Large Scale Image Understanding with Non-convex Multi-task Learning

Liang Li*, Chenggang Yan†, Xing Chen†††, Shuqiang Jiang§, Seungmin Rho ‡, Jian Yin ‡‡, Baochen Jiang ‡ and Qingming Huang*††

*University of Chinese Academy of Sciences, Beijing, China
†Department of Automation, Tsinghua University, Beijing, China
‡National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences(CAS), Beijing, China
††Academy of Mathematics and Systems Science, CAS, Beijing, China
§Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
‡‡Department of Multimedia, Sungkyul University, Korea
∥Department of Computer, Shandong University, Weihai, China

Abstract—Large scale image understanding is drawing more and more attention from the researchers and industry. Inspired by the game theory and machine learning algorithm, this paper proposes a semantic dictionary to solve the key problem of visual polysemia and concept polymorphism in the large scale image understanding. The semantic dictionary characterizes the probability distribution between visual appearances and semantic concepts, and the learning of semantic dictionary is formulated into a minimization problem of the payoffs, where the players adjudge their strategies (i.e. the probability distribution) at each iteration. Non-convex multi-task learning is introduced to solve the above optimization problem. Finally, the wide applications of semantic dictionary are validated in our experiments, including the large scale semantic image search and image annotation.

Index Terms—Game theory, machine learning, semantic dictionary, payoffs minimization, large scale systems

I. INTRODUCTION

With the rapid development of internet and multimedia technology, large scale image understanding becomes a hot research topic recently due to its wide applications in our daily life. However the phenomenon of Visual Polysemia and Concept Polymorphism (VPCP) has still been a great challenge in this area. Visual polysemia reveals the fact that one certain visual appearance may have different semantic explanations, and concept polymorphism represents the truth that one concept may have many visual appearances under different examples. Particularly in web-scale conditions, there exist the more complex connections between visual appearances and semantic concepts so that the VPCP problem becomes graver: On the hand of VP, one visual appearance may occur in thousands of web concepts so that it is extremely difficult to infer its exact semantic; On the other hand of CP, one concept has various of different instances, where there are diverse visual appearances. In one word, the VPCP problem is a very complex and important issue in the large scale environment.

To solve the above problem, researchers mainly proposed their approaches at the perspective of the multimedia and computer vision, including image classification, image annotation, object and scene recognition, and image search, etc. In details, Weinberger et al. [1] proposed the large margin nearest neighbor classification on the base of distance metric learning model. Boiman et al. [2] introduced an Image-To-Class distance metric learning method for image classification by learning per-class Mahalanobis metric. Qi et al. [3] studied a technology for cross-category transfer learning for the classification task. Bucak et al. [4] introduced an algorithm for multi-label multiple kernel learning to recognize the objects. But none of them solve the problem of VPCP directly, either the VP problem or the CP problem. One main reason is that the relationship between image visual appearances and semantic information has not been individualized.

In the last few years, game theory [5]–[8] is widely used in many hot domains, such as economics, biology, computer science philosophy and so on. Inspired by the incomplete static game theory [9]–[11] and machine learning algorithm [12]–[14], we model a semantic dictionary (Fig. 1) to solve the problem of visual polysemia and concept polymorphism. In the viewpoint of the mathematics, this dictionary is a matrix where each column depicts the relationship between one concept with all the visual appearances while each row represents the relationship between one visual appearance with all the concepts. In the viewpoint of the game theory, each column can be regarded as one “player”, the ”strategy” of each player

Fig. 1. The semantic dictionary built between visual appearances and semantic concepts.
is how to assign the weights which reveal the relationships with every visual appearance. The objective function is to minimize the “payoffs” (the reconstruction loss) when the players ceaselessly take the different strategy about the semantic dictionary.

In short, this paper learns a semantic dictionary for large scale image understanding, which is to characterize the membership distribution between each visual appearance and each semantic concept. With its help, the images can be represented into a description of the intuitive semantic, rather than the incomprehensible visual information. Inspired by the game theory, we propose a new respective for modeling the semantic dictionary, i.e. the learning of semantic dictionary is formulated into the payoffs minimization problem. For solve this optimization, we introduce a non-convex multi-task learning algorithm to formulate the payoffs. In this algorithm, the players need to adjudge their strategy at each iteration along a sub-gradient direction, and the convergence guarantees the payoffs to achieve the minimal payoffs. Finally, the wide applications of semantic dictionary are validated in our experiments, including large scale semantic image search and image annotation.

The rest of this paper is organized as follows: Section II introduces the visual appearance representation method, and interprets the semantic dictionary. Section III details the learning procedure of semantic dictionary and represents the interprets the semantic dictionary. Section IV shows the experimental results of different tasks on both the standard benchmark and the large scale image database. Finally, Section V concludes the ideas of this paper.

II. SEMANTIC DICTIONARY

As mentioned above, semantic dictionary is designed to bridge the image visual appearances and the semantic concepts. With its help, the images can be represented into a description of the semantic, rather than the visual information. In this section, firstly, we introduce the popular image visual representation bag-of-visual-words (BOV) model [15], and then represent an intuitive interpretation about semantic dictionary, such as its structure, concept set source, etc.

A. Image Visual Representation Model

The de-facto image visual representation method in the multimedia domain is based on the BOV model [15], which is motivated by the bag-of-words from the information retrieval domain. In this model, an image is characterized as a collection of visual appearance descriptors, which are extracted from local patches and quantized into discrete visual words, and then a compact histogram representation is computed for further image applications.

The classical BOV model is based on the $k$-means algorithm. Let $X = [x_1; \ldots; x_n] \in \mathbb{R}^{n \times d}$, depicts a collection of $n$ visual descriptors, and each descriptor is a feature vector with $d$-dimensionality, e.g. the most popular local descriptor-SIFT [16]. The $k$-means algorithm is to minimize the reconstruction error:

$$\min_B \left( \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|x_i - b_j\|^2 \right)$$ (1)

where $B = [b_1; \ldots; b_k] \in \mathbb{R}^{k \times d}$ is called dictionary, where there are $k$ clustering centers and every cluster center is a visual word. $\| \cdot \|$ is the $\ell_2$-norm. The formulation of Equ. 1 can be transferred to a matrix factorization problem:

$$\min_{A,B} \sum_{i=1}^{n} \min_{a_i} \|x_i - a_iB\|^2$$ (2)

subject to $\|a_i\|_0 = 1, \|a_i\|_1 = 1, a_i \geq 0, \forall i$

where $[a_1; \ldots; a_m] \in \mathbb{R}^{n \times k}$ is the cluster membership indicators. $\|a_i\|_0 = 1$ is a cardinality constraint, meaning that only one element of $a_i$ is nonzero, $a_i > 0$ means the nonnegative property of $a_i$, and $\|a_i\|_1$ is the $\ell_1$-norm, the sum of the absolute value of each element in $a_i$. After the optimization, the index of the only nonzero element in $a_i$ depicts the corresponding visual word of $x_i$.

The BOV model has become the most popular image representation model in the multimedia domain because of its several advantages: firstly, as a result of the local salient and the invariant information of rotation and scale, the visual words are very discriminating. Secondly, the BOV model can provide a compact and discriminative description with a collection of visual words, so that it is easy to be stored and searched. Finally, based on the BOV, the distance among images can be computed quickly through some fast and simple operators, e.g. dot-product.

B. Semantic Dictionary Description

As illustrated in Fig. 1, semantic dictionary represents the between concept set and visual set, where $k$ is the number of local visual appearances in the whole dictionary and $m$ is the number of semantic concepts. The probability value in the Fig. 1 denotes the degree of the relationship between the corresponding visual appearance and the corresponding concept. To efficiently make use of this structure, we detail the semantic dictionary:

1) **Visual Set:** local visual descriptor is used to represent the image. In details, SIFT [16], a robust local feature with invariant ration and scale, is located by Difference of Gaussian and consists of 128-dimensionality histogram. Finally, SIFT is quantized into visual words by hierarchical $k$-means algorithm.

2) **Concept Set:** The concepts in real world are not independent but closely related. To provide the more discriminative semantic concepts and cover the wider objects and scenes, we follow the structure in [17], which is the most current image understanding database in the multimedia and computer vision domains. Here, we simplify the original concept structure with a flat representation.
Before learning the semantic dictionary, a short interpretation is represented as follows. Suppose having a visual dictionary $VD$ with $k$ visual words and a semantic concept collection $SC$ with $m$ concepts, a $k \times m$ membership distribution (semantic dictionary) can be jointly learned between each concept and each visual appearance. In other words, each concept in $SC$ has a corresponding $k$-bin membership distribution histogram (i.e. each player makes the strategy about the $k$ weights with each visual appearance in $VD$), and each visual appearance in $VD$ has a corresponding $m$-bin membership distribution histogram.

III. SEMANTIC DICTIONARY LEARNING WITH NON-CONVEX MULTI-TASK LEARNING

Semantic dictionary is to represent the typical relationship between visual appearances and semantic concepts, where each column can be regarded as one "player", the "strategy" of each player is how to assign the weights which reveal the relationships with every visual appearance. The objective function is to minimize the "payoffs" when the players ceaselessly take the different strategy about the semantic dictionary. A non-convex multi-task learning algorithm is introduced to formulate the payoffs. The learning procedure can be formulated into the following payoffs optimization problem:

$$
J(D) = \min_{D} \left( \sum_{i=1}^{m} \frac{1}{mn_i} \| y_i - X_i d_i \|^2 + \gamma \sum_{j=1}^{k} \min(|d_j|, \theta) \right) 
$$

subject to $W_{i,j} \geq 0, \forall i,j$ (3)

where $X_i \in \mathbb{R}^{n_i \times k}$ is the feature description matrix of images with the $i$-th concept, and each row is a sample with a $k$-dimensionality vector $y_i \in \mathbb{R}^{n_i}$ is the corresponding labels of images with the $i$-th concept. $n_i$ is the number of samples for the $i$-th concept. The matrix $D = [d_1, \cdots, d_m] \in \mathbb{R}^{k \times m}$ is the semantic dictionary, which actually consists of the weighted vector for the $m$ concepts. $d_j$ is the $j$-th row of the semantic dictionary $D$. $\| \cdot \|$ denotes the $\ell_2$-norm, and $| \cdot |$ denotes the $\ell_1$-norm. The first term is the least square loss function, which is often used in the machine learning literature, and it measures the reconstruction quality. The second term is the regularization penalty, which measures the complexity of semantic dictionary $D$. The parameter $\gamma (> 0)$ balances the effects of these two terms, and restricts the sparsity of semantic dictionary. The parameter $\theta (> 0)$ is a titleholding parameter, and it controls the inter-impact among the players.

Because the above formulation of Equ. III is a non-convex problem, it is very difficult to solve. We introduce an iteration algorithm to solve the optimization problem, which is detailed in Algorithm 1. In the view of the game theory, the total payoffs of the objective function will become smaller and smaller as the iteration is carried out continuously, and finally it converges to a fixed value. In essence, at each iteration, the players need to adjudge their strategy along a sub-gradient direction. The objective function in Equ. 4 can be decomposed into a differential loss term and a non-differential penalty term.

### Algorithm 1: Semantic Dictionary Learning

1. Initialize $\gamma_j^{(0)} = \gamma_j$.
2. for $\tau = 1, 2, \cdots$ do
3. Let $D^\tau$ be a good strategy of the players, the global payoffs can be transferred as following:
4. $$
\min_{D \in \mathbb{R}^{k \times m}} \left( \ell(D) + \gamma_j^{(\tau - 1)} \sum_{j=1}^{k} \min(|d_j|, \theta) \right) 
$$
5. Let $\ell(D) = \sum_{i=1}^{m} \frac{1}{mn_i} \| y_i - X_i d_i \|^2$, and $\gamma_j = \gamma I(|(D^\tau)_j| < \theta)(j = 1, \cdots, k)$, where $(D^\tau)_j$ is the $j$-th row of $D^\tau$ and $I(\cdot)$ denotes the $\{0, 1\}$ indicator function.
6. end

Then the iterative shrinkage algorithm [18] is introduced to solve this key sub-problem. The basic idea of this algorithm is to build a regularization of the linearized differentiable function part of the objective at each iteration.

Firstly, we define the following functions for simplifying the notations:

$$
p(D) : \mathbb{R}^{k \times m} \mapsto \mathbb{R}^k, p(D) = |d_1|, \cdots, |d_k| \mathrm{T},
$$

$$
q(v) : \mathbb{R}_+ \mapsto \mathbb{R}_+, q(v) = \sum_{j=1}^{k} \min(v_j, \theta).
$$

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

$$
\min_{D \in \mathbb{R}^{k \times m}} \left\{ \ell(D) + \gamma q(p(D)) \right\}
$$

Assuming $g^\tau$ is a sub-gradient of $q(v)$ at $v = p(D^\tau)$, according to the definition of the sub-gradient, an upper bound of $q(p(D))$ can be constrained under the linear approximation:

$$
q(p(D)) \leq q(p(D^\tau)) + \langle g^\tau, p(D) - p(D^\tau) \rangle
$$

where $\langle \cdot \rangle$ denotes the inner product. When the ideal strategy $D^\tau$ at the $\tau$-th iteration has been assigned, an upper bound of the payoffs in Equ. 6 can also be obtained:

$$
\ell(D) + \gamma q(p(D)) \leq \ell(D) + \gamma q(p(D^\tau)) + \gamma |g^\tau|, p(D) - p(D^\tau) \rangle
$$

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

$$
g^\tau = [I((D^\tau)_1| < \theta), \cdots, I(|(D^\tau)_k| < \theta)] \mathrm{T}
$$

Next, we define the conjugate function [12] of the concave function $q(v)$:

$$
q^*(u) = \inf_v \left\{ u^T v - q(v) \right\}
$$

Based on the theory of Bertsekas at al. [13], the following theory can be obtained:

$$
q(v) = \inf_u \left\{ v^T u - q^*(u) \right\}
$$
Thus, we can rewritten the objective function of Equ. 6:

$$
\min_{D,u}\{\ell(D) + \gamma u^T p(D) - \gamma q^*(u)\}
$$

The above minimazation problem can be solved with the block coordinate descent [14]:

- Fix $D = \tilde{D}$:
  $$
  \tilde{u}^T = \arg\min_u\{\ell(\tilde{D}) + \gamma u^T p(\tilde{D}) - \gamma q^*(u)\} = \arg\min_u\{u^T p(\tilde{D}) - q^*(u)\}
  $$

According to the Danskin’s Theory [13], one feasible solution of the above equation is the sub-gradient of $q(v)$ at $v = p(\tilde{D})$, i.e. $\tilde{u}^T = g^T$ in Equ. 9.

- Fix $u = \tilde{u}$ as $\tilde{u}^T = [I(|(\tilde{D}^T)^1| < \theta), \ldots, I(|(\tilde{D}^T)^k| < \theta)]^T$:
  $$
  \tilde{D}^{(\tau+1)} = \arg\min_D\{\ell(D) + \gamma (\tilde{u})^T p(D) - \gamma q^*(\tilde{u})\} = \arg\min_D\{\ell(D) + \gamma (\tilde{u})^T p(D)\}
  $$

Through the above learning, we can obtain the semantic dictionary $D \in \mathbb{R}^{k \times m}$. For any image $i$, we firstly describe it with the BOV model, and marked it by $x_i \in \mathbb{R}^k$ where $k$ is the dimensionality of the BOV representation. The semantic representation of the image $i$ can be computed by the fast inner product,

$$
SR(i) = x_i \cdot D,
$$

And the semantic distance between image $i$ and image $j$ can be obtained by the cosine distance.

IV. EXPERIMENTS

In this section, we first introduce the database and the relevant experimental settings, and then validate the effectiveness of the proposed semantic dictionary on the public test sets for the common image understanding tasks: large scale semantic image search and image annotation.

A. Database and Experimental Setting

**Database:** The most popular image database-ImageNet [17] is used as our training data, which is organized by a semantic hierarchy. The original data has over one thousand categories and over one million images, and we simply the data and choose a frequently-used 217 concepts and there are 267k images together (named by ImageNet267K). Furthermore, we re-label 120 × 217 images from the ImageNet267K as the training set (named by ImageNet25K). Another standard image database is the (Corel5K) [19], which is from 50 Corel Stock Photo CDs and consists of 5000 images. Besides, we collect 0.8 million Flickr image as the distracter images (named by Flickr800K).

**Experimental settings:** Semantic dictionary is learnt from the ImageNet25K. The visual set consists of 131072 visual words, which are obtained by the hierarchical $k$-means clustering, and the concept set includes 217 concept from the ImageNet.

B. Large Scale Semantic Image Search

In this paragraph, we validate its efficiency of semantic dictionary on large scale image database, where there are one million images (ImageNet267K+Flickr800K).

**Comparisons Methods:** (1) The classical BOV approach [20] is compared as the baseline approach, where the size of visual words is 0.2 million. (2) SoftBOV [21], a improvement version of BOV, where each descriptor is encoded with soft assignment of 4 nearest neighbors. (3) VLAD [22] (vector of locally aggregated descriptors), which is a state-of-the-art method and derive from the Fisher kernel. The parameter is the same with that in [22]. (4) In our semantic dictionary approach, $\gamma = 0.85 \times 10^{-4}$ and $\theta = 100 \times \gamma$, which is obtained by the cross-validation. $\gamma$ restricts the sparsity of semantic dictionary. $\theta$ is a titleholding parameter, and it controls the inter-impact among the players.

Fig. 2 shows the error of parameter estimation under the parameter $\gamma = 0.85 \times 10^{-4}$ and $\theta = 100 \times \gamma$, which also reflects the convergence trend of the global payoffs. After about 16 iterations (i.e. the players adjudge the strategy at the 16-th time), the error (payoff) achieves a stable value, although the convergent value is not zero.

For the image retrieval task, we follow [20]–[22] and take the mean average precision (MAP) as the evaluation metric. In details, 250 representative images are chosen from the ImageNet267K as the query images. For each query result, the precision-recall curve is computed and the average precision (AP) is obtained by summing the area below the precision-recall curve. Finally, MAP is the meaning of AP from all the query images. Fig. 3 shows the comparisons of the above approaches on the ImageNet sets with different number of images. We can find: firstly, our method improves the MAP sharply than the classical BOV and SoftBOV, and compared with the VLAD, our method also boosts the MAP with 5.7%, the good performance of our method benefits from the consideration of the VPCP problem when we design the semantic dictionary schema. Secondly, as the number of
images increases, the retrieval performance decreases. But the decreasing rate of our approach is slow, which demonstrates that our method is less sensitive to the increment of the data scale.

Besides the obvious improvements of MAP, it is necessary to point out that our approach is also efficient. Table I shows the average time of search an image by different methods. Experimental results show that our method is comparable and satisfied for the real application. The lowness of SoftBOV is mainly rooted in the expensive image soft expansion. The above experiments are carried out on a server with 32 GB memory and 8-core 2.13Ghz processor. In short, the semantic dictionary approach shows significant advantages on both accuracy and efficiency.

C. Image Annotation on ImageNet267K

Image annotation is the most important task in the image understanding, and effective image annotation approach can help the images from the Internet to obtain more accurate label result, which can be used for further applications, such as the data mining, text-based retrieval and so on. To evaluate the performance of image annotation, semantic dictionary is tested and compared on the real web data-ImageNet267K.

**Comparisons Methods:** (1) Binary SVM (BSVM) [23], as an one-vs.-all classification model, we have to separately train the SVM classifiers for each of 217 concepts. One hundred positive images and two hundred negative image are selected from the ImageNet25K. (2) TinyImage [24], this is one idea of nearest neighbor voting. Following the work [24], firstly, the images in ImageNet267K database are down sampled to 32 × 32, then the top-100 nearest neighbors for the query are searched based on SSD pixel distance, and at last the label of query image is replaced by the concept with the most votes. (3) Supervised Multi-class Labeling (SML) [25], as the setting of [25], images are represented by the bags of localized features, and Gaussian mixture model (GMM) consists of 64 components, which are obtained by the separate training.

**Evaluation Rule:** The average precision (AP) is employed as the evaluation metric. Following the ImageNet [17], every method give a top-5 annotation result for the query image, and this annotation is valid if one of the top-5 results is the same with the benchmark. The final AP is averaged over 100 query images from the ImageNet267K.

The average precision of different approaches is shown in the Fig. 4, where we can observe that our model provides the better ability of image annotation for most of the query images. In details, the mean AP of our method is 55.43%, this is a satisfied precision in the large scale annotation task, and after all there are usually less than 0.01% web images with useful label. Besides, the performances of BSVM on the "bike" and TinyImage "nut" are better than our method (1.9% and 0.1%), the reason is that there are few salient visual appearances on this two categories while the semantic descriptions are transferred from the visual information of images, and in this situation the description power of semantic is not robust.

D. Image Annotation on Corel5K

To evaluate the transferring performance of our method, semantic dictionary is evaluated on another standard benchmark (Corel5K). This database contains 5000 images with 260 concepts, where we find that the 217 concepts from ImageNet25K cover some of the Corel5K database. In this paragraph, we complement the annotation task with the semantic dictionary learnt from ImageNet25K.

**Comparisons Methods:** (1) Binary SVM (BSVM) [23]. Similar to the above procedure, we train binary SVM clas-
sifiers separately. In the training stage, the Corel5K is decomposed into three image subsets: the training set of 4,000 images, the validation set of 500 images, and the test set of 500 images. (2) Supervised Multi-class Labeling (SML) [25], the parameter setting is the same with that in Sec. IV-C.

**Evaluation Rule:** average precision (AP) is used as the measurement. Different from the Sec. IV-C, we take a more strict precision evaluation, where every method gives a top-3 annotation for the query image, and this annotation is valid if one of the top-3 results is the same with the benchmark. 10 common categories are chosen as the test images.

Fig. 5 shows the average precision comparisons on the Corel5K database, the performance of our method unexpectedly exceeds the BSVM and SML with a 4.25% and 3.21% improvement separately, and in fact our method was trained not on the Corel5K database but on the ImageNet25K. This means that our semantic dictionary has potential to be directly used in the other image sets, but it does not need the training process. As the concept set of our semantic dictionary covers more concepts in our daily life, the robust transferring ability of our model can bring in some significant applications, such as the real-time object annotation on the mobile devices, etc. We also notice that our method does not work well on the "bird" category because there are too many species of the bird in the world and the birds in the Corel5K rarely appear in our daily life.

**V. Conclusion**

In this paper, we proposed a semantic dictionary to solve the problem of visual polysemia and concept polymorphism in the large scale image understanding, which characterizes the probability distribution between visual appearances and semantic concepts. Inspired by the game theory and machine learning algorithm, the learning of semantic dictionary is formulated into a minimization problem of the payoffs, where the players adjudge the strategy (i.e. the weights assigned to every visual appearance) at each iteration. Non-convex multi-task learning is introduced to solve the above optimization problem, and the convergence guarantees the payoffs to approximate the minimum of the global payoffs. In the future, we will focus on the scalability of semantic dictionary for web scale image database, which covers usually ten thousand concepts.

**Acknowledgements**

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61332016, 61322212, 61402431, 61472203, 11301517 and 71101139.

**References**