Hindawi
Advances in Multimedia
Volume 2018, Article ID 6153607, 11 pages
https://doi.org/10.1155/2018/6153607

Research Article

Combining Convolutional Neural Network and Markov Random Field for Semantic Image Retrieval

Haijiao Xu,[1] **Changqin Huang**,[1,2] **Xiaodi Huang**,[3] **Chunyan Xu**,[4] **and Muxiong Huang**[1]

[1] *School of Information Technology in Education, South China Normal University, Guangzhou, China*
[2] *Guangdong Engineering Research Center for Smart Learning, South China Normal University, Guangzhou, China*
[3] *School of Computing and Mathematics, Charles Sturt University, Albury, NSW, Australia*
[4] *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China*

Correspondence should be addressed to Changqin Huang; cqhuang@scnu.edu.cn

Received 4 May 2018; Accepted 12 June 2018; Published 1 August 2018

Academic Editor: Yong Luo

With the rapidly growing number of images over the Internet, efficient scalable semantic image retrieval becomes increasingly important. This paper presents a novel approach for semantic image retrieval by combining Convolutional Neural Network (CNN) and Markov Random Field (MRF). As a key step, image concept detection, that is, automatically recognizing multiple semantic concepts in an unlabeled image, plays an important role in semantic image retrieval. Unlike previous work that uses single-concept classifiers one by one, we detect semantic multiconcept by using a multiconcept scene classifier. In other words, our approach takes multiple concepts as a holistic scene for multiconcept scene learning. Specifically, we first train a CNN as a concept classifier, which further includes two types of classifiers: a single-concept fully connected classifier that is best suited to single-concept detection and a multiconcept scene fully connected classifier that is good for holistic scene detection. Then we propose an MRF-based late fusion approach that is able to effectively learn the semantic correlation between the single-concept classifier and multiconcept scene classifier. Finally, the semantic correlation among the subconcepts of images is cought to further improve detection precision. In order to investigate the feasibility and effectiveness of our proposed approach, we conduct comprehensive experiments on two publicly available image databases. The results show that our proposed approach outperforms several state-of-the-art approaches.

1. Introduction

With the rapid development of information technique, a large number of multimedia objects such as images are available on the Web. Given a semantic query, how to effectively find relevant images from such a scalable Web database remains a challenge. For semantic image retrieval, image concept detection is a vital step. To address this issue, many approaches have been proposed, such as Markov random walk [1], group sparsity [2], ensemble learning [3], and multiview semantic learning [4]. Although effective, these approaches work in the case of single-concept-based image retrieval. This means that each semantic query is supposed to contain only one semantic concept, restricting its practice usability.

In this paper, we specifically consider the problem of multiconcept-based image retrieval. This paradigm allows users to employ multiple semantic concepts to search relevant images. Its critical step is image multiconcept detection, *that is,* identifying multiple semantic concepts in an unseen image. Most previous studies [5, 6] utilize multiple and independent single-concept classifiers to detect such a semantic multiconcept scene. Nonetheless, this method may be ineffective, since a visual multiconcept scene (*e.g.,* "grass, person, soccer, and sports") is hard to be detected solely by a single-concept classifier. Therefore, further studies on image multiconcept detection are necessary.

In recent years, CNNs have achieved the state-of-the-art performance in many image tasks, such as single-concept-based image retrieval [7, 8], face recognition [9], image segmentation [10], and image reconstruction [11]. This indicates that a CNN can learn robust visual features by capturing semantic structures of images. A natural idea is to devise a

specific CNN for image multiconcept detection. For a task of image multiconcept scene detection, most conventional CNNs focus only on single-concept detection of images. As a result, they perform suboptimally on images with multiconcept scenes. We hence design a specific CNN that suits holistic scene detection, with two kinds of fully connected classifiers: a single-concept classifier and a multiconcept scene classifier. The former suits single-concept detection, while the latter is for holistic scene detection. Differing from the existing works that use single-concept classifiers, our method employs a multiconcept scene classifier to detect a semantic multiconcept scene, regarding multiple concepts as a holistic scene for multiconcept scene learning. Using our proposed MRF-based fusion method, we model the semantic correlation between single-concept classifier and multiconcept scene classifier and estimate the relevance score for an image multiconcept scene. The semantic link among the subconcepts presented in the images is further used to improve detection accuracy. Experimental results on MIR Flickr 2011 [12] and NUS-WIDE [13] datasets demonstrate the effectiveness of our proposed method. The major contribution of this paper is twofold:

(1) Combining CNN and MRF, we propose a unified, novel CNN framework for image multiconcept scene detection.

(2) We model the semantic link between a single-concept classifier and a holistic scene classifier in a way that effectively detects the semantic multiconcept scene in an unlabeled image.

The remainder of this paper is organized as follows. Section 2 briefly reviews some related works. Section 3 details our proposed approach. Section 4 reports our experiments with setup, results, and analysis, and Section 5 concludes this paper with some remarks on further studies.

2. Related Work

Clustered in terms of discriminative, generative, and nearest-neighbor methods, image concept detection is a vital step for semantic image retrieval. A discriminative method learns a classifier that projects visual images to semantic concepts, *that is,* Stochastic Configuration Networks (SCN) [14], while a generative method (*e.g.,* a feature-word-topic model [15]) concentrates on learning the correlation between visual images and semantic concepts. By a majority vote of nearest neighbors of an image, a nearest-neighbor method assigns a semantic concept to this image. An influential work is the TagProp [6], which employed a weighted nearest-neighbor graph to learn semantic concepts of unseen images, achieving competitive learning performance. These above-mentioned methods lose sight of the valuable semantics latently embedded in image concepts so as to simplify the design of the system and related calculation. Alternatively, some others effectively integrate the semantics information under a unified learning framework, achieving the sound performance of concept detection. In [16], the Google semantic distance was proposed to extract the semantics of semantic concepts and phrases. In [17], a semantic ontology-based hierarchical pooling method was proposed to improve the coverage or diversity of the training images.

In the research field of image retrieval, MRF-based methods are also widely used, achieving promising performance. Laferte et al. [18] proposed a discrete MRF approach, which employed the maximum a posteriori estimation on the quadtree so as to reduce the computational expense. Metzler et al. [19] proposed a MRF-based query expansion approach that provided an effective mechanism for modeling semantic dependencies of image concepts. In [20], a potential function was proposed for parameter estimation and model inference, which empowered the learning ability for a concept classifier. Kawanabe et al. [1] utilized Markov random walks on graphs of textual tags to improve the performance of image retrieval. Lu et al. [21] utilized maximum-likelihood estimation to train a spatial Markov model and then employed this model for image concept detection. Dong et al. [22] proposed a sub-Markov random walk approach with concept prior to image retrieval, which can be interpreted as a conventional random walker on a graph with added auxiliary nodes. Most traditional methods concentrate on single-concept-based image retrieval. For an image multi-concept query, they employ a combination of single-concept classifiers [5, 6] to detect image multiconcept scene.

CNN-based deep learning has recently achieved state-of-the-art performance in single-concept-based image tasks. Simonyan et al. [23] trained a deep CNN termed VGG, achieving competitive performance on the large-scale dataset ImageNet. Szegedy et al. [7] proposed a deeper CNN architecture termed GoogLeNet, achieving better learning performance on ImageNet. To improve performance of image retrieval, Hoang et al. [24] proposed three masking schemes to select a representative subset of local convolutional features. Girshick et al. [8] proposed a scalable object detection approach, Regions with CNN features (R-CNN), which applied high-capacity CNNs to bottom-up region proposals. Ren et al. [25] proposed a Region Proposal Network (RPN) that shared full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. In [26], a Multi-Loss regularized Deep Neural Network (ML-DNN) framework was proposed, which exploited multiple loss functions with different theoretical motivations to mitigate overfitting during semantic concept learning. He et al. [27] proposed a residual learning framework to alleviate the training of neural networks. Wang et al. [28] proposed a deep ensemble learning approach for large-scale data analytics. Huang et al. [29] proposed a Dense convolutional Network (DenseNet) that connected each layer to every other layer in a feed-forward fashion, strengthening feature propagation and reducing training expense. Despite effectiveness, these methods are confined to cope with single-concept-based image retrieval, limiting its practical usability. This motivates us to devise a new model to resolve this issue.

3. Proposed Approach

Our approach, called CMMR, aims to combine CNN and MRF for the multiconcept-based image retrieval. Suppose that L and U stand for a training set and a test set, respectively. Each image I in L or U is represented as a low-level visual feature vector. Given a vocabulary V with K unique semantic

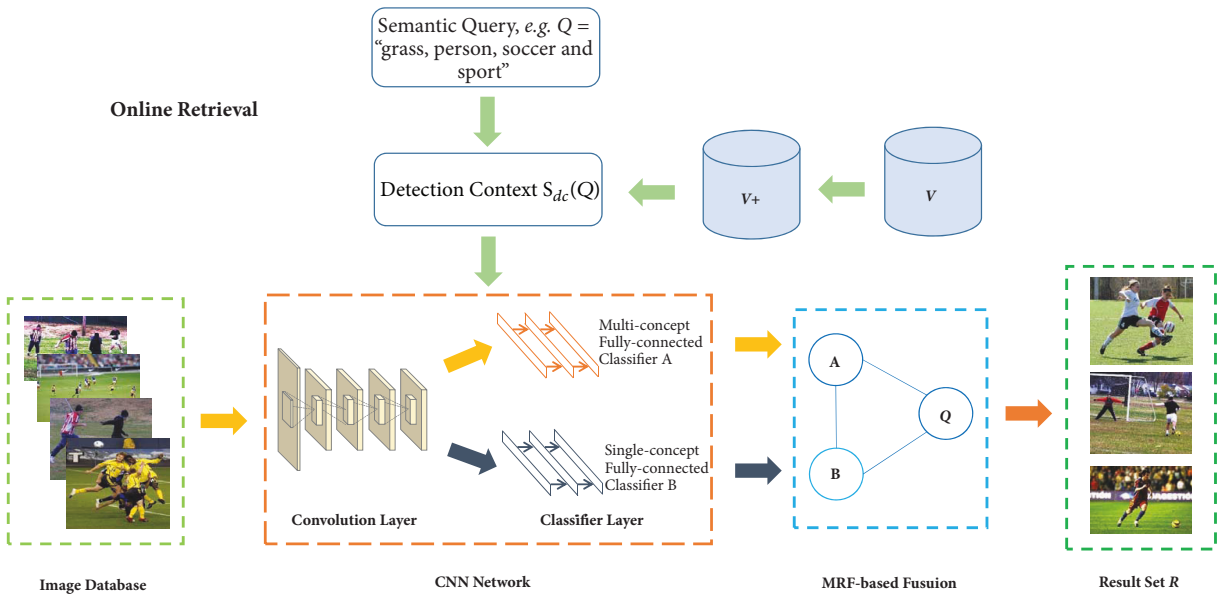


Figure 1: The proposed CMMR framework.

concepts, each concept c in V is a single concept, *for example,* "grass" or "person." Each image in the training set L is labeled with several semantic single concepts c, while the images in the test set U have no concept labels. Each semantic scene with the multiconcept C, *for example,* "clouds, sky, and sunset," is an element of the power set of V, *that is,* $C \in 2^V$ or $C \subseteq V$. Given a multiconcept query $Q \in 2^V$ (*e.g.*, "grass, person, soccer, and sports") and the target set U, our goal is to find a result set $R \subseteq U$ with relevant images. The result set R satisfies the following conditions: (1) each relevant image I in R includes all target single concepts $c \in Q$; and (2) $\forall I \in R$ and $I' \in U - R$, $f_s(Q, I) > f_s(Q, I')$, where $f_s(Q, I)$ and $f_s(Q, I')$ stand for the relevance scores for Q.

Figure 1 shows our proposed CMMR framework with working mechanisms. Our CMMR framework consists of three main components: CNN framework, MRF-based fusion, and online retrieval. CMMR aims to learn concept classifiers. Normally the last layer of CNN is a single-concept classifier. We replace it with two types of classifiers: a single-concept fully connected classifier for single-concept detection and a multiconcept scene fully connected classifier for holistic scene detection. The MRF-based fusion component learns the semantic correlation between such two types of classifiers and produces the ultimate semantic score for a given multiconcept query with a semantic scene Q. Online retrieval obtains the search result for this Q by taking four steps. First, CMMR generates the detection context $S_{dc}(Q)$ by using a semantic neighbor approach. The proposed CNN then learns a single-concept classifier and a multiconcept scene classifier. Third, the use of MRF-based fusion approach learns the ultimate semantic scores of Q. Finally, CMMR employs the learned semantic scores to perform semantic image retrieval.

3.1. Multiconcept Vocabulary Generation. CMMR regards each multiconcept $C \in 2^V$ as a whole, *that is,* one concept of a holistic scene. In order to avoid meaningless concept permutation, CMMR chooses the meaningful multiconcept C to generate a multiconcept vocabulary V^+ according to the following cooccurrence rule over the training set L:

$$|C| \leq m, \tag{1}$$

$$f_c(C) \geq n \tag{2}$$

where $|C|$ is the cardinality of C, *for example,* $|\text{grass, person, soccer, sports}| = 4$, and $f_c(C)$ is the multiconcept frequency of C. If the size of V^+ is too large, we can adjust the thresholds m and n to reduce the computational expense. In this way, V^+ containing multiconcepts C is generated.

3.2. CNN Network of Our Proposal. Normally a CNN has multiple convolutional layers followed by fully connected classifier layers. The functionality of the convolutional layers is to learn and extract robust visual features, while the classifier layers learn a concept classifier. Any CNNs for image tasks can be incorporated into our framework. Without loss of generality, we choose an influential model, GoogLeNet [7], to build our convolutional layer.

Image concept detection serves as a critical step in semantic image retrieval. Most conventional CNNs concentrate on image single-concept detection, thus performing suboptimally on image multiconcept scene detection. Furthermore, an original CNN (*e.g.*, GoogLeNet) aims to predict one concept label of an unseen image, whereas in our case each image is labeled with multiple concepts. Therefore, we modify the GoogLeNet so as to fit multiconcept scene detection.

First, we design a specific fully connected classifier layer that suits holistic scene detection, comprising two kinds of classifiers: a multiconcept scene classifier and a single-concept classifier. They share one convolutional layer, since this convolutional layer generates a general visual representation. Second, we follow [30] to define our softmax loss function f_{loss} of multiconcept learning. With this definition, the normalized prediction $p(C_j \mid I_i)$ of the image I_i in the jth multiconcept C_j is calculated as

$$p\left(C_j \mid I_i\right) = \frac{e^{\phi(I_i, C_j)}}{\sum_{j=1}^{N_2} e^{\phi(I_i, C_j)}} \tag{3}$$

where C_j (*e.g.*, "grass, person, soccer, and sports") is one holistic scene concept, $\phi(I_i, C_j)$ is the activation function, and N_2 is the number of multiconcepts. Following [30], we use a rectified linear unit as our nonlinear activation function. We minimize the Kullback-Leibler divergence between the prediction and the ground truth; f_{loss} is defined as

$$f_{loss} = -\frac{1}{N_1} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \overline{p}\left(C_j \mid I_i\right) \log\left(p\left(C_j \mid I_i\right)\right) \tag{4}$$

where N_1 is the number of images and $\overline{p}(C_j \mid I_i)$ is the ground truth in the image I_i in the jth multiconcept C_j. It is obvious that we have $\overline{p}(C_j \mid I_i) = 1$ if C_j appears in I_i and $\overline{p}(C_j \mid I_i) = 0$ otherwise.

3.3. CNN Training. Training a CNN is a two-stage process: convolution layer training and classifier layer training. The former extracts deep feature, while the latter learns a reasonable concept classifier. This process is time-consuming, especially for training on large image databases. Therefore, a publicly released pretrained GoogLeNet is employed to accelerate training. This procedure includes three steps. After being initialized with the pretrained GoogLeNet, our CNN model is able to extract deep features. Next, these deep features are fed into the classifier layer, which is then well trained. Finally, the CNN is well retrained by freezing the bottom convolution blocks, as well as by fine-tuning the top convolution block and the classifier.

For learning multiconcept of a scene C_j, the positive sample set $Po(C_j)$ and the negative sample set $Ne(C_j)$ are built as follows:

$$\begin{aligned} Po\left(C_j\right) &= \left\{I_i \mid C_j \subseteq An\left(I_i\right)\right\}, \\ Ne\left(C_j\right) &= \left\{I_i \mid I_i \notin Po\left(C_j\right)\right\}, \end{aligned} \tag{5}$$

where $An(I_i)$ is the annotation set for training image I_i. Based on above positive samples and negative samples, we train the multiconcept classifier. For traditional single-concept classifier training, the images labeled with the concept c_j are employed as positive samples and the rest as negative samples.

3.4. Detection Context Generation. Given a multiconcept query with a semantic scene Q, K_1 concept neighbors participate the concept detection and output the relevance scores. These concept neighbors are tightly linked to Q and hence can be taken as the detection context, denoted as $S_{dc}(Q)$. Some details on the procedure of generating the detection context are given below.

First, we generate a semantic neighbor set $S(Q) \subset V^+$ by choosing neighbor concepts C_j with probabilities $p(Q \mid C_j) > 0$. This symmetric semantic probability $p(Q \mid C_j)$ measures the interdependency between two concepts Q and C_j, which is represented as

$$p\left(Q \mid C_j\right) = \frac{2 f_c\left(Q \cup C_j\right)}{f_c(Q) + f_c\left(C_j\right)} \tag{6}$$

where $f_c(Q)$ and $f_c(C_j)$ are the occurrence frequency of Q and C_j, respectively, and $f_c(Q \cup C_j)$ is the number of images simultaneously including two multiconcepts Q and C_j. Each multiconcept C_i is seen as its own semantic neighbor and hence $p(C_j \mid C_j) = 1$.

Second, we assign all subconcepts $C_s \subseteq Q$ into the context set $S_{dc}(Q)$. Finally, we assign top-r related concepts C_r into the context set $S_{dc}(Q)$ from the rest of $S(Q)$. Thus, the detection context $S_{dc}(Q)$ is generated, with K_1 elements. The interdependency probability $p(Q \mid C_j)$ should be normalized as follows:

$$p\left(Q \mid C_j\right) = \begin{cases} \dfrac{p\left(Q \mid C_j\right)}{\sum_{i=1}^{K_1} p\left(Q \mid C_i\right)} & \text{if } C_i, C_j \in S_{dc}(Q), \\ 0 & \text{elsewhere.} \end{cases} \tag{7}$$

3.5. MRF-Based Fusion for Multiconcept Scene Learning. With our CNN, the concept classifier has been learned. This concept classifier projects visual images to semantic concepts. If a semantic concept and its related concepts frequently appear in images, the relevance prediction of this semantic concept will be boosted in our model. Given a multiconcept query with the semantic scene Q, all concepts C_j in the detection context $S_{dc}(Q)$ are used for estimating the relevance. The relevance prediction $p(Q \mid I; A)$ is estimated as follows:

$$p(Q \mid I; A) = \sum_{j=1}^{K_1} p\left(Q \mid C_j\right) p\left(C_j \mid I\right), \quad C_j \in S_{dc}(Q). \tag{8}$$

The relevance prediction $p(C_j \mid I)$ predicted by a multiconcept classifier A is seen as an evidence of C_j in an image I, while the semantic correlation $p(Q \mid C_j)$ is treated as a weight of this relevance prediction. In view of the promising performance in single-concept learning reported in [6, 7], a single-concept classifier B is integrated into the classifier layer of our CNN. Following [6], this single-concept prediction $p(Q \mid I; B)$ between Q and I can be estimated as follows:

$$p(Q \mid I; B) = \prod_{j=1}^{q} p\left(c_j \mid I\right), \quad c_j \in Q, \tag{9}$$

where q is the cardinality of Q and c_j is a conventional single concept that is predicted by a single-concept classifier B.

As a graphic model, MRF provides a basis for modeling contextual constraints in image retrieval. Hence, we employ MRF to analyze the semantic link between two types of classifiers mentioned above and produce the ultimate semantic score for Q. We first construct a specific MRF for the two types of classifiers and the query concept, *that is,* $\{A, B, Q\}$, so as to model their correlation. Then we infer the MRF-based fusion method for image concept detection.

Given a set of random variables $X = \{X_1, \ldots, X_N\}$ on an MRF graph, the joint probability of MRF is a Gibbs distribution [31]:

$$p(X) = \frac{e^{-E(X)}}{Z}, \tag{10}$$

where Z is a normalization factor and $E(X)$ is the energy function, *that is,* the sum of clique potentials over all possible cliques. If using random variable $y_{Qi} \in \{0, 1\}$ represents absence or presence of a multiconcept Q for an image I, the joint probability of the random variable set $\{A, B, y_{Qi}\}$ can be defined as

$$p(A, B, y_{Qi}) = \frac{e^{-E(A,B,y_{Qi})}}{Z}, \tag{11}$$

where

$$E(A, B, y_{Qi}) = V_1(A, y_{Qi}) + V_2(B, y_{Qi}). \tag{12}$$

We define the potential functions as

$$V_1(A, y_{Qi}) = \alpha_1 p(Q \mid I; A) \tag{13}$$

$$V_2(B, y_{Qi}) = \alpha_2 p(Q \mid I; B) \tag{14}$$

where $\boldsymbol{\alpha}_Q = [\alpha_1, \alpha_2]$ are the CMMR parameters to be estimated and $s.t.\ \alpha_1 + \alpha_2 = 1$.

3.6. Parameter Optimization. A widely used technique for parameter optimization is a maximum likelihood, which chooses the parameters that maximize the joint probabilities over the training set. As such, we maximize the log-likelihood function $\mathscr{L}_Q$ of the query Q. The final relevance prediction $p(y_{Qi} \mid I; A, B)$ of the image I is given by

$$p(y_{Qi} \mid I; A, B) = \frac{p(A, B, y_{Qi})}{p(A, B, Q) + p(A, B, \overline{Q})}, \tag{15}$$

$$= \frac{e^{-E(A,B,y_{Qi})}}{e^{-E(A,B,Q)} + e^{-E(A,B,\overline{Q})}}, \tag{16}$$

where Q and $\overline{Q}$ are equivalent to $y_{Qi} = 1$ and $y_{Qi} = 0$, respectively. Therefore, $\mathscr{L}_Q$ is written as

$$\mathscr{L}_Q = \sum_{i=1}^{N_1} \log p(y_{Qi} \mid I; A, B). \tag{17}$$

By substituting $p(y_{Qi} \mid I; A, B)$ in (17) with (15)-(16) and (11)-(14), we obtain the following log-likelihood function $\mathscr{L}_Q$:

$$\mathscr{L}_Q = \sum_{i=1}^{N_1} \left\{ -E(A, B, y_{Qi}) - \log\left(e^{-E(A,B,Q)} + e^{-E(A,B,\overline{Q})} \right) \right\}. \tag{18}$$

By using the gradient descent method [32], the log-likelihood $\mathscr{L}_Q$ for optimizing $\boldsymbol{\alpha}_Q$ is maximized. The gradient of $\mathscr{L}_Q$ with respect to α_i $(i \in \{1, 2\})$ can be expressed as the following form:

$$\frac{\partial \mathscr{L}_Q}{\partial \alpha_i} = \sum_{i=1}^{N_1} \left\{ -\varphi(y_{Qi}) + p(y_{Qi} \mid I; A, B)\, \varphi(Q) + (1 - p(y_{Qi} \mid I; A, B))\, \varphi(\overline{Q}) \right\}, \tag{19}$$

where $\varphi(y_{Qi}) = [p(y_{Qi} \mid I; A), p(y_{Qi} \mid I; B)]$.

3.7. Online Retrieval. CMMR concentrates on semantic image retrieval, including single-concept-based image retrieval and multiconcept-based image retrieval. A user employs multiple concepts to search for top-K semantically similar images from a database. In a word, we perform four steps for semantic image retrieval.

Step 1. Employ a semantic neighbor method to build the detection context $S_{dc}(Q)$.

Step 2. Learn a multiconcept scene classifier A and a single-concept classifier B by our proposed CNN.

Step 3. Learn the final relevance score of Q by using MRF-based fusion.

Step 4. Perform semantic image retrieval by using the learned relevance scores. Higher relevance score ranks higher.

The detailed process of semantic image retrieval is presented in Algorithm 1. From Algorithm 1, we conduct complexity analysis of time and space. Computing a set V^+ of multiconcept scene is an offline process, costing $O(1)$ time. Training a CNN is also an offline process, including deep feature extracting and classifier layer learning. This consumes $O(mn)$ time, where m and n are the trainable parameter number of CNN networks and the size of image set, respectively. By initializing our CNN with a pretrained GoogLeNet and using a very small classifier layer, the number m is substantially reduced, boosting training efficiency. Computing the detection context $S_{dc}(Q)$ is an online process, with $O(1)$ time and $O(1)$ space. For each test image, time and space complexity of computing predictions and fusing predictions are all $O(1)$. Therefore, all test images spend $O(n)$ time and $O(n)$ space. Ultimately, ranked images are returned through heap sort, consuming $O(n \log n)$ time and $O(1)$ space. Hence, the complexities of time and space of Algorithm 1 are $O(n \log n)$ and $O(n)$, respectively.

Input: training set L with label vocabulary V, test set U and query with multi-concept scene Q
Output: ranked search result R
Compute a set V^+ of multi-concept scene by Eqs. (1) and (2);
Train our CNN and obtain multi-concept scene classifier A and single-concept classifier B;
Construct detection context $S_{dc}(Q)$;
for *each* $I \in U$ **do**
    Detect image concepts using classifier A and compute relevance prediction $p(Q \mid I; A)$ by Eq. (8);
    Detect image concepts using classifier B and compute relevance prediction $p(Q \mid I; B)$ by Eq. (9);
    Perform relevance prediction fusion of A and B, and compute final prediction $p(y_{Qi} \mid I; A, B)$ by Eqs. (15) and (16);
end
Perform heap sort over all predictions $p(y_{Qi} \mid I; A, B)$ for obtaining top-M images;
Output the image list $\{I(1), I(2), \ldots, I(M)\}$ that stands for the search result R;

Algorithm 1: Semantic image retrieval process.

4. Experiments

Our experiments on semantic image retrieval include multiconcept-based image retrieval and single-concept-based image retrieval.

4.1. Datasets. We conducted the comprehensive experiments of our approach on two public datasets: MIR Flickr 2011 and NUS-WIDE. Since they include large vocabularies, we chose them to evaluate the performance of multiconcept-based image retrieval. These two datasets are publicly available, containing images and ground truth for single-concept task evaluation. MIR Flickr 2011 contains 18,000 images labeled with 99 semantic concepts. We split it into 8000 training images and 10,000 test images. NUS-WIDE is comprised of 269,648 images with a vocabulary of 81 semantic concepts. We downloaded 230,708 images in total for our experiments. This dataset is randomly divided into two sets: 138,375 images for training and the rest of 92,333 images for test.

On MIR Flickr 2011, we follow literature [33], by using GIST, HOG, SIFT, and RGB histograms as visual features. To compare two features, we employ L_2 distance for GIST, *HI* for HOG, χ^2 for SIFT, and L_1 for RGB. On NUS-WIDE, we use six visual features [13]. Similarly, we employ L_2 distance for wavelet texture, *HI* for an edge direction, χ^2 for SIFT, and L_1 for LAB and HSV, which are used in [33].

The average number of images associated with a concept is around 940 in MIR Flickr 2011 and 5381 in NUS-WIDE. The average number of concepts associated with an image is approximately 11 in MIR Flickr 2011 and about 3 in NUS-WIDE. The label vocabularies consist of dozens of label concepts, and around two-thirds of the semantic concepts have frequencies less than the mean concept frequency. Hence, semantic scene retrieval on these imbalanced datasets is challenging.

4.2. Evaluation Measures. Given a query with semantic scene Q, the ground truth for Q is defined as follows: if an image depicts all $|Q|$ target concepts $c_j \in Q$, it is considered to be relevant; and it is irrelevant otherwise. To evaluate the performance of semantic retrieval, we use three evaluation measures: Mean Average Precision (MAP), Precision at n (P@n), and Precision-Recall (PR) curve. For each semantic query, Average Precision (AP) can be computed as $AP = \sum_i \varepsilon(i)p(i)/r$, where r is the total number of relevant images in the test set U, i is the rank in the retrieved image list R, $\varepsilon(i)$ is an indicator function that equals 1 if the ith image is relevant to Q and equals 0 otherwise, and $p(i)$ is the precision at cut-off i in R, which is defined as a ratio between r and the number of retrieved images. MAP is the mean value of APs on all the queries. For Q, the correctness of high ranking retrieved image counts more. Clearly, the higher the MAP the better the retrieval performance. P@n is a variant of precision, where only the top-n ranked images are considered. Higher P@n means better retrieval performance. Besides MAP and p@n, we employ PR curve to measure semantic retrieval performance.

4.3. Experimental Configurations. In (1) and (2), m and n, respectively, control concept cardinality and concept frequency. Since training images with 11 and 3 concepts appear the most frequently, we set $m = 11$ for MIR Flickr 2011 and $m = 3$ for NUS-WIDE, respectively. To reduce computational cost, the size of V^+ is limited to an acceptable one. This means that if the frequency of a concept exceeds n, it is put into V^+; otherwise it is discarded. We set $n = 200$ for MIR Flickr 2011 and $n = 50$ for NUS-WIDE in our experiments. Thus, V^+ contains 15,970 and 2084 multiconcepts, respectively. In (7), K_1 is used to control the size of $S_{dc}(Q)$, which is determined by 5-fold cross-validation. By testing K_1 from a candidate set $\{2 * i \mid i = 1, \ldots, 20\}$, we observe that the best performance is achieved when setting $K_1 = 10$ on MIR Flickr 2011 and $K_1 = 4$ on NUS-WIDE, respectively. Therefore, we set their values accordingly. In addition, all the parameters in the compared methods are turned to the best performance reported in the relevant literatures.

The basic structure of the convolution layer we use is the same as the one used in [7]. For the classifier layer, it starts by a densely connected layer with the output size of 1024, followed by a 20% dropout. For all layers, rectified linear unit is employed as the nonlinear activation function. The optimization of the whole CNN is achieved by the stochastic gradient descent method with the mini-batch size of 128 at a 0.9 momentum. At the beginning, the CNN learning rate is

Table 1: MAPs (%) and P@10s (%) of semantic image retrieval over all 1599 semantic queries on MIR Flickr 2011. MAP scores and P@10 scores are given in the format MAP/P@10.

Method	All concepts	Query concept length $\|Q\|$		
		2	3	4
TagProp	12.8/27.3	15.2/32.1	10.4/22.6	9.9/21.9
FastTag	13.5/27.6	15.8/31.9	10.9/23.2	10.4/22.1
VGG	17.1/32.8	19.8/37.9	14.0/27.7	13.7/27.3
DBM	17.8/35.7	20.5/40.6	15.0/30.8	14.4/30.1
GoogLeNet	19.0/35.8	21.9/40.6	15.4/30.5	15.2/30.5
Ours	**21.3/39.2**	**24.6/45.2**	**17.7/34.0**	**16.8/31.8**

Table 2: MAPs (%) and P@10s (%) of semantic image retrieval over all 1581 semantic queries on NUS-WIDE. MAP scores and P@10 scores are given in the format MAP/P@10.

Method	All concepts	Query concept length $\|Q\|$		
		2	3	4
TagProp	5.5/13.8	5.5/14.9	5.2/12.8	4.5/9.8
FastTag	9.8/27.7	9.9/28.7	9.0/25.4	8.1/22.3
VGG	14.4/36.4	14.8/38.3	13.4/34.7	12.3/30.0
DBM	14.2/36.9	14.3/38.9	13.2/34.7	12.2/30.1
GoogLeNet	15.6/37.0	16.4/39.1	14.4/34.9	13.1/31.1
Ours	**17.8/42.6**	**18.0/44.3**	**16.5/41.1**	**15.4/35.8**

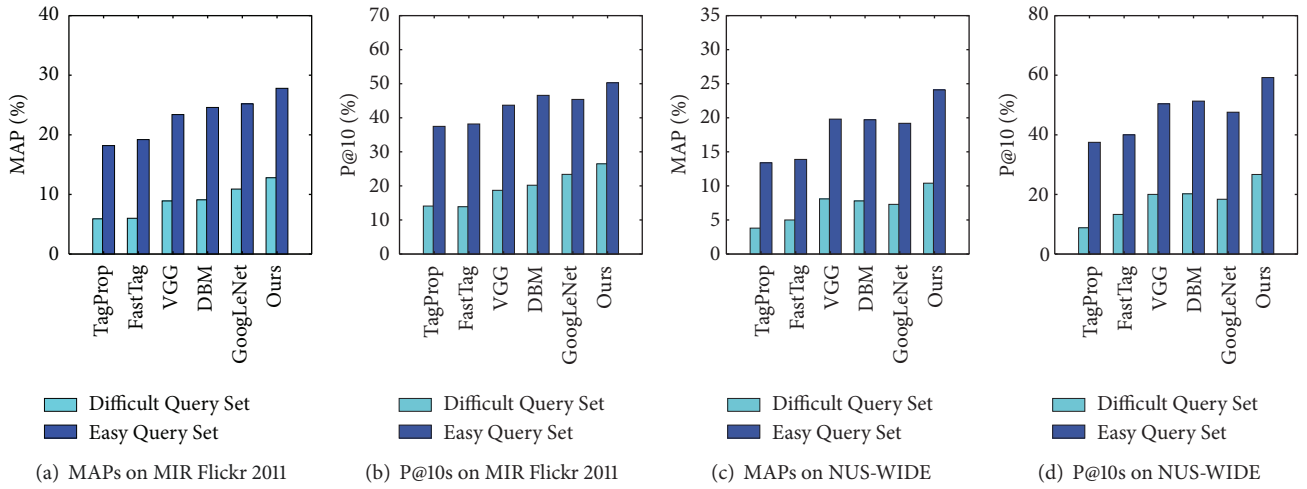
adjusted to 0.01. After 20 epochs, a staircase weight decay is used.

4.4. Comparisons. Our method is compared with several state-of-the-art concept-based methods, including TagProp [6], FastTag [34], VGG [23], DBM [35], and GoogLeNet [7]. As a classical nearest-neighbor method, TagProp uses single-concept techniques to resolve multiconcept-based image retrieval. FastTag learns two linear classifiers coregularized in a joint convex loss function that can be efficiently optimized in closed form on large-scale datasets. The others are influential single-concept-based deep learning methods. After experimenting with TagProp on the large-scale dataset NUS-WIDE, we found that this method is difficult to scale up to a large-scale dataset due to its $O(n^2)$ time and space complexity. As such, we perform TagProp experiments by using 25,000 examples on NUS-WIDE. In addition, following literature [6], we use (9) to compute relevance prediction, given a query with multiconcept scene Q.

4.5. Experiments on Semantic Image Retrieval. To evaluate retrieval performance, we construct a test query set $\mathcal{Q}$, by following two steps. First, all single-concept queries $c \in V$ are added to $\mathcal{Q}$. Then 1500 randomly generated queries with multiconcept scenes $C_j \in 2^V$ are put into $\mathcal{Q}$, with 500 2-concepts, 500 3-concepts, and 500 4-concepts, where i-concept is a multiconcept with cardinality i. In this way, $\mathcal{Q}$ is built. On MIR Flickr 2011, $\mathcal{Q}$ is comprised of 1500 multiconcepts and 99 single concepts, while $\mathcal{Q}$ contains 1500 multiconcepts and 81 single concepts on NUS-WIDE. The MAPs and P@10s are used for evaluation on semantic image retrieval with varying query lengths. Tables 1 and 2 report MAP scores and P@10 scores, where MAP scores and P@10 scores are given in the format MAP/P@10.

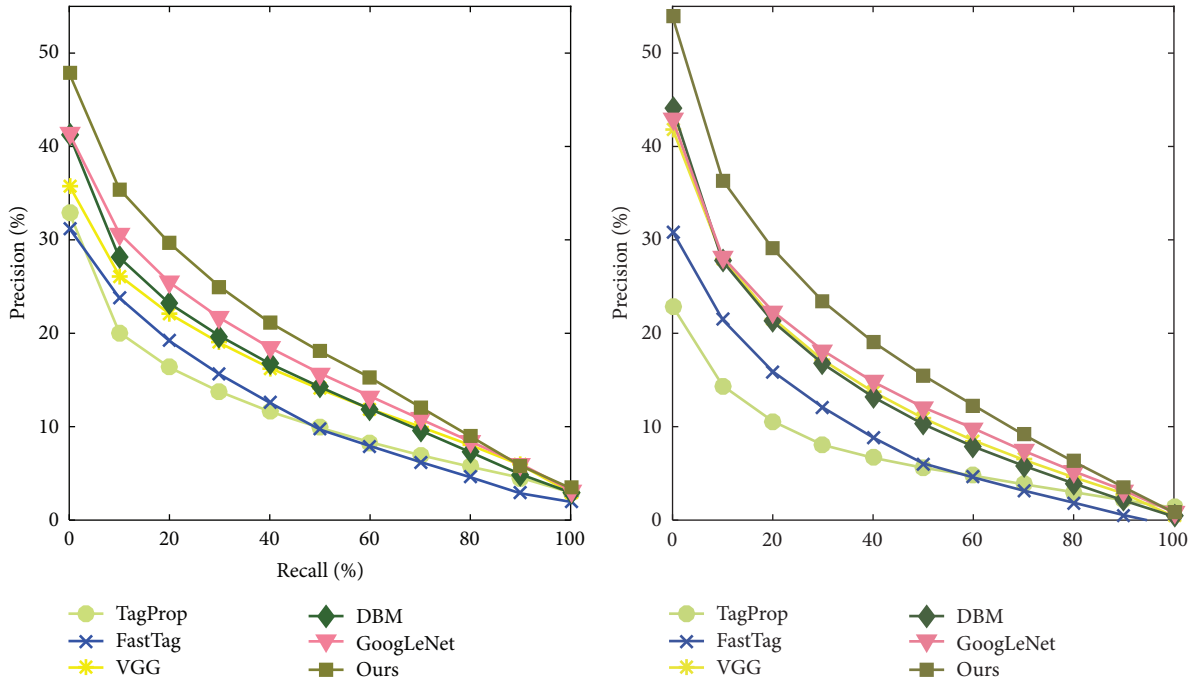
From the results, we can see that our method, CMMR, is better than other methods. Clearly, multiconcept queries perform much worse than single-concept queries on both datasets. This is because detecting a multiconcept scene is more difficult than detecting a single-concept one. A multiconcept scene may have the characteristic visual appearance, while the goal of traditional single-concept models is to achieve precise results of single-concept detection. To search for a holistic scene, traditional methods use a combination of single-concept technologies. However, in some cases, this may lose some semantics latently embedded in the holistic scene. Therefore, only using single-concept classifiers is difficult to detect a sophisticated multiconcept scene. This observation motivates us to jointly consider the multiconcept scene classifier and the single-concept classifier in devising our CNN. Moreover, the MRF-based fusion method can effectively learn the semantic correlation of multiconcept scene classifier and single-concept classifier, boosting the detection accuracy of a semantic scene.

We further conduct the comparisons with different experiment settings. More specifically, we construct a group of comparative evaluation, that is, a difficult query set with less than 100 relevant images and an easy query set with more than 100 relevant images. The experimental results are shown in Figure 2. We can find out that our method still leads the search results. Figure 3 shows the PR curves of all compared methods on two datasets, illustrating the precision variation with the varying recall. As can be seen, our method CMMR has the better precision than compared methods at every level of recall.



(a) MAPs on MIR Flickr 2011 (b) P@10s on MIR Flickr 2011 (c) MAPs on NUS-WIDE (d) P@10s on NUS-WIDE

Figure 2: Semantic retrieval performance (MAPs % and P@10s %) over the comparative group: a difficult query set and an easy query set on two datasets.

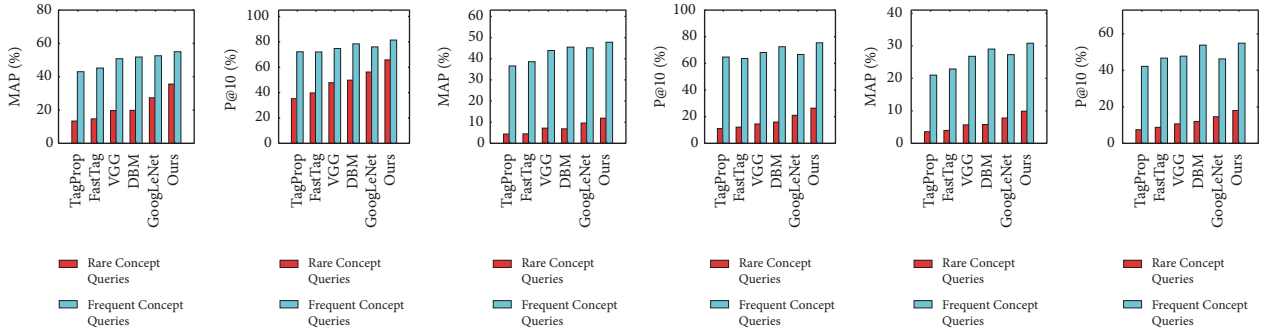


(a) MIR Flickr 2011 (b) NUS-WIDE

Figure 3: The PR curves on two datasets.

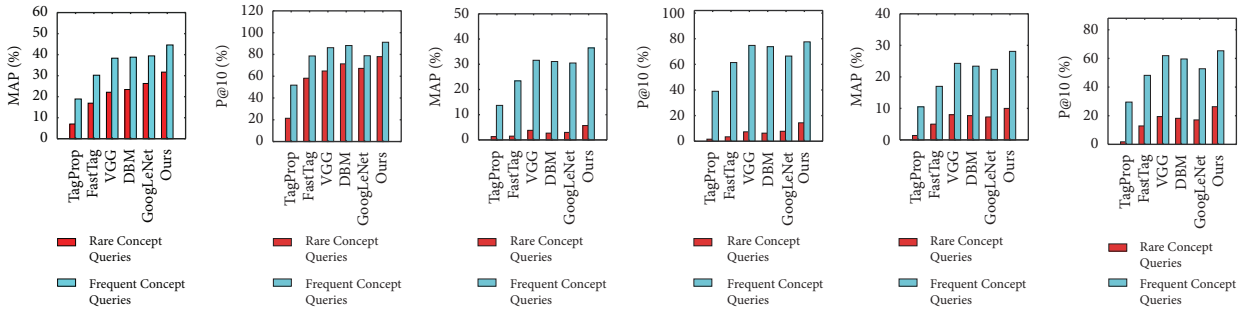
4.6. Experiments on Rare Concept Queries. Most existing approaches assume balanced concept distributions or equal misclassification costs. Nevertheless, a real-world dataset is commonly highly imbalanced [36]. When presented with complex imbalanced datasets, these methods fail to properly represent the distributive characteristics of the data and resultantly provide unfavorable precision. On MIR Flickr 2011 and NUS-WIDE, the frequencies of most concepts are below average, leading a concept classifier to overclassify the frequent concepts with high occurrence frequencies in the learning stage. This makes it hard to derive a proper model for rare concepts with low occurrence frequencies. In such situations, a concept classifier commonly has the good performance on frequent concepts but very poor performance on rare concepts. This observation suggests that, for developing a classifier, we should consider varying frequencies of concepts.

Two groups of experiments are devised: rare concept queries and frequent concept queries. In the first group, the top-50 rare single concepts, the top-100 rare 2-concepts, and the top-100 rare 3-concepts from $\mathcal{Q}$ are selected as three respective sets of the single-concept rare queries, the 2-concept rare queries, and the 3-concept rare queries,



(a) MAPs on single-concept R_1 and F_1 (b) P@10s on single-concept R_1 and F_1 (c) MAPs on 2-concept R_2 and F_2 (d) P@10s on 2-concept R_2 and F_2 (e) MAPs on 3-concept R_3 and F_3 (f) P@10s on 3-concept R_3 and F_3

Figure 4: MAPs and P@10s (%) of semantic image retrieval for rare concepts and frequent concepts on MIR Flickr 2011.



(a) MAPs on single-concept R_1 and F_1 (b) P@10s on single-concept R_1 and F_1 (c) MAPs on 2-concept R_2 and F_2 (d) P@10s on 2-concept R_2 and F_2 (e) MAPs on 3-concept R_3 and F_3 (f) P@10s on 3-concept R_3 and F_3

Figure 5: MAPs and P@10s (%) of semantic image retrieval for rare concepts and frequent concepts on NUS-WIDE.

respectively, denoted by R_1, R_2, and R_3. In the second group, the top-50 frequent single concepts, the top-100 frequent 2-concepts, and the top-100 frequent 3-concepts from $\mathcal{Q}$ are, respectively, chosen as the set F_1 of single-concept frequent query, a set F_2 of the 2-concept frequent query, and a set F_3 of the 3-concept frequent query.

As shown in Figures 4 and 5, concept classifiers achieve the higher MAPs and P@10s on the frequent concept sets F_1, F_2, and F_3 but far lower MAPs and P@10s on the rare concept sets R_1, R_2, and R_3, significantly impacting retrieval performance and user experience. For the rare concept sets R_1, R_2, and R_3 on MIR Flickr 2011, our approach outperforms the compared methods, with the better improved 30%, 24%, and 26% over the second best method in terms of MAP, respectively. On NUS-WIDE, a similar improvement is also observed. During rare concept detection with semantic scene Q, a group of weighted concept classifiers of its detection context $S_{dc}(Q)$ take part in concept detection through MRF-based fusion method. Among these concepts from $S_{dc}(Q)$, some concepts $C_j \in S_{dc}(Q)$ may be frequent concepts, which significantly boosts the relevance prediction of Q and makes the rare concept Q easier to be detected. Moreover, our maximization of the log likelihood of semantic concepts compensates for the varying frequencies of concepts. Consequently, our approach can remit the issue of concept imbalance, thus boosting retrieval performance.

5. Conclusion

Searching semantic images with high accuracy turns to be significant nowadays because of a vast number of real-world applications such as cognitive educational resource retrieval. As a key step, image scene detection plays an important role in semantic image retrieval. In this paper, we have presented a novel CNN framework for semantic image retrieval, which combines CNN and MRF in a novel way that enhances the capacity of multiconcept scene detection. Compared with previous methods, our CNN framework seamlessly incorporates three components: single-concept classifier, multiconcept scene classifier, and semantics. The combination of these three components can enhance the capability of CNN for detecting semantic scenes. We have conducted the comprehensive experiments on two public datasets. The favorable results indicate that our proposed method outperforms the compared approaches.

For future work, we intend to develop a better learning and fusion method for multiconcept scene detection.

Additionally, we would also like to explore the links among concepts, *for example,* concept graph or semantic hierarchy to boost the retrieval performance.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 61370229 and 61702388), the GDUPS (2015), the CSC (no. 201706755023), and China Postdoctoral Science Foundation (nos. 2016M600657 and 2017T100637).

References

[1] M. Kawanabe, A. Binder, C. Müller, and W. Wojcikiewicz, "Multi-modal visual concept classification of images via Markov random walk over tags," in *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pp. 396–401, 2011.

[2] S. Zhang, J. Huang, H. Li, and D. N. Metaxas, "Automatic image annotation and retrieval using group sparsity," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 3, pp. 838–849, 2012.

[3] S. Tang, Y.-D. Zhang, Z.-X. Xu, H.-J. Li, Y.-T. Zheng, and J.-T. Li, "An efficient concept detection system via sparse ensemble learning," *Neurocomputing*, vol. 169, pp. 124–133, 2015.

[4] Z. Guan, L. Zhang, J. Peng, and J. Fan, "Multi-View Concept Learning for Data Representation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3016–3028, 2015.

[5] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1371–1384, 2008.

[6] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proceedings of the International Conference on Computer Vision*, pp. 309–316, 2009.

[7] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[8] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: unsupervised fine-tuning with hard examples," in *Proceedings of the European Conference on Computer Vision*, pp. 3–20, 2016.

[9] G. Hu, X. Peng, Y. Yang, T. M. Hospedales, and J. Verbeek, "Frankenstein: learning deep face representations using small data," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 293–303, 2018.

[10] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proceedings of the International Conference on Computer Vision*, pp. 1377–1385, 2015.

[11] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 38, pp. 295–307, 2016.

[12] S. Nowak, K. Nagel, and J. Liebetrau, "The CLEF 2011 photo annotation and concept-based retrieval tasks," in *Proceedings of the CLEF Conference and Labs of the Evaluation Forum*, pp. 1–25, 2011.

[13] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: a real-world web image database from National University of Singapore," in *Proceedings of International Conference on Image and Video Retrieval*, pp. 48–56, 2009.

[14] D. Wang and M. Li, "Stochastic Configuration Networks: Fundamentals and Algorithms," *IEEE Transactions on Cybernetics*, vol. 47, no. 10, pp. 3466–3479, 2017.

[15] C.-T. Nguyen, N. Kaothanthong, T. Tokuyama, and X.-H. Phan, "A feature-word-topic model for image annotation and retrieval," *ACM Transactions on the Web (TWEB)*, vol. 7, no. 3, pp. 1–24, 2013.

[16] R. L. Cilibrasi and P. M. B. Vitányi, "The google similarity distance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, 2007.

[17] S. Zhu, C.-W. Ngo, and Y.-G. Jiang, "Sampling and ontologically pooling web images for visual concept learning," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1068–1078, 2012.

[18] J.-M. Laferte, P. Perez, and F. Heitz, "Discrete Markov image modeling and inference on the quadtree," *IEEE Transactions on Image Processing*, vol. 9, no. 3, pp. 390–404, 2000.

[19] D. Metzler and W. B. Croft, "Latent concept expansion using Markov random fields," in *Proceedings of the ACM International SIGIR Conference*, pp. 311–318, 2007.

[20] Y. Xiang, X. Zhou, T.-S. Chua, and C.-W. Ngo, "A revisit of generative model for automatic image annotation using markov random fields," in *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pp. 1153–1160, June 2009.

[21] Z. Lu and H. H. S. Ip, "Spatial Markov kernels for image categorization and annotation," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 41, no. 4, pp. 976–989, 2011.

[22] X. Dong, J. Shen, L. Shao, and L. Van Gool, "Sub-Markov random walk for image segmentation," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 516–527, 2016.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," https://arxiv.org/abs/1409.1556v6.

[24] T. Hoang, T.-T. Do, D.-K. Le Tan, and N.-M. Cheung, "Selective deep convolutional features for image retrieval," in *Proceedings of the 25th ACM International Conference on Multimedia, MM 2017*, pp. 1600–1608, October 2017.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[26] C. Xu, C. Lu, X. Liang et al., "Multi-loss Regularized Deep Neural Network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 12, pp. 2273–2283, 2016.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 770–778, July 2016.

[28] D. Wang and C. Cui, "Stochastic configuration networks ensemble with heterogeneous features for large-scale data analytics," *Information Sciences*, vol. 417, pp. 55–71, 2017.

[29] G. Huang, Z. Liu, L. v. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, Honolulu, Hawaii, USA, July 2017.

[30] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," https://arxiv.org/abs/1312.4894v2.

[31] C. Wang, N. Komodakis, and N. Paragios, "Markov Random Field modeling, inference and learning in computer vision and image understanding: a survey," *Computer Vision and Image Understanding*, vol. 117, no. 11, pp. 1610–1627, 2013.

[32] X. Li, "Preconditioned Stochastic Gradient Descent," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1454–1466, 2018.

[33] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid, "Image annotation with tagprop on the MIRFLICKR set," in *Proceedings of the 2010 ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2010*, pp. 537–546, USA, March 2010.

[34] M. Chen, A. Zheng, and K. Weinberger, "Fast image tagging," in *Proceedings of the International Conference on Machine Learning*, pp. 1274–1282, 2013.

[35] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *Journal of Machine Learning Research*, vol. 15, pp. 2949–2980, 2014.

[36] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.