This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2903448, IEEE Transactions on Multimedia

Cross-modality Bridging and Knowledge Transferring for Image Understanding

Chenggang Yan*, Liang Li*, Chunjie Zhang, Bingtao Liu, Yongdong Zhang, IEEE Senior Member and Qionghai Dai, IEEE Senior Member

***Abstract*—The understanding of web images has been a hot research topic in both artificial intelligence and multimedia content analysis domains. The web images are composed of various complex foregrounds and backgrounds, which makes the design of an accurate and robust learning algorithm a challenging task. To solve the above significant problem, firstly, we learn a cross-modality bridging dictionary for the deep and complete understanding of vast quantity of web images. The proposed algorithm leverages the visual features into the semantic concept probability distribution, which can construct a global semantic description for images while preserving the local geometric structure. To discover and model the occurrence patterns between intra- and inter-categories, the multi-task learning is introduced for formulating the objective formulation with Capped-ℓ_1 penally, which can obtain the optimal solution with a higher probability and outperform the traditional convex function based methods. Secondly, we propose a knowledge-based concept transferring algorithm to discover the underlying relations of different categories. This distribution probability transferring among categories can bring the more robust global feature representation, and enable the image semantic representation to generalize better as the scenario becomes larger. Experimental comparisons and performance discussion with classical methods on the ImageNet, Caltech-256, SUN397 and Scene15 datasets show the effectiveness of our proposed method at three traditional image understanding tasks.**

***Index Terms*—Object and scene recognition, image semantic search, cross-modality bridging, multi-task learning, knowledge transferring.**

I. Introduction

Benefitting from the rapid development of social media and smart phones, vast amount of web images are produced everyday on the Internet, and the imperious demands of images automatic analysis make web image understanding a hot research topic [1]–[6]. However as there are natural semantic gap between vision and language [7]–[9], cross-modality understanding problem is still far from being solved. Fig. 1 demonstrates the implicit semantic gap between visual and language, which relationship is very deep and complex. As the amount of images goes larger, the semantic gap problem appears and brings a great influence to cross-modality understanding. In details, on the one hand, one visual appearance may be found in thousands of web images with different categories, that is to say, there are some common appearances among different categories. On the other hand, one concept has thousands of instances, and each instance could consist of some visual appearances. Cross-modality bridging [10]–[12] is the job of pairing visual appearance with the most accurate concept, and it is becoming more challenging as the amount of web images grows larger.

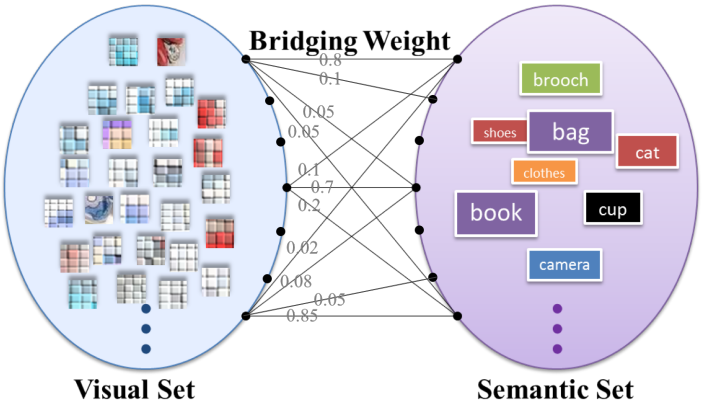

Fig. 1. The illustration of cross-modality bridging between the visual modality and semantic modality.

Many related works have been proposed to solve the problem of cross-modality bridging: (1) Latent topic model, such as discriminative Latent Dirichlet Allocation model [13], patch-based latent variable modeling [14], cross-view learning [7], etc. (2) Middle-level themes learning, such as semantic multinomial (SMN) model [15], local category co-occurrences [16], etc. (3) Distance metric learning, such as cross-category transfer learning [17] , Gaussian mixture model [18] , etc. (4) Semantic classifier model, such as object bank [19] , shared latent semantic space learning [9], etc. (5) Deep learning, such as deep convolutional neural network [20] , image broadening

*The first two authors contributed equally to this work.

Chenggang Yan and Bingtao Liu is with the Institute of Information and Control, Hangzhou Dianzi University, Hangzhou, China

Liang Li is the corresponding author, and he is with the Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS; He is also with the College of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, 100190, China, email: iang.li@ict.ac.cn

Chunjie Zhang are with the Institute of automation, CAS; Beijing, China

Yongdong Zhang are with the Advanced Computing Research Laboratory, Institute of Computing Technology, CAS; Beijing, 100190, China

Qionghai Dai is with the Department of Automation, Tsinghua University, Beijing, China

This work was supported by National Basic Research Program of China(973-Program): 2015CB351802, in part by National Natural Science Foundation of China: 6140243, 61572488, 61620106009, U1636214, 61650202 and 61672497, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013.

The fourth revised version was received on December 27, 2018; The third revised version was received on October 11, 2018; The second revised version was received on April 6, 2018; The first revised version was received on Jan 14, 2018; The original manuscript was received on June 19, 2017.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2903448, IEEE Transactions on Multimedia

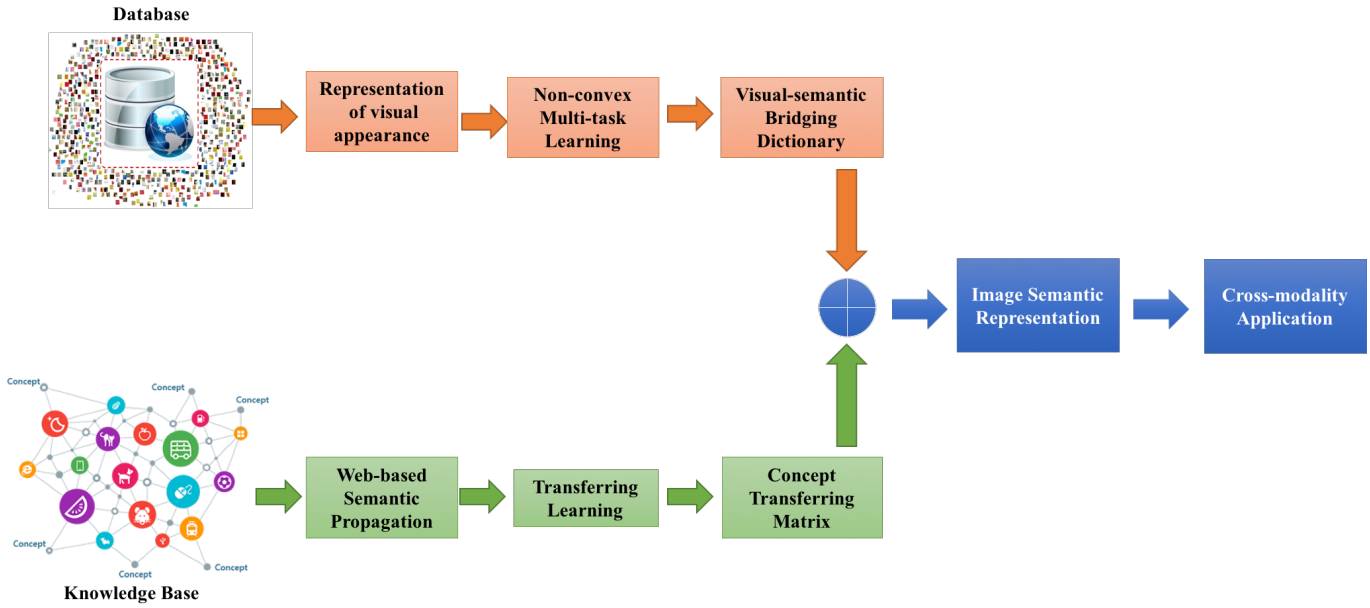

Fig. 2. The flowchart of the proposed scheme for cross-modality understanding.

based convolutional neural network [21], recurrent neural networks [22], a two-branch neural network with multiple layers of linear projections [8], and so on.

Although deep model based methods achieved great success on image understanding, the underlying mechanism of neural work is still not fully understood. The cross-modality bridging still needs further research. In history, feature learning [23]–[30] has been a successful type of methods for various computer vision tasks. Regularization of corresponding constraint is the key for feature learning, and the ℓ_1 norm is mostly used. It is a continuous and convex surrogate to loosely approximate the ideal constraint of ℓ_0 norm, but the ℓ_1 norm may brings suboptimal solutions because of the over-penalized problem. All the above methods give us some insight on solving the cross-modality bridging problem.

In this paper, the core of image understanding is regarded as a bridging problem between text labels and original visual images. Here we solve the above problem by learning an cross-modality bridging dictionary under the multi-task feature learning framework, where visual appearances are interpreted into the probability distribution of semantic categories. The cross-modality bridging dictionary is a matrix which bridges the columns of semantic categories and the rows of visual appearances, and it records the co-occurrences between the semantic set and the visual set. The processing of multi-task feature learning are forked column wisely. Each job tries to discover the co-occurrence and discriminative patterns within and between different concept categories. The objective function with the $Capped-\ell_1$ penalty outperforms the traditional convex function. The non-intuitive visual appearance can be encoded into an accurate semantic description using the visual-semantic dictionary.

Further, we introduce a knowledge-based semantic propagation to transfer the probability distributions for related categories, and this can boost the robustness of the final global image semantic description. In details, a categories transferring matrix is learnt to improve the generalization of image semantic description, and the comparison experiments shows this semantic propagation procedure can bring a significant performance promotion. As the flowchart shown in Fig. 2, our image understanding algorithm consists two parts: the first part is the learning of visual-semantic bridging weight, and the second is the learning of concept transferring matrix. After the above necessary parameters are learnt, the visual appearance feature of a given image is calculated using Bag-of-visual-words model. Then, the semantic representation is computed as the inner product of visual appearance feature and the visual-semantic bridging weight. Thirdly, the concept transferring matrix is used to transfer the weight to related categories, which improves the generalization capability of the semantic representation. The semantic representation can be further used by many applications, such as scene description, image classification and semantic image retrieving.

Our main contributions can be summarized as following,

- Cross-modality bridging dictionary is proposed to solve the image understanding, which characterizes the probability distribution of semantic categories for the visual appearances.
- Knowledge-based semantic propagation is introduced to transfer the probability distributions for related categories, which upgrade the robustness of the final global image semantic description.
- Experimental comparisons with state-of-the-art methods on four public datasets evaluate the effectiveness of the proposed method. Particularly, the performance of our approach on large scale image search outperforms the traditional shallow models and the deep models.

The following of the paper contains: Section II introduces

the related works about cross-modality modeling and knowledge transferring. Section III explains the explicit visual-semantic dictionary. Section IV introduces the dictionary learning process. Section V introduces the semantic distance metric and explains the learning of the transferring matrix. Section VI discusses the experimental results. At last, Section VII concludes the paper.

II. Related Works

A. Cross-modality Bridging

Cross-modality bridging is the job of pairing visual appearance with the most accurate concept, and it it becoming more challenging as the amount of the web images grows larger.many related works have been proposed to solve the problem of cross-modality bridging: (1) *latent topic model*, [13] propose a visual contexts based discriminative Latent Dirichlet Allocation framework; [7] introduce a cross-modality learning approach by jointly minimizing the distance between the mappings of query and image in the latent subspace, which can efficiently preserve the inherent structure in each original space. (2) *Middle-level themes learning*, [15] propose a representation using semantic multinomial (SMN) to model context; [16] further research on local category co-occurrences in the SMN. (3) *Distance metric learning*, [17] proposed crosscategory transfer learning for classification. (4) *Semantic classifier model*, [19] introduced object bank (OB), which encodes the visual appearance and relative location of objects in images. (5) *Deep learning*, [20] use the deep convolutional neural network to classify the images in the ImageNet. [22] use convolutional neural networks on image regions, recurrent neural networks on sentences to generate image region descriptions. Peng et al. [31]–[33] propose a serious of cross-modality analysis methods on base of deep learning, including cross-modal correlation learning, semi-supervised cross-media feature learning, etc. Although deep model based methods achieved great success on image understanding, the underlying mechanism of neural work is still not fully understood. The cross-modality bridging still needs further research.

B. Transfer Learning

Transfer learning [34] as a new machine learning paradigm has gained increasing attention lately For now, typical application of the transferring learning mainly contains text classification, image classification, emotional classification, coordination filtering and artificial intelligence planning and so on. Dai et al. [35] propose a coclustering based classification (CoCC) algorithm to learn from the in-domain and apply the learned knowledge to out-of-domain. Authors [36] also estimate the initial probabilities under a distribution Dl of one labeled data set, and then use an EM algorithm to revise the model for a different distribution Du of the unlabeled test data for text classification. Gu et al. [37] propose a multi-task clustering, which performs multiple related clustering tasks together and utilizes the relation of these tasks to enhance the clustering performance. [38] used lexical prior knowledge in the form of domain-independent sentiment-laden terms and domain-dependent unlabeled data to increace the effectiveness of real-world sentiment prediction tasks. Inimage processfields, Dai et al. [39] proposed a translated learning framework for classifying target data using data from another feature space. Zhu et al. [34] propose the heterogeneous transfer learning method for image classification. To bridge text documents and images, they use tagged images and create a semantic view for each target image by using collective matrix factorization technique. Raina et al. [40] present a new machine learning framework called "self-taught learning" for using unlabeled data in supervised classification tasks. Peng et al. [41] propose a novel hybrid transfer network for cross-modal common representation learning. Pan et al. [42] propose a new deep architecture of incorporating the transferred semantic attributes into the CNN plus RNN framework. In conclusion, more and more transfer learning work has been applied to various applications and has achieved remarkable results in the research. However, the existing algorithms can not meet the actual application requirements in the era of big data because of the algorithmic complexity and limited amount of data.

III. An Introduction about cross-modality bridging dictionary

Cross-modality bridging dictionary is a direct probability bridge between the visual appearance features and the semantic categories of images. In other words, visual appearances are interpreted into the probability distribution of semantic categories. The cross-modality bridging dictionary is a matrix which bridges the columns of semantic categories and the rows of visual appearances, and it records the co-occurrences between the semantic set and the visual set. The learning of the dictionary is formulated into a multi-task feature learning problem. The explicit semantic of an image can be calculated from its visual appearance using the cross-modality bridging dictionary. In this section, we explain the semantic categories and structure of the proposed cross-modality bridging dictionary.

A. Image Visual Representation Model

The BOV model is the mostly used method for image visual representation. It is originated from the *bag-of-words* method for information retrieving. BOV model firstly extracts a bunch of visual appearance descriptors from the local patches of the image and then computes a compact histogram representation, which can be used by further applications.

Recently, deep convolutional neural networks, such as Alexnet, GoogleNet, Inception, VGG, ResNet, is becoming the main way of feature learning. However, the features from the output of Deep Networks usually have high semantic than traditional features. The deep networks need more training dataset and stronger computing devices.The above demands bring a high threshold to new research. Moreover, these deep features are direct for the classification and recognition tasks, and are limited in the image search task. So here we still use the traditional visual appearances as the base description for further learning.

B. Cross-modality Bridging Dictionary

As shown in Fig. 1, the explicit cross-modality bridging dictionary records the co-occurrences between the semantic set and the visual set, where *k* denotes the local visual appearance count and *m* denotes the semantic category count. The dictionary entry is the co-occurrence weight of the corresponding visual appearance and semantic category.

For example, the corresponding $k \times m$ relations between each pair of the visual set $\boldsymbol{V}$ with *k* visual appearances and semantic set $\boldsymbol{S}$ with *m* categories can be learnt by the explicit cross-modality bridging dictionary. Each category in $\boldsymbol{S}$ has *k*-bin membership histogram and each visual appearance in $\boldsymbol{V}$ has *m*-bin membership distribution histogram from the columns and rows of the dictionary respectively. The learning procedure of the bridging dictionary is detailed in the next section.

IV. Cross-modality Bridging Dictionary with Multi-task Learning

The rows and columns of the explicit visual-semantic dictionary depict the relationship between semantic categories and visual appearances. The entries of the dictionary are filled with a multi-task learning algorithm which employs the $Capped-\ell_1$ penalty [24]. The co-occurring probability distribution between visual appearances and semantic categories is concurrently learnt by multiple tasks. Each task processes a batch of co-occurring visual appearances, which are learnt by the above algorithm automatically. Related tasks may share some common appearances while tasks from different groups have different batches of visual appearances. The amount of common visual patterns between related tasks is constrained by the penalty item. The following is the definition of the objective function.

$$\boldsymbol{J(D)} = \min_{\mathbf{D}} \left(\sum_{i=1}^{m} \frac{1}{mn_i} \parallel \mathbf{y}_i - X_i \mathbf{d}_i \parallel^2 + \gamma \sum_{j=1}^{k} \min(|\mathbf{d}^j|, \theta) \right)$$
$$subject\ to\ \mathbf{d}_{i,j} \geq 0, \forall i, j \tag{1}$$

where $\mathbf{X}_i \in \Re^{n_i \times k}$ is the feature description matrix of the *i*-th category and images, and each row is a sample with a *k*-dimensionality vector $\mathbf{y}_i \in \Re^{n_i}$ is the corresponding labels of images with the *i*-th category. n_i is the number of samples for the *i*-th category. The visual-semantic dictionary $\mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_m] \in \Re^{k \times m}$ actually consists of the weighted vector for the *m* categories. $\mathbf{d}^j$ is the *j*-th row of the dictionary $\mathbf{D}$. $\| \cdot \|$ denotes the ℓ_2-norm, and $| \cdot |$ denotes the ℓ_1-norm. The first term is the least square loss function that measures the quality of reconstruction. The second term is the regularization function, which constrains the complexity of weight matrix $\mathbf{D}$. The parameter $\gamma(>0)$ controls the balance between two terms, which also restricts the dictionary sparsity. $\theta(>0)$ is a thresholding parameter and controls the inter-impact of the dictionary.

Theoretically, the formulation of Equ. 2 is very difficult to solve as it's a non-convex problem. The optimization problem can be approximated using an iteration algorithm [28] detailed in Algorithm 1. The loss of the objective function will be gradually minimized as the iteration goes continuously, and finally reaches an acceptable convergence. The objective function in Equ. 2 consists of a differential loss term and a non-differential penalty term. The key sub-problem is solved with an iterative shrinkage algorithm [43] which builds a regularization of the linearized differentiable function part of the objective at each iteration.

Algorithm 1: Cross-modality Bridging Dictionary Learning

Initialize $\gamma_j^{(0)} = \gamma$;
for $\tau = 1, 2, \cdots$ **do**
  Let $\overline{\mathbf{D}}^{\tau}$ be a feasible solution of the dictionary, the global objective function can be transferred as following:

$$\min_{\mathbf{D} \in \Re^{k \times m}} \left(\ell(\mathbf{D}) + \gamma_j^{(\tau-1)} \sum_{j=1}^{k} \min(|\mathbf{d}^j|, \theta) \right) \tag{2}$$

  Let $\ell(\mathbf{D}) = \sum_{i=1}^{m} \frac{1}{mn_i} \parallel \mathbf{y}_i - X_i \mathbf{d}_i \parallel^2$, and $\gamma_j^{\tau} = \gamma I(|(\overline{\mathbf{D}}^{\tau})^j| < \theta)(j = 1, \cdots, k)$, where $(\overline{\mathbf{D}}^{\tau})^j$ is the *j*-th row of $\overline{\mathbf{D}}^{\tau}$ and $I(\cdot)$ denotes the {0,1} indictor function.
end

Firstly, the following functions are defined for simplification:

$$\begin{aligned} p(\mathbf{D}) &: \Re^{k \times m} \mapsto \Re_+^k, p(\mathbf{D}) = [|\mathbf{d}^1|, \cdots, |\mathbf{d}^k|]^T, \\ q(\mathbf{v}) &: \Re_+^k \mapsto \Re_+, q(\mathbf{v}) = \sum_{j=1}^{k} \min(v_j, \theta). \end{aligned} \tag{3}$$

where $[x]_+ = \max\{0, x\}$, Equ. 2 can be rewritten as,

$$\min_{\mathbf{D} \in \Re^{k \times m}} \{\ell(\mathbf{D}) + \gamma q(p(\mathbf{D}))\} \tag{4}$$

Assuming g^{τ} is a sub-gradient of $q(\mathbf{v})$ at $\mathbf{v} = p(\overline{\mathbf{D}}^{\tau})$, $\langle \cdot \rangle$ denotes the inner product, according to the definition of the sub-gradient, an upper bound of the objective function in Equ. 2 can also be obtained:

$$\ell(\mathbf{D}) + \gamma q(p(\mathbf{D})) \leq \ell(\mathbf{D}) + \gamma q(p(\overline{\mathbf{D}}^{\tau})) + \gamma \langle g^{\tau}, p(\mathbf{D}) - p(\overline{\mathbf{D}}) \rangle \tag{5}$$

From the above formulation, one sub-gradient of $q(\mathbf{v})$ at $\mathbf{v} = p(\overline{\mathbf{D}}^{\tau})$ can be computed:

$$g^{\tau} = [I(|(\overline{\mathbf{D}}^{\tau})^1| < \theta), \cdots, I(|(\overline{\mathbf{D}}^{\tau})^k| < \theta)]^T \tag{6}$$

Next, we define the conjugate function [44] of the concave function $q(\mathbf{v})$:

$$q^{\star}(\mathbf{u}) = \inf_{\mathbf{v}} \{\mathbf{u}^T \mathbf{v} - q(\mathbf{v})\} \tag{7}$$

The following theory can be deducted with theory of [45]:

$$q(\mathbf{v}) = \inf_{\mathbf{u}} \{\mathbf{v}^T \mathbf{u} - q^{\star}(\mathbf{u})\} \tag{8}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2903448, IEEE Transactions on Multimedia

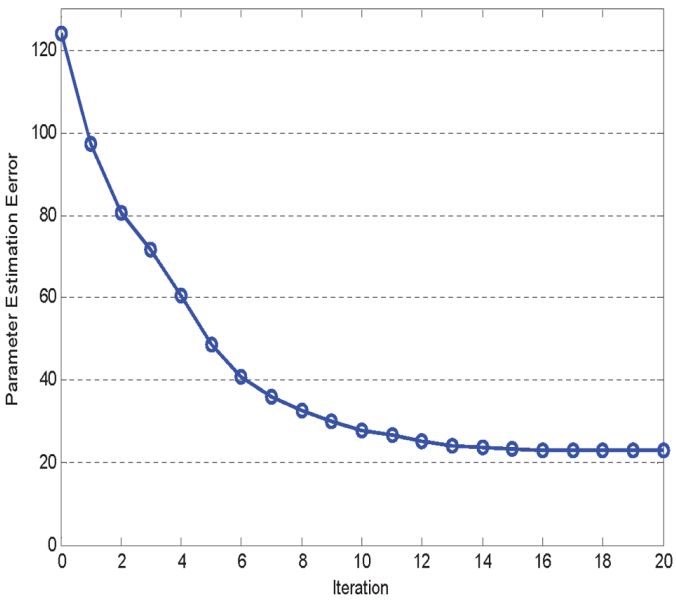

Fig. 3. Parameter estimation error $\|\bar{D} - D\|_{2,1}$.

Thus, the objective function can be rewritten as Equ. 4:

$$\min_{\mathbf{D},\mathbf{u}} \{\ell(\mathbf{D}) + \gamma \mathbf{u}^T p(\mathbf{D}) - \gamma q^\star(\mathbf{u})\} \tag{9}$$

The above minimazation problem can be solved with the block coordinate descent [46]:

- Fix $\mathbf{D} = \overline{\mathbf{D}}^\tau$:

$$\overline{\mathbf{u}}^\tau = \arg\min_{\mathbf{u}} \{\mathbf{u}^T p(\overline{\mathbf{D}}^\tau) - q^\star(\mathbf{u})\} \tag{10}$$

  According to the Danskin's Theory [45], one feasible solution of the above equation is the sub-gradient of $q(\mathbf{v})$ at $\mathbf{v} = p(\overline{\mathbf{D}}^\tau)$, i.e. $\overline{\mathbf{u}}^\tau = g^\tau$ in Equ. 6.

- Fix $\mathbf{u} = \overline{\mathbf{u}}^\tau = [I(|(\overline{\mathbf{D}}^\tau)^1| < \theta), \cdots, I(|(\overline{\mathbf{D}}^\tau)^k| < \theta)]^T$:

$$\overline{\mathbf{D}}^{(\tau+1)} = \arg\min_{\mathbf{D}} \{\ell(\mathbf{D}) + \gamma (\overline{\mathbf{u}}^\tau)^T p(\mathbf{D})\} \tag{11}$$

The parameter estimation error is shown in Fig. 3 under the setting $\gamma = 0.85 \times 10^{-3}$ and $\theta = 100 \times \gamma$. The convergence trend can also be seen from Fig. 3, and the error settles to a small value after about 16 iterations which means a convergence is reached.

The explicit visual-semantic dictionary $\mathbf{D} \in \Re^{k \times m}$ is learnt through the above algorithms. Any image i is firstly represented using the BOV model, and then marked as $x_i \in \Re^k$ where k is the dimensionality of the BOV representation. Using the inner product, the semantic representation of the image i can be computed,

$$SR(i) = x_i \cdot \mathbf{D} \tag{12}$$

V. Semantic Propagation

After calculating the semantic representations, the semantic distances between all images can be further computed. As the semantic descriptions of the images are a bunch of categories which are normally not independent, it is necessary to learn the transferring matrix between categories and to improve the generalization capability of the semantic descriptions.

Algorithm 2: Semantic Propagation

Input: $\langle x, y \rangle$: a pair of categories from the transferring matrix;
Φ_{isA}: the isA relationship of Probase;
Ψ: the synonym set in the Probase;

Output: The semantic category transferring matrix $\Omega \in \Re^{m \times m}$, Ω is a symmetric matrix, and m is the number of semantic set **S** in the transferring matrix;

```
for τ = 1, ··· , m do
    x=S(τ);
    Collect all the super-categories of x from Φ_isA as the set Υ^x;
    for μ = τ, ··· , m do
        y=S(μ);
        Collect all the super-categories of y from Φ_isA as the set Υ^y;
        According to the synonym set Ψ, let Υ_c = {Υ^x ∩ Υ^y} indicates the common super-categories set;
        Ω⟨x, y⟩=max{P(x|Υ_1^c) · P(y|Υ_1^c), ··· , P(x|Υ_n^c) · P(y|Υ_n^c)}, here, n = |r^c|_0;
    end
end
```

Firstly, the relations between the categories are modeled using a probabilistic knowledge base known as Probase [47], which records isA relations between semantic objects. The knowledge base is constructed from 2 years of Microsoft Bing search log and 1.68 billion web pages [48]. For example, "Camry is a car", where "Camry" is a subordinate category, and "car" is a superordinate category. In addition, Probase has other properties:

Conditional probability $P(x|z)$ and $P(z|x)$ are provided for each isA relation (x isA z) to measure the typicality, also known as typically scores, which derives from the co-occurrences:

$$P(x|z) = \frac{\text{occurrences of } (x, z) \text{ in Hearst extraction}}{\text{occurrences of } z \text{ in Hearst extraction}}$$

All possible superordinate categories for any category in the dictionary are searched in the Probase. Fig. 4 shows two examples "Cat" and "Dog", which provides the category distribution of text labels with basic-level conceptualization. Here, some common measures for conceptualization including P(c|e), MI, P(e|c), NPMI, PMI^k and BLC.

Then, for each category pair in the dictionary, max-pooling of probability is applied to the common super-category set which calculates the transferring weight between the category pair.

At last, the transferring matrix Ω is filled by the complete traversal of the above step. The whole process of semantic propagation is shown in Algorithm 2 . The transferring matrix measures the amount of relevancy between different

	Score by P(c|e)		Score by MI		Score by P(e|c)		Score by NPMI		Score by PMI^K		Score by BLC	
cat	animal	0.359	animal	0.303	house broken animal	0.1	pet	0.114	pet	0.197	pet	0.388
	pet	0.217	pet	0.276	nominal term	0.1	household pet	0.104	household pet	0.126	domestic animal	0.114
	domestic animal	0.088	domestic animal	0.105	domestic pet specie	0.1	companion animal	0.104	companion animal	0.123	companion animal	0.103
	mammal	0.067	domesticated animal	0.056	town animal	0.1	domestic pet	0.102	domestic pet	0.107	household pet	0.095
	specie	0.062	mammal	0.054	simple printing command	0.1	domestic animal	0.101	domestic animal	0.091	animal	0.087
	predator	0.056	companion animal	0.053	housebroken animal	0.1	house broken animal	0.096	house broken animal	0.078	domestic pet	0.073
	domesticated animal	0.047	predator	0.045	strict or true carnivore	0.1	domesticated animal	0.095	nominal term	0.072	domesticated animal	0.06
	companion animal	0.039	household pet	0.044	ground fault circuit interrupt	0.1	common household pet	0.095	visually similar word	0.07	introduced predator	0.028
	household pet	0.032	domestic pet	0.036	animal familiar	0.1	nominal term	0.095	common household pet	0.069	pet animal	0.026
	small animal	0.032	small animal	0.028	ordinary material particular	0.1	visually similar word	0.094	domesticated animal	0.066	common household pet	0.026
dog	animal	0.418	animal	0.37	average domestic pet	0.1	pet	0.113	pet	0.185	pet	0.333
	pet	0.188	pet	0.234	panama quarantine pet	0.1	companion animal	0.106	companion animal	0.143	domestic animal	0.136
	domestic animal	0.09	domestic animal	0.108	nicaragua quarantine pet	0.1	domestic animal	0.104	household pet	0.11	animal	0.134
	mammal	0.08	mammal	0.067	nonhuman animal reservoir	0.1	household pet	0.102	domestic animal	0.11	companion animal	0.117
	specie	0.063	domesticated animal	0.057	belize quarantine pet	0.1	domestic pet	0.101	domestic pet	0.105	household pet	0.074
	domesticated animal	0.048	companion animal	0.053	creation animal	0.1	domesticated animal	0.098	domesticated animal	0.078	domesticated animal	0.07
	companion animal	0.039	household pet	0.036	trotting quadrupedal mammal	0.1	average domestic pet	0.094	average domestic pet	0.07	domestic pet	0.066
	household pet	0.027	domestic pet	0.031	domestic companion	0.1	panama quarantine pet	0.094	panama quarantine pet	0.068	pet animal	0.024
	predator	0.024	specie	0.025	additionally certain animal	0.1	nicaragua quarantine pet	0.094	nicaragua quarantine pet	0.066	mammal	0.023
	domestic pet	0.023	carnivore	0.019	wilderness river	0.1	common household pet	0.094	nonhuman animal reservoir	0.064	common household pet	0.021

Fig. 4. The typical measures for conceptualization including P(c|e), MI, P(e|c), NPMI, PMI^k and BLC.

categories. So the semantic propagation re-ranks the semantic representation based on semantic distance, different from the traditional visual rank re-ranking model [49].

The primary image semantic representation is calculated using the Equ. 12, and then the category transferring matrix Ω is calculated through the semantic expansion. Finally, the image semantic description is calculated as follows.

$$Des(i) = SR(i) + SR(i) \cdot \Omega, \tag{13}$$

where the first term is the primary semantic description of the image i derived from the explicit visual-semantic dictionary, while the second term is from the transferring of semantic categories.

Our semantic representation describes the image using the probability distribution of categories, and the cosine function can be used to serve as the image semantic distance metric.

$$\begin{aligned} Sim^{cosine}(A,B) &= \frac{|A \cdot B|}{\|A\| \cdot \|B\|} \\ &= \frac{\sum_{i=1}^{m} A_i \times B_i}{\sqrt{\sum_{i=1}^{m}(A_i)^2} \times \sqrt{\sum_{i=1}^{m}(B_i)^2}} \end{aligned} \tag{14}$$

VI. Experiments

Our model is evaluated with public benchmarks on three traditional understanding tasks: large scale semantic image search, object and scene recognition.

Database: (1)ILSVRC2010 [50], which contains 1000 categories and 1,461,406 images. (2) Caltech-256 [51], which has 256 categories and 29,780 images with each category contains at least 80 images. (3) Scene-15 [52], has 15 categories and 4485 images with an average of 200-400 images for each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2903448, IEEE Transactions on Multimedia

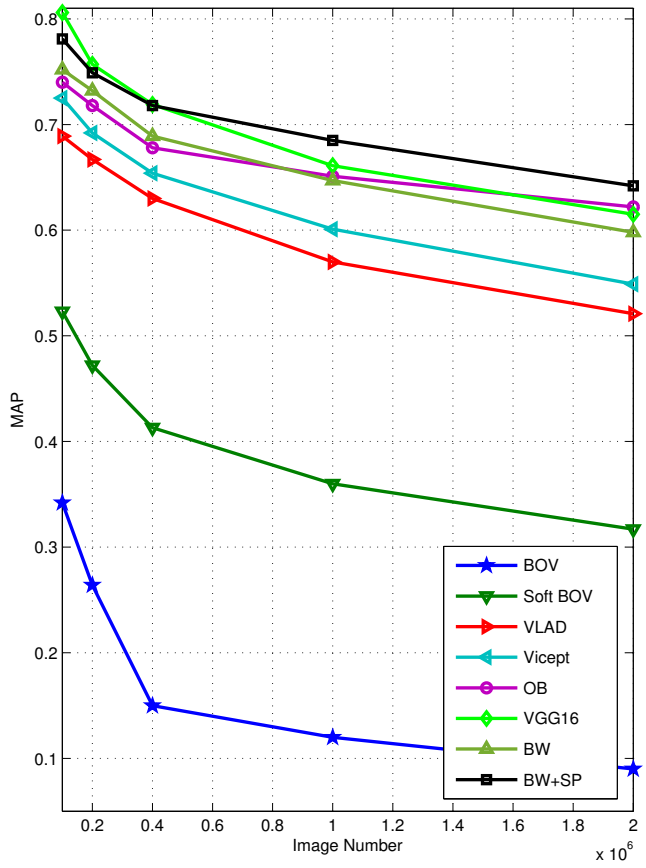

Fig. 5. Comparisons of different methods using MAP with different scale of image database.

category. (4) SUN397 [53], consists of 397 categories and 108762 images. (5) Flickr600k, which are collected from the Flickr to function as the distracters.

A. Large Scale Semantic Image Search

The effectiveness of our method is validated with two-million-image-size large scale databases (ILSVRC2010+Flickr600k).

Comparison Methods: (1) The BOV model [54] with 0.2 million visual words is set as the baseline. (2) SoftBOV [55] is a revised version of BOV with descriptors encoded with soft assignment of 4 nearest neighbors. (3)VLAD [56] derives from Fisher kernel and is a state-of-the-art method. Its parameter is set the same as [56]. (4) Vicept [57] adopts a mixed-norm regularization learning. (5) Object Bank (OB) [19] is a category detector using semantic descriptions. (6) The 16 layers VGG network [58] is pretrained on ImageNet dataset. We adopt the Adam to optimize the model in an end-to-end manner. Learning rate is scheduled as 10^{-3} and a staircase weight decay is applied after 10 epochs. β_1, β_2 in Adam are set to 0.8 and 0.999. We fix the VGG network at the beginning and fine-tune it after 20 epochs with a relatively small learning rate 10^{-5}. We set the length of candidate glimpse sequence $T = 10$ and the number of sampling times in emission indicator $K = 4$, which are learned through cross-validation. (7) Our bridging weight model using only explicit visual-semantic dictionary (BW) with parameters $\gamma = 0.85 \times 10^{-3}$ and $\theta = 100 \times \gamma$ obtained by the cross-validation. (8) Further add the semantic propagation to BW (BW+SP).

For image retrieving, the mean average precision (MAP) is chosen as the evaluation metric following [54]–[56]. Query images are 1000 representative images selected from ILSVRC2010. The precision-recall curve is drawn for each query and the average precision (AP) is calculated by summing the area under the curve. MAP is the average of APs from all the queries.

Fig. 5 shows the results of the comparison methods with different scale datasets from ILSVRC2010. Several observations can be found: first, our two methods (BW and BW+SP) outperform classical BOV and SoftBOV models with higher MAPs, which benefit from the strong cross-modality bridging dictionary learning and the robust knowledge-based semantic propagation. Second, compared to the state-of-the-art methods VLAD and OB, MAP values are also boosted to 9.3% and 2.1% by our methods, which can be attributed to the adoption of the multi-task learning to approximate the optimal solution. Third, compared with traditional semantic description (Object Bank), our proposed BW+SP has a significant improvement to prove its semantic representation power. Moreover, compared with the popular deep feature VGG16, our method is not as good as VGG16 in the small datasets, but still has a better performance in the large scale dataset (the number of images is more than 1 million). In the real-world applications, the processing ability on the large scale data plays the more important role. Four, through the comparison of BW and BW+SP, the consideration of the relevancy between categories benefits the description and the semantic propagation improves the performance. Fifth, the retrieval performance decreasing rate of our methods is slower than other approaches as the image number scales, which means that the proposed model is robust and scalable.

B. Object Recognition on Caltech-256

Comparisons Methods: (1) Binary SVM (BSVM) [59] is a one-v.s.-all classification model and we train the SVM classifiers for each categories with 50 positive images and two hundred negative image; (2) TinyImage [60] is based on nearest neighbor voting. The images are firstly down-sampled to 32×32, and the votes are gathered from the top-100 nearest neighbors; (3) Supervised Multi-class Labeling (SML) [61] uses the original setting. Images are represented by the bags of localized features. Gaussian mixture model consists of 64 separately trained components; (4) KSPM [51], spatial pyramid matching with kernel SVM; (5) Object Bank (OB) [19]; (6) LSS [62], low-dimensional semantic spaces with weak supervision; (7) S3R [63], sub-semantic space representation; (8) SMN [15]; (9) SR-LSR [14]; (10)FV [18], 128K-dimension Fisher Vector, where we use only SIFT descriptors and the case where we use both SIFT and LCS descriptors (again with a

TABLE I
COMPARISON WITH RELATED WORK ON THE CALTECH-256

Methods	Average Precision
BSVM [59]	0.290
TinyImage [60]	0.235
SML [61]	0.355
KSPM [51]	0.328
ObjectBank [19]	0.391
LSS [62]	0.334
S3R [63]	0.435
SMN [15]	0.394
SR-LSR [14]	0.482
LDR [21]	0.456
FV(SIFT) [18]	0.474
FV(SIFT+LCS) [18]	0.494
BW	0.477
BW+SP	**0.495**

TABLE II
COMPARISON WITH RELATED WORK ON THE SUN397

Dataset	Methods	Accuracy(%)
SUN397	BSVM [59]	31.2
	ObjectBank [19]	37.6
	LSS [62]	34.4
	SR-LSR [14]	44.6
	EMFS [16]	40.7
	LDR [21]	42.6
	FV(SIFT) [18]	43.3
	FV(SIFT+LCS) [18]	47.2
	VGG16 [58]	48.2
	VGG16+SP	**49.1**
	BW	43.6
	BW+SP	48.3

simple weighted linear combination); (11)LDR [21], which is trained on 1.2 million images of ImageNet; (12) The proposed method, BW and BW+SP.

Table I lists the average precision of different approaches on the Caltech-256 database. We can find the following conclusions, first, the proposed BW+SP approach outperforms the KSPM [51] by about 6.7%, which demonstrates the usefulness of visual-semantic bridging. Second, compared with BW, which is without the knowledge-based semantic propagation, the BW+SP method has a further precision improvement. This means the semantic propagation is necessary. Third, BW+SP model outperforms the state-of-the-art semantic-based method SR-LSR [14], which justifies the superiority of our method. Four, compared with the traditional state-of-the-art visual feature FV(SIFT) and FV(SIFT+LCS) [18], our method is comparable with these, which benefit from the use of semantic propagation. Fifth, our BW+SP approach has a 3.9% performance promotion than the deep network LDR [21], and this verify the effectiveness of the proposed method.

C. Scene recognition on SUN397

Scene recognition is an important image understanding task, and it can be used in plenty of real-world applications, such as automatic driving. The performance of scene recognition on SUN397 is evaluated in this subsection.

Comparisons Methods: (1) Binary SVM (BSVM) [59]; (2) Object Bank (OB) [19]; (3) LSS [62]; (4) SR-LSR [14]; (5) EMFS [16]; (6) LDR [21] ;(7)FV [18], where we use only SIFT descriptors and the case where we use both SIFT and LCS descriptors; (8) The proposed method, BW and BW+SP.

The result of scene recognition on the SUN397 dataset is shown in Table II. Three observations can be found as follows, first, compared with the classical semantic-based description ObjectBank [19] and SR-LSR [14], our method has a 10.7% and3.7% accuracy improvement, which show the strong semantic representation power of the proposed method. Second, compared with the traditional state-of-the-art visual feature FV(SIFT), FV(SIFT+LCS) [18] and the BW method (without semantic propagation) , our method also achieve a better performance, which benefit from the use of semantic propagation. Third, our BW+SP approach has a accuracy promotion than the deep network LDR [21] , and it has a comparable performance with the deeper network VGG16 [58], thus this verify the robust semantic consistency representation of our model. Moreover, we also extend our semantic propagation (SP) to VGG16 network, which brings a 0.9% performance improvement. This also prove the significance and expandability of the proposed semantic propagation.

D. Scene recognition on Scene15

In this subsection we evaluate the performance of different methods on Scene15 at the scene recognition task.

Comparisons Methods: (1) Binary SVM (BSVM) [59]; (2) KSPM [51]; (3) Object Bank (OB) [19]; (4) LSS [62]; (5) SMN [15]; (6) SR-LSR [14]; (7) EMFS [16]; (8) LDR [21];(9) VGG16 [58] (10) The proposed method, BW and BW+SP (the BW with semantic propagation).

TABLE III
COMPARISON WITH RELATED WORK ON THE SCENE15

Dataset	Methods	Accuracy(%)
Scene15	BSVM [59]	68.8
	KSPM [51]	80.4
	ObjectBank [19]	80.9
	LSS [62]	72.1
	SMN [15]	71.7
	SR-LSR [14]	86.1
	EMFS [16]	85.7
	LDR [21]	84.2
	VGG16 [58]	**88.4**
	BW	87.2
	BW+SP	87.5

The result of scene recognition on the Scene15 is shown in Table III. Through the comparisons, we can find the following views. First, compared with the semantic-based description methods, such as ObjectBank [19] and SR-LSR [14], our method has a better accuracy improvement, which show the strong semantic representation power of the proposed method. Second, compared with the traditional multi-task learning methods, KSPM [51], LSS [62] and SMN [15], our BW+SP model has a at least 6.6% performance promotion, and this prove the effectiveness of the multi-task learning with Capped penalty. Third, compared with the BW method (without semantic propagation) , our BW+SP method achieve

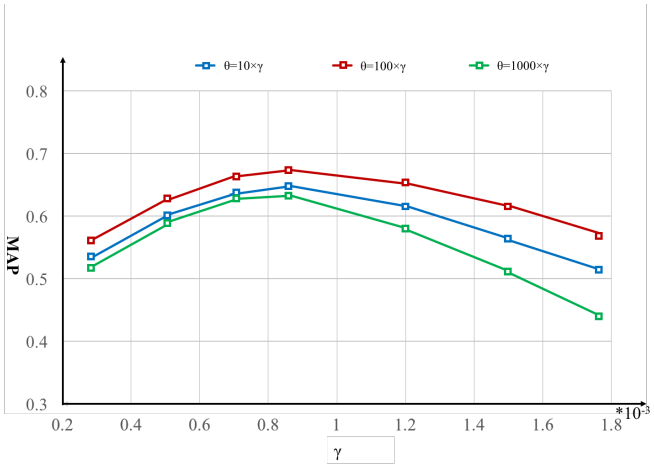

Fig. 6. Paremeter discussion of γ and θ at the 1 million image search task.

a near accuracy, which indicates that there is weak semantic relationship among different scenes in the Scene15. Four, compared with the deep networks, i.e. LDR [21] and VGG16 [58], our method has a higher accuracy than the LDR [21], but it is not as good as the VGG16 method. This results from the deeper network structure of VGG16.

E. Parameter Discussion

Here we discuss the impact of parameters γ and θ in the Equ. 2 in the large scale image search (1 million images). For the γ, we evaluate the different values, while for the θ, we set $\theta = 10 \times \gamma$, $\theta = 100 \times \gamma$ and $\theta = 1000 \times \gamma$ according to the experiences. Fig. 6 shows the variation tendency of MAP under the different γ and θ. First, as to the γ, With it becomes larger, the MAP first has some promotion, and then begins to weaken. Moreover, the above variation tendency keep similar under different θ. Second, as to the θ, according to the experiences, the θ usually is set to the times of γ. We can find the MAP is better when the $\theta = 100 \times \gamma$, and when the $\theta = 1000 \times \gamma$, the performance declines very much.

VII. Conclusion

This paper introduces an explicit visual-semantic dictionary model for cross-modality understanding. where images are represented as the probability distribution of semantic categories. The dictionary is learnt with multi-task learning, which models the co-occurrence and discriminative patterns within and between categories. Further, we propose a knowledge-based semantic propagation to improve the generalization capability of the semantic representation as the scale goes larger.

References

[1] T. Rui, P. Cui, and W. Zhu, "Joint user-interest and social-influence emotion prediction for individuals," *Neurocomputing*, vol. 230, pp. 66–76, 2017.

[2] L. Sun, X. Wang, Z. Wang, H. Zhao, and W. Zhu, "Social-aware video recommendation for online social groups," *IEEE Transactions on MultiMedia*, vol. 19, no. 3, pp. 609–618, 2017.

[3] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Transactions on MultiMedia*, vol. 19, no. 3, pp. 598–608, 2017.

[4] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multitask shared sparse regression," *IEEE Transactions on MultiMedia*, vol. 19, no. 3, pp. 632–645, 2017.

[5] Y. Wu, N. Cao, D. Gotz, Y.-P. Tan, and D. A. Keim, "A survey on visual analytics of social media data," *IEEE Transactions on MultiMedia*, vol. 18, no. 11, pp. 2135–2148, 2016.

[6] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Transactions on MultiMedia*, vol. 19, no. 6, pp. 1234–1244, 2017.

[7] Y. Pan, T. Yao, T. Mei, H. Li, C.-W. Ngo, and Y. Rui, "Click-through-based cross-view learning for image search," in *ACM SIGIR*, 2014, pp. 717–726.

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

[9] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE TIP*, vol. 25, no. 11, pp. 5427–5440, 2016.

[10] M. Shao and Y. Fu, "Cross-modality feature learning through generic hierarchical hyperlingual-words," *IEEE Transactions on NEURAL NETWORKS AND LEARNING SYSTEMS,*, vol. 28, no. 2, pp. 451–464, 2017.

[11] J. Wu, S. Zhao, V. S. Sheng, J. Zhang, C. Ye, P. Zhao, and Z. Cui, "Weak-labeled active learning with conditional label dependence for multilabel image classification," *IEEE Transactions on MultiMedia*, vol. 19, no. 6, pp. 1156–1168, 2017.

[12] A. Chadha and Y. Andreopoulos, "Voronoi-based compact image descriptors: Efficient region-of-interest retrieval with vlad and deep-learning-based descript," *IEEE Transactions on MultiMedia*, vol. 19, no. 7, pp. 1596–1608, 2017.

[13] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Context aware topic model for scene recognition," in *IEEE CVPR*, 2012, pp. 2743–2750.

[14] X. Li and Y. Guo, "Latent semantic representation learning for scene classification," in *ICML*, 2014, pp. 1–8.

[15] N. Rasiwasia and N. Vasconcelos, "Holistic context models for visual recognition," *IEEE Trans. on PAMI*, vol. 34, no. 5, pp. 902–917, 2012.

[16] X. Song, S. Jiang, and L. Herranz, "Joint multi-feature spatial context for scene recognition in the semantic manifold," in *IEEE CVPR*, 2015, pp. 1312–1320.

[17] G. Qi, C. Aggarwal, Y. Rui, Q. Tian, S. Chang, and T. Huang, "Towards cross-category knowledge propagation for learning visual concepts," in *IEEE CVPR*, 2011, pp. 897–904.

[18] J. Snchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *IJCV*, vol. 105, no. 3, p. 222?245, 2013.

[19] L. Li, H. Su, Y. Lim, and F. Li, "Object bank: An object-level image representation for high-level visual recognition," *IJCV*, vol. 107, no. 1, pp. 20–39, 2013.

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

[21] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014, pp. 487–495.

[22] K. Andrej and F. Li, "Deep visual-semantic alignments for generating image descriptions," in *IEEE CVPR*, 2015, pp. 664–676.

[23] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[24] T. Zhang, "Multi-stage convex relaxation for feature selection," in *Bernoulli*, 2012, pp. 2277–2293.

[25] J. Ye and J. Liu, "Sparse methods for biomedical data," *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 15, 2012.

[26] L. Li, S. Q. Jiang, and Q. M. Huang, "Learning hierarchical semantic description via mixed-norm regularization for image understanding," *IEEE Transactions on MultiMedia*, vol. 14, no. 5, pp. 1401–1413, 2012.

[27] Y. Guo, "Convex subspace representation learning from multi-view data," in *Proceedings of the Association for the Advancement of Artificial Intelligence(AAAI)*, 2013, pp. 387–393.

[28] P. Gong, J. Ye, and C. Zhang, "Multi-stage multi-task feature learning," *JMLR*, vol. 14, pp. 2979–3010, 2013.

[29] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *Proceedings of the Association for the Advancement of Artificial Intelligence(AAAI)*, 2015, pp. 2750–2756.

[30] D. Wang, P. Cui, M. Ou, and W. Zhu, "Learning compact hash codes for multimodal representations using orthogonal deep structure," *IEEE Transactions on MultiMedia*, vol. 17, no. 9, pp. 1404–1416, 2015.
[31] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multi-grained fusion by hierarchical network," *IEEE TMM*, vol. 20, no. 2, pp. 405–420, 2018.
[32] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges," *IEEE TCSVT*, 2017.
[33] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, "Semi-supervised cross-media feature learning with unified patch graph regularization," *IEEE TCSVT*, vol. 26, no. 3, pp. 583–596, 2016.
[34] Y. Zhu, Y. Chen, and Z. Lu, "Heterogeneous transfer learning for image classification," in *AAAI*, 2011, pp. 1304–1309.
[35] W. Dai, G. rong Xue, and Q. Yang, "Co-clustering based classification for out-of-domain documents," in *ACM KDD*, 2007, pp. 210–219.
[36] W. Dai, G.-R. Xue, and Q. Yang, "Transferring naive bayes classifiers for text classification," in *AAAI*, 2007, pp. 540–545.
[37] Q. Gu and J. Zhou, "Learning the shared subspace for multi-task clustering and transductive transfer classification," in *IEEE ICDM*, 2009, pp. 159–168.
[38] T. Li, Y. Zhang, and V. Sindhwani, "A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge," in *ACL*, 2009, pp. 244–252.
[39] D. W, C. Y, and X. G. R, "Translated learning: transfer learning across different feature spaces," in *NIPS*, 2008, pp. 353–360.
[40] R. R, B. A, L. H, P. B, and N. AY, "Self-taught learning: Transfer learning from unlabeled data," in *ICML*, 2007, pp. 759–766.
[41] X. Huang, Y. Peng, and M. Yuan, "Cross-modal common representation learning by hybrid transfer network," in *IJCAI*, 2017, pp. 1893–1900.
[42] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *IEEE CVPR*, 2017.
[43] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
[44] R. Rockafellar, "Convex analysis," *Princeton University Press (Princeton, NJ)*, 1970.
[45] D. Bertsekas, "Nonlinear programming," *Athena Scientific*, 1999.
[46] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear gauss-seidel method under convex constraints," *Operations Research Letters*, vol. 26, no. 3, pp. 127–136, 2000.
[47] W. Wu, H. Li, H. Wang, and K. Zhang, "Probase: a probabilistic taxonomy for text understanding," in *SIGMOD*, 2012, pp. 481–492.
[48] M. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *COLING*, 1992, pp. 539–545.
[49] Y. Jing and S. Baluja, "Visualrank: Applying page-rank to large-scale image search," *IEEE Trans. PAMI*, vol. 30, no. 11, pp. 1877–1890, Nov 2008.
[50] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009, pp. 248–255.
[51] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," in *Technical Report,CalTech*, 2009.
[52] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features:spatial pyramid matching for recognizing natural scene categories," in *IEEE CVPR*, 2006, pp. 2169–2178.
[53] J. Xiao, J. Hayes, K. Ehringer, A. Olivia, and A. Torralba, "Sun database: Large scale scene recognition from abbey to zoo," in *IEEE CVPR*, 2010, pp. 3485–3492.
[54] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE CVPR*, 2006, pp. 2161–2168.
[55] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *IEEE CVPR*, 2008, pp. 1–8.
[56] H. J*é*gou, M. Douze, C. Schmid, and *é*rez P., "Aggregating local descriptors into a compact image representation," in *IEEE CVPR*, 2010, pp. 3304–3311.
[57] L. Li, S. Jiang, and Q. Huang, "Learning hierarchical semantic description via mixed-norm regularization for image understanding," *IEEE Trans. on Multimedia*, vol. 14, no. 5, pp. 1401–1413, 2012.
[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[59] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.
[60] A. Torralba, R. Fergus, and W. Freeman, "80 million tiny images: a large dataset for nonparametric object and scene recognition," *IEEE Trans. PAMI*, vol. 30, no. 11, pp. 1958–1970, Sep 2008.
[61] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. PAMI*, vol. 29, no. 3, pp. 394–410, 2007.
[62] N. Rasiwasia and N. Vasconcelos, "Scene classification with low-dimensional semantic spaces and weak supervision," in *IEEE CVPR*, 2008, pp. 1–6.
[63] C. Zhang, J. Cheng, J. Liu, J. Pang, C. Liang, Q. Huang, and Q. Tian, "Object categorization in sub-semantic space," *Neurocomputing*, vol. 142, pp. 248–255, 2014.