Image Retrieval by Dense Caption Reasoning

Xinru Wei, Yonggang Qi, Jun Liu, Fang Liu

School of Communication and Information Engineering, BUPT, Beijing, China

***Abstract*—Humans tend to understand image scene by recognizing visual elements, then conjecturing and inferring based on them, hence are able to search relevant images. In this paper, we study the problem of complex image retrieval by reasoning image dense captions, which is similar to the way of human perception for searching images. Specifically, we transform the problem of complex image retrieval into a dense captioning and scene graph matching issue by using structured language descriptions for retrieval. Experimental results on a novel proposed large-scale content-based image retrieval dataset demonstrate the effectiveness of our proposed method.**

***Index Terms*—Image Retrieval, Dense Caption Reasoning, Captioning, Scene Graph Matching, Deep Learning**

I. Introduction

Retrieving images by visual query is one of the most attracting vision problem, which aims to search for images by reasoning about the visual elements of query image. It is a very challenging problem since an ideal retriever should be able to not only understand the whole scene but also the describing contents in details. Plenty of previous arts exist for addressing this task.

Traditional methods for content-based image retrieval often utilize low-level visual feature representations such as color, shape and appearance by means of SIFT[1], HOG[2], Fisher vector[3], etc. Meanwhile, many also rely on richer representations to work, e.g., bags of features[4], spatial pyramids[5]. However, there is an obvious drawback of the above efforts, in which semantic gap exists between the hand-crafted features extracted and the profusion of high-level human perceptions in regards to the stimuli images. There are mainly two reasons behind this: (i) the visual variation is quite large in real images which low-level features can hardly handle with, (ii) people often search images after inferring, that is, people tend to conjecture different visual concepts to be relevant, e.g., "food, forks, knives and plates" might be evidence of inferring "kitchen", "restaurant" or even "family gathering".

Recently, there has been much interest in dealing with CBIR(Content Based Image Retrieval) by matching the visual elements of images in forms of natural language, where image captioning plays the key role. Image captioning[6][7][8] achieves convincing performance due to the power of deep learning techniques. It significantly expands the complexity of the label space from a fixed small set of categories to sequences of words, which are able to express significantly richer visual concepts contained in images. Inspired by this, we treat CBIR as a caption generation and matching problem in this paper.

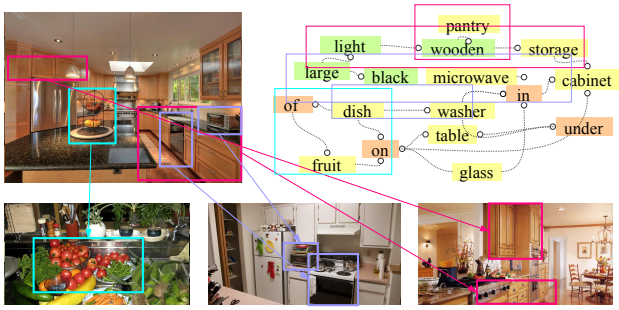

Fig. 1. **Upper Left:** Query image. **Upper Right:** Part of its scene graph to query image. **Below:** Example output relevant images, which contains very similar visual concepts like "fruits", "food", "dish washer", "microwave" and "light wooden storage" to query image.

Caption matching is quite critical for ranking images given the produced query and candidate captions. It is a text matching problem in the field of NLP(Natural Language Processing), and traditional method for text matching involves string-based method[9], corpus-based method[10] and knowledge-based method[10]. However, these methods are not designed for image caption matching, which concentrates on matching the structured visual elements in images, i.e., objects, interactions between objects and the attributes of objects. Therefore, a scene graph construction and matching strategy are presented to handle this problem.

In this paper we deal with the problem of image retrieval by generating and matching image captions (see Fig. 1). Specifically, for a given image: (i) a dense set of descriptions across regions are generated, (ii) a scene graph is constructed by structuring the produced natural languages, which involves objects, relationships and attributes, (iii) images are ordered according to their scene graph similarities given by using visual concept embeddings, which is capable to calculate semantic distance between any pair of concepts. In addition, we proposed a novel large-scale CBIR dataset. For that existing CBIR datasets either comes from classification dataset, e.g., VOC challenge dataset[11], which only concentrates on simple scene, or the dataset[12] contains complex scene images but without explicit annotation of their similarities. Therefore, to facilitate CBIR in complex image scenes, we select 10,000 real images from Visual Genome dataset[13], and for each of the images we manually labeled 100 of it's most similar images

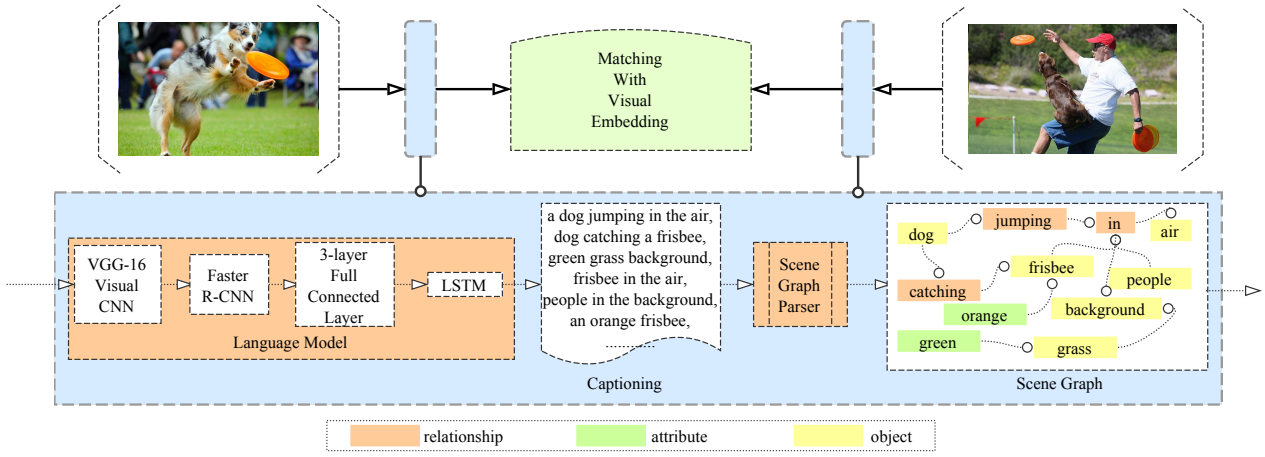

Fig. 2. Model overview. An input image is first processed by a CNN. Then the output is proposed by faster R-CNN which is followed by a three-layer fully-connected recognition network and described with an LSTM language model to generate image captions. The captions are processed with a scene graph parser and scene graph matching. Finally we get the similarity of two images based on visual embedding.

according to human's understanding of the scene in images. The contribution of this paper is three-fold:

(i) We introduce a dense caption reasoning strategy for image retrieval, which involves two stages: (a) dense caption generation, (b) scene graph construction and reasoning. Dense captioning allows us to describe the contents of images in a region manner thoroughly, and scene graph makes captions structured, hence they can be used for matching images.

(ii) A novel method for scene graph matching is presented. Instead of matching captioning words in a hard manner, we utilize visual concepts embedding, similar to word embedding, which is able to evaluate the semantic distance between visual concepts of scene graph involving objects, relationships and attributes.

(iii) A novel CBIR dataset is proposed, which contains totally 10,000 real images, and for each query image, there are 100 relevant images annotated according to people's understanding of the image contents. To our best knowledge, it is the first time that such a large-scale dataset for complex scene image retrieval has been proposed.

II. Searching Images by Caption Reasoning

Given a query image q and a set of candidate images $I = \{i_1, i_2, \ldots, i_m\}$, our goal is to rank candidate images according to their content relevant to query image q. The overview of our approach is shown in Fig. 2. There are mainly three steps: (i) A deep network is pre-trained to generate the dense descriptions of query image q and candidate images in I. (ii) We then parse the image captioning to a structured scene graph. (iii) Score the similarity between pair of images by scene graph matching. We describe the details as follows.

Image to Captions. Following[8], VGG-16 is used to obtain deep features for any image of shape $3 \times W \times H$ and gives rise to a tensor of features of shape $C \times W' \times H'$, where $C = 512$, $W' = \lfloor \frac{W}{16} \rfloor$, and $H' = \lfloor \frac{H}{16} \rfloor$. Then the tensor forms the input to the localization layer which has the same structure of Faster R-CNN. It internally selects B regions of interest and returns three output tensors giving information about region coordinates(a matrix of shape $B \times 4$), region scores(a vector of length B giving a confidence score for each output region) and region features(a tensor of shape $B \times C \times X \times Y$). Afterwards, we apply a recognition network, which is a three-layer fully-connected neural network that processes region features and produces a matrix of shape $B \times D (D = 4096)$ for next layer. Finally, a LSTM language model is adopted for captioning.

Caption to Scene Graph. To facilitate CBIR, a scene graph[12] is used to structure the language described contents of image, including object instances, attributes, and relationships between objects. Formally, caption c can be parsed to a scene graph as:

$$G(c) =< O(c), E(c), K(c) > \tag{1}$$

where $O(c) \subseteq C$ is the set of objects contained in c, $E(c) \subseteq O(c) \times R \times O(c)$ is the set of hyper-edges representing relations between objects, and $K(c) \subseteq O(c) \times A$ is the set of attributes associated with objects. More specifically, we adopt a variant of the rule-based version of the Stanford Scene Graph Parser[14], where a Probabilistic Context-Free Grammar (PCFG) dependency parser is utilized to generate its scene graph rely on linguistic rules.

Scene Graph Matching. Scene graph matching can be typically treated as a problem of tuple matching, which aggregates similarities among all the tuple pairs between two scene graphs. Formally, a function T is defined to obtain logical tuples from a scene graph:

$$T(G(c)) = O(c) \cup E(c) \cup K(c) \tag{2}$$

For example, the tuples for sentence "an orange frisbee in the air" are $\langle \textbf{\textit{frisbee}} \rangle$, $\langle \textbf{\textit{air}} \rangle$, $\langle frisbee, \textbf{\textit{in}}, air \rangle$ and $\langle frisbee, \textbf{\textit{orange}} \rangle$. Previous work[15] performs tuple matching in a hard manner, which means that only the precisely

same tuple pair makes accounts. However, it might not be an optimal solution for searching images. For example, $k1 = \langle laptop, on, desk \rangle$, $k2 = \langle table, under, computer \rangle$ may describe the same scene, while not be a valid matching in[15].

To make it suitable for image retrieval, we proposed a soft matching method. First of all, we train a word2vec model by human generated captions in order to measure visual concept descriptions commonly appearing in images. Thus, we are able to convert the descriptions of tuples into word vectors and measure the semantic similarity between tuples by calculating their distance. Consequently, searching images turns to tuples matching. Equation is defined as follows:

$$similarity(q, i) = \frac{\sum_{n=1}^{N} \sum_{m=1}^{M} \|t_{q_n} - t_{i_m}\|}{N \times M} \quad (3)$$

Where N is the number of elements in the collection $T(G(q))$, M is the number of elements in the collection $T(G(i))$, and $t_{q_n} \in T(G(q))$, $t_{i_m} \in T(G(i))$.

III. Experiment

In this section, we present experimental results of our approach for searching images, and compare with several baseline arts on a novel large-scale CBIR dataset.

A. Dataset

We select 10,000 pictures from Visual Genome dataset[13] which contains complex image scenes. For each of these images, we manually choose 100 most similar images to query images according to human understanding of images. To our best knowledge, there is no existing dataset, and it's the first time that such a large-scale dataset for complex scene image retrieval is proposed. Example ground truth images is shown in Fig. 3.

B. Image Retrieval Settings

Here we show how our method can be used for two tasks: image and natural language respectively as input to query images.

Image As Query. We randomly choose 100 images from our proposed dataset as queries. For each query and all candidate images, our proposed method is used to generate the captions and parse captions into scene graph. Afterwards, we rank all the candidate images by scene graph matching as shown in Eq. 3. We repeat this processing 10 times.

Natural Language As Query. Our method also could be used for searching images by natural language descriptions. The process of experiment is almost the same as above, the only difference is that instead of using image query (i.e., auto-generated captions), human-generated text description is directly used for searching relevant images. Human-generated descriptions come from ground truth captions of Visual Genome dataset.

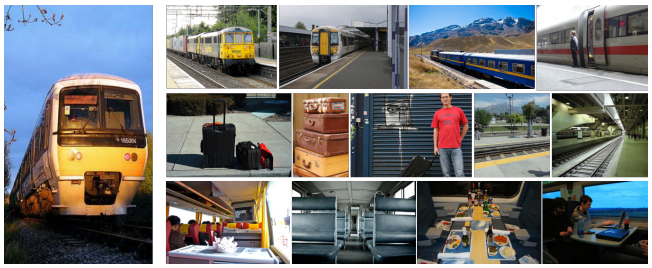

Fig. 3. **Dataset example.** The image on the left is a query image. The rest are ground truth images. We can observe that not only exactly the same scene of "train" can be retrieved, but also scenes that humans tend to link with are involved, like *"luggage"*, *"people with luggage"*, *"railway"*, *"seat on train"*, *"dinner on train"* or *"people on train"*.

C. Competitors

We compare with two low-level feature based methods, two deep-based approaches, and two different caption matching strategies to validate our proposed method. Following are the details:

SIFT: Scale Invariant Feature Transform(SIFT) is used to extract query and candidate image features, and rank them according to the L2-norm of their SIFT features.

SIFT+Bag of visual word: We adopt a BoW strategy based on SIFT features to construct a visual vocabulary by training data and quantize SIFT features as the final representation.

VGG-16: Convolutional neural network is proved to be successful on various vision tasks, including classification. Hence here we choose Vgg-16, one of the state-of-the-art, to serve as feature extractor for matching images.

Text-Visual Concept Matching: To validate our scene graph matching strategy, we compare with a state-of-the-art CBIR deep model[8] on a text query image experiment. Specifically, we use the same image retrieval model as the one in[8], where given the query image, corresponding captions are produced, and thus to match the deep visual features captures.

Scene Graph Hard matching: Our proposed method introduces a soft scene graph matching strategy after captioning. So we compare with hard captioning matching which applies scene graph matching method in[15].

Captioning+TF-IDF: Besides hard matching, TF-IDF, which is often used for text feature descriptor, is utilized to calculate the similarity between image generated captions, hence the evidence for ranking.

D. Results & Analysis

Quantitative and qualitative results are shown in Table I and Fig. 4 respectively. Specifically, Table I reports the top k accuracy of image retrieval, which is the rate of correct images appearing in the top k retrieval results, i.e., recall at k noted as R@k, and $k \in \{10, 50, 100\}$ in our setting.

We can observe from Table I that our proposed method outperforms all the other competitors. Specifically, for the task "Image as query", our proposed method overwhelms all two low-level feature based methods (SIFT and SIFT+BoW), and offers an over 3-fold improvement compared to deep-model

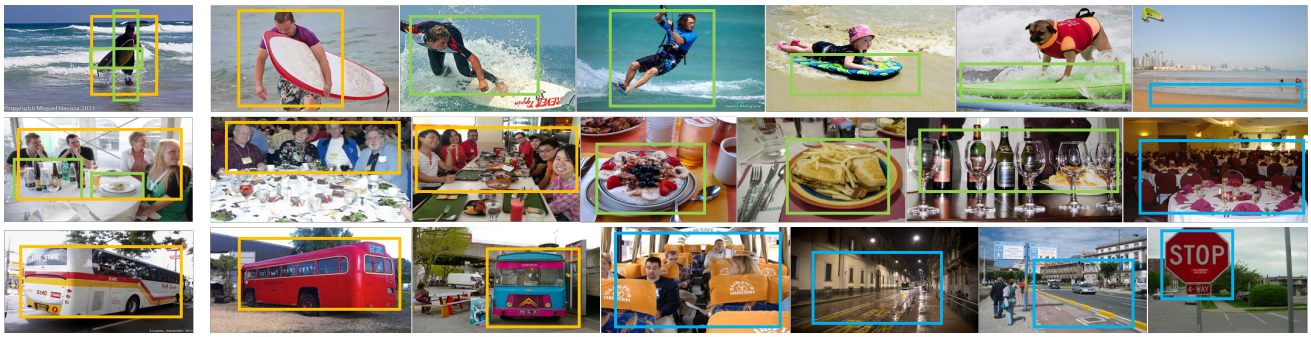

Fig. 4. **Example image retrieval results**. Images in the first colomn are query images, and the rest are their top retrieval results. Orange boxes in result images indicate exactly the same whole scene to query images. Green boxes indicate the same tuple appear in both query and retrieved images. Blue boxes indicate image contents unseen from query but could be reasonable inferred. For the query image *"a man holding a surfing board in sea"*, our method is able to return pictures about *"man holding surfing board"*, *"man surfing"*, *"baby on surfing board"*, *"dog surfing"*, and reason out *"beach"*.

TABLE I
TOP K RESULTS (R@K, HIGHER IS BETTER) FOR IMAGE RETRIEVAL.

Image As Query	R@10	R@50	R@100
SIFT	0.0	0.08	0.08
SIFT+BoW	0.12	0.08	0.21
VGG-16	0.21	0.14	0.15
Scene Graph Hard Matching	0.33	0.32	0.43
Captioning+TF-IDF	0.31	0.23	0.57
DCR-Image Query (**Our**)	**0.42**	**0.48**	**0.67**
Text As Query			
Text-Visual Concept Matching	0.01	0.13	0.24
DCR-Text Query (**Our**)	**0.04**	**0.17**	**0.28**

VGG-16. Moreover, compared with all the other matching strategies (Scene Graph Hard Matching and Captioning+TF-IDF), our approach achieves better performance which validates the effectiveness. For the task of "Text as query", our proposed method also shows better performance compared with state-of-the-art image retrieval algorithm. Fig. 4 shows some sample image retrieval results.

IV. CONCLUSION

We introduced a novel approach for complex image retrieval. In particular, given an image, a captioning network is utilized for dense caption generation, then a scene graph is constructed by using the dense captions, hence a graph matching algorithm is applied to calculate the similarity between images. In addition, we proposed a novel CBIR dataset which contains 10,000 images. Experimental results over several baseline methods validated the effectiveness of our proposed approach.

ACKNOWLEDGE

This work is supported by the National Natural Science Foundation of China (61601042, 61671078), the Fundamental Research Funds for the Central Universities under Grant No.2016RCGD09, and 111 Project of China (B08004, B17007). This work is conducted on the platform of Center for Data Science of Beijing University of Posts and Telecommunications.

REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.
[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.
[3] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR 2007. IEEE Conference on*. IEEE, 2007, pp. 1–8.
[4] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *CVPR 2010. IEEE Conference on*. IEEE, 2010, pp. 2559–2566.
[5] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.
[6] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR 2015. IEEE Computer Society Conference on*, 2015, pp. 3128–3137.
[7] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR 2015. IEEE Computer Society Conference on*, 2015, pp. 2625–2634.
[8] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *CVPR 2016. IEEE Computer Society Conference on*, 2016, pp. 4565–4574.
[9] A. Barrón-Cedeno, P. Rosso, E. Agirre, and G. Labaka, "Plagiarism detection across distant language pairs," in *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 37–45.
[10] R. Mihalcea, C. Corley, C. Strapparava *et al.*, "Corpus-based and knowledge-based measures of text semantic similarity," in *AAAI*, vol. 6, 2006, pp. 775–780.
[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.
[12] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *CVPR 2015. IEEE Computer Society Conference on*, 2015, pp. 3668–3678.
[13] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, and D. A. Shamma, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, no. 1, pp. 32–73, 2016.
[14] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in *Proceedings of the Fourth Workshop on Vision and Language*, 2015, pp. 70–80.
[15] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.