of every single image corresponding to the different attention computed through the LSTM. For retrieval we have to find the best matching between these two sets. On the other hand, in order to train in a supervised way the attention model we have to align the set of bounding boxes of the salient objects with the result of the n steps of the LSTM.

  We have solved both problems using the same framework formulating them as a bi-partite graph matching problem, and using a variation of the Hungarian loss introduced in [32]. In this way, we compute the loss as the minimum cost assignment between every pair of elements in both sets. Given a cost matrix between every pair of elements, the computation of the minimum cost assignment is done by the Hungarian algorithm [21] in polynomial time.

  We use two different cost functions for each of the two problems. The cost between the query features q and the computed image features x is given by the cosine dissimilarity between the query and the feature.

$$C_{\text{sim}}(\mathbf{q}_i, \mathbf{x}_j) = 1 - \cos(\mathbf{q}_i, \mathbf{x}_j)$$

For supervising the attention model, we have used the binary cross entropy as a cost function between each of the ground truth masks β and each of the generated attention maps α.

$$C_{\text{attn}}(\alpha_i, \beta_j) = -(\alpha_i \log(\beta_j) + (1 - \alpha_i) \log(1 - \beta_j))$$

4   Experimental Results

4.1   Implementation Details

We have implemented our method on PyTorch framework. For all experiments, the image features are extracted using the feature module of the pre-trained VGG-16 network model. This feature representation is particularly appropriate for our task as it can effectively capture high-level semantic information from the images and at the same time it naturally retains most spatial information. For sketches, we used the same VGG-16 model but replacing the last layer in the classifier module with a specific layer to accommodate the regression of the features extracted from the sketch to the word2vec space. The output is a feature vector of 1000 dimensions, which is then mapped to a common joint

neural embedding space. Its jointly trained by freezing the feature extraction part on both the query and the image side. The salient object detector based on supervised attention model was implemented using the mask generated from the bounding boxes of the objects in MS-COCO images. In case of the Sketchy database, we used the bounding box of the sketches in the image mask. This was possible because the dataset was designed for fine grained SBIR. We first compare the proposed method with several previous SBIR methods for single object, including hand-crafted features: GF-HOG [8], S-HELO [27], LSK [28]; and deep learning based: Siamese CNN [23], Sketch-a-Net (SaN) [43], GN Triplet [29], 3D shape [33], DSH [15] as shown in the Table 2. For all the methods mentioned above we follow the same protocol and evaluation metrics as in [15]. In the case of [15] we used the trained model for 128-bits provided in the author's github repository.

Fig. 3: Images of (a) person with pizza and (g) person with surfboard; (b)-(f) and (h)-(l) are respective n ($n = 5$) attention maps for image (a) and (b) obtained by our LSTM-based image model.

4.2 Datasets

Sketchy Dataset [29]: This is a large collection of sketch-photo pairs. The dataset consists of images belonging to 125 different classes, each having 100 images. After having these total $125 \times 100 = 12,500$ images, crowed workers were employed for sketching the objects that appear in these $12,500$ images, which resulted in $75,471$ sketches.

TU-Berlin Dataset [3]: The TU-Berlin dataset contains 250 categories with a total of $20,000$ sketches. We also utilize the extended set provided in [15], with natural images corresponding to the sketch classes with a total size of $204,489$.

MS-COCO Dataset [13]: Originally it is a large scale object detection, segmentation, and captioning dataset. We use the MS.COCO dataset for constructing a database of images containing multiple objects. As the label number for each image also varies considerably, rendering MS-COCO is even more challenging. We use the class names of the Sketchy dataset and take all possible combinations by taking two, three, four, five class names. Afterwards, we download the images belonging to these combined classes, and use them for training and retrieval. Few combined classes having less than 10 images are eliminated, leaving 125 number of combined classes for the experiment with at least 900 images.

Fig. 4: Qualitative results obtained by our proposed method: eight example queries consisting texts as well as sketches with their top-10 retrieval results on the Sketchy, TU-Berlin and MS-COCO dataset. Red boxes indicates false positives. (Best viewed in pdf)

4.3   Ablation Analysis

In this subsection, we perform some experiments to carefully analyze the contribution of the critical components of our proposed model. For this study we used single object retrieval from the Sketchy [29] using sketch as a query modality. In an effort to evaluate each of our contributions, we trained every variation of the system with exactly the same training data.

In the first row of Table 1 we show the results that we obtain if we replace the VGG-16 network that regresses the word2vec representation of sketches by another VGG-16 network that just outputs a one-hot encoding representation of the word that represents the class of the sketch. The sketch features obtained from the one-hot encoding produce reasonable retrieval performances, but the figures are still clearly lower than our original design in the last row of the table (12% less in mAP). We speculate this is because hidden representations and knowledge within the trained neural network is able to store more information by correlating the data points that are semantically close.

We also introduce two other variations by replacing the supervised attention module with a global average pooling of image features (second row of Table 1) and a standard attention model without mask level supervision (third row) in order to show the impact of the supervised attention model in detecting salient objects. The comparison reveals the fact that supervised attended regions have a relevant impact on the results and therefore, are capable of discovering the discriminating regions, which facilitates the task of multi-label image classification. The global pooling basically provides no segregated object information and fails
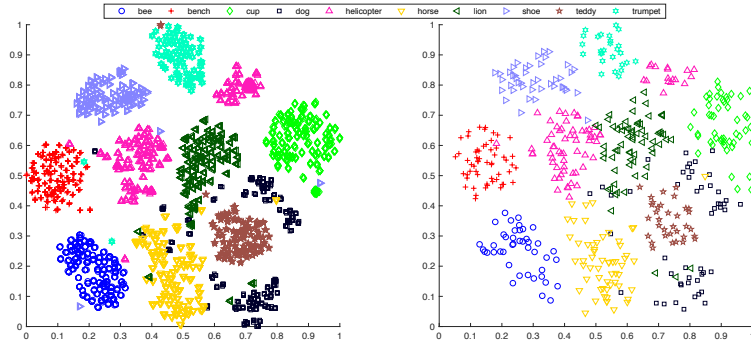
Fig. 5: *t-SNE* visualization of our image and sketch embeddings of 10 representative categories from the Sketchy dataset. After embedding the images as well as the sketches to the common space by our model, natural images and sketch queries from most of the categories are almost scattered into the same clusters. (Best viewed in pdf)

to generalize knowledge from the seen objects together to the unseen ones. The general attention is also suboptimal in exploring the object features individually.

Description	Sketchy
Regressed vector $\rightarrow$ one hot vector	0.688
Supervised attention $\rightarrow$ global pooling	0.726
Supervised attention $\rightarrow$ general attention	0.754
Full model	**0.809**

Table 1: Ablation analysis

Finally, in Figure 6(d), we elucidate t-SNE [17] visualization of the image features and sketch features corresponding to 10 different categories. We can see that the distributions of the features in both domains are very similar reflecting that the network is capable to align features among both modalities.

4.4   Results and Discussions

Single Object: In Table 2, we report the comparison of mAP and precision@200 over all SBIR methods on two datasets. Generally, deep learning-based methods can achieve much better performance than handcrafted methods and the results on Sketchy are higher than those on TU-Berlin since the data in Sketchy is relatively simpler with fewer categories. The corresponding precision-recall (P-R) curves for both the datasets are illustrated in Figure 6(a) and (b). Our method leads with significant improvements over the best-performing comparison methods on the two datasets, respectively in both mAP and precision@200. We argue

Methods	Sketchy		TU-Berlin	
	mAP	Precision @ 200	mAP	Precision @ 200
GF-HOG [8]	0.135	0.167	0.114	0.158
S-HELO [27]	0.161	0.181	0.121	0.153
LKS [28]	0.193	0.231	0.151	0.172
Siamese CNN [23]	0.587	0.745	0.489	0.623
SaN [43]	0.208	0.292	0.178	0.182
GN Triplet [29]	0.651	0.797	0.597	0.782
3D shape [33]	0.161	0.181	0.123	0.147
DSH [15]	0.620	0.694	0.556	0.743
Proposed (Sketch)	**0.809**	**0.886**	**0.653**	**0.796**
Proposed (Text)	**0.802**	**0.881**	**0.641**	**0.727**
Proposed (Sketch + Text)	**0.803**	**0.882**	**0.645**	**0.735**

Table 2: Image retrieval performance

that this is because our architecture is specifically designed to handle visual alignment of the query to the images using the Hungarian Loss.
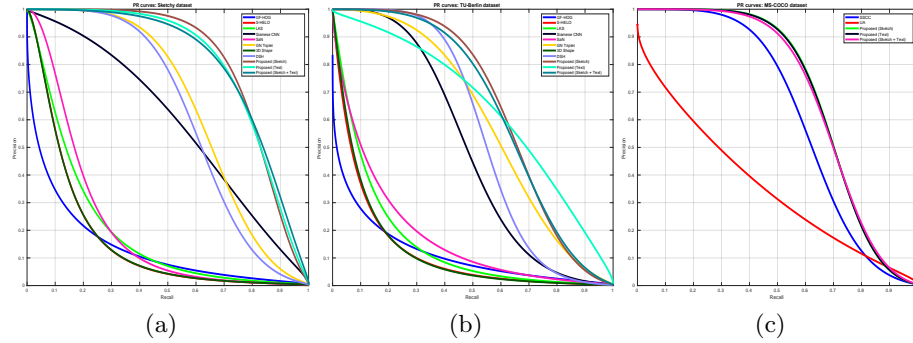

(a)   (b)   (c)

Fig. 6: Precision-recall curves obtained by different methods on (a) Sketchy, (b) TU-Berlin, (c) MS-COCO datasets. (Best viewed in pdf)

Multiple Objects: For retrieving images with multiple objects, we have considered the MS-COCO dataset as mentioned above. Two existing methods are considered for comparison: SSCC [18] and UA [2]. However, it is worth mentioning that none of the above two methods works with multi-modal queries; as they allow retrieving images with multiple queries, we slightly modified these methods to accept multi-modal queries. The multi-modal multi-object image retrieval mean average precision (mAP@all) of our proposed method and the two baselines

are reported in Table 3 and the corresponding precision-recall (P-R) curves are shown in Figure 6(c). The performance margins between our proposed method and the selected state-of-the-art methods are significant, suggesting the existing cross-modal image retrieval methods fail to handle the multi-modal multi-object image retrieval task. SSCC [18] attains relatively better results. A possible reason for this is allowance of multiple queries and a relatively simple model for query processing. However, this method is designed only for text modality and also deals with semantic constraints, which can be a reason of the worse performance than our proposed system. Some qualitative results of retrieving images using multi-modal and multi-object queries are Figure 4. It can be seen that our proposed method is able to produce acceptable retrieval results. Albeit some false alarms are produced, they mostly have some visual similarity with the actual retrieval.

Methods	MS-COCO	
	mAP	Precision @ 200
SSCC [18]	0.623	0.667
UA [2]	0.354	0.413
Proposed (Sketch)	0.697	0.753
Proposed (Text)	0.696	0.751
Proposed (Sketch + Text)	**0.693**	**0.749**

Table 3: Multiple objects

5   Conclusions and Future Work

In this paper, we have proposed a common neural network model for sketch as well as text based image retrieval. One of the most important advantages of our framework is that it allows to retrieve images queried by terms of multiple modalities (text and sketch). We have designed an image attention mechanism based on LSTM that allows to put attention on the specific zones of the images depending on the inter related objects which usually co-occur in nature. This has been learned by our model from the images in the training set. We have tested our proposed framework on the challenging Sketchy dataset for single object retrieval and on a collection of images from the COCO dataset for multiple object retrieval. Furthermore, we have compared our experimental results with three state-of-the-art methods. We have found that our method performs satisfactorily better than the considered state-of-the-art methods on all the two datasets with some cases of failure with justifiable reasons. For the future we plan to investigate on more efficient training strategies, as few shot learning approaches that learn from a small amount of training data in human-centered scenarios that allow users to search in their own databases in a more efficient and effortless manner.

Acknowledgement

References

1. Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D.: Human attention in visual question answering: Do humans and deep networks look at the same regions? CVIU **163**, 90–100 (2017)
2. Dey, S., Dutta, A., Ghosh, S.K., Valveny, E., Lladós, J., Pal, U.: Learning cross-modal deep embeddings for multi-object image retrieval using text and sketch. In: ICPR. pp. 916–921 (2018)
3. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: Sketch-based image retrieval: Benchmark and bag-of-features descriptors. IEEE TVCG **17**(11), 1624–1636 (2011)
4. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: A deep visual-semantic embedding model. In: NIPS. pp. 2121–2129 (2013)
5. Gordo, A., Almazán, J., Murray, N., Perronin, F.: Lewis: Latent embeddings for word images and their semantics. In: ICCV. pp. 1242–1250 (2015)
6. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: ECCV. pp. 241–257 (2016)
7. Gordo, A., Larlus, D.: Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In: CVPR. pp. 5272–5281 (2017)
8. Hu, R., Barnard, M., Collomosse, J.: Gradient field descriptor for sketch based retrieval and localization. In: ICIP. pp. 1025–1028 (2010)
9. Hu, R., Collomosse, J.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval. CVIU **117**(7), 790–806 (2013)
10. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV pp. 32–73 (2017)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1097–1105 (2012)
12. Lan, T., Yang, W., Wang, Y., Mori, G.: Image retrieval with structured object queries using latent ranking svm. In: ECCV. pp. 129–142 (2012)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
14. Liu, C., Mao, J., Sha, F., Yuille, A.L.: Attention correctness in neural image captioning. In: AAAI. pp. 4176–4182 (2017)
15. Liu, L., Shen, F., Shen, Y., Liu, X., Shao, L.: Deep sketch hashing: Fast free-hand sketch-based image retrieval. In: CVPR. pp. 2862–2871 (2017)
16. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV. pp. 1150–1157 (1999)

17. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. JMLR pp. 2579–2605 (2008)
18. Mai, L., Jin, H., Lin, Z., Fang, C., Brandt, J., Liu, F.: Spatial-semantic image search by visual feature synthesis. In: CVPR. pp. 1121–1130 (2017)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (2013)
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS. pp. 3111–3119 (2013)
21. Munkres, J.: Algorithms for the assignment and transportation problems. JSIAM **5**(1), 32–38 (1957)
22. Paulin, M., Mairal, J., Douze, M., Harchaoui, Z., Perronnin, F., Schmid, C.: Convolutional patch representations for image retrieval: an unsupervised approach. IJCV **121**, 149–168 (2017)
23. Qi, Y., Song, Y.Z., Zhang, H., Liu, J.: Sketch-based image retrieval via siamese convolutional neural network. In: ICIP. pp. 2460–2464 (2016)
24. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: CVPRW. pp. 512–519 (2014)
25. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: CVPR. pp. 49–58 (2016)
26. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV pp. 211–252 (2015)
27. Saavedra, J.M.: Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In: ICIP. pp. 2998–3002 (2014)
28. Saavedra, J.M., Barrios, J.M., Orand, S.: Sketch based image retrieval using learned keyshapes (lks). In: BMVC. vol. 1, pp. 1–10 (2015)
29. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: Learning to retrieve badly drawn bunnies. ACM SIGGRAPH (2016)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv **abs/1409.1556** (2014)
31. Singh, S., Hoiem, D., Forsyth, D.: Learning to localize little landmarks. In: CVPR. pp. 260–269 (2016)
32. Stewart, R., Andriluka, M., Ng, A.Y.: End-to-end people detection in crowded scenes. In: CVPR. pp. 2325–2333 (2016)
33. Wang, F., Kang, L., Li, Y.: Sketch-based 3d shape retrieval using convolutional neural networks. In: CVPR. pp. 1875–1883 (2015)
34. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification. In: CVPR. pp. 2285–2294 (2016)
35. Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L.: A comprehensive survey on cross-modal retrieval. arXiv **1607.06215** (2016)
36. Wei, Y., Xia, W., Lin, M., Huang, J., Ni, B., Dong, J., Zhao, Y., Yan, S.: Hcp: A flexible cnn framework for multi-label image classification. PAMI pp. 1901–1907 (2016)
37. Xiao, C., Wang, C., Zhang, L., Zhang, L.: Sketch-based image retrieval via shape words. In: ACM ICMR. pp. 571–574 (2015)
38. Xie, J., Dai, G., Zhu, F., Fang, Y.: Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval. In: CVPR. pp. 3615–3623 (2017)
39. Xu, H., Wang, J., Hua, X.S., Li, S.: Interactive image search by 2d semantic map. In: ACM ICWWW. pp. 1321–1324 (2010)

40. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)
41. Yang, H., Tianyi Zhou, J., Zhang, Y., Gao, B.B., Wu, J., Cai, J.: Exploit bounding box annotations for multi-label object recognition. In: CVPR. pp. 280–288 (2016)
42. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: CVPR. pp. 4651–4659 (2016)
43. Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe. In: CVPR. pp. 799–807 (2016)
44. Zhu, F., Xie, J., Fang, Y.: Learning cross-domain neural networks for sketch-based 3d shape retrieval. In: AAAI. pp. 3683–3689 (2016)
45. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: CVPR. pp. 4995–5004 (2016)