2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Learning Cross-Modal Deep Embeddings for Multi-Object Image Retrieval using Text and Sketch

Sounak Dey, Anjan Dutta, Suman K. Ghosh, Ernest Valveny, Josep Lladós
Computer Vision Center, Computer Science Department
Autonomous University of Barcelona
Barcelona, Spain
Email: {sdey, adutta, sghosh, ernest, josep}@cvc.uab.es

Umapada Pal
CVPR Unit
Indian Statistical Institute
Kolkata, India
Email: umapada@isical.ac.in

***Abstract*—In this work we introduce a cross modal image retrieval system that allows both text and sketch as input modalities for the query. A cross-modal deep network architecture is formulated to jointly model the sketch and text input modalities as well as the the image output modality, learning a common embedding between text and images and between sketches and images. In addition, an attention model is used to selectively focus the attention on the different objects of the image, allowing for retrieval with multiple objects in the query. Experiments show that the proposed method performs the best in both single and multiple object image retrieval in standard datasets.**

I. Introduction

Image retrieval systems aim to retrieve images from large databases that are relevant to a given query. Most existing systems allow users to query by image (Content Based Image Retrieval) or by text (Text Based Image Retrieval). Though these two alternatives provide an effective way to interact with a retrieval system by providing the query as an example image or text, they also pose some constraints in situations where either an example image is not available or text is not sufficient to describe the query. In this type of scenarios sketches arise as an alternative mode to provide the query that can handle the limitations of text and image. As sketches can efficiently and precisely express the shape, pose and fine-grained details of the target images, they are becoming popular as an alternative form of query. Though in many scenarios sketches are more intuitive to express the query they still present some limitations. Drawing a sketch can be tedious for some users not skilled in drawing, and a sketch based interface may need special hardware (e.g. stylus) which might not be always available. Thus a SBIR (Sketch Based Image Retrieval) system can not entirely replace traditional text based retrieval which has its own convenience (e.g. use of keyboard vs stylus), but can effectively supplement and/or complement the traditional text based querying in many cases. Thus, a multimodal image retrieval system can be envisioned, where the query can be either a sketch or a text. The popular use of smartphones with touch screen interfaces where people stores many personal pictures will bring innovative search services based in this multimodal input. Although recently many approaches [1] [2] [3] for sketch based image retrieval have been proposed, none of them permit to use text as an additional or complementary input modality. On the other hand traditional text based retrieval systems only permit to use text as their query modality.

Another significant limitation of most existing image retrieval pipelines is that they can only deal with scenarios where only one salient object is significant (see Fig. 1 to better understand the limitations of current methods). To the best of our knowledge none of the sketch based image retrieval methods can deal multi-object scenarios, however few methods [4] based on textual queries are able to retrieve relevant images containing multiple objects by learning a common subspace between text description and image. Nevertheless, these methods need a detailed description (caption) about the images to be retrieved, sometime which is unfeasible to provide. Additionally these methods rely on co-learning of text and images in a semantic space, and often limited to a closed dictionary. Hence, we propose a unified image retrieval method, which can take both sketch or text as query. The method obtains different representations for text, sketch and

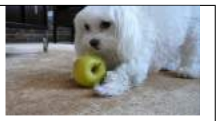
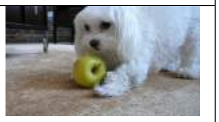
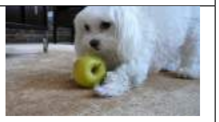
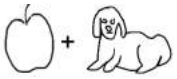
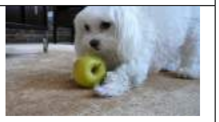
Modalities	Input	Retrievals
Sketch		
Text	"apple"	
Description	"Dog with apple"	
Sketch + Sketch		
Text + Text	"dog"+"apple"	

Fig. 1. Input modality vs retrieval results: examples shown in the last two rows are addressed by our proposed method.

image and then learns a common space where it is more meaningful to compute distances between text and images and between sketches and images. Additionally, the method includes an attention mechanism that permits to focus retrieval on those parts of the image relevant to the query. In this way, our approach can also perform multi-object image retrieval.

A. Related Work

As we are proposing a retrieval system where input queries can be of different modalities, our work is related to multimodal retrieval approaches like [5]. However, none of the existing works in multimodal retrieval actually propose to combine text and sketch, but there are several image retrieval systems for each of these two modalities, which are in a way related to our work. Thus, in this section we will review those works related to multimodal retrieval of images focusing on sketch based image retrieval and text based image retrieval approaches.

Sketch Based Image Retrieval (SBIR) is challenging because free hand sketches drawn by humans have no references but focus only on the salient object structures. In comparison to natural images, sketches are usually distorted. In recent years some studies are made to bridge the domain gap between sketches and natural images, in order to deal with this problem. These methods can be grouped into two categories namely hand-crafted methods and cross domain deep learning-based methods. Most of the hand-crafted SBIR methods first generate an approximate sketch by extracting edge or contour maps from the natural images. Then hand-crafted features (e.g. SIFT [6], HOG [7], gradient field HOG [8], histogram of edge local orientations (HELO) [1] and Learned Key Shapes(LKS) [9]) are extracted from both the sketches and edge maps of natural images. Sometimes these features are further clustered using 'Bag-of-Words'(BoW) methods to generate an embedding for SBIR. 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 [10] have recently been used to learn domain-transformable features from sketches and images with an end-to-end framework [2], [3], [11]. Both category [9], [12] [1] and fine grained SBIR [2], [11], [13] tasks achieve higher performance with deep methods which better handles the domain gap. All the current deep SBIR methods tend to perform well only in a single object scenario with a simple contour shape on a clean background.

On the other hand few works exist which jointly leverage images and natural text for different computer vision tasks. Zero shot learning [14], language generation [15], multimedia retrieval [16], image captioning [17] and Visual Question Answering [18] build a joint space embedding for textual and visual cues to compare both modalities directly in that space. The first category of methods are based on Canonical Correlation Analysis (CCA) for obtaining a joint embedding [19]. There are recent methods that also use deep CCA for such embedding [20]. Alternatively to CCA there are other methods that learn a joint space embedding using ranking loss. A linear transformation of visual and textual features with a single-directional ranking loss is presented in WSABIE [21] and DeViSE [22]. Bidirectional ranking loss [23] with possible constraint is seen in [24]. Cross-modal queries are often done in these joint image and text embedding i.e. retrieve image with textual queries and vice-versa [24]. In many of these works learning the joint embedding is by itself, the final objective. In contrast to these works we try to leverage both the joint space embedding and a sequential attention model to retrieve images having multiple objects.

SBIR suffers the major drawback of varying drawing skills amongst users. Certain visual characteristics can be cumbersome to sketch, yet straightforward to describe in text. We hypothesize that these two input modalities can be modelled jointly with respect to the third for a successful retrieval system as semantically they represent the same. In experiments we will show that using a neural embedding we can jointly learn embedding spaces, where image and the query (sketch and/or text) can be represented as a vector, thus a simple nearest neighbour can be used for retrieval.

From the methodological point of view in order to learn different representation for different objects(in image), we exploit the recent advances in deep learning and, in particular, attention mechanisms [25] to select features from salient regions. Attention models allow to select a subset of features by an elegant weighting mechanism. Soft attention has been widely used in machine translation [25], image captioning [26]. Recently attention has also been used for multi object detection in [27]. In this work we rely on an attention module to do a soft detection of different objects by giving more weights to the salient regions of the image corresponding to the query object. Thus our attention module is responsible for doing an object detection in place.

B. Contribution

Thus in summary we propose the following contributions through this work:

- A unified framework for cross-modal image retrieval that can perform both sketch to image or text to image retrieval based on learning a common embedding between text and images and between sketches and images.
- A retrieval framework for images having multiple salient objects that leverages an attention model to select the subset of image features relevant to each query object.

The rest of the paper is organized as follows: Section II describes our proposed cross-modal and multi-object image retrieval framework with all details on text, sketch and image models. In Section III, we describe the datasets and the experimental protocols we have used to show the effectiveness of the method and present the results of the experiments. Finally, Section IV concludes the paper and some future directions of the present work are mentioned.

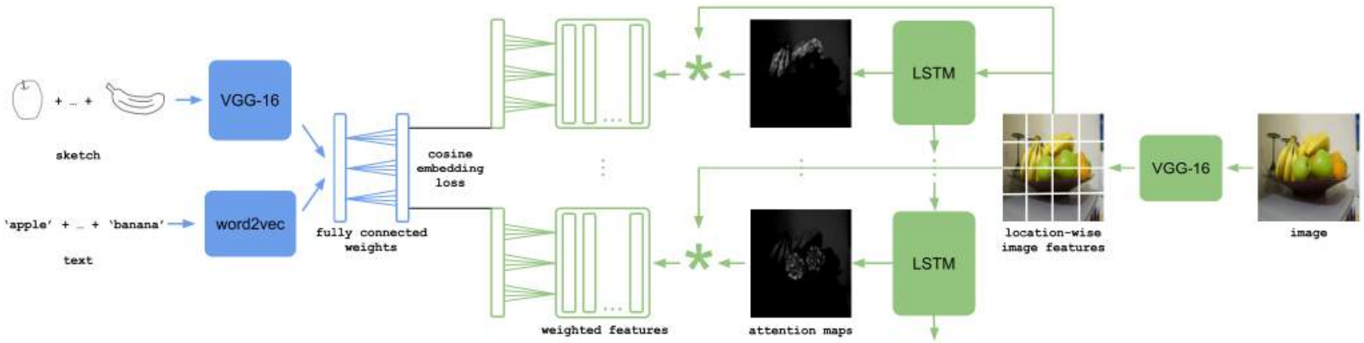


Fig. 2. Overall architecture of our framework

II. Cross-modal and multi-object retrieval framework

In this section, we introduce our proposed framework (see Fig. 2), which is a common neural network structure that allows to query image databases with sketch as well as text inputs. For that, on one side, we have separately designed sketch and text models that respectively permit to obtain a feature representation of sketches and text. On the other hand, we use an image model that processes input images and outputs a set of image features weighted by the attention model. Finally, the network learns a common embedding between text/sketch features and image features. Below we provide the details of each part of the framework.

A. Sketch representation

For the sketch representation, we adopt a modified version of the VGG-16 network [28]. We replace the input layer to accept a single channel sketch image and the last fully connected layer to produce class wise probability for 125 classes. Henceforth, we retrain the entire network as a classification framework on the sketch images from the Sketchy dataset [2] having 125 sketch classes. Once trained we remove the last fully connected (FC) and softmax layer, for obtaining the sketch representation.

B. Text representation

For word representation, we have used the standard word2vec [29] 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.

C. Image representation

The image representation relies on the VGG-16 network [28] pre-trained on ImageNet. However, for obtaining the image representation the top FC layers are removed, and we take the output of the last convolutional layer. In this way, we extract a mid-level M-dimensional visual feature representation on a grid of spatial locations. Specifically, we use $P = \{p_l | p_l \in \mathbb{R}^M; l = 1, \ldots, L\}$ to represent the spatial CNN features at each of the L grid locations. Given a query with n objects (either sketches or text descriptions), we compute n different sequentially dependent attention maps for a single image utilizing an LSTM. Considering, α_{i_l}, for $l = 1, 2, \ldots, L$ to be the weights on the L spatial grids for the ith query, we impose the following condition for obtaining a statistically meaningful feature description for each image and each query object.

$$\alpha_{i_l} \propto \exp(F(p_{i_l})) \text{ s.t. } \sum_{l=1}^{L} \alpha_{i_l} = 1 \tag{1}$$

where F is the neural network model used for obtaining the weights on the spatial grids of the image. Practically, F is designed with an LSTM. Hence the final image representation for ith query results in as:

$$\mathbf{f}_i = \sum_{l=1}^{L} \alpha_{i_l} p_{i_l} \tag{2}$$

Each image feature $\mathbf{f}_i$ is the weighted average of image features over all the spatial locations $l = \{1, 2, \ldots, L\}$. In practice, the convolutional layers produce image features of $7 \times 7 \times 512$ dimensions, which after incorporating the attention weights results in 512 dimension feature vector delineating an image for a particular query object. Our LSTM based attention map generator is a soft attention mechanism, that remembers the attended regions through hidden vector which negates the possibility of attention at the same object multiple times. The LSTM takes the CNN-based features as well as the hidden representation as input to generate the attention maps at each time step.

D. Joint neural embedding

Given the query (either text or sketch) and image representation as above, the goal is to project the respective representations into a common space. For doing so, in each case of sketch and text representation, we employ a non-linear transformation containing two fully connected layers. In case of image representation, it is done by augmenting a non-linear transformation consisting a single MLP layer on the weighted set of image features. We adjust the sizes of these respective non-linear transformations accordingly, to produce a 512 dimensional feature vector as the final representation, for each modality. The goal is to learn mappings of different modalities to make the embedding "close" for a given same query-image pair, and "afar" for different query-image pair.

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

The training of this system is done by including the cosine embedding loss as follows:

$$L_{\mathtt{cos}}((\mathbf{q},\mathbf{f}),y) = \begin{cases} 1-\mathtt{cos}(\mathbf{q},\mathbf{f}) & \text{if } y=1 \\ \mathtt{max}(0,\mathtt{cos}(\mathbf{q},\mathbf{f})-m) & \text{if } y=-1 \end{cases} \tag{3}$$

where $\mathbf{q}$ and $\mathbf{f}$ are respectively the learned query (either text or sketch) and image representation, $\mathtt{cos}$ denotes normalized cosine similarity and m is the margin. For training through this paradigm, we generate positive ($y = 1$), as well as, negative examples ($y = -1$), which respectively correspond to the query-image pairs belonging to the same and different classes.

E. Multiple objects queries

Our framework permits querying by multiple objects represented with the same modality, which is particularly useful for retrieving images containing multiple objects. In this case, the loss is computed in a cumulative manner over all the query objects:

$$\sum_{i=1}^{n} L_{cos}((\mathbf{q}_i,\mathbf{f}_i),y)$$

where n is the number of queries and $\mathbf{q}$ is the representation of any query object. Although, our method supports querying with multiple objects, in practice, we consider querying at most with 2 different objects. This is mainly because of the unavailability of the appropriate datasets needed for training and retrieval. While querying with multiple objects the sum of distances between the queries and the image is considered for ranking the retrievals.

III. Experimental results

A. Datasets

1) Sketchy: The Sketchy dataset [2] is a large collection of sketch-photo pairs. The dataset consists of images belonging to 125 different classes, each having 100 photos. After having these total $125 \times 100 = 12500$ images, crowed workers are employed for sketching the objects that appear in these 12500 images, which resulted in 75471 sketches. The Sketchy database also gives a fine grained correspondence between particular photos and sketches. Furthermore, the dataset readily contains various data augmentations very useful for deep learning based methods.

2) COCO: Originally the COCO dataset [30] is a large scale object detection, segmentation, and captioning dataset. We use the COCO dataset for constructing a database of images containing multiple objects. We use the class names of the Sketchy dataset and take all possible combinations by taking two class names. Afterwards, we download the images belonging to these combined classes, and use them for training and retrieval. Few combined classes having very less number (less than 10) images are eliminated, leaving 365 number of combined classes for the experiment.

B. Experimental protocol

1) Single object: For single object query, we use the Sketchy dataset [2].

a) Sketch-based image retrieval: During the training step, the positive examples are fabricated considering the fine grained information. This follows that for an image I, we consider all the corresponding fine grained sketches drawn by different sketchers. Let n_s be the total number of sketches that correspond to the image I according to the fine grained information. Therefore, we construct n_s different positive training instances using the same image I. The negative training examples for all the n_s sketches are created by randomly choosing images from the classes other than that of I. In this way, we randomly select 80% of the images from each class for training and the rest for testing. However, during the test phase, we obtain the sketch and image representation from the respective models in independent manner, and the retrieval is done according to the distances between them. Here it is to be noted that the fine grained information is not considered for ranking the retrievals.

b) Text-based image retrieval: In case of text-based image retrieval as well, we randomly select 80% of the images from each class for training and the rest for testing. In this case, the positive training instances are created by considering the class label, the image, which creates training examples equal to the number of training images. Equal number of negative instances are created by randomly selecting images belonging to the class different from the texts.

2) Multiple objects: For multiple objects, we consider the database derived by us from the COCO dataset [30].

a) Sketch-based image retrieval: As we mentioned before, in practice, we consider only 2 different objects, which is due to the unavailability of appropriate dataset for the particular task. In this case as well, 80% images of each combined class are selected for training set, and the rest of the images from each class are kept for the testing phase. To comply with diverse sketch instance, style and appearance, an image I is reconsidered n_m times. These n_m instances of the same image is framed with n_m different combinations of sketches that belong to the individual classes creating the combined class. The negative examples are created by associating each combined sketch query an image that does not belong to the same combined class.

b) Text-based image retrieval: Creating training pairs for text-based image retrieval is relatively straight forward. As in the previous case, 80% images from each combined class are selected for training set, and the rest of the images from each class are kept for the testing purpose. For creating the positive training examples each image that belongs to a combined class is associated with the text labels of the individual classes. The negative examples are created by bracketing each combined text query an image that belongs to a different combined class.

For training our proposed neural network, we have used the Adam [31] optimization algorithm with a learning rate of 0.01 and a learning rate decay of 0.6. All the learnable parameters used in the model are initialized from a standard normal distribution $\mathcal{N}(0, 1)$. The results shown in Table I are produced after training the respective text and sketch models for 50 epochs.

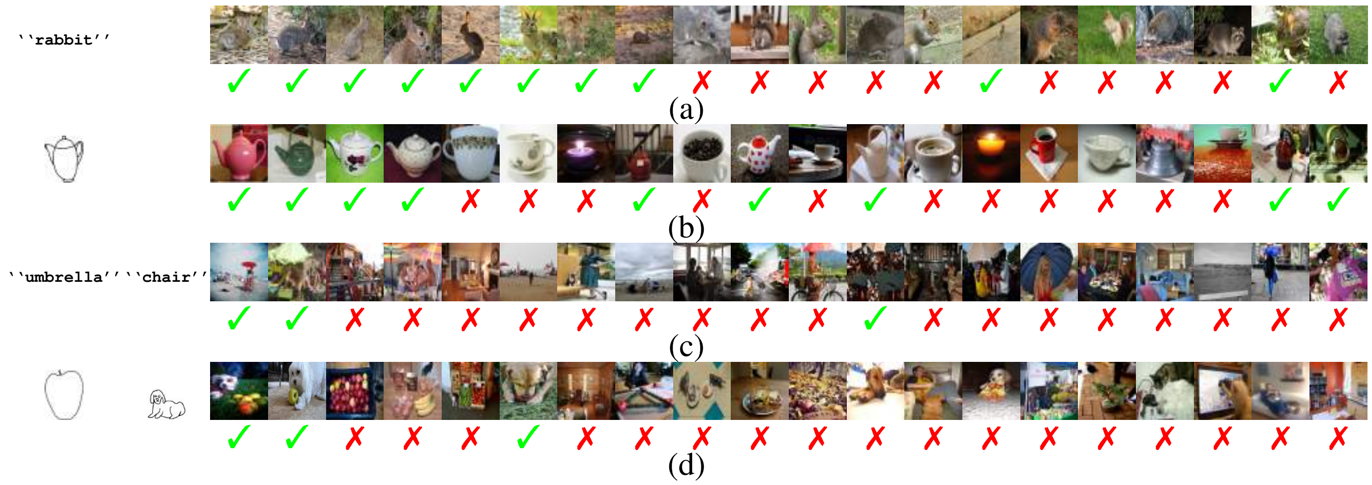


Fig. 3. Qualitative results obtained by our proposed method while querying the database with a combination of texts and sketches: (a) text query ``rabbit'', (b) sketch query *teapot*, (c) combination of text queries ``umbrella'' and ``chair'' and (d) combination of sketch queries *apple* and *dog*. (Best viewed in pdf)

TABLE I
RESULTS OBTAINED BY OUR PROPOSED METHOD AND COMPARISON WITH STATE-OF-THE-ART METHODS. (MAP)

Methods	Single Object		Multiple (two) Objects	
	Text	Sketch	Text	Sketch
S-HELO [1]	–	16.10	–	3.07
Sketch-a-Net [32]	–	20.80	–	3.39
GN Triplet [2]	–	65.18	–	4.23
Proposed method	**76.28**	**68.81**	**18.65**	**12.06**

C. Discussions and comparisons

In Table I, we have presented the results obtained by our proposed framework, and compared them with three state-of-the-art methods: S-HELO [1], Sketch-a-Net [32], and GN (Google Net) Triplet [2]. All these three methods proposed some kind of sketch-based image retrieval strategy either based on handcrafted or learning based feature representation. We have used either their technique or trained model to extract the final representation of the sketches and the images from our test set, and have employed them for retrieving images based on a sketch query. This step produces a mean average precision (mAP) [33] score for each of these methods, which are shown in Table I. In case of querying by multiple sketches, we first obtain individual representation of each sketch query, used those representations to compute multiple (equal to the number of sketches) distance matrices, and take the average of them to calculate the final retrieval list. As these methods do not allow querying by text, we do not use them for text based image retrieval procedure. Although, in literature, there are some methods that allow querying by caption [17] they use a detailed text description of the query which is not our case. This is why we have not compared our method with any other text-based image retrieval method. Here, it is worth reminding that our method can retrieve images based on the combination of words describing the principal objects that appear in the images.

From the results shown in Table I, it is evident that in case of single object images, our proposed sketch-based image retrieval method has performed the best. In this case, GN Triplet has performed quite closely to us. In case of multiple objects images, all three state-of-the-art methods have performed quite poorly. Although used by us for retrieving images based on multiple queries, these state-of-the-art methods had not been designed for doing this specific task. Therefore, averaging the distances over multiple queries might have lost substantial information, which can explain these poor results. We have observed that our text-based image retrieval system performed considerably better than the sketch-based system, both in case of single and multiple object scenarios. This is probably because the spaces for sketch and image are apart than the ones from text and image.

In Fig. 3(a) and Fig. 3(b), we present two qualitative results respectively by querying with a text and a sketch input for single object. From the retrieved images shown in the figures, it is clear that the quality of the retrieval is quite satisfactory, as the first few images in each case belong to the same class as the query. Furthermore, the wrong ones have the major amount of context similar to the correct retrievals. For example, in Fig. 3(a), together with "rabbit" some images of "squirrel" also appear because they share substantial amount of context like grass, soil etc. This phenomena also appeared to be true in case of sketch based retrieval (see Fig. 3(b)). The qualitative results of retrieving images based on multiple text and sketch queries are respectively shown in Fig. 3(c) and Fig. 3(d). In these cases as well, the first few retrieved images are mostly true images. However, the number of true images retrieved are much less than the single object case. The majority of the false retrieved images contain objects belonging to one of the queried class, which is quite justified.

IV. 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 advantage of our framework is that it allows one to retrieve images queried by multiple objects of the same modality. 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 has performed satisfactorily better than the considered state-of-the-art methods on all the two datasets with some cases of failure with justifiable reasons.

One of the future directions of this work will obviously focus on improving the retrieval performance on both type of datasets (single or multiple objects). For this purpose, we plan to investigate on more efficient training strategies. Furthermore, our framework can potentially allow to query by multiple modalities at the same time. We will also explore this possibility which will allow the users to query a database in a more efficient and effortless manner.

Acknowledgement

This work has been partially supported by the European Union's Marie Skłodowska-Curie grant agreement No. 665919 (H2020-MSCA-COFUND-2014:665919:CVPR:01), the Spanish projects TIN2015-70924-C2-2-R, TIN2014-52072-P, and the CERCA Program of Generalitat de Catalunya. The Titan X and the Titan Xp used for this research were donated by the NVIDIA Corporation.

References

[1] J. M. Saavedra, "Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo)," in *ICIP*, 2014, pp. 2998–3002.

[2] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM SIGGRAPH*, 2016.

[3] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu, "Sketch-based image retrieval via siamese convolutional neural network," in *ICIP*, 2016, pp. 2460–2464.

[4] A. Gordo and D. Larlus, "Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval," in *CVPR*, 2017, pp. 5272–5281.

[5] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," *arXiv preprint arXiv:1607.06215*, 2016.

[6] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999, pp. 1150–1157.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886–893.

[8] R. Hu and J. Collomosse, "A performance evaluation of gradient field hog descriptor for sketch based image retrieval," *CVIU*, vol. 117, no. 7, pp. 790–806, 2013.

[9] J. M. Saavedra, J. M. Barrios, and S. Orand, "Sketch based image retrieval using learned keyshapes (lks)." in *BMVC*, vol. 1, no. 2, 2015, p. 7.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[11] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy, "Sketch me that shoe," in *CVPR*, 2016, pp. 799–807.

[12] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "An evaluation of descriptors for large-scale image retrieval from sketched feature lines," *CG*, vol. 34, no. 5, pp. 482–498, 2010.

[13] K. Li, K. Pang, Y.-Z. Song, T. Hospedales, H. Zhang, and Y. Hu, "Fine-grained sketch-based image retrieval: The role of part-aware attributes," in *WACV*, 2016, pp. 1–9.

[14] M. Bucher, S. Herbin, and F. Jurie, "Improving semantic embedding consistency by metric learning for zero-shot classiffication," in *ECCV*, 2016, pp. 730–746.

[15] J. Ah-Pine, G. Csurka, and S. Clinchant, "Unsupervised visual and textual information fusion in cbmir using graph-based methods," *ACM TOIS*, vol. 33, no. 2, p. 9, 2015.

[16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[17] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt *et al.*, "From captions to visual concepts and back," in *CVPR*, 2015, pp. 1473–1482.

[18] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "Vqa: Visual question answering," *IJCV*, vol. 123, no. 1, 2017.

[19] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *NC*, vol. 16, no. 12, pp. 2639–2664, 2004.

[20] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *CVPR*, 2015, pp. 4437–4446.

[21] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *IJCAI*, vol. 11, 2011, pp. 2764–2770.

[22] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, "Devise: A deep visual-semantic embedding model," in *NIPS*, 2013, pp. 2121–2129.

[23] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[24] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *CVPR*, 2016, pp. 5005–5013.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[26] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *CoRR*, vol. abs/1502.03044, 2015.

[27] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *ICCV*, 2017.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR*, 2013.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*, vol. abs/1412.6980, 2014.

[32] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. Hospedales, "Sketch-a-net that beats humans," *CoRR*, vol. arXiv:1501.07873, 2015.

[33] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," in *CVPR*, 2017, pp. 2298–2307.