SIP (2017), vol. 6, e11, page 1 of 13 © The Authors, 2017.

doi:10.1017/ATSIP.2017.12

INDUSTRIAL TECHNOLOGY ADVANCES

Advances in deep learning approaches for image tagging

JIANLONG FU AND YONG RUI

The advent of mobile devices and media cloud services has led to the unprecedented growth of personal photo collections. One of the fundamental problems in managing the increasing number of photos is automatic image tagging. Image tagging is the task of assigning human-friendly tags to an image so that the semantic tags can better reflect the content of the image and therefore can help users better access that image. The quality of image tagging depends on the quality of concept modeling which builds a mapping from concepts to visual images. While significant progresses are made in the past decade on image tagging, the previous approaches can only achieve limited success due to the limited concept representation ability from hand-crafted features (e.g., Scale-Invariant Feature Transform, GIST, Histogram of Oriented Gradients, etc.). Further progresses are made, since the efficient and effective deep learning algorithms have been developed. The purpose of this paper is to categorize and evaluate different image tagging approaches based on deep learning techniques. We also discuss the relevant problems and applications to image tagging, including data collection, evaluation metrics, and existing commercial systems. We conclude the advantages of different image tagging paradigms and propose several promising research directions for future works.

Keywords: Image tagging, Deep learning

Received 30 November 2016; Revised 24 August 2017

I. INTRODUCTION

Recent years have witnessed the emergence of mobile devices (e.g., smart phones, digital cameras, etc.) and cloud storage services, which has led to an unprecedented growth in the number of personal media resources, such as photos and videos. For example, people take photos using their smart devices every day and everywhere. It is reported that Flickr[1] has 1.7 million photos uploaded every day and Instagram[2] claims 40 million photos per day in 2015. Such a great number of images demand effective image-accessing techniques. As evidenced by the success of commercial search engines (e.g., Google[3], Bing[4]), one of the most effective ways for common users to access both web images and personal photos is through text. Therefore, image tagging has become an active research topic in the recent few years, which aims to label an image with human-friendly *concepts*[5].

Microsoft Research, No. 5, Dan Ling Street, Haidian District, Beijing, P. R. China

Corresponding author:
J. Fu
Email: jianf@microsoft.com

[1] https://www.flickr.com
[2] http://instagram.com
[3] http://www.google.com
[4] http://www.bing.com
[5] "concept" and "tag" are considered as interchangeable terms, and we do not differentiate them in this paper.

In general, the task of effective image tagging typically consists of two stages, which involve initial image tagging and subsequent tag refinement. Before diving into the details of various image tagging/tag refinement techniques, we will first define some key terminologies used throughout the paper.

- *Image tagging* attempts to label an image with one or more human-friendly textual concepts to reflect the visual content of the image [1, 2]. The resultant tags constitute the tag list for this image. Note that image tagging can be done manually by a human, or automatically by an algorithm [3], or by combining the both [4]. However, initial tag lists (e.g., the middle tag list in Fig. 1) are often imperfect so that a post process is usually needed to refine the result.
- *Image tag refinement* aims to remove imprecise tags and supplement incomplete tags, since the tags in a tag list may be imprecise for that image, and some relevant tags may be missing from the tag list. If we refer to Fig. 1, the input to image tag refinement is an image with its initial image tagging list, and the output is the refined tag list. Note that the output tag list in the right is more refined than the initial tag list in the middle.

While image tagging and tag refinement are the two key technologies to help users access images, how to do it accurately is not easy. Most previous image tagging/tag refinement approaches depend on hand-crafted features,

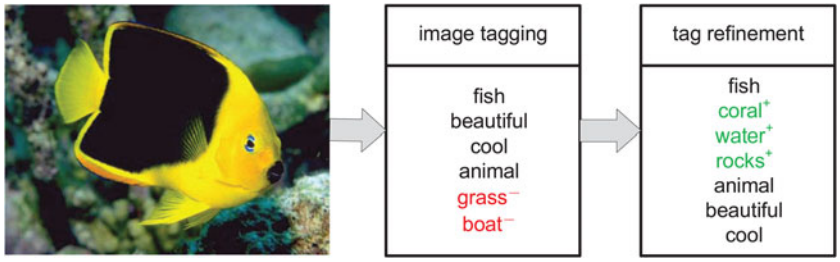

Fig. 1. Examples of image tagging and tag refinement results. A red tag with a "−" superscript indicates the imprecise tags, which should be removed from initial image tagging results, and a green tag with a "+" superscript indicates the enriched tags by image tag refinement approaches. All the tags are ranked according to relevance scores to the image.

e.g., Scale-Invariant Feature Transform (SIFT) [5], GIST [6], Histogram of Oriented Gradients (HOG) [7], and so on. Based on these low-level feature descriptors, visual representation algorithms (e.g., bag-of-word features [8] or spatial pyramid features [9]) have been proposed to describe image content and associate the content with natural language-based keywords. However, these hand-crafted feature descriptors are designed to capture low-level visual patterns by pre-defined feature types (e.g., color, shape, or texture). Although promising results have been achieved by combining multi-type feature representations [10, 11], these features are still inadequate to detect and describe high-level semantic concepts (e.g., the "coral" and "rocks" in Fig. 1).

With the recent success in many research areas, deep learning techniques have attracted great attention [12]. There are two main paradigms in deep learning research, i.e., supervised learning and unsupervised learning. The former paradigm can automatically learn hierarchical deep networks from raw pixels for pattern analysis and image classification. For example, convolutional neural networks (CNNs) have achieved a winning top-5 test error rate of 15.3%, compared with the 26.2% achieved by the second-best approach, which combines scores from many classifiers trained by a set of hand-crafted features [13]. Recently, by using the promising deep residual nets, the top-5 image classification error has been reduced to 3.57% on ImageNet[6] test set, which has achieved the superior performance even than humans. Another breakthrough has been achieved by the second paradigm (i.e., unsupervised learning), in which algorithms can automatically learn the concept representations, such as cat faces and human bodies from unlabeled data by unsupervised methodologies [14]. Besides, unsupervised learning has been demonstrated its superior ability on deep learning model pre-training, which can help supervised methods in achieving a better local minimum and fast convergence [13].

In this paper, we survey previous works on image tagging along two major dimensions in the literature, i.e., model-based and model-free approaches, according to whether tag an image by tagging model training or instance-based tag neighbor voting. We mainly focus on the works, which leverage the state-of-the-art deep learning technologies. The aims of this paper are twofold. First, we summarize related works from literature to investigate the research development along the above two dimensions and compare the advantages and limitations in theory. Second, we summarize a comprehensive experimental protocol, including widely-used experimental datasets for image tagging and well-acknowledged evaluation measurements for showing the superior image tagging results achieved by deep learning techniques.

Note that although typical image classification networks by deep learning techniques, which are trained on corresponding image classification datasets, can be directly adopted to image tagging tasks, the results are far from the requirements in real-world applications due to the limited vocabularies [15, 16]. For example, CIFAR-10 and CIFAR-100 datasets [13] provide 10 and 100 categories, which are still difficult to describe the various image content from real scenarios. Although significant progresses have been made by introducing ImageNet dataset, some studies have shown that the vocabularies from ImageNet are so professional and fine-grained that are not very suitable for image tagging tasks on user photos [17]. Therefore, one of the major challenges for image tagging is to acquire sufficient and high-quality training data for a large vocabulary, which is often too expensive to obtain by human labors [18]. With the success of commercial image search engines, a large number of works have been proposed to learn image tagging models from web images (e.g., user-contributed photos from social network or noisy web images from commercial search engines) [10, 19–21]. In this paper, we will conduct detailed introduction and summarization on these approaches. The survey of image classification on carefully labeled data can be found in [22].

The rest of this paper is organized as follows. Section II describes brief introduction of deep learning techniques. Section III introduces the methodology on deep learning-based image tagging. Section IV provides comprehensive datasets and performance metrics. Section V further introduces the selected typical approaches and detailed evaluation results. Finally, Section VI presents several practical commercial image tagging systems on both cloud and client, followed by the conclusion and future challenges in Section VII.

[6]ImageNet is an image database organized according to the noun hierarchy from WordNet, in which each node of the hierarchy is depicted by hundreds or even thousands of images. Details can be found in http://www.image-net.org

https://doi.org/10.1017/ATSIP.2017.12

II. BRIEF OVERVIEW ON DEEP LEARNING TECHNIQUES

Most shallow-structured architectures in machine learning have achieved promising results in existing research. These architectures generally consist of a single layer of nonlinear feature transformation. Examples of the shallow-structured architectures include Gaussian Mixture Models (GMM) [23], Hidden Markov Models (HMM) [24], Conditional Random Fields (CRF) [25], Maximum Entropy Models (MEM) [26], Support Vector Machine (SVM) [27], Logistic Regression (LR) [28], Multi-layer Perceptron (MLP) [29], and so on. Although shallow architectures have shown effective performance in simple or well-constrained scenarios, their limited modeling and representation capabilities have difficulties in dealing with the complicated real-world applications, such as human speech recognition, natural image, and scene understanding. Significant progresses have been made after efficient and effective deep learning algorithms are developed.

A) Categorization on deep neural networks

Deep leaning techniques refer to a class of machine learning techniques, where many layers of information processing stages in hierarchial architectures are exploited for unsupervised feature learning and for supervised pattern analysis/classification. The core procedure of deep learning is to compute hierarchical features or representations from the observed data, where the higher-level features or factors are defined from lower-level data structure. Deep learning techniques can be generally divided into three dimensions, i.e., generative deep architectures, discriminative deep architectures and hybrid deep architectures [30]. Typical architectures include Deep Belief Network (DBN) [31], Boltzmann Machine (BM) [32], Restricted Boltzmann Machine (RBM) [33], Deep Auto-Encoder (DAE) [34], CNN [13], and so on. Among these architectures, CNNs have shown superior performance on the learning of discriminative image feature representation, and thus are widely-used in image classification and annotation tasks. In the following, we will specify the network structure of a typical CNN.

B) Convolutional neural networks

A typical CNN often consists of several convolutional and fully-connected layers. The exact number of layers generally depends on the requirement of network capacity and memory cost for a specific classification task (e.g., eight layers for AlexNet, and 22 layers for GoogleNet). In particular, let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ be the matrix of image training data, where $\mathbf{x}_i$ is the feature vector of the ith image. N is the total number of images. Denote $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T \in \{0,1\}^{N\times K}$, where $\mathbf{y}_i \in \{0,1\}^{K\times 1}$ is the category indicator vector for $\mathbf{x}_i$. K is the number of categories. Suppose there are M layers in total and $\mathbf{W} = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(M)}\}$ are the model parameters. In each layer, we absorb the bias term into the weights and denote them as a whole. $\mathbf{W}^{(m)} = [\mathbf{w}_1^{(m)}, \ldots, \mathbf{w}_{d_m}^{(m)}]^T \in R^{d_m \times d_{m-1}}$, where $\mathbf{w}_i^{(m)} \in R^{d_{m-1}}$, d_{m-1} is the dimension of the $(m-1)$th feature map. $\mathbf{Z}^{(m)}(\mathbf{X}) = [\mathbf{z}^{(m)}(\mathbf{x}_i), \ldots, \mathbf{z}^{(m)}(\mathbf{x}_N)]^T \in R^{N\times d_m}$ denotes the feature map produced by the mth layer.

Given an image $\mathbf{x}_i$, convolutional layers first task the image as input. The extracted deep representations are denoted as $\mathbf{W} * \mathbf{x}_i$, where $*$ denotes a set of operations of convolution, polling, and activation, and $\mathbf{W}$ denotes the overall parameters for convolutional and pooling operations. The pooling layer in a CNN is designed to summarize the outputs of neighboring groups of neurons in the same kernel map, which can be conducted by mean-pooling or max-pooling. Some earlier works adopt the neighborhood summarization strategy by adjacently pooling without overlaps. To make it more clear, a pooling layer can be considered as consisting of a grid of pooling units spaced s pixels apart. Each pooling unit summarizes a neighborhood of size $r \times r$ centered at the location in a CNN. If $s = r$, we can obtain the non-overlapping polling. If $s < r$, we can obtain overlapping pooling. Extensive studies have shown that the overlapping pooling are difficult to overfit [13].

For the activation function, the standard way to activate a neuron's output is through tan or sigmoid function, which are considered as saturating nonlinear functions. However, for the training optimization with gradient descent, these saturating nonlinear functions show much slower convergence than the non-saturating nonlinear function, such as $\max(0, x)$. The nonlinear function is referred to Rectified Linear Units (ReLUs). It has been demonstrated that deep CNNs with ReLUs can be trained several times faster than their equivalents with tan unites, which is crucial for the training of large models on large datasets.

Given the output from convolutional layers, we further feed it into a series of fully-connected layers. The output of the last fully-connected layer is considered as an input to a softmax classifier, which can generate a distribution over the final category labels, which is given by:

$$\mathbf{p}(\mathbf{x}_i) = f(\mathbf{W} * \mathbf{x}_i), \tag{1}$$

where $f(\cdot)$ represents fully-connected layers to map convolutional features to a feature vector that could be matched with the categories entries, as well as includes a softmax layer to further transform the feature vector to probabilities. The goal is to minimize the following objective function in the form of a softmax regression with weight decay, which is given by:

$$\mathcal{L}(\mathbf{W}) = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{K} \mathbf{1}_{Y_{ij}}(j) \log p(Y_{ij} = 1 | \mathbf{x}_i; \mathbf{W})\right] + \frac{\beta}{2}||\mathbf{W}||_F, \tag{2}$$

where Y_{ij} is the (i, j)th entry of $\mathbf{Y}$. $\mathbf{1}_{Y_{ij}}(j)$ is the indicator function such that $\mathbf{1}_{Y_{ij}}(j) = 1$ if $Y_{ij} = 1$, otherwise zero. β is the coefficient of weight decay, which is designed to reduce model complexity and thus to prohibit overfitting.

https://doi.org/10.1017/ATSIP.2017.12

Table 1. The comparison of different CNN architectures on model size, classification error rate, and model depth.

Model	Size (M)	Top-1/top-5 error (%)	# layers	Model description
AlexNet	238	41.00/18.00	8	5 conv + 3 fc layers
VGG-16	540	28.07/9.33	16	13 conv + 3 fc layers
VGG-19	560	27.30/9.00	19	16 conv + 3 fc layers
GoogleNet	40	29.81/10.04	22	21 conv + 1 fc layers
ResNet-50	100	22.85/6.71	50	49 conv + 1 fc layers
ResNet-152	235	21.43/3.57	152	151 conv + 1 fc layers

"conv" and "fc" indicates convolutional and fully-connected layers, respectively.

The probability p can be calculated by a softmax function, which is shown as:

$$p(Y_{ij} = 1|\mathbf{x}_i; \mathbf{W}) = \frac{\exp(\mathbf{z}_j^{(M-1)})}{\sum_{k=1}^{K} \exp(\mathbf{z}_k^{(M-1)})}. \quad (3)$$

To optimize the above objective function, the derivatives to $\mathbf{w}_j^{(M)}$ in the output layer can be calculated as:

$$\frac{\partial \mathcal{L}(\mathbf{W})}{\partial \mathbf{w}_j^{(M)}} = -\frac{1}{N} \sum_{i=1}^{N} \mathbf{z}^{(M-1)}(\mathbf{x}_i) \Big[\mathbf{1}_{Y_{ij}}(j) - p\left(Y_{ij} = 1|\mathbf{z}^{(M-1)}(\mathbf{x}_i); \mathbf{w}_j^{(M)}\right) \Big] + \beta \mathbf{w}_j^{(M)}. \quad (4)$$

Parameters in other layers can be calculated by the back propagation algorithm (BP) [35]. Different CNN architectures have different numbers of convolutional layers and fully-connected layers. Detailed performance comparisons for different CNN model architectures on ImageNet challenges are shown in Table 1.

III. METHODOLOGY FOR IMAGE TAGGING WITH CNNS

The extensive research on image tagging can be basically divided into model-based and model-free approaches. The model-based approaches heavily rely on pre-trained classifiers with machine learning algorithms [36–39], while the model-free approaches propagate tags through the tagging behavior of visual neighbors [40, 41]. The two streams of approaches both assume that there should be a well-labeled training image database (i.e., source domain) with the same or at least a similar data distribution as testing data (i.e., target domain), so that the labeled database can ensure good generalization abilities for both model training and tag propagation. Besides, there are also emerging works focusing on solving the problems when the training image database are not available, which is called zero-shot learning. Since zero-shot learning-based approaches often adopt model-based methodology for the training of image tagging model, we categorize zero-shot learning into model-based approaches in this paper. We will survey the two types of paradigms in the following sections.

A) Model-free image tagging with deep representation

The state-of-the-art model-free image tagging approaches adopt the powerful CNN features as image representations, which specifically replace traditional hand-crafted features with features that are automatically learned with a CNN. The CNN feature can be either extracted from pre-trained deep neural networks (e.g., GoogleNet [42], VGG [43], or ResNet [44]) on ImageNet, or fine-tuned on the target tag vocabularies from target domain. Specifically, the feature vector in the last fully-connected layer, which is activated by a rectified activation function (e.g., ReLU), usually constitutes the final compact representations (e.g., 4096 dimensions) for an image.

Based on the deep representation, most of works adopt the idea of "collective intelligence," which constitutes the tagging hypotheses by neighbor voting from the training instances. The specific procedure generally refers to tag propagation from visual-similar images. The instance voting-based models are non-parametric and even do not involve any training process. The complexity of this type of approaches depends on the amount of training data instances. A typical image tag refinement framework by instance neighbor voting is shown in Fig. 2. For example,

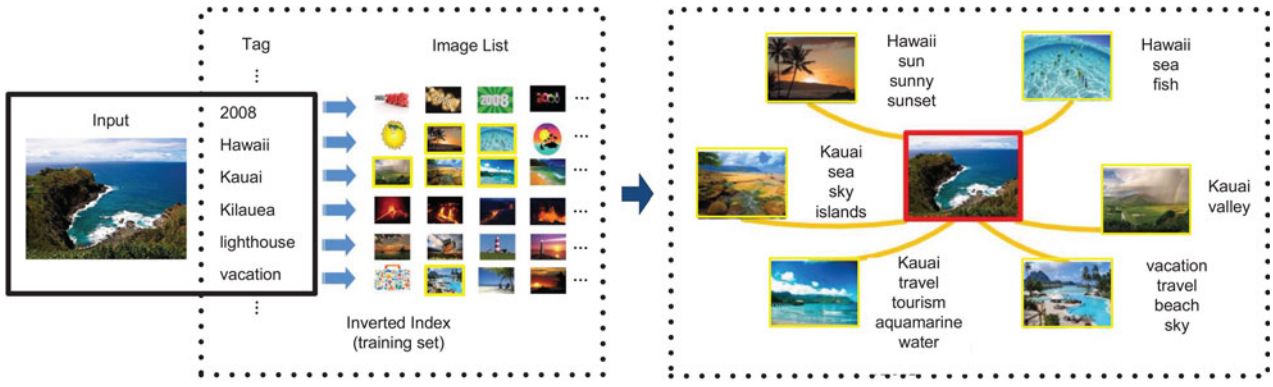

Fig. 2. An example of model-free image tag refinement framework. For an input image on the left, we first find its semantic-related images in training set by searching with its initial image tagging results. Second, we build a star-graph from the semantic-related images based on visual similarity on the right. Both the semantic-related and visual-similar nearest neighbor images for the input image are marked by yellow rectangles in the left image list and the right star graph. The final tagging list for the input image is generated from the voting of those nearest-neighbor images marked by yellow rectangles. [Best viewed in color].

https://doi.org/10.1017/ATSIP.2017.12

Li *et al.* [40, 45], Wang *et al.* [41] and Guillaumin *et al.* [46] propose to accumulate the neighbor votes received from the visual similar images of the input image for each tag. Later, to estimate the relevance between a tag and an input image, the approaches of content-based tag refinement [47] and tag ranking [37] propose to use Gaussian function to measure the similarity between the input image and training images, which have already been annotated by the tag. This type of approaches also has a very intuitive explanation that tags of the input image can be determined by the soft voting from its visual neighbors from training set. Image retagging [48] and low-rank-based image tag refinement approaches [49], though effective by imposing some reasonable constraints (e.g., low rank, error sparsity, etc.), still rely on the idea of the instance neighbor voting.

In summary, based on the strong representation ability of CNN features, the model-free based algorithms can outperform those model-based approaches if the number of training instances is large enough. However, since the training data are often collected from user-contributed image-tag pairs (e.g., Flickr), the assumption of model-free approaches that the tags provided by users are reasonably accurate is not justifiable. Previous research has reported that half of the user-contributed tags from Flickr are noises to image content [50]. Therefore, due to the limited size and heavy noises of the training set in real applications, incomplete and imprecise tags still hinder user from accessing to the images after the tagging/tag refinement process by these model-free approaches.

B) Model-based image tagging by deep neural network

Since the model-free approaches can only achieve limited performance on the limited training instances in real applications. To enhance the generalization ability of image tagging models, the second paradigm of image tagging approaches proposes to learn parameterized models from training data. One of the representative works from the early research of model-based concept representation is visual synset [38], which shares similar motivation to the work of view-dependent concept representations [10]. Visual synset applies multi-class one-versus-all linear SVM [51], which is learned from the automatically generated visual synsets from a large collection of web images. Compared with the visual synset, view-dependent approach [10] can discriminate different views and represent them with proper view-dependent representations. Specifically, model-based components for generic views are designed to enhance generalization ability, while model-free components for specific views act as complements since they are too sparse and diverse. Besides, view-dependent approach [10] is easily paralleled as the model-based and model-free components are learned per concept, with no dependency on each other. The view-dependent approach is more scalable, which can add new concept representations into the existing vocabulary rather than re-training the visual models on the whole vocabulary again compared to the methods using discriminative models, such as SVM. Moreover, the view-dependent approach collects larger vocabularies for image tagging with nearly 30% increases compared to visual synset.

To further relieve the problem of limited training instances, some pioneer works propose to leverage the integration of transfer learning schemes and deep learning techniques. Transfer deep learning targets at the transfer of knowledge from a source domain to a target domain using deep learning algorithms. A typical model-based image tagging framework by transfer deep learning is shown in Fig. 3. In [52], transfer learning problems are divided into two categories. One class tries to improve the classification

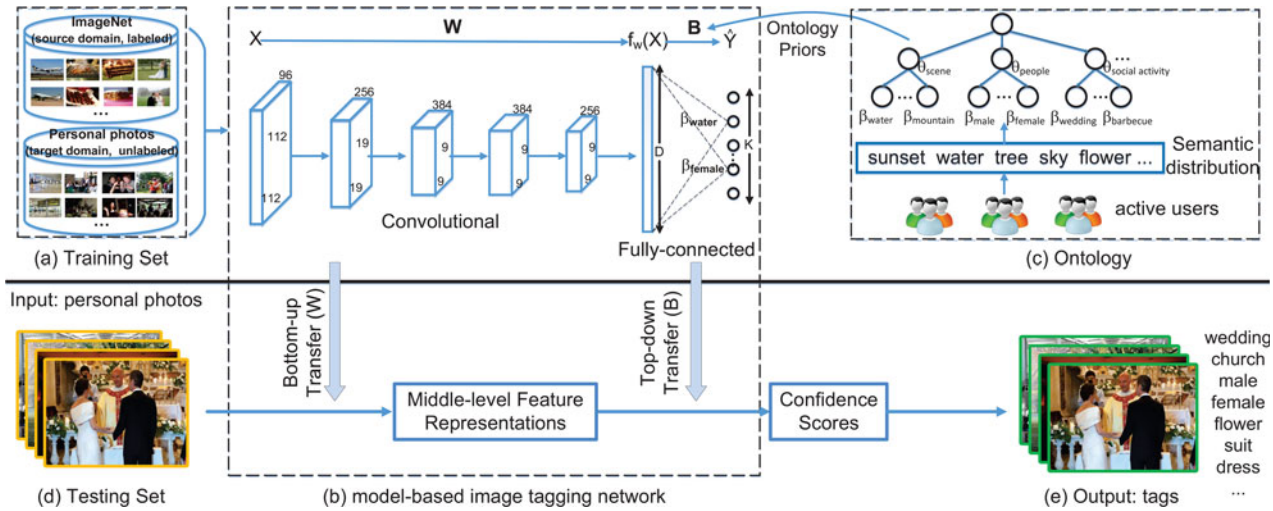

Fig. 3. An example of model-based image tagging framework. (a) The training set contains the labeled source images and the unlabeled target images. (b) The network of the transfer deep learning with ontology priors. It is first trained on both ImageNet (the source domain) and personal photos (the target domain) by pre-training and fine-tuning for discovering shared middle-level feature abstractions across domains. Once the shared feature abstractions are learned, the top layer with ontology priors is further trained. In the testing stage, the resultant parameters **W** and **B** can be transferred to the target domain to obtain the middle-level feature representations (a bottom-up transfer) and high-level confidence scores (a top-down transfer). (c) An illustration of the ontology collecting scheme. (d) The input, in the testing stage, is highly flexible which can either be a single photo or a photo collection. (e) The tagging result.

https://doi.org/10.1017/ATSIP.2017.12

performance of categories that have deficit training examples by discovering other similar categories and transferring knowledge among them [53]. The second class of works aims to solve the different data distribution problems, so as to effectively transfer the knowledge learned from the source domain to the target [17, 52].

However, although promising performance has been achieved from the earlier model-based approaches, the data noise problems existed in training instances are not considered for further improvement. To simultaneously learn better generalization model and noise-robust model, there are two schemes to adapt deep learning algorithms to noisy data. One type of approaches proposes to conduct preprocessing procedure to remove the noisy data before training. Other works propose to enable the deep learning network itself to identify the authenticity of labels and thus a robust model can be learned. The preprocessing methods can be implemented either by the conventional outlier detection, or by the pre-training strategy in deep learning frameworks. First, the specific methods in outlier detection include Principle Component Analysis (PCA), Robust PCA [54], Robust Kernel PCA [55], probabilistic modeling [56] and one-class SVM [57], and so on. These methods regard the outliers as those "few and different" samples, which usually locates far away from the sample centers in some learned embedding subspaces. Specifically, the samples which have a large discrepancy with other samples in the same category is considered to be noises. However, the challenge of these methods is to distinguish "hard samples" from the truly noisy samples. Those "hard samples" are correct image samples under a category, yet with large visual variances with the majority. Second, recovering the clean training samples by a layer-wise auto-encoder or denoising auto-encoder [58] in pre-training and then initializing a deep network by the pre-trained model parameters is an effective method to remove global noises, which has been used and extensively evaluated in face parsing [59]. However, these methods are mainly designed for the cases where noises are contained from parts of images (e.g., background noises), while web images (e.g., from Flickr) are often completely mislabeled.

To train a robust deep learning model on noisy training data, Larsen *et al.* propose one of the pioneer works which adds noise modeling into the neural networks [60]. However, they make a symmetric label noise assumption, which is often not true in real applications. Mnih *et al.* propose to label aerial images from noisy data, where only a binary classification is considered [61]. One of the promising works is proposed by Sukhbaatar *et al.* who introduce an extra noise layer as a part of training process in multi-class image classification [62]. They first train a base model on noisy training data for several iterations, then activate the extra noise layer to absorb the noise from the learned base model. Recently, a similar work to [62] has been proposed to handle the noise problems [21]. This work starts from embedding the feature map of the last deep layer into a new affinity representation. Second, by adopting the "few and different" assumption about the noises, this work minimizes the discrepancy between the affinity representation and its low-rank approximation. Third, this discrepancy is further transformed into the objective function to give those "few and different" noisy samples low-level authorities in training. Extensive experiments have shown superior performance than previous research.

Unlike above model-based learning, which learns concept representations for image tagging by cleaning image samples from noisy training instances, zero-shot learning takes one step further, which aims to learn tagging model from seen categories and tag an image from unseen categories [63–68]. Attributes sharing and word embedding are the two main streams in zero-shot learning approaches. One of the typical approaches is from [69], where the authors consider pre-fixed unseen tags with the number of U and learn a multi-label model to annotate the 2^U possible combinations between them. The limitation of this approach derives from the small number of unseen tags. FastoTag is recently proposed to enrich the family of zero-shot learning by zero-shot multi-label classification [70]. They directly model the labels and propose to assign or rank many unseen tags for unseen images. To achieve better performance, deep learning features which are extracted from CNN model are indispensable in image modeling.

IV. DATASETS AND PERFORMANCE METRICS

A few benchmark datasets have recently been proposed to fairly compare different image tagging methods. We review the following representative datasets, because: (1) they are collected from real-world data sources; (2) they have large-scale image-tag pairs, which are open to research community; and (3) they have been widely-used in research areas. MIRFlickr-25K and NUS-WIDE-270K datasets are two representative image tagging datasets, which are widely-used in many research papers, and MIT-Adobe-FiveK and MSCOCO are two emerging but challenging datasets for image tagging evaluation. Also, we introduce ImageNet dataset, since it has been widely-used to pre-train deep learning models for subsequent image tagging. Detailed statistics for different datasets are shown in Table 2.

– **MIT-Adobe-FiveK** dataset is collected from 5000 photographs taken with SLR cameras by a set of different photographers [71]. These photographs cover a broad range of scenes, subjects, and lighting conditions. The

Table 2. The statistics of the number of tags and images for different datasets for image tagging.

Datasets	# tags	# Image
MIT-Adobe-FiveK [71]	6	5 k
MIRFlickr-25K [72]	1386	25 k
NUS-WIDE-270K [50]	5018	270 k
MSCOCO [73]	80	120 k
ImageNet [22]	5247	3.2 million

https://doi.org/10.1017/ATSIP.2017.12

specific semantic annotation includes six general categories, such as indoor/outdoor and main subjects (e.g., people, nature, man-made objects, etc.)
- **MIRFlickr-25K** dataset is collected by the Leiden University [72], which contains 25 000 images with 1386 unique tags, where 14 tags of them are manually annotated as ground truth.
- **NUS-WIDE-270K** dataset is a larger open benchmark dataset, which is collected by National University of Singapore in 2009 [50]. The dataset contains 269 648 images with 5018 unique tags, where 81 tags of them are manually annotated as ground truth. Besides, the dataset provides low-level visual features for each image (e.g., color, texture, and clustered visual words).
- **MSCOCO** dataset is an emerging dataset with object labels, segmentation information and image captions [73]. In image tagging tasks, only images and object tags are used, with 82 783 image samples for training and 40 504 image samples for testing. The number of tags with ground truth is 80, and the average number of tags is 2.95 per image sample.
- **ImageNet** dataset uses the hierarchical structure of WordNet[7], which covers 12 subtrees with 5247 synsets and 3.2 million images in total [22]. A subset from ImageNet forms the training/testing data for ImageNet challenge, which focuses on 1 K categories with an average of 500 to 1000 clean and full resolution images.

To evaluate the performance of different image tagging/tag refinement approaches, several evaluation measurements have been proposed in research community, including F-score, Average Precision (AP), and Normalized Discounted Cumulative Gain (NDCG).

F-score: The F-score is calculated by the harmonic mean of precision and recall, which is given by:

$$F = \frac{2 \times P \times R}{P + R}, \tag{5}$$

where P and R indicate precision and recall, respectively.

Average precision: The AP for each tag is calculated by counting the number of relevant images from the retrieved image list, which is given by:

$$AP(t) = \frac{1}{N} \sum_{i=1}^{N} \frac{n_i}{i} \delta(x_i, t), \tag{6}$$

where N is the total number of images associated with tag t, n_i is the number of relevant images from the top i ranked images. $\delta(x_i, t) = 1$ if the image x_i is relevant to tag t, and otherwise zero. The mean value over all testing tags calculated by averaging $AP(t)$ is called Mean Average Precision (MAP).

Normalized discounted cumulative gain: The NDCG measures multi-level relevance and assumes the relevant tags are more useful if appearing higher in a ranked list. This metric at the position l in the ranked list is defined by:

$$NDCG@l = Z_l \sum_{i=1}^{l} \frac{2^{r^i} - 1}{log(1 + i)}, \tag{7}$$

where 2^{r^i} is the relevance level of the ith tag and Z_l is a normalization constant so that $NDCG@l = 1$ for the perfect ranking.

[7] https://wordnet.princeton.edu

V. SELECTED APPROACHES AND DETAILED EVALUATION

In this section, we select typical approaches from both model-free and model-based paradigms to evaluate the image tagging performance, and analyze the advantages of each paradigm. Our criteria to select the compared approaches are: (1) widely-used in research community and (2) strong performance under the same experiment settings.

A) Model-free image tagging

We first select three typical model-free image tagging models for comparison, which include:

- **KNN** [74]: **K**-**N**earest-**N**eighbor based approaches generate a tag list by calculating the relevance of a tag to a given image. The procedure is conducted by first searching the top-K nearest neighbor images which have been annotated by the tag, and further ranking a tag list by tag frequency.
- **TagVoting** [40]: Compared with KNN, **Tag Voting** based approaches leverage user-side information, and constrain each user has at most one image in the neighbor set. Besides, this approach considers tag prior frequency to limit the occurrence of frequent tags.
- **TagProp** [46]: **Tag Prop**agation based approaches employ neighbor voting with distance metric learning scheme into image tagging.

Table 3 shows that the tagging performance of different model-free image tagging approaches, by using either traditional hand-crafted features with bag-of-word (BoW) representations [8] or deep learning features. We can observe from the table that deep learning based representation can achieve consistently better performance than hand-crafted features for the three typical model-free approaches, which

Table 3. Image tagging performance (measured by MAP) on MIRFlickr-25K and NUS-WIDE-270K for different comparison approaches.

Approaches	MIRFlickr	NUS-WIDE
BoW+KNN [74]	0.34	0.19
BoW+TagVoting [40]	0.33	0.18
BoW+TagProp [46]	0.34	0.19
CNN+KNN [74]	0.63	0.39
CNN+TagVoting [40]	0.64	0.40
CNN+TagProp [46]	0.65	0.42

https://doi.org/10.1017/ATSIP.2017.12

demonstrates the powerful representation ability of deep learning techniques to image tagging. For different model-free tagging models, we can observe that KNN [74] and TagVoting [40] can achieve comparable performance. TagProp [46] can achieve better results, due to the proposed distance metric learning scheme. Note that the deep features used here are extracted from the pre-trained VGG-19 network [43] on ImageNet. Specifically fine-tuned features for this image tagging application can achieve further improvement.

B) Model-based image tagging

We select the following typical model-based image tagging approaches for comparison:

- **CNN**: Convolutional Neural Network adopts the state-of-the-art network architecture with several convolutional layers and fully-connected layers [13].
- **RPCA+CNN**: Robust Principle Component Analysis + CNN first removes samples with large reconstruction errors by RPCA [54], and conducts CNN training on the cleaned samples.
- **CAE+CNN**: Convolutional Auto-Encoder + CNN proposes to reduce the noise effect in CNN training by the layer-wise pre-training and fine-tuning strategy [59].
- **NA+CNN**: Noise Adaption layer + CNN proposes to add an additional bottom-up noise-adaption layer into the traditional CNN architecture for noise removal [62].
- **NR+CNN**: Noise Robust layer + CNN designs an objective function to assign those "few and different" noisy samples with low-level authorities in training, and thus the noise effect can be reduced [21].

Each model is trained and evaluated on NUS-WIDE-270K [50], since the dataset contains the largest vocabulary size among different image tagging datasets.

Since the annotated tags for each image is limited (only 81 tags have ground truth), a randomly-selected 1000 photos from NUS-WIDE-270K [50] are used as the test set for extensive evaluation. Each method produces top five categories with the highest prediction scores as a tagging list. 25 human-labelers are employed to evaluate each tag with three levels: 2–Highly Relevant; 1–Relevant; 0–Non Relevant. We adopt both the mean average precision (MAP) and the Normalized Discounted Cumulative Gain (NDCG) as the metric to evaluate the tagging performance. Table 4 shows the tagging performance. We can observe that simple image classification architecture CNN [13] cannot achieve good performance, since there are usually noises in image tagging datasets. For different noise-robust CNN architectures, CAE+CNN [59] proposes to reduce the noise effect by auto-encoder, which can achieve a slightly better result than simple CNN. Compared with CAE+CNN [59], RPCA+CNN [54], and NA+CNN [62] can achieve significant improvement with a large margin in terms of NDCG. NR+CNN [21] achieves the best performance due to the effective noise robust layer proposed in their paper. Exemplary tagging results of NR+CNN [21] are shown in Fig. 4. We can also observe the vocabulary difference between image classification and image tagging, where the underlined tags in Fig. 4 are even missing from ImageNet categories, yet important to reflect the diverse visual content in image tagging.

Table 4. Tagging performance in terms of both mean average precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) for the 1000 testing photos from NUS-WIDE-270K dataset.

	CNN [13]	CAE+ CNN [59]	RPCA+ CNN [54]	NA+ CNN [62]	NR+ CNN [21]
MAP	0.28	0.32	0.33	0.47	0.53
NDCG@1	0.08	0.11	0.23	0.24	0.32
NDCG@3	0.18	0.25	0.32	0.33	0.41
NDCG@5	0.26	0.34	0.39	0.41	0.46

VI. APPLICATIONS

Extensive research on both academic and industrial fields have focused on image tagging as well as its related applications, such as photo search, and photo storytelling.

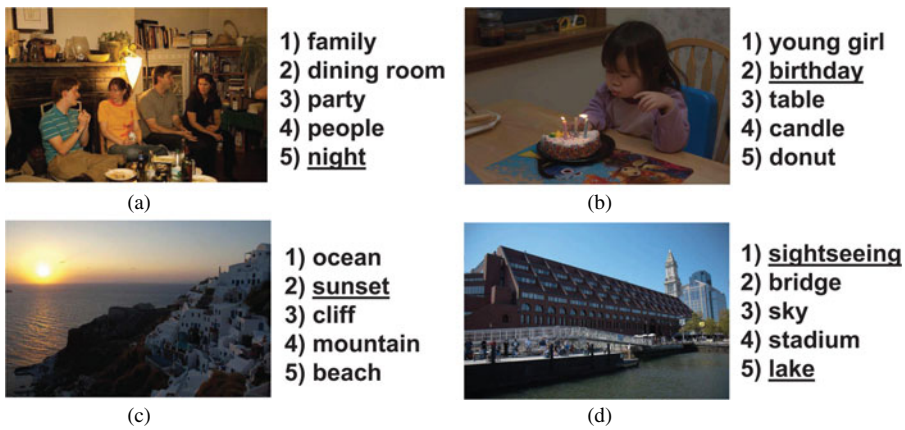

Fig. 4. Image tagging results by a typical model-based approach [21]. Note that the underlined tags are missing from ImageNet categories, which shows the vocabulary difference between image classification and image tagging.

https://doi.org/10.1017/ATSIP.2017.12

Although most of the images are generated and stored on clients (e.g., mobile phones, tablet, etc.), existing works mainly rely on cloud-based solutions, which usually consist of image data transmitting from client to cloud, image semantic tagging and online image search on cloud. Typical commercial products include PhotoTime[8], Google Photos[9] and Microsoft OneDrive[10], which enable effective photo indexing and search not only by time and location, but also by face grouping and object/scene understanding, and thus photos can be found by entering human-friendly tags (e.g., "beach," "sunset") in the search box. However, the cloud-based solution requires sending each photo to a remote server on the cloud, which hardly guarantees instant photo indexing and search on the phone, due to network latency. With the increasing computing capacity on mobile clients, designing effective client-based photo search systems on the phone becomes a fundamental challenge for better user experience. Further improvement has been observed from the recently-released photo management system on iOS 10, which can provide efficient photo grouping by face, tag, time and location on the phone. Once photos have been tagged and indexed on the phone, the system further supports semantic photo search by typing a single word from a predefined vocabulary. In the following, we will survey several existing commercial products which are built on image tagging techniques.

A) PhotoTime

PhotoTime gives the first attempt to provide the features of sorting and organizing personal photos by image tagging techniques. The app is only available for iOS users now, with the total app size of 33 MB. With the help of cutting-edge deep neural networks, PhotoTime can automatically tag and categorize photos and conduct face grouping, hence users can effortlessly search photos by the provided features. Besides, PhotoTime can organize photos not only on the phone, but also those from social media platforms (e.g., iCloud, Facebook, Instagram, Twitter, Dropbox, Flickr, and Google+), without additional storage space needed. The versions released after June 2016 have made Amazon Cloud embedded for more flexible storage.

PhotoTime sends photo thumbnails, which are much smaller than the original photos, to cloud servers for image tagging, and keep the amount of data transmission minimal. When the tagging procedure is finished, PhotoTime sends back the auto-generated tags instead of photos or thumbnails. The specific tagging speed depends on network latency, with the highest efficiency of 500 photos per minute. Users can add personal tags in the detailed photo view, and all the new tags can become searchable tags.

[8] https://phototime.com
[9] https://photos.google.com
[10] https://onedrive.live.com

B) Google Photos

Google Photos is a photo sharing and storage service developed by Google. Mobile apps for Android and iOS allow users to upload and access pictures on mobile. There are also desktop tools for Windows and MacOS. Google Photos grows very fast, with 200 million active users in the first month after its release, and it also allows users to back up an unlimited number of photos and videos. Google Photos organizes photo library using smart groups based on time and location. When users upload photos, Google Photos can automatically analyze them and add them to specific categories for quick access. Hence, users are able to conduct keyword-based search to find a photo. The keyword could be object, time point, or location.

Google Photos makes use of a CNN architecture, which is similar to the architecture proposed by Geoffrey Hinton and used in the ImageNet Large Scale Visual Recognition Competition. In contrast to the 1000 visual classes used in the competition, Google uses more visual classes based on the popular labels from users. Google Photos also makes use of free base entities, which are the basis for knowledge graph in Google search. Besides, users can browse into any image albums and display it on screen as slideshow. In addition, the images could be remixed to create videos for sharing on social media. Users are also allowed to label persons with their names, which can be used for searching by name in the future.

C) Microsoft OneDrive

OneDrive provides a file-hosting service developed by Microsoft, which is a part of Microsoft suite of online services. OneDrive allows users to store files as well as other personal data like Windows settings or BitLocker recovery keys in the cloud. When mobile users from Android, iOS, and Windows Phone take a photo by their phones, the photo will be automatically uploaded to OneDrive. OneDrive organizes photo libraries based on date and several tags, like "mountain," "outside," "hand," etc. It can also automatically recognizes objects in a photo and creates tags for them by the state-of-the-art deep neural networks with over 6 K tag categories, so users are able to search photos with the tags provided. Furthermore, users can also remove or edit the tags for a photo.

D) Photo app on iOS 10

The photo app embedded on iOS 10 provides a client-based way to organize photos. Photos are categorized by date and location, and two new album views are provided, which includes "People" and "Places." Each album contains photos and videos organized by faces or location information. Apple now uses deep neural networks to recognize faces, and then organizes these photos into mini albums inside the "People" album. Users can add names to them, which allows users to organize all the photos by person. The "Places" album records a map of user photos, which helps users to browse exactly where they recorded a video or snapped a

https://doi.org/10.1017/ATSIP.2017.12

picture on their trips. The photo app embedded on iOS 10 allows users to search for photos using keyword-based natural language which can be people, objects, locations, and other characteristics. However, the system usually suffers from low recall due to the limited vocabulary of queries that can be searched. Therefore, designing a system with the capability of instantly finding arbitrary photos of what users want is still worthy to explore in the future.

VII. CONCLUSIONS AND PERSPECTIVES

In this paper, we presented a survey of various methods of image tagging using deep learning. We have summarized the research into two paradigms, i.e., model-free image tagging models with deep representation and model-based image tagging models by deep neural network. Although significant progresses have been made in the recent years, there are still emerging topics that deserve further investigation. We summarize the future challenges as follows.

Combination of model-based/model-free approaches: as observed by many existing approaches, on the one hand, the model-free approaches which are generally based on nearest neighbor voting, can achieve superior performance if training samples are in large size with good diversity, but suffer from high computation costs in both training and testing. On the other hand, model-based approaches are generally efficient with adequate modeling capability on few training samples. Designing a framework that can combine the advantages of two types of models is a possible direction. In such a manner, the model-based components can ensure a good generalization ability when we cannot find the similar enough or duplicate visual neighbors in training set, while the model-free components complement the loss of the discriminative power of the model-based components.

Multi-label image tagging: most existing work focus on single-label image tagging by using deep neural networks, where each image is assumed to be associated with one image tag. Since the diversity of image content and the complexity of user search intentions, single tag is not adequate to reflect the multiple semantics conveyed from an image. Therefore, multi-label image tagging is an important research topic for managing the great number of user photos. The main challenges derive from the huge human efforts to obtain fully-annotated image sets and the difficulties to model and predict multiple semantics from deep neural networks. Although initial efforts have been made [75], great efforts are still necessary to design networks by leveraging dense image understanding and representation technologies, such as object detection and image segmentation.

Visual attention-based image tagging: a large part of image tagging approaches leverage large textual corpus for expanding the tagging list for higher recall. For example,"bridge" and "water," "photo frame," and "wall" are likely to appear in the same image. Therefore, if "photo frame" has been detected and tagged for an image, and then "wall" will have higher probability to be tagged according to the prior knowledge. Although the tag of "wall" can increase the recall of a tagging system, it is not much important for users. Because "wall" is not likely to be an attention area, compared with the foreground "photo frame." How to automatically learn user attention to an image, and tag those objects that can be searched is indispensable for an intelligent image tagging system. Learning from the recent-proposed image captioning and visual question answering datasets [73] can be good choices, because there are explicit user intentions existed in this type of datasets.

Fine-grained image tagging: although existing tagging system can provide users with good experience for photo browsing and search to some extend. It is not sufficient for common users to find very fine-grained and personalized photos. For example, tagging different bird species [76, 77], flower types [78, 79], car models [80, 81], or even human sentiment [82] have attracted extensive attention. This task is very challenging as some fine-grained categories (e.g., eared grebe, and horned grebe) can only be recognized by domain experts. Different from general image tagging, fine-grained image tagging should be capable of localizing and representing the very marginal visual differences within subordinate categories [83, 84]. Besides, tagging an image with "girl" or "woman" is not very existing, compared with tagging them with more personalized tags (e.g., "daughter") based on the contextual information from a user album and user-provided meta-data.

Photo storytelling: beyond image tagging, automatic generation of natural language description for individual images (a.k.a. image captioning) has attracted extensive research attention [85–88]. Some works even take one step further to investigate the generation of a paragraph to describe a photo stream for the purpose of storytelling [89–91]. This task is even more challenging than individual image description due to the difficulty in modeling the large visual variance in an ordered photo collection and in preserving the long-term language coherence among multiple sentences. Although challenging, storytelling can better benefit user experience to recall their memory, which is worthy to discover in the future.

REFERENCES

[1] Jin, Y.; Khan, L.; Wang, L.; Awad, M.: Image annotations by combining multiple evidence and WordNet, in *Proc. ACM Multimedia*, 2005, 706–715.

[2] Wang, M.; Ni, B.; Hua, X.-S.; Chua, T.-S.: Assistive tagging: a survey of multimedia tagging with human–computer joint exploration. *ACM Comput. Surv.*, **44** (4) (2012), 25:1–25:24.

[3] Wang, X.-J.; Zhang, L.; Liu, M.; Li, Y.; Ma, W.-Y.: Arista - image search to annotation on billions of web photos, in *Conf. on Computer Vision and Pattern Recognition*, 2010, 2987–2994.

[4] Tang, J.; Chen, Q.; Wang, M.; Yan, S.; Chua, T.-S.; Jain, R.: Towards optimizing human labeling for interactive image tagging. *ACM Trans. Multimed. Comput. Commun. Appl.*, **9** (4) (2013), 29:1–29:18.

[5] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, **60** (2) (2004), 91–110.

https://doi.org/10.1017/ATSIP.2017.12

[6] Oliva, A.; Torralba, A.: Building the gist of a scene: the role of global image features in recognition. *Prog. Brain Res.* **155** (2006), 23–36.

[7] Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D.: Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, **32** (9) (2010), 1627–1645.

[8] Sivic, J.; Zisserman, A.: Video Google: a text retrieval approach to object matching in videos, in *Int. Conf. on Computer Vision*, 2003, 1470–1477.

[9] Lazebnik, S.; Cordelia, S.; Jean, P.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in *Conf. on Computer Vision and Pattern Recognition*, 2006, 2169–2178.

[10] Fu, J.; Wang, J.; Rui, Y.; Wang, X.-J.; Mei, T.; Lu, H.: Image tag refinment with view-dependent concept representations, in *IEEE Transactions on Circuits and Systems for Video Technology*. IEEE, 2014.

[11] Fu, J.; Wang, J.; Li, Z.; Xu, M.; Lu, H.: Efficient clothing retrieval with semantic-preserving visual phrases, in *Asian Conf. on Computer Vision*, 2013, 420–431.

[12] Bengio, Y.: Learning deep architectures for AI. *Found. Trends Mach. Learn.*, **2** (1) (2009), 1–127.

[13] Krizhevsky, A.; Sutskever, I.; Hinton, G.E.: Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems*, 2012, 1106–1114.

[14] Le, Q.V. *et al.*: Building high-level features using large scale unsupervised learning, in *Int. Conf. on Machine Learning*, 2012.

[15] Graham, B.: Spatially-sparse convolutional neural networks, 2014. arXiv preprint arXiv:1409.6070.

[16] Deng, J.; Berg, A.C.; Li, K.; Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? in *European Conf. on Computer Vision*, 2010, 71–84.

[17] Fu, J.; Mei, T.; Yang, K.; Lu, H.; Rui, Y.: Tagging personal photos with transfer deep learning, in *Proc. World Wide Web*, 2015, 344–354.

[18] Li, X.; Uricchio, T.; Ballan, L.; Bertini, M.: Snoek, C.G.M.; Bimbo, A.D.: Socializing the semantic gap: a comparative survey on image tag assignment, refinement and retrieval, in *CoRR*, arXiv: abs/1503.08248, 2015.

[19] Chen, X.; Shrivastava, A.; Gupta, A.: Neil: extracting visual knowledge from web data, in *Int. Conf. on Computer Vision*, 2013.

[20] Divvala, C.S.K.; Farhadi, A.: Learning everything about anything: Webly-supervised visual concept learning, in *Conf. on Computer Vision and Pattern Recognition*, 2014.

[21] Fu, J.; Wu, Y.; Mei, T.; Wang, J.; Lu, H.; Rui, Y.: Relaxing from vocabulary: robust weakly-supervised deep learning for vocabulary-free image tagging, in *IEEE Int. Conf. on Computer Vision*, 2015.

[22] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L.: ImageNet: a large-scale hierarchical image database, in *Conf. on Computer Vision and Pattern Recognition*, 2009, 248–255.

[23] Stauffer, C.; Grimson, W.E.L.: Adaptive background mixture models for real-time tracking, in *Conf. on Computer Vision and Pattern Recognition*, vol. **2**, 1999, 246–252.

[24] Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE*, **77** (2) (1989), 257–286.

[25] Lafferty, J.D.; McCallum, A.; Pereira, F.C.N.: Conditional random fields: probabilistic models for segmenting and labeling sequence data, in *Int. Conf. on Machine Learning*, vol. **1**, 2001, 282–289.

[26] Jaynes, E.T.: Information theory and statistical mechanics. *Phys. Rev.*, **106** (4) (1957), 620.

[27] Cortes, C.; Vapnik, V.: Support-vector networks. *Mach. Learn.*, **20** (3) (1995), 273–297.

[28] Hosmer, D.W. Jr.; Lemeshow, S.; Sturdivant, R.X.: *Applied Logistic Regression*, 3rd ed. John Wiley & Sons, 2013.

[29] Rumelhart, D.E.; Hinton, G.E.; Williams, R.J.: Learning internal representations by error propagation. DTIC Document, Technical Report, 1985.

[30] Deng, L.: A tutorial survey of architectures, algorithms, and applications for deep learning, in *APSIPA Transactions on Signal and Information Processing*, January 2014.

[31] Hinton, G.E.; Osindero, S.; Teh, Y.-W.: A fast learning algorithm for deep belief nets. *Neural Comput.*, **18** (7) (2006), 1527–1554.

[32] Ackley, D.H.; Hinton, G.E.; Sejnowski, T.J.: A learning algorithm for Boltzmann machines. *Cognit. Sci.*, **9** (1) (1985), 147–169.

[33] Smolensky, P.: Information processing in dynamical systems: foundations of harmony theory. DTIC Document, Technical Report, 1986.

[34] Bengio, Y.; Lamblin, P.; Popovici, D.; Larochelle, H.: Greedy layer-wise training of deep networks, in *Advances in Neural Information Processing Systems*, 2007, 153–160.

[35] Rumelhart, D.; Hintont, G.; Williams, R.: Learning representations by back-propagating errors. *Nature*, **323** (1986), 533–536.

[36] Li, T.; Mei, T.; Kweon, I.-S.; Hua, X.-S.: Contextual bag-of-words for visual categorization. *IEEE Trans. Circuits Syst. Video Technol.*, **21** (4) (2011), 381–392.

[37] Liu, D.; Hua, X.-S.; Yang, L.; Wang, M.; Zhang, H.-J.: Tag ranking, in *Proc. of World Wide Web*, 2009, 351–360.

[38] Tsai, D.; Jing, Y.; Liu, Y.; Rowley, H.A.; Ioffe, S.; Rehg, J.M.: Large-scale image annotation using visual synset, in *Int. Conf. on Computer Vision*, 2011, 611–618.

[39] Wu, P.; Hoi, S.C.-H.; Zhao, P.; He, Y.: Mining social images with distance metric learning for automated image tagging, in *Proc. of Web Search and Data Mining*, 2011, 197–206.

[40] Li, X.; Snoek, C.G.; Worring, M.: Learning tag relevance by neighbor voting for social image retrieval, in *Proc. ACM Int. Conf. on Multimedia Information Retrieval*, 2008, 180–187.

[41] Wang, X.-J.; Zhang, L.; Jing, F.; Ma, W.-Y.: Annosearch: image auto-annotation by search, in *Conf. on Computer Vision and Pattern Recognition*, 2006, 1483–1490.

[42] Szegedy, C. *et al.*: Going deeper with convolutions, in *Conf. on Computer Vision and Pattern Recognition*, 2015.

[43] Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition, in *Int. Conf. on Learning Representations*, 2015, 1409–1556.

[44] He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep residual learning for image recognition, in *Conference on Computer Vision and Pattern Recognition*, 2016, 770–778.

[45] Li, X.; Snoek, C.G.M.; Worring, M.: Learning social tag relevance by neighbor voting. *IEEE Trans. Multimed.*, **11** (7) (2009), 1310–1322.

[46] Guillaumin, M.; Mensink, T.; Verbeek, J.J.; Schmid, C.: Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation, in *Int. Conf. on Computer Vision*. IEEE, 2009, 309–316.

[47] Wang, C.; Jing, F.; Zhang, L.; Zhang, H.-J.: Content-based image annotation refinement, in *Conf. on Computer Vision and Pattern Recognition*, 2007.

[48] Liu, D.; Hua, X.-S.; Wang, M.; Zhang, H.-J.: Image retagging, in *Proc. ACM Multimedia*, 2010, 491–500.

[49] Zhu, G.; Yan, S.; Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity, in *Proc. ACM Multimedia*, 2010, 461–470.

https://doi.org/10.1017/ATSIP.2017.12

[50] Chua, T.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore, in *Proc. ACM Conf. on Image and Video Retrieval*, 2009.

[51] Cortes, C.; Vapnik, V.: Support-vector networks. *Mach. Learn.*, **20** (3) (1995), 273–297.

[52] Maxime, O.; Leno, B.; Ivan, L.; Josef, S.: Learning and transferring mid-level image representations using convolutional neural networks, in *Conf. on Computer Vision and Pattern Recognition*, 2014, 1717–1724.

[53] Srivastava, N.; Salakhutdinov, R.: Discriminative transfer learning with tree-based priors, in *Advances in Neural Information Processing Systems*, 2013, 2094–2102.

[54] Candès, E.J.; Li, X.; Ma, Y.; Wright, J.: Robust principal component analysis? *J. ACM*, **58** (3) (2011), 11:1–11:37.

[55] Xu, H.; Caramanis, C.; Mannor, S.: Outlier-robust PCA: the high-dimensional case. *IEEE Trans. Inf. Theory*, **59** (1) (2013), 546–572.

[56] Kim, J.; Scott, C.D.: Robust kernel density estimation. *Mach. Learn.*, **13** (2012), 2529–2565.

[57] Liu, W.; Hua, G.; Smith, J.R.: Unsupervised one-class learning for automatic outlier removal, in *Conf. on Computer Vision and Pattern Recognition*, 2014.

[58] Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *Mach. Learn.*, **11** (2010), 3371–3408.

[59] Luo, P.; Wang, X.; Tang, X.: Hierarchical face parsing via deep learning, in *Conf. on Computer Vision and Pattern Recognition*, 2012, 2480–2487.

[60] Larsen, J.; Andersen, L.N.; Hintz-madsen, M.; Hansen, L.K.: Design of robust neural network classifiers, in *Int. Conf. on Acoustics, Speech and Signal Processing*, 1998, 1205–1208.

[61] Mnih, V.; Hinton, G.E.: Learning to label aerial images from noisy data, in *Int. Conf. on Machine Learning*, 2012.

[62] Sukhbaatar, S.; Fergus, R.: Learning from noisy labels with deep neural networks, 2014. arXiv:1406.2080.

[63] Akata, Z.; Perronnin, F.; Harchaoui, Z.; Schmid, C.: Label-embedding for attribute-based classification, in *Conf. on Computer Vision and Pattern Recognition*, 2013.

[64] Jayaraman, D.; Grauman, K.: Zero shot recognition with unreliable attributes, in *Advances in Neural Information Processing Systems*, 2014, 3464–3472.

[65] Lampert, C.H.; Nickisch, H.; Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, **36** (3) (2014), 453–465.

[66] Norouzi, M. *et al.*: Zero-shot learning by convex combination of semantic embeddings, in *CoRR*, arXiv: abs/1312.5650, 2013.

[67] Palatucci, M.; Pomerleau, D.; Hinton, G.E.; Mitchell, T.M.: Zero-shot learning with semantic output codes, in *Advances in Neural Information Processing Systems*, 2009, 1410–1418.

[68] Socher, R.; Ganjoo, M.; Manning, C.D.; Ng, A.Y.: Zero-shot learning through cross-modal transfer, in *Advances in Neural Information Processing Systems*, (2013), 935–943.

[69] Fu, Y.; Yang, Y.; Hospedales, T.M.; Xiang, T.; Gong, S.: Transductive multi-label zero-shot learning, in *British Machine Vision Association*, 2015.

[70] Zhang, Y.; Gong, B.; Shah, M.: Fast zero-shot image tagging, in *Computer Vision and Pattern Recognition*, 2016.

[71] Bychkovsky, V.; Paris, S.; Chan, E.; Durand, F.: Learning photographic global tonal adjustment with a database of input/output image pairs, in *Conf. on Computer Vision and Pattern Recognition*, 2011, 97–104.

[72] Huiskes, M.J.; Lew, M.S.: The MIR flickr retrieval evaluation, in *Proc. ACM Int. Conf. on Multimedia Information Retrieval*, 2008.

[73] Lin, T. *et al.*: Microsoft COCO: common objects in context, in *CoRR*, arXiv: abs/1405.0312, 2014.

[74] Makadia, A.; Pavlovic, V.; Kumar, S.: Baselines for image annotation. *Int. J. Comput. Vis.*, **90** (1) (2010), 88–105.

[75] Gong, Y.; Jia, Y.; Leung, T.; Toshev, A.; Ioffe, S.: Deep convolutional ranking for multilabel image annotation, in *CoRR*, arXiv: abs/1312.4894, 2013.

[76] Branson, S.; Horn, G.V.; Belongie, S.J.; Perona, P.: Bird species categorization using pose normalized deep convolutional nets, in *British Machine Vision Conf.*, 2014.

[77] Zhang, X.; Xiong, H.; Zhou, W.; Lin, W.; Tian, Q.: Picking deep filter responses for fine-grained image recognition, in *Conf. on Computer Vision and Pattern Recognition*, 2016, 1134–1142.

[78] Nilsback, M.-E.; Zisserman, A.: A visual vocabulary for flower classification, in *Conf. on Computer Vision and Pattern Recognition*, 2006, 1447–1454.

[79] Reed, S.E.; Akata, Z.; Schiele, B.; Lee, H.: Learning deep representations of fine-grained visual descriptions, in *Conf. on Computer Vision and Pattern Recognition*, 2016.

[80] Krause, J.; Jin, H.; Yang, J.; Li, F.: Fine-grained recognition without part annotations, in *Conf. on Computer Vision and Pattern Recognition*, 2015, 5546–5555.

[81] Lin, T.-Y.; RoyChowdhury, A.; Maji, S.: Bilinear CNN models for fine-grained visual recognition, in *Int. Conf. on Computer Vision*, 2015, 1449–1457.

[82] Wang, J.; Fu, J.; Mei, T.; Xu, Y.: Beyond object recognition: visual sentiment analysis with deep coupled adjective and noun neural networks, in *Int. Joint Conf. on Artificial Intelligence*, 2016.

[83] Fu, J.; Zheng, H.; Mei, T.: Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition, in *Conf. on Computer Vision and Pattern Recognition*, 2017.

[84] Zheng, H.; Fu, J.; Mei, T.; Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition, in *Int. Conf. on Computer Vision*, 2017.

[85] Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-RNN), in *Int. Conf. on Learning Representations*, 2015.

[86] Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D.: Show and tell: a neural image caption generator, in *Conf. on Computer Vision and Pattern Recognition*, 2015.

[87] Johnson, J.; Karpathy, A.; Fei-Fei, L.: Densecap: fully convolutional localization networks for dense captioning, in *Conf. on Computer Vision and Pattern Recognition*, 2016.

[88] Yu, D.; Fu, J.; Mei, T.; Rui, Y.: Multi-level attention networks for visual question answering, in *Conf. on Computer Vision and Pattern Recognition*, 2017.

[89] Park, C.C.; Kim, G.: Expressing an image stream with a sequence of natural sentences, in *Advances in Neural Information Processing Systems*, 2015.

[90] Huang, T.H. *et al.*: Visual storytelling, in *Conf. of the North American Chapter of the Association for Computational Linguistics*, 2016.

[91] Liu, Y.; Fu, J.; Mei, T.; Chen, C.W.: Let your photos talk: generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks, in *The Association for the Advance of Artificial Intelligence*, 2017, 1445–1452.

Jianlong Fu received the B.S. degree from University of Science and Technology Beijing, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He is currently an associate researcher at Microsoft Research. His research interests include computer vision and multimedia content analysis, especially on fine-grained image recognition, weakly-supervised and vocabulary-free deep image tagging, visual question answering, visual sentiment analysis, personal photo experience of browsing, searching, sharing and storytelling.

Yong Rui (SM'04–F'10) received the B.S. degree from Southeast University, the M.S. degree from Tsinghua University, and the Ph.D. degree from the University of Illinois at Urbana-Champaign. He is currently the Chief Technology Officer and Senior Vice President of Lenovo Group. He is responsible for overseeing Lenovo's corporate technical strategy, research and development directions, and Lenovo Research organization, which covers intelligent devices, big data analytics, artificial intelligence, cloud computing, 5G and smart lifestyle-related technologies. He has authored 2 books, 12 book chapters, and 260 refereed journal and conference papers. With over 19,000 citations, and an h-Index of 59, his publications are among the most referenced. He holds 62 issued U.S. and international patents. He is a recipient of many awards, including the 2016 IEEE Computer Society Technical Achievement Award, the 2016 IEEE Signal Processing Society Best Paper Award and the 2010 Most Cited Paper of the Decade Award from Journal of Visual Communication and Image Representation. He is the Editor-in-Chief of IEEE MultiMedia magazine, an Associate Editor of ACM Trans. on Multimedia Computing, Communication and Applications (TOMM), and a founding Editor of International Journal of Multimedia Information Retrieval. He was an Associate Editor of IEEE Trans. on Multimedia (2004–2008), IEEE Trans. on Circuits and Systems for Video Technologies (2006–2010), ACM/Springer Multimedia Systems Journal (2004–2006), International Journal of Multimedia Tools and Applications (2004–2006), and IEEE Access (2013–2016). He is a Fellow of IEEE, IAPR and SPIE, a Distinguished Member of the ACM.

https://doi.org/10.1017/ATSIP.2017.12