

Unsupervised Joint Mining of Deep Features and Image Labels for Large-scale Radiology Image Categorization and Scene Recognition

Xiaosong Wang, Le Lu, *IEEE Senior Member*, Hoo-chang Shin, Lauren Kim,
Mohammadhadi Bagheri, Isabella Nogues, Jianhua Yao, Ronald M. Summers
Department of Radiology and Imaging Sciences, National Institutes of Health Clinical Center,
10 Center Drive, Bethesda, MD 20892

{xiaosong.wang, le.lu, hoo-chang.shin, lauren.kim, mohammad.bagheri, isabella.nogues}@nih.gov,
jyao@cc.nih.gov, rms@nih.gov

Abstract

The recent rapid and tremendous success of deep convolutional neural networks (CNN) on many challenging computer vision tasks largely derives from the accessibility of the well-annotated ImageNet and PASCAL VOC datasets. Nevertheless, unsupervised image categorization (i.e., without the ground-truth labeling) is much less investigated, yet critically important and difficult when annotations are extremely hard to obtain in the conventional way of “Google Search” and crowd sourcing. We address this problem by presenting a looped deep pseudo-task optimization (LDPO) framework for joint mining of deep CNN features and image labels. Our method is conceptually simple and rests upon the hypothesized “convergence” of better labels leading to better trained CNN models which in turn feed more discriminative image representations to facilitate more meaningful clusters/labels. Our proposed method is validated in tackling two important applications: 1) Large-scale medical image annotation has always been a prohibitively expensive and easily-biased task even for well-trained radiologists. Significantly better image categorization results are achieved via our proposed approach compared to the previous state-of-the-art method. 2) Unsupervised scene recognition on representative and publicly available datasets with our proposed technique is examined. The LDPO achieves excellent quantitative scene classification results. On the MIT indoor scene dataset, it attains a clustering accuracy of 75.3%, compared to the state-of-the-art supervised classification accuracy of 81.0% (when both are based on the VGG-VD model).

1 Introduction

Deep Convolutional Neural Networks (CNNs) have demonstrated remarkable success on many challenging computer vision tasks of object recognition, detection,

segmentation and scene recognition using public image datasets (e.g., Pascal VOC [16], ImageNet ILSVRC [12, 48], MS COCO [38]), with significantly superior performance than previous arts, especially those non-deep methods built upon hand-crafted image features. However, the good efficacy of CNNs often comes at the cost of large amounts of annotated training data. ImageNet pre-trained deep CNN models [26, 30, 37] serve an indispensable role to be bootstrapped or fine-tuned [24] for all externally-sourced data exploitation tasks [36, 5].

In the medical imaging domain, nevertheless, no large-scale labeled image dataset comparable to ImageNet exists (except the one in [49] which is not directly comparable). Modern hospitals store vast amounts of radiological images/reports in their Picture Archiving and Communication Systems (PACS). The main challenge now lies in how to obtain or compute the ImageNet-like semantic labels given a large collection of medical images. Conventional means of collecting image labels (e.g., Google image search using the terms from WordNet ontology hierarchy [40], SUN/PLACE databases [61, 66] or NEIL knowledge base [6]; followed by crowd-sourcing [12]) are not applicable, due to 1) unavailability of a high quality or large capacity medical image search engine, and 2) the formidable difficulties of medical annotation tasks for annotators with no clinical training. Additionally, even for well-trained radiologists, this type of “assigning labels to images” task is not aligned with their diagnostic routine work so that drastic inter-observer variations or inconsistency are expected. The protocols of defining image labels based on visible anatomic structures (often multiple), or pathological findings (possibly multiple) or both cues may have intrinsically high ambiguities.

Recent semi-supervised image feature learning and self-taught image recognition techniques [53, 46, 35, 27, 11] have advanced both supervised image classification and unsupervised clustering processes, demonstrating some promising results. Common image patches [53, 34], ob-

ject parts [27], prototypes [11, 10] or spatial context [14] can be first mined amongst images of the same theme (e.g., the same indoor scene class [44]) and then concatenated to serve as the discriminative image representations for classification. All of these methods, however, require image labels in order to learn class-specific informative image representations.

In this paper, we present the **Looped Deep Pseudo-task Optimization** framework (LDPO) for joint mining of image features and labels, with no prior knowledge of the image categories. The “true” image category labels are assumed to be latent and not directly observable. The main idea is to learn and train CNN models using pseudo-task labels (since human-annotated labels are unavailable) and iterate this process with the expectation that pseudo-task labels will gradually resemble the real image categories. This looped optimization algorithm flow starts with deep CNN feature extraction and image encoding using domain-specifically (e.g., CNN trained on radiology images and text report-derived labels [49]) or generically initialized CNN models. Afterwards, the CNN-encoded image feature vectors are clustered to compute and refine image labels, then we feed the newly clustered labels to fine-tune the current CNN models. Next, the obtained more task-specific and representative deep CNN will serve as the deep image encoder in the successive iteration. This looped process will halt until a stopping criterion is met. For medical image annotation, LDPO generated image clusters can be further interpreted by a natural language processing (NLP) based text mining system and/or a clinician [65].

Our contributions are three-fold. **1)** The unsupervised joint mining of deep image features and labels via LDPO is conceptually simple and based on the hypothesized “convergence” of better labels lead to better trained CNN models which in turn, offer more effective deep image features to facilitate more meaningful clustering/labels. *This looped property is unique to deep CNN classification-clustering models since other types of classifiers do not learn better image features simultaneously.* **2)** We apply our method to the large-scale medical image auto-annotation. To the best of our knowledge, this is the first work exploiting to integrate unsupervised deep feature clustering and supervised deep label classification for self-annotating a large-scale radiology image database where the conventional means of image annotation may not be quite feasible. Our best converged model obtains the Top-1 classification accuracy of 0.8109 and Top-5 accuracy 0.9412 with 270 formed image categories. **3)** LDPO framework is also validated through the scene recognition task where the ground-truth labels are available (only for the validation purpose). We report the 67-class clustering accuracy of 75.3% on the MIT-67 indoor scene dataset [44] that doubles the performance from the baseline methods (of using k-means or agglomerative clus-

tering on the ImageNet-pretrained deep image features via AlexNet [30]) and is strongly close to the fully-supervised deep classification result of 81.0% [8].

2 Related Work

Image Categorization or Auto-annotation: Image auto-annotation task is addressed via multiple instance learning [59] but the target domain is restricted to a small subset (only 25 out of 1000 classes) of ImageNet [12] and SUN [61]. [58] introduces a hierarchical set of unlabeled data clusters (spanning a spectrum of visual concept granularities) that can be efficiently labeled to produce high performance classifiers (thus less label noises than the instance-level labeling). [49] first extract the sentences that depict disease referencing key images (analogous to “key frames in videos”) via NLP from a total collection of $\sim 780K$ patients’ radiology text reports, and 215,786 key images of 61,845 unique patients are found. Then, image categorization labels are computed using unsupervised hierarchical Bayesian document clustering, i.e., latent Dirichlet allocation (LDA) topic modeling [1], to form 80 classes. The text-computed category information offers some coarse level of radiology semantics but appears to be limited in two aspects: 1) The classes are *highly unbalanced*, in which one dominating category contains 113,037 images while other classes contain a few dozens. 2) Some classes can be highly incoherent among their within-the-class image instances.

Unsupervised and Semi-supervised Learning: Dai *et al.* [11, 10] study the semi-supervised image classification and clustering on problems of texture [31], small- to middle-scale object categories (e.g., Caltech-101 [18]) and scene recognition [45]. Ensemble projections (EP) as a rich set of visual prototypes are derived as the new image representation for clustering and recognition. Graph based approaches [39, 29] are used to link the unlabeled image instances to labeled ones (which are served as anchors) and propagate labels by the graph topology and connectiveness weights. In an unsupervised manner, Coates *et al.* [9] employ k-means to mine image patch filters and utilize the resulted filters for feature computation. Surrogate classes are obtained by augmenting each image patch with its geometrically transformed versions and a CNN is trained on top of these surrogate classes to generate features, as studied in [15]. [64] integrates the hierarchical agglomerative clustering process into a recurrent neural network by merging the clusters (as groups of images) iteratively toward the predefined cluster number and simultaneously updating the CNN activations for image representation.

Our looped optimization method shares a similar concept with [64] in the joint learning of image clusters and image representations. However, it differs significantly in the following respects: 1) an unlabeled image collection can be initialized with either randomly-assigned labels or labels obtained by a pseudo-task (e.g., text topic modeling gen-

erated labels [49]); 2) Our framework has the flexibility of working with any clustering function. Particularly, it employs Regularized Information Maximization (RIM [21]) clustering to perform clustering the image (like k-means) with model selection on finding the optimal number of clusters *whereas only agglomerative clustering loss [22] can be integrated into the neural network model in [64]*. 3) The empirical convergence process of our LDPO method is observable and quantifiable, as described in Sec. 3.2.

Mid-level Image Representation: Since the seminal work on discriminative image patch discovery [53], mid-level visual elements based image representation has been explored intensively and found being effective on boosting the performance of many visual computing tasks, particularly scene recognition [53, 13, 27, 54, 33, 2, 11, 34, 60]. A variety of mid-level visual elements can be harvested, e.g., image patches [53, 13, 34, 60], parts/segments [27, 54, 2], prototypes [11], attributes [51, 7] through different learning and mining techniques, e.g., iterative optimization [53, 27], classification and co-segmentation [54], Multiple Instance Learning (MIL) [33], random forests [2], ensemble projection[11] and association rule mining [34]. Nonetheless, these methods require that images are grouped before their representations are mined inside each group, which is a form of weakly supervised learning (WSL).

Our work is partly related to the iterative optimization in [53, 27] that seeks to identify discriminative local visual patterns as parts and reject others, while our goal is to jointly mine better deep image representations and the labels for all images, towards iterative auto-annotation. We can integrate the association rule mining technique [34] to extract the frequent image parts (that are further used to encode image representation) into our LDPO pipeline, and report excellent unsupervised scene recognition accuracy of 75.3% on MIT indoor scene dataset [60, 8, 44].

3 Joint Mining of Deep Features and Labels

Supervised or semi-supervised learning paradigms (as described in Sec. 2) usually require (at least partial) image labels as a prerequisite. These lines of work, at the era of “deep learning”, would necessitate a huge amount of data annotation efforts. For medical imaging applications, well-trained clinical professionals or physicians are in need for data labeling, instead of Amazon Mechanical Turkers in computer vision. Employing and converting the medical records stored in the PACS into image labels or tags is a highly non-trivial and unsolved NLP problem with high labeling uncertainties, observed by [50]. Our approach exploits unsupervised category discovery using empirical image cues for grouping or clustering, through an iterative optimization process of 1) deep image feature extraction and clustering; and 2) deep CNN model fine-tuning (i.e., using new labels from clustering), to update deep feature extraction in the next round.

Without loss of generality, our method is first employed in the scenario of medical image categorization. We highlight the problem-specific settings for scene recognition task when they are different. As illustrated in Fig. 1, the iteration begins by extracting the deep CNN image feature using either a domain-specific [49] or generic ImageNet [30] CNN model (Sec. 3.1). Next, the clustering on deep feature with k -means or k -means followed by RIM is exploited (Sec. 3.2). By evaluating the purity and mutual information between formed clusters in consecutive rounds, the system either terminates the current iteration (and yields converged clustering outputs); or uses the newly refined image cluster labels to train or fine-tune the CNN model in the next iteration. For medical image categorization (dashed box in Fig. 1), LDPO-generated image clusters can be further fed into text processing. The system can extract semantically meaningful text words for each formed cluster. Furthermore, the hierarchical category relationship is built using the class confusion measures of the final converged CNN classification models (Sec. 3.3).

3.1 Deep CNN Image Representation & Encoding

A variety of CNN models can be used in our method. We analyze the CNN activations from layers of different depths in AlexNet [30], VGGnet[52] and GoogLeNet [55]. Pre-trained models on the ImageNet ILSVRC data are obtained from Caffe Model Zoo [26]. We also employ the Caffe CNN implementation [26] to perform fine-tuning on CNNs using the key image database [49, 50]. AlexNet is a popular 7-layer CNN architecture and the extracted features from its convolutional or fully-connected layers have been broadly investigated [20, 47, 28, 41]. In our experiments we harness image feature activations of the 5th convolutional layer *Conv5* and 7th fully-connected (FC) layer *FC7*, suggested by [8, 3]. GoogLeNet [55] is a much deeper CNN architecture that comprises 9 inception modules and each module is a set of convolutional layers with multiple window sizes of $1 \times 1, 3 \times 3, 5 \times 5$. We utilize the deep features from the last inception layer *Inception5b* and the final pooling layer *Pool5*. For the scene recognition task, very deep VGGNet (VGG-VD) [52] is also employed, in addition to AlexNet. The extracted features from VGG-VD’s last fully-connected layers are used for the patch-mining based image encoding. Table 1 illustrates the detailed CNN layers and their activation dimensions.

Deep image features extracted from the last convolution layer preserve their overall spatial locations or image layouts while the fully-connected CNN layer will lose spatial information. We adopt to encode the last convolutional layer outputs (as feature activation maps) in a form of dense pooling via Fisher Vector (FV) [43] and Vector Locally Aggregated Descriptor (VLAD) [25], before feeding them to the fully-connected layer. The dimensions of FV or VLAD encoded deep features are much higher than those of the FC

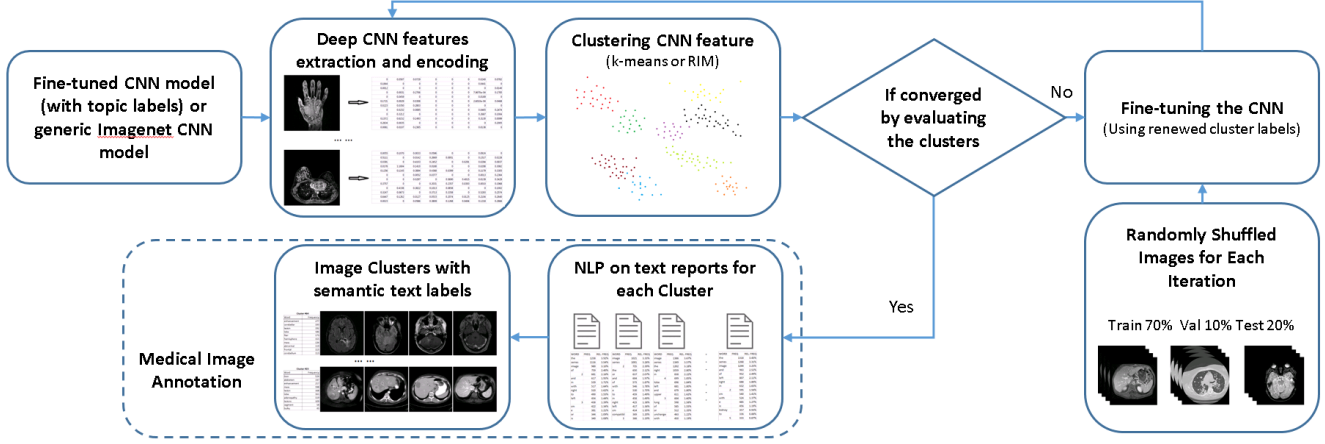


Figure 1. The overview of looped deep pseudo-task optimization (LDPO) framework.

Table 1. Configurations of CNN output layers and encoding schemes (Output dimension is 4096, except Pool5 as 1024).

CNN model	Layer	Activations	Encoding
Medical Image Categorization			
AlexNet	FC7	4096	—
AlexNet	Conv5	(13, 13, 256)	FV+PCA
AlexNet	Conv5	(13, 13, 256)	VLAD+PCA
GoogLeNet	Pool5	1024	—
GoogLeNet	Inception5b	(7, 7, 1024)	VLAD+PCA
Scene Recognition			
AlexNet	FC7	4096	PM+PCA
VGG-VD	FC7	4096	PM+PCA

layers. Since there is redundant information from the encoded deep features, Principal Component Analysis (PCA) is performed to reduce the feature dimensionality to 4096 (same to the FC dimension [30, 55]) that makes different encoding schemes more comparable.

Mined mid-level visual elements based image encoding has proven to be a more discriminative representation in natural scene recognition [53, 13, 27, 34, 60]. Visual elements are expected to be common amongst the images with same label but seldom occur in other categories. The association rule mining technique is integrated into our looped optimization method flow (similar to [34]) to automatically discover mid-level image patches for encoding. We conjecture that discriminative patches can be discovered and gradually improved through the LDPO iterations even if the initialization image labels are not accurate.

CNN activation based encoding: Given a pre-trained (generic or domain-specific) CNN model (e.g., Alexnet or GoogLeNet), an input image I is resized to fit the model definition and feed into the CNN model to extract features $\{f_{i,j}^L\}$ ($1 \leq i, j \leq s^L$) from the L -th convolutional layer with dimensions $s^L \times s^L \times d^L$, e.g., $13 \times 13 \times 256$ of *Conv5* in AlexNet and $7 \times 7 \times 1024$ of *Pool5* in GoogLeNet. For the Fisher Vector implementation, we use the settings as suggested in [8]: 64 Gaussian components are adopted to

train the Gaussian mixture Model (GMM). The dimension of resulted FV features is significantly higher than $FC7$'s, i.e. $32768(2 \times 64 \times 256)$ vs 4096. After PCA, the FV representation per image is reduced to a 4096-component vector. A list of deep image features, the encoding methods and output dimensions are provided in Table 1. To be consistent with the setting of FV encoding, we initialize the VLAD encoding of convolutional features by k -means clustering ($k = 64$). Thus the resulted dimensions of VLAD descriptors are $16384(64 \times 256)$ of *Conv5* in AlexNet and $65536(64 \times 1024)$ of *Inception5b* in GoogLeNet, both reduced to 4096 via PCA.

Patch mining based encoding: We adopt a procedure similar to that in [34] to extract mid-level elements for image representation. Our method, however, unlike [34], does not require prior knowledge of the image categories. For each image I in the dataset, we first extract a set of patches from multiple spatial scales and compute the CNN activation for each patch. Among all activations (e.g., 4096-D vectors on FC7), only indexes of top k maximal activations are recorded and used to form a transaction (e.g., $\{1024, 3, 24, 4096\}$, $k = 4$) [34]. Each image contains a set of transactions, which appears on the image. Instead of retrieving patches in a class-specific fashion ([34] with known labels), we employed association rule mining inside the sets of either randomly grouped images (for the first iteration) or image clusters computed by “clustering on CNN features”. The top 50 mined patterns (which cover the maximum numbers of patches) per image cluster are further merged across the entire dataset to form a consolidated vocabulary of visual elements. Detailed global merging procedures are elaborated in Algorithm 1. Compared to [34], we find that our global merging strategy effectively reduces redundancy and offers more discriminative image features for both clustering and classification tasks (see details in Sec. 4). Finally, the “bag-of-elements” image representations are computed as the same process in [34].

Algorithm 1: Global merging of patch clusters

Input: A set of mined patterns from each of K image clusters, i.e. $\mathcal{V} = \{v_i\}$, $|\mathcal{V}|=50*K$, set of patches $p_i \in \mathcal{P}$ for each pattern and LDA detectors [23] $d_i \in \mathcal{D}$ trained on associated patch set p_i .
Output: A set of merged patterns $\mathcal{V}' = \{v_n\}$ and associated patch LDA detectors $\mathcal{D}' = \{d_n\}$
for each set of $\{v_i, p_i, d_i\}$ **do**
 Compute $S_{ij} = 1/|p_j| \sum_{x \in X_{p_j}} d_i^T x$ (X_{p_j} is a set of CNN activations of patches in p_j).
end
if both S_{ij} and $S_{ji} > a$ predefined threshold **then**
 Merge $\{v_i, p_i, d_i\}$ and $\{v_j, p_j, d_j\}$ and train new LDA detector d_n based on $p_i \cup p_j$.
end
return $\mathcal{V}', \mathcal{D}'$;

3.2 Image Clustering and LDPO Convergence

Image clustering plays an indispensable role in our LDPO framework. We hypothesize that the newly generated clusters driven by looped deep pseudo-task optimization have incrementally improved quality than previous ones, in the following measurements: 1) Images in each cluster are visually more coherent and discriminative from instances in other clusters; 2) The image counts among all clusters are approximately balanced; 3) The number of clusters is self-adaptive by model selection. Two clustering methods are exploited, i.e., standalone k -means; or an over-segmented k -means (where K is much larger than the first setting, e.g., 1000) followed by RIM [21] for model selection and parameter optimization.

k -means is an efficient clustering algorithm provided that the number of clusters is known. For scene recognition application, we use k -means clustering to initialize the patch mining procedure and generate new image labels for the next iteration, while the underlying cluster number is unknown for the medical image categorization problem. Therefore we first utilize k -means clustering to initialize the RIM clustering with a considerably large k ; then RIM will perform model selection to optimize on k . RIM works without the assumption that the cluster number is known as a priori and is designed for discriminative clustering, by maximizing the mutual information between data distribution and the resulted categories via a regularization term on model complexity. The objective function is defined as

$$f(\mathbf{W}; \mathbf{F}, \lambda) = I_{\mathbf{W}}\{c; \mathbf{f}\} - R(\mathbf{W}; \lambda), \quad (1)$$

where $c \in \{1, \dots, K\}$ is a category label, \mathbf{F} is the set of image features $\mathbf{f}_i = (f_{i1}, \dots, f_{iD})^T \in \mathbb{R}^D$. $I_{\mathbf{W}}\{c; \mathbf{f}\}$ is an estimation of the mutual information between the feature vector \mathbf{f} and the label c under the conditional model $p(c|\mathbf{f}, \mathbf{W})$. $R(\mathbf{W}; \lambda)$ is the complexity penalty and specified according to $p(c|\mathbf{f}, \mathbf{W})$. We adopt the unsupervised multilogit regression cost as [21]. The conditional model

and the regularization term are subsequently defined as

$$p(c = k|\mathbf{f}, \mathbf{W}) \propto \exp(w_k^T \mathbf{f} + b_k) \quad (2)$$

$$R(\mathbf{W}; \lambda) = \lambda \sum_k w_k^T w_k, \quad (3)$$

where $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_K, b_1, \dots, b_K\}$ is the set of parameters and $\mathbf{w}_k \in \mathbb{R}^D, b_k \in \mathbb{R}$. Maximizing the above objective function is equivalent to solving a logistic regression problem. R is the L_2 regulator of weight $\{w_k\}$ and its power is controlled by λ . Large λ values enforce reduction of the total number of categories or clusters by imposing no penalty on unpopulated categories [21]. This characteristic enables RIM to attain the optimal number of categories coherent with the data distribution. λ is fixed to 1 in all our experiments.

Before using the newly-generated clustering labels of image to fine-tune the deep CNN model in the next iteration, the LDPO framework is designed to evaluate the current clustering quality to decide if a convergence has been reached. Two convergence measurements have been adopted from [56], i.e., Purity and Normalized Mutual Information (NMI). We take these two criteria as the forms of empirical similarity examination between two clustering outcomes from adjacent LDPO iterations. When the similarity measure is above a certain threshold, we consider that the optimal clustering-based data categorization is reached. It has been empirically found that the final category numbers (from the RIM process) in later LDPO iterations stabilize around a constant. The convergence on classification plots is also observable through the increased top-1, top-5 classification accuracy values in the first few LDPO rounds and eventually stabilize around a constant.

NLP Text Processing: The category discovery of medical images entails clinically-semantic labeling of the medical images. From the optimized clusters (obtained after Sec. 3.2), we collect the associated text reports and assemble each image cluster's text reports together into a group. Next, NLP is performed on each unit of radiology reports to find highly recurring words that may serve as informative key words per cluster by counting and ranking the frequency of each word. Common words to all clusters are first removed from the list. The resulting key words and randomly sampled exemplary images for each cluster or category are compiled for reviewing by board-certified radiologists. This process shares some analogy to the human-machine collaborative image database construction [65, 58].

3.3 Hierarchical Category Relationship

ImageNet [12] is constructed according to the WordNet ontology hierarchy [40]. In this work, our converged CNN classification model can be further extended to explore the hierarchical class relationship in a tree representation. First, the pairwise class similarity or affinity score $A_{i,j}$ between classes (i,j) is modeled via an adapted measurement of CNN

classification confusion.

$$A_{i,j} = \frac{1}{2} \left(\text{Prob}(i|j) + \text{Prob}(j|i) \right) \quad (4)$$

$$= \frac{1}{2} \left(\frac{\sum_{I_m \in C_i} \text{CNN}(I_m|j)}{|C_i|} + \frac{\sum_{I_n \in C_j} \text{CNN}(I_n|i)}{|C_j|} \right) \quad (5)$$

where C_i, C_j are the image sets for class i, j respectively, $|\cdot|$ is the cardinality function, $\text{CNN}(I_m|j)$ is the CNN classification score of image I_m (from class C_i) according to class j that is directly obtained by the N-way CNN softmax. Here $A_{i,j} = A_{j,i}$ is symmetric by averaging $\text{Prob}(i|j)$ and $\text{Prob}(j|i)$.

The Affinity Propagation algorithm [19] (AP) is then invoked to perform a “tuning parameter-free” partitioning on this pairwise affinity matrix $\{A_{i,j}\} \in \mathbb{R}^{K \times K}$. This process can be executed recursively to generate a hierarchically-merged category tree. Without loss of generality, we assume that at level L, classes i^L, j^L are formed by merging classes at level L-1 through AP clustering. The new affinity score A_{i^L, j^L} is computed as follows.

$$A_{i^L, j^L} = \frac{1}{2} \left(\text{Prob}(i^L|j^L) + \text{Prob}(j^L|i^L) \right) \quad (6)$$

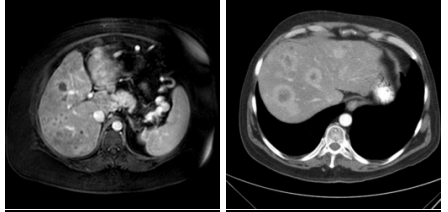
$$\text{Prob}(i^L|j^L) = \frac{\sum_{I_m \in C_{i^L}} \sum_{k \in j^L} \text{CNN}(I_m|k)}{|C_{i^L}|} \quad (7)$$

where the L-th level class label j^L includes all merged original classes (i.e., 0-th level before AP is called) $k \in j^L$ obtained thus far. N-way CNN classification scores only need to be evaluated once at the beginning of AP. The consequent value of any A_{i^L, j^L} at any merged level is the sum of the 0-th level confusion scores. The modeled category hierarchy can alleviate the highly uneven visual separability among discovered image categories [63].

4 Experimental Results

Datasets: We experiment on the same medical image dataset as in [49] that contains totally 215,786 2D key-images and the associated radiology reports of 61,845 unique patients. Key-images are resized to 256×256 bitmap images (from 512×512). The intensity ranges are rescaled using the default “optimal” window settings stored in the DICOM header files (Intensity rescaling improves the CNN classification accuracy by $\sim 2\%$ comparing to [49]). Patient-sensitive information in radiology reports is removed for privacy reasons. Furthermore, we quantitatively evaluate our LDPO framework on three widely-reported scene recognition benchmark datasets: 1). I-67 [44] of 67 indoor scene classes with 15620 images; 2). B-25 [62] of 25 architectural styles from 4794 images; 3). S-15 [32] of 15 outdoor and indoor mixed scene classes with 4485 images. For scene recognition, the ground truth (GT) labels are only used to validate the final quantitative LDPO clustering results (where cluster-purity becomes classification accuracy). The cluster number is assumed to be known to

Cluster #23	
Word	Frequency
liver	524
abdomen	337
enhancement	217
mass	198
lesion	168
lobe	161
adenopathy	119
lesions	109
segment	58
bulky	45



Cluster #64	
Word	Frequency
enhancement	277
cerebellar	193
lesion	192
lobe	186
flair	173
hemisphere	155
mass	134
abnormal	119
frontal	115
cerebellum	113

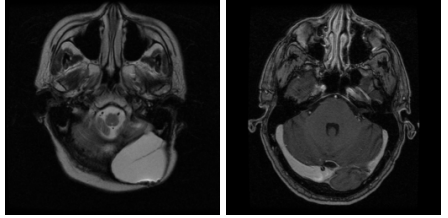


Figure 2. Sample images of two unsupervisedly discovered image clusters with associated clinically semantic key words, containing (likely appeared) anatomies, pathologies, their attributes and imaging protocols or properties.

Table 2. Classification Accuracy of Converged CNN Models

CNN setting	Cluster #	Top-1	Top-5
AlexNet-FC7-Topic	270	0.8109	0.9412
AlexNet-FC7-ImageNet	275	0.8099	0.9547
AlexNet-Conv5-FV	712	0.4115	0.4789
AlexNet-Conv5-VLAD	624	0.4333	0.5232
GoogLeNet-Pool5	462	0.4109	0.5609
GoogLeNet-Inc.5b-VLAD	929	0.3265	0.4001

LDPO during clustering (Sec. 3.2) for a fair comparison. Thus the model selection RIM module is dropped.

In each LDPO round, 1) the image clustering step (Sec. 3.2) is applied on the entire image dataset in order to assign a cluster label to each image, 2) for CNN model fine-tuning (Sec. 3.1), images are randomly reshuffled into three subsets of training (70%), validation (10%) and testing (20%) at each iteration. This ensures that LDPO convergence will generalize to the entire image database. The CNN model is fine-tuned at each LDPO iteration once a new set of image labels is generated from the clustering stage. We use Caffe [26] implementation of CNN models. The softmax loss layer (i.e., ‘FC8’ in AlexNet and ‘loss3/classifier’ in GoogLeNet) is more significantly modulated by 1) setting a higher learning rate than all other CNN layers; and 2) updating the (varying but converging) number of category classes from the clustering results.

4.1 Unsupervised Medical Image Categorization

We first investigate the convergence issue of the LDPO method under different system configurations and then report the CNN classification performance on the discovered categories.

Clustering Method: As shown in Fig. 3 (a), RIM can estimate unsupervised category numbers consistently well under different image representations (deep CNN feature configurations + encoding schemes). Standalone k -

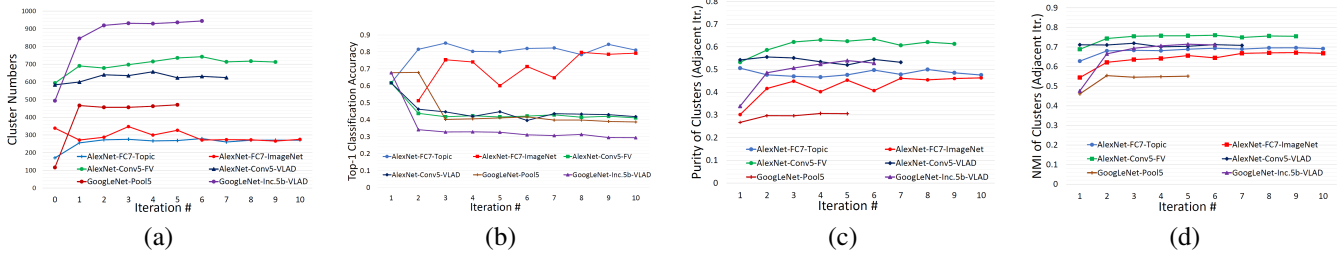


Figure 3. LDPO Performance using RIM clustering under different image encoding methods (i.e., FV and VLAD) and CNN Architectures (i.e., AlexNet and GoogLeNet). The number of clusters discovered, Top-1 accuracy of trained CNNs, the purity and NMI measurements of clusters from adjacent iterations are illustrated in (a, b, c, d), respectively.

means clustering enables LDPO to converge quickly with high classification accuracies whereas RIM based model selection module produces more balanced and semantically meaningful clustering results (see more in Sec. 4.1). This advantage is probably due to RIM’s two unique properties: 1) less restricted geometric assumptions in the clustering feature space; 2) the capacity to attain the optimal number of clusters by maximizing the mutual information between input data and the induced clusters via a regularized term.

Pseudo-Task Initialization: Both generic and domain-specific CNN models [30, 55, 49] are employed for LDPO initialization. Fig. 3 illustrates the performance of LDPO using two CNN variants – AlexNet-FC7-ImageNet and AlexNet-FC7-Topic. AlexNet-FC7-ImageNet yields noticeably slower LDPO convergence than its counterpart of AlexNet-FC7-Topic, as the latter has already been fine-tuned by the report-derived category information on the same radiology image database [49]. Nevertheless, the final clustering outcomes are similar after convergence from AlexNet-FC7-ImageNet or AlexNet-FC7-Topic. At ~ 10 iterations, two different initializations result in similar cluster numbers, purity/NMI scores and even classification accuracies (Table 2).

Deep CNN Feature and Image Encoding: Different configurations of image representation can affect the performance of medical image categorization, as shown in Fig. 3. Deep images features are extracted at different layers of depth from two CNN models (i.e., AlexNet, GoogLeNet) and may present the depth-specific visual information. Different image feature encoding schemes (FV or VLAD) add further options or variations into this process. The numbers of clusters range from 270 (AlexNet-FC7-Topic with no explicit feature encoding scheme) to 931 (the more sophisticated GoogLeNet-Inc.5b-VLAD with VLAD encoding). The numbers of clusters discovered by RIM are expected to reflect the amount of knowledge or information complexity stored in the PACS database.

Unsupervised Categorization: Our category discovery clusters are generally visually coherent within the cluster and size-balanced across clusters. However, image clusters formed only based on text information (of $\sim 780K$ radiol-

ogy reports) are highly unbalanced [49], with three clusters inhabiting the majority of images. Note that our method imposes no explicit constraint on the number of instances per cluster. Fig. 2 shows sample images and their top-10 associated key words from two randomly selected clusters (more results are provided in the supplementary material). The LDPO clusters are found to be clinically or semantically related to the corresponding key words, which describe presented anatomies, pathologies (e.g., adenopathy, mass), their associated attributes (e.g., bulky, frontal) and imaging protocols or properties.

Categorization Recognizable? We validate the following hypothesis: a high quality unsupervised image categorization scheme will generate labels that can be more easily recognized by any supervised CNN model. From Table 2, AlexNet-FC7-Topic has the **Top-1** classification accuracy of 0.8109 and **Top-5** accuracy 0.9412 with 270 formed image categories while AlexNet-FC7-ImageNet achieves the accuracies of 0.8099 and 0.9547, from 275 discovered classes. In contrast, [49] reports the **Top-1** accuracies of 0.6072, 0.6582 and **Top-5** as 0.9294, 0.9460 from only 80 classes using AlexNet [30] or VGGNet-19 [52], respectively. The classification accuracies shown in Table 2 are computed using the final LDPO-converged CNN models and the testing dataset. Markedly better accuracies (especially on **Top-1**) on classifying higher numbers of classes (that are generally more challenging) also demonstrate the advantages of the LDPO discovered image clusters or labels over those in [49], under the same radiology database. Upon evaluation by two board-certified radiologists, AlexNet-FC7-Topic of 270 categories and AlexNet-FC7-ImageNet of 275 classes are considered the best of total six model-feature-encoding setups. Interestingly, both models have no external feature encoding schemes built-in and preserve gloss image layouts (without spatially unordered FV or VLAD encoding modules [8, 25]). Refer to supplementary material for more results on radiologists’ evaluation.

4.2 Scene Recognition

We use three scene recognition datasets to quantitatively evaluate the proposed LDPO-PM method (with patch min-

Table 3. Clustering performance of LOM and other methods on 3 scene recognition datasets. The last Column presents the state-of-the-art fully-supervised scene Classification Accuracy (CA) for each dataset, produced by [8, 42, 66] respectively.

Dataset	KM [57]	LSC [4]	AC [22]	EP [10]	MDPM [34]	LDPO-A-FC	LDPO-A-PM	LDPO-V-PM	Supervised
Clustering Accuracy (%)									CA (%)
I-67 [44]	35.6	30.3	34.6	37.2	53.0	37.9	63.2	75.3	81.0 [8]
B-25 [62]	42.2	42.6	43.2	43.8	43.1	44.2	59.2	59.5	59.1 [42]
S-15 [32]	65.0	76.5	65.2	73.6	63.4	73.1	90.2	84.0	91.6 [66]
Normalized Mutual Information									
I-67 [44]	.386	.335	.359	-	.558	.389	.621	.759	-
B-25 [62]	.401	.403	.404	-	.424	.407	.588	.546	-
S-15 [32]	.659	.625	.653	-	.596	.705	.861	.831	-

ing) based on two metrics: 1) clustering based scene recognition accuracy and 2) supervised classification (e.g., Liblinear) on image representations learned in an unsupervised fashion. The purity and NMI measurements are computed between the final LDPO clusters and GT scene classes where **purity becomes the classification accuracy against GT**. The LDPO cluster numbers are set to match the GT class numbers of (67, 25, 15), respectively. We compare the LDPO scene recognition performance to those of several popular clustering methods, such as **KM** [57]: *k*-means; **LSC** [4]; **AC** [22]: Agglomerative clustering; **EP** [10]: Ensemble Projection + kmeans; and **MDPM** [34]: Mid-level Discriminative Patch Mining + kmeans. Both EP and MDPM use mid-level visual elements based image representations. Three variants of our method (i.e., **LDPO-A-FC7**: FC7 feature on AlexNet, **LDPO-A-PM**: FC7 feature on AlexNet with patch mining, and **LDPO-V-PM**: FC7 feature on VGG-VD with patch mining) are exploited. On all three datasets, the **LDPO-A-PM** and **LDPO-V-PM** achieve significantly higher purity and NMI values than the previous clustering methods (cf. Table 3). Especially for the MIT-67 indoor scene dataset [44], our best model **LDPO-V-PM** achieves the unsupervised scene recognition accuracy of 75.3%, which nearly doubles the performances of KM and AC on FC7 features of an ImageNet pretrained AlexNet [30, 3]. Note that the state-of-the-art supervised classification accuracy on MIT-67 is 81.0% [8] and our unsupervised method is comparatively close to that. VGG-VD – a deeper CNN model – empirically boosts the recognition performance from **LDPO-A-PM** of 63.2% to **LDPO-V-PM** at 75.3% on MIT-67. However this performance gain is not observed on two other smaller datasets.

Next, we evaluate the supervised discriminative power of LDPO-PM learned image representation. We measure its classification accuracy using the MIT-67 dataset and its standard partition [44], i.e., 80 training and 20 testing images per class. As in [53, 54, 13, 34, 60, 8], we use the Liblinear classification toolbox [17] on the **LDPO-V-PM** image representation (noted as **LDPO-V-PM-LL**), under 5-fold cross validation. The supervised and unsupervised scene recognition accuracy results from previous state-of-the-art work and variants of our method are listed in Ta-

Table 4. Scene Recognition Accuracy on MIT-67 Dataset [44]

Method	Accuracy (%)
D-patch [53]	38.1
D-parts [54]	51.4
DMS [13]	64.0
MDPM-VGG [34]	77.0
MetaObject [60]	78.9
FC (VGG)	68.9
CONV-FV (VGG) [8]	81.0
LDPO-V-PM-LL	72.5
LDPO-V-PM	75.3

ble 4. The one-versus-all Liblinear classification in **LDPO-V-PM-LL** does not noticeably improve upon purely unsupervised **LDPO-V-PM**. This may indicate that the LDPO-PM image representation is sufficient to adequately separate images from different scene classes. Last, we experiment the clustering convergence issue with two different initializations: random initialization or image labels obtained from k-means clustering on FC7 features of an ImageNet pretrained AlexNet. While the clustering accuracy of the LDPO-PM with random initialization increases rapidly during its first iterations, both schemes ultimately converge to similar performance levels. This suggests that the LDPO convergence is insensitive to the chosen initialization.

5 Conclusion

In this paper, we present a Looped Deep Pseudo-task Optimization framework for unsupervised joint mining of image features and labels. Our method is validated through two important problems: 1) discovery and exploration of semantic categories from a large-scale medical image database and 2) unsupervised scene cognition on three public datasets. Extensive experiments demonstrate excellent quantitative and qualitative results on both tasks. The measurable LDPO “convergence” makes the ill-posed image auto-annotation problem better constrained.

Acknowledgements This work was supported by the Intramural Research Program of the NIH Clinical Center. This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>).

References

- [1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.
- [2] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101—mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer, 2014.
- [3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.
- [4] X. Chen and D. Cai. Large scale spectral clustering with landmark-based representation. In *AAAI*, 2011.
- [5] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *Proc. of ICCV*, 2015.
- [6] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Proc. of ICCV*, 2013.
- [7] J. Choi, M. Rastegari, A. Farhadi, and L. S. Davis. Adding unlabeled samples to categories by learned attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 875–882, 2013.
- [8] M. Cimpoi, S. Maji, and A. Vedaldi. Deep filter banks for texture recognition and segmentation. *Proc. of IEEE CVPR*, 2015.
- [9] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. *AI and Statistics*, 2011.
- [10] D. Dai and L. Van Gool. Ensemble projection for semi-supervised image classification. In *Proc. of ICCV*, 2013.
- [11] D. Dai and L. Van Gool. Unsupervised high-level feature learning by ensemble projection for semi-supervised image classification and image clustering. Technical report, arXiv:1602.00955, 2016.
- [12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.
- [13] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *Advances in Neural Information Processing Systems (NIPS)*, pages 494–502, 2013.
- [14] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.
- [15] A. Dosovitskiy, J. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. *NIPS*, 2014.
- [16] M. Everingham, S. M. A. Eslami, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.
- [17] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.
- [18] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. *Proc. of IEEE CVPR workshop*, 2004.
- [19] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.
- [20] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015.
- [21] R. Gomes, A. Krause, and P. Perona. Discriminative clustering by regularized information maximization. *NIPS*, 2010.
- [22] K. C. Gowda and G. Krishna. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern recognition*, 10(2):105–112, 1978.
- [23] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *European Conference on Computer Vision*, pages 459–472. Springer, 2012.
- [24] M. Huh, P. Agrawal, and A. A. Efros. What makes imagenet good for transfer learning? In *arXiv preprint: arXiv:1608.08614*, 2016.
- [25] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1704–1716, Sept 2012.
- [26] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
- [27] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. *CVPR*, pages 923–930, 2013.
- [28] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *Proc. of IEEE CVPR*, pages 3128–3137, 2015.
- [29] D. Kingma, S. Mohamed, D. Rezende, and M. Welling. Semi-supervised learning with deep generative models. *NIPS*, 2014.
- [30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [31] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1265–1278, 2005.
- [32] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 2, pages 2169–2178. IEEE, 2006.
- [33] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 851–858, 2013.
- [34] Y. Li, L. Liu, C. Shen, and A. van den Hengel. Mid-level deep pattern mining. In *CVPR*, pages 971–980, 2015.

- [35] Y. Li and Z. Zhou. Towards making unlabeled data never hurt. *ICML*, 2011.
- [36] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan. Computational baby learning. In *Proc. of ICCV*, 2015.
- [37] M. Lin, Q. Chen, and S. Yan. Network in network. In *Proc. of ICLR*, 2015.
- [38] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [39] W. Liu, J. He, and S. Chang. Large graph construction for scalable semi-supervised learning. *ICML*, 2010.
- [40] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.
- [41] J. Y. Ng, F. Yang, and L. S. Davis. Exploiting local features from deep networks for image retrieval. *CoRR*, abs/1504.05133, 2015.
- [42] K.-C. Peng and T. Chen. A framework of extracting multi-scale features using multiple convolutional neural networks. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2015.
- [43] F. Perronnin, J. Snchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Computer Vision ECCV 2010*, volume 6314 of *Lecture Notes in Computer Science*, pages 143–156. Springer Berlin Heidelberg, 2010.
- [44] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 413–420. IEEE, 2009.
- [45] A. Quattoni and A. Torralba. Recognizing indoor scenes. *Proc. of IEEE CVPR*, 2009.
- [46] R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng. Self-taught learning: transfer learning from unlabeled data. *ICML*, 2007.
- [47] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *ArXiv:1403.6382*, 2014.
- [48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014.
- [49] H. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. Summers. Interleaved text/image deep mining on a large-scale radiology database. *Proc. of IEEE CVPR*, 2015.
- [50] H. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. Summers. Interleaved text/image deep mining on a large-scale radiology image database for automated image interpretation. *Journal of Machine Learning Research*, pages 17(107): 1–31, 2016.
- [51] A. Shrivastava, S. Singh, and A. Gupta. Constrained semi-supervised learning using attributes and comparative attributes. In *European Conference on Computer Vision*, pages 369–383. Springer, 2012.
- [52] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learn. Repr.*, 2015.
- [53] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *European Conference on Computer Vision*, 2012.
- [54] J. Sun and J. Ponce. Learning discriminative part detectors for image classification and cosegmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3400–3407, 2013.
- [55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *IEEE Conf. on Computer Vision and Pattern Recognition*, arXiv:1409.4842, 2015.
- [56] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine. Unsupervised object discovery: A comparison. *International Journal of Computer Vision*, 2009.
- [57] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.
- [58] M. Wigness, B. Draper, and J. Beveridge. Efficient label collection for unlabeled image datasets. *Proc. of IEEE CVPR*, 2015.
- [59] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep multiple instance learning for image classification and auto-annotation. *Proc. of CVPR*, pages 3460–3469, 2015.
- [60] R. Wu, B. Wang, and Y. Yu. Harvesting discriminative meta objects with deep cnn features for scene classification. In *Proc. of ICCV*, 2015.
- [61] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.
- [62] Z. Xu, D. Tao, Y. Zhang, J. Wu, and A. C. Tsoi. Architectural style classification using multinomial latent logistic regression. In *European Conference on Computer Vision*, pages 600–615. Springer, 2014.
- [63] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu. Hd-cnn: Hierarchical deep convolutional neural network for large scale visual recognition. *Proc. of ICCV*, 2015.
- [64] J. Yang, D. Parikh, and D. Batra. Joint unsupervised learning of deep representations and image clusters. *arXiv preprint arXiv:1604.03628*, 2016.
- [65] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv:1506.03365*, 2015.
- [66] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014.