Exploring Implicit Image Statistics for Visual Representativeness Modeling

Xiaoshuai Sun[1]*, Xin-Jing Wang[2], Hongxun Yao[1], Lei Zhang[2]
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]Microsoft Research Asia, Beijing, China
[1]{xiaoshuaisun, h.yao}@hit.edu.cn, [2]{xjwang, leizhang}@microsoft.com

Abstract

In this paper, we propose a computational model of visual representativeness by integrating cognitive theories of representativeness heuristics with computer vision and machine learning techniques. Unlike previous models that build their representativeness measure based on the visible data, our model takes the initial inputs as explicit positive reference and extend the measure by exploring the implicit negatives. Given a group of images that contains obvious visual concepts, we create a customized image ontology consisting of both positive and negative instances by mining the most related and confusable neighbors of the positive concept in ontological semantic knowledge bases. The representativeness of a new item is then determined by its likelihoods for both the positive and negative references. To ensure the effectiveness of probability inference as well as the cognitive plausibility, we discover the potential prototypes and treat them as an intermediate representation of semantic concepts. In the experiment, we evaluate the performance of representativeness models based on both human judgements and user-click logs of commercial image search engine. Experimental results on both ImageNet and image sets of general concepts demonstrate the superior performance of our model against the state-of-the-arts.

1. Introduction

Measuring representativeness is an important basis for solving heuristic problems such as "How to choose five words to describe one of your friend?" or "How to pick out one image that best illustrates a concept?". In the early studies of representativeness heuristic, Kahneman and Tversky [1] expressed representativeness according to which the subjective probability of an item is determined by the degree to which it is similar in essential characteristics to its parent set. Based on this expression, the representativeness of an item in a given data set could be quantitatively measured by their similarity [2, 3, 4]. In recent cognitive psychology research, Tenenbaum and Griffiths [5] proposed a rational model of representativeness based on Bayesian analysis of what makes an observation a good sample of a category or of a process, in which the most representative sample is the one that best provides evidence for the target process or concept relative to possible alternatives. This model can explain why people think HHTHT more likely than HHHHH to be produced by flipping a fair coin, even though the two patterns are equal likely. Abbott et al. [6] extended the Bayesian measure which defines the representativeness of a sample from a distribution [5] to define a measure of the representativeness of an item to a set.

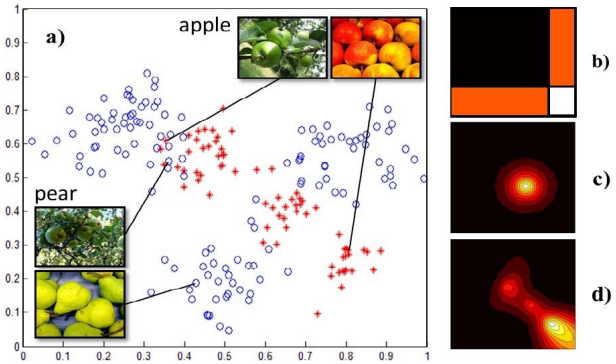

Figure 1. An illustration of representativeness function for b) Bayesian measure [6], c) naive prototypes, and d) our proposed model. Taking the prototypes as a middle-level representation, our model characterizes visual representativeness based on not only the visible data(*) but also the potential negatives (o) inferred from ontological knowledge bases. The paper focuses on three issues: 1) Given a set of images as positives, how to automatically acquire semantic reliable negatives; 2) How to discover all possible prototypes without knowing the precise number; 3) How to measure representativeness based on the customized image knowledge.

During the development of modern computer vision and multimedia technologies, especially for intelligent image browsing systems and search engines, people also investigate computational definition of representativeness to rank the retrieved images under keywords or example queries. Simon et al. [7] selected representative views of city scenes,

*This work was performed at Microsoft Research Asia

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.73

IEEE computer society

noted as canonical views, based on clustering techniques and visual-textual likelihood. Kennedy and Naaman [8] proposed a novel approach to automatically generate representative and diverse views of landmarks by leveraging the community-contributed collections of rich media on the web. Given a set of images with the same landmark tag, they first cluster the images into different views and then rank the clusters according to four heuristic criteria including the number of users, visual coherence, cluster connectivity and variability in dates. In [9], a graph-based representation, namely ImageKB, was proposed to efficiently organize semantic categories, entities and billions of images. Given an <Entity, Category> pair in ImageKB, the representativeness of an image is measured using semantic similarity, category relevance and representative confidence of its K nearest neighbors. This measurement is closely related to the ideas of visual coherence and cluster connectivity, yet much more general in query types and data scale. Doersch et al. [10] proposed a novel framework which detects representative elements for cities such as "Paris" using weak geographic supervised information from a large amount of Google street-view photographs. They adopted a discriminative clustering approach to automatically locate local geo-informative features which appear frequently in one locale while rarely in others. Although the geographic supervision cannot be served as a general principle, the promising results demonstrate the effectiveness of the discriminative framework. Generally, most of the related works are consistent with ([8]-[14]) or directly derived from ([6, 15]) the classic Bayesian definition of representativeness [5]:

$$R(d,h) = \log \frac{P(h|d)}{P(h)} = \log \frac{P(d|h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}, \quad (1)$$

where d is an observed item, $\mathcal{H}$ denotes a set of alternative hypotheses that might explain d. The representativeness $R(d,h)$ is the ratio of posterior to prior probability, which characterizes the extent to which the observation of d could increase the probability of h. Although the above definition has a very simple form, we still face multiple challenges to generalize it to the visual world. The most important issue for implementing Eq. 1 is to define a reasonable hypotheses set $\mathcal{H}$ for visual signals. Abbott et al. [6] assumed an independent Bernoulli distribution with one single parameter θ for each binary visual feature. As shown in Figure 1, such statistical assumption leads to very coarse middle-level representations of the items, and will possibly fail to provide an accurate representativeness function because the visual world is undoubtedly much more complex.

Main Approach In this paper, we proposed a hybrid statistical model for visual representativeness by integrating the prototype theory with the classic Bayesian measure and adopting semantic knowledge from the image ontology datasets to avoid modeling the visual world with certain types of statistical distributions. The formal definition of our visual representativeness is as follows. Given $\mathbf{x}^*$ as a test image, and $\mathcal{D}_r$ a reference data set containing multiple images with similar semantic concepts. The representativeness of $\mathbf{x}^*$ for $\mathcal{D}_r$ is defined as:

$$\mathrm{R}(\mathbf{x}^*, \mathcal{D}_r) = \frac{p(\mathbf{x}^*|\mathcal{D}_r)}{p(\mathbf{x}^*|\mathcal{D}_n)} = \frac{\sum_{\mathbf{r}} p(\mathbf{x}^*|\mathbf{r})p(\mathbf{r}|\mathcal{D}_r)}{\sum_{\mathbf{n}} p(\mathbf{x}^*|\mathbf{n})p(\mathbf{n}|\mathcal{D}_n)}, \quad (2)$$

where $\mathbf{r}$ denotes a prototype of $\mathcal{D}_r$, and $\mathbf{n}$ a prototype of $\mathcal{D}_n$. This definition has two advantages: 1) Instead of all possible alternatives, it only focuses on the reference data and its negatives; 2) Probability inferences are conducted within a small yet compact prototype space. In the implementation phase, we focus on two main issues: 1) How to find negative references with reasonable semantic meanings; 2) How to discover all possible prototypes without knowing the precise number. For the first issue, we apply image ontology datasets, e.g. ImageNet [16] and ImageKB [9], to get the required semantic structure and relevant image instances and then group them together to form a negative reference set. For the other issue, we proposed a prototype discovery algorithm which automatically determines the number of prototypes based on an efficient statistical unimodality test. In summary, this paper makes the following contributions:

- We proposed a semantic embedded visual representativeness model. As a hybrid model derived from prototype theory and the Bayesian measure of representativeness, our model has a solid foundation from cognitive research on representativeness and use dynamic prototypes to get a flexible mid-level representation.
- Ontological knowledge bases are adopted to embed visual semantics in the proposed framework which avoids modeling the complex visual world with fixed types of statistical distribution and limited number of parameters. Together with the prototype theory, semantic embedding ensures a more accurate and meaningful measure of representativeness for general visual concepts.
- The proposed model can be applied to many vision-related applications such as detecting representative regions of a photograph, finding a group of representative images that best describe an entity, or more practically ranking images based on the search results of commercial image search engines.

The structure of the paper is as follows. In Section 2, we provide some background information, including semantic ontology databases and the prototype theory. Section 3 then delivers the details of our semantic embedded representativeness model. We present the experimental results in Section 4 and give a discussion in Section 5 focusing on the relationship between the proposed representativeness and other visual properties such as saliency. Finally, we conclude this paper in Section 6.

2. Related Works

Semantic Knowledge Base Semantic knowledge bases provide the relationships between words (WordNet[17]) or meaningful semantic concepts (NeedleSeek [18]) which establish the foundation of automatic semantic analysis for natural language processing and information retrieval. With respect to vision research, Deng et al.[16] proposed a large-scale hierarchical image database named ImageNet based on the structure of WordNet which serves as an image ontology base containing 21,841 WordNet synsets and over 14 million highly selected images (2011 Fall release). Compared to previous image datasets, ImageNet inherits the semantic hierarchy of WordNet and meanwhile provides high resolution images that are manually verified to contain the relevant concepts. Based on ImageNet, Deselaers and Ferrari [19] built a novel distance function between images by not only estimating the visual similarity as in tractional works but also assessing whether they are from the same basic-level category. The ImageNet distance exploits semantic similarity measured through the ImageNet semantic hierarchy, which outperforms and goes beyond direct visual distances in traditional vision research. The promising results [16, 19, 20] inspired us to embed semantic ontological knowledge into visual representativeness computation to make our model more reasonable and consistent with both semantic and appearance aspects of the visual world.

Prototype Theory In cognitive science, the prototype theory [21, 22] states that categories tends to be defined in terms of prototypes or prototypical instances that contain the attributes most representative of items inside and least representative of items outside a category. Sufficiently specific categories can be defined as a single prototype represented by typical shapes and attributes [2, 19]. As shown in Figure 1, using prototypes as an intermediate representation has two advantages: 1) consistent with the cognitive understanding of semantic categories and 2) leads to a continuous and well-bounded measure function.

3. The Model

The framework of our proposed model is shown in Figure 2. Our model takes three kinds of input: 1) keywords that represent a concept, 2) keywords + the related images, and 3) unlabeled images. For case 3), we use Google annotation service to label the input images and select the most related keywords to represent the underlying concept. Taking the keyword as a seed, we build a customized image ontology based on large-scale semantic ontology databases and image search engines. The customized image ontology contains images for both the input concept as well as the confusable semantic neighbors (negative references). Potential prototypes are mined by a dynamic prototype discovery algorithm which is designed for arbitrary data distributions and can effectively guarantee the unimodality of each prototype. Finally we estimate the representativeness of related images by Eq. 2.

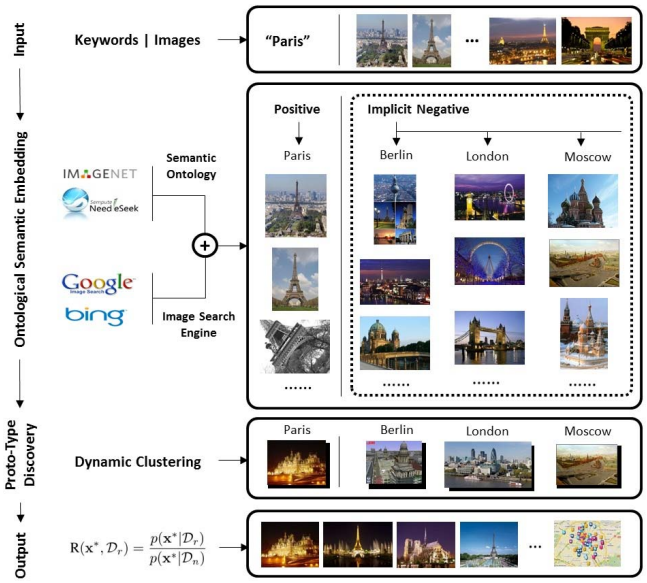

Figure 2. The proposed visual representativeness model.

3.1. Embedding Ontological Visual Semantics

We embed visual semantics in the proposed model by customizing a small image ontology from semantic knowledge bases such as WordNet [17] and NeedleSeek [18]. The customized image ontology is built and organized like ImageNet containing both the semantic entities and relevant images. However, the attached images are left unverified since we don't want to involve any human interactions. Note that ImageNet has covered over twenty thousand synsets of WordNet and most of its attached images have been manually verified, we can directly build our customized image ontology from ImageNet as long as the query entity is available. Given an image set, the construction of the customized image ontology can be done by the following steps:

Step 1: Locating the Semantic Concept If the concept (represented as keyword) of the image set is given, we directly go to Step 2. If the concept is not specified, we obtain the keyword by simply applying the on-line annotation service from Google Image Search.

Step 2: Finding the Negative References Given the keyword, we search for the most related and confusable concepts on either WordNet or NeedleSeek. Empirically, we use the sibling nodes of the given concept on NeedleSeek as the negative references for implementation simplicity.

Step 3: Building Customized Image Ontology For those concepts that are available on ImageNet, we directly attach the corresponding images as a part of our customized image ontology. Instances for other concepts are obtained by crawling images through Google Image Search.

3.2. Dynamic Prototype Discovery

In the above section, we have constructed a customized image ontology which consists of two groups of images including $\mathcal{D}_r$ (the original input image set) and $\mathcal{D}_n$ (the negative reference image set). The goal of this section is to discover the prototypes from $\mathcal{D}_r$ and $\mathcal{D}_n$ as a middle-level representation for our representativeness model. Since we do not know the exact number of prototypes, this task can be regarded as an incremental clustering problem which requires a careful estimation of the number of clusters. A common solution is to begin with a small k, and iteratively increase it by splitting the intermediate clusters rejected by a specific statistical test [23, 24].

To ensure the robustness of the discovery algorithm, we follow the proposal of [25, 26] and use a general statistical test named *Dip-dist* to estimate the number of prototypes. Note that, we select *Dip-dist* because it is efficient ($O(bnlogn + n^2)$) and robust to various kinds of data distributions. Algorithm 1 shows the dynamic prototype discovery algorithm in details. We set $k_{init} = 1, m = 2$, the dip-test[25] parameters $\alpha = 0, b = 1,000$ and the unimodality threshold $v_{thd} = 0.01$ in all the experiments. Practically, if the input dataset for dis-test is too large(> 400), we replace the input by its subset, which contains 400 randomly selected items, to boost the processing speed. The test result on a small group of images is shown in Figure 3.

Algorithm 1: Dynamic prototype discovery

Input: Dataset $X = \{x_i\}_{i=1}^N$, the initial number of prototypes k_{init}, a splitting number m, a statistic significance level α for the unimodality test, threshold v_{thd} for spliting the candidate prototype.

Output: Prototypes $P = \{p_j\}_{j=1}^k$ and the conditional probability $p(p|X)$

$k \leftarrow k_{init}$;
Run k-means on X to obtain cluster centers $C = \{c_j\}_{j=1}^k$;
Initialize the prototype set by: $P \leftarrow C$;
repeat
  $k' \leftarrow k$
  for $j = 1, ..., k$ **do**
    $score_j \leftarrow$ unimodality test (p_j, α, v_{thd})[26];
  end
  if $\max_j(score_j) > 0$ **then**
    $k \leftarrow k + 1, d \leftarrow \text{argmax}_j(score_j)$;
    $X_{sub} = \{x_i | x_i \in X, |x_i - p_d|_2 = \min_j |x_i - p_j|_2\}$;
    Run k-means on X_{sub} to get m prototypes D;
    $P = P \bigcup D - \{p_d\}$;
  end
until $k' = k$;
for $j = 1, ..., k$ **do**
  $p(p_j|X) = \frac{|\{x_i | x_i \in X, |x_i - p_j|_2 = \min_{p_n \in P} |x_i - p_n|_2, \}|}{|X|}$;
end
return $P, p(p|X)$;

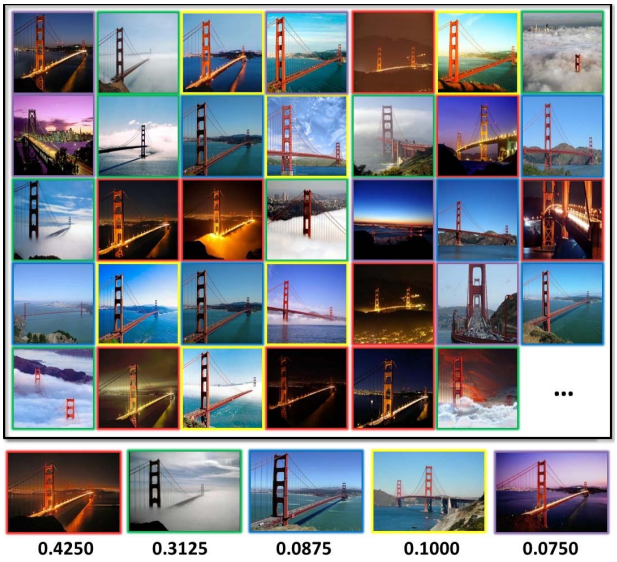

Figure 3. Typical results of dynamic prototype discovery. Top: image set of the Golden Gate Bridge; Bottom: the discovered prototypes with conditional probability. Our algorithm is able to incrementally discover semantically meaningful prototypes without knowing the exact number of potential topics. In this case, the prototypes summarize the image set by environmental conditions.

3.3. Computing Representativeness

We obtain the customized image ontology in Section 3.1, and the prototypes with conditional probabilities in Section 3.2. The last step is to define the relationship between the items and prototypes. For simplicity, the conditional probability $p(\mathbf{x}^*|\mathbf{r})$ is defined as the similarity between item $\mathbf{x}^*$ and the underlying prototype $\mathbf{r}$:

$$p(\mathbf{x}^*|\mathbf{r}) = \exp\{-\lambda|\mathbf{x}^* - \mathbf{r}|_2\}, \quad (3)$$

where λ is a scaling constant to keep $p(\mathbf{x}^*|\mathbf{r})$ in a reasonable interval. Based on all pre-computed terms, the final score for representativeness is computed according to Eq. 2.

4. Experiments

In this section, we show how our model performed in two applications including 1) improving the quality of ImageNet and 2) mining representative images for general concepts. Images used in our experiment are downloaded from ImageNet and Google Image Search. For each image, we construct a 5096-D compact BoW feature by combining a 1,000-D SIFT BoW vector [27] with a 4096-D RGB color histogram. Besides subjective evaluations [5, 6], we propose to use the number of user-clicks as another ground-truth for image representativeness. Note that, the user-click data are acquired from a different image search engine, specifically the Bing Image Search, in order to eliminate the potential ranking bias caused by the usage of query association data.

Evaluation Metrics Given a list of images ranked by an image search engine, web-users usually click on those images which they believe to be attractive/representative to their queries. By taking user-clicks as votes for the representativeness of tested images, we define another evaluation metric based on the choices made by Web-users:

$$S_W(R) = \sum_{i=1}^{N} UC_i/R_i, \tag{4}$$

where i is the image index, R_i denotes the ranking position of image i, and UC_i is the number of user clicks for test image i recorded in the query association log of Bing Search[1]. Since S_W has a potential bias problem (see Sec. 5), we propose a subjective metric S_S as a complement for Section 4.2. Similar as [6], we asked human subjects to label the representativeness score ranging from 1 to 10 (10 for the best) for each image used in the experiment. The subjects are asked to carefully study the test concepts before starting their labeling work. Based on the labels, the subjective ranking score S_S can be defined as:

$$S_S(R) = \sum_{i=1}^{N} UR_i/R_i, \tag{5}$$

where UR_i is the average subjective score of image i.

Baseline Models To demonstrate the effectiveness of our method, we compare our results with two baseline approaches: *Bayesian Model* and *Naive Prototype Model*.

Bayesian Model - The Bayesian Model [6] is a natural generalization of the cognitive theory of representativeness [5] and implemented based on Bayesian Sets [11] which is a statistical technique initially proposed for measuring how appropriately a new sample can fit into a given set of data. Formally, given a data set $\mathcal{D}$ and a subset of samples $\mathcal{D}_s = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \subset \mathcal{D}$ representing a conceptual group, Bayesian Sets measures the representativeness of a given sample $\mathbf{x}^* \in \{\mathcal{D} \backslash \mathcal{D}_s\}$ by the following equation:

$$\text{Bscore}(\mathbf{x}^*, \mathcal{D}_s) = \frac{p(\mathbf{x}^*, \mathcal{D}_s)}{p(\mathbf{x}^*)p(\mathcal{D}_s)}. \tag{6}$$

In the original work [6], $\mathbf{x_i} \in \mathcal{D}$ is represented as a binary feature vector $(x_{i1}, ..., x_{iJ})$ where $\mathbf{x}_{ij} \in \{0, 1\}$. It assumes that each element of $\mathbf{x}_i$ has an independent Bernoulli distribution parameterized by θ : $p(\mathbf{x}|\theta) = \Pi_{j=1}^{J} \theta_j^{x_{ij}} (1 - \theta_j)^{x_{ij}}$. By combining the Beta distribution $p(\theta|\alpha, \beta) = \Pi_j \frac{\Gamma(\alpha_j+\beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \theta_j^{\alpha_j - 1}(1 - \theta_j)^{\beta_j - 1}$ as the conjugate prior for θ, Eq. 6 could be reformulated as:

$$\text{Bscore}(\mathbf{x}^*) = \prod_j \frac{\alpha_j + \beta_j}{\alpha_j + \beta_j + N} \left(\frac{\tilde{\alpha}_j}{\alpha_j}\right)^{x_{*j}} \left(\frac{\tilde{\beta}_j}{\beta_j}\right)^{1 - x_{*j}}, \tag{7}$$

[1] http://www.bing.com

where $\tilde{\alpha}_j = \alpha_j + \sum_{n=1}^{N} x_{nj}$ and $\tilde{\beta}_j = \beta_j + N - \sum_{n=1}^{N} x_{nj}$. As suggested by Abbott *et al* [6], we compute the logarithm of Eq. 7 as the final score for representativeness:

$$\mathbf{BM}(\mathbf{x}_*, \mathcal{D}_s) = \log \text{Bscore}(\mathbf{x}^*) = c + \sum_j s_j x_{*j}, \tag{8}$$

where $c = \sum_j \log(\alpha_j + \beta_j) - \log(\alpha_j + \beta_j + N) + \log \tilde{\beta}_j - \log \beta_j$, and $s_j = \log \tilde{\alpha}_j - \log \alpha_j - \log \tilde{\beta}_j + \log \beta_j$. In the experiment, the hyperparameters α and β are empirically set as $\alpha = k\mathbf{m}, \beta = k(1 - \mathbf{m})$, where $\mathbf{m}$ is the mean value of the binary features over all samples in the test dataset and k is a scaling factor which controls the effect of $\mathbf{m}$ in model initialization. In the experiment, we set $k = 1$, and followed the procedure of [6] to generate binary features from the compact BoWs. For each element of the BoW vector, if it is positively skewed, we assign the value **1** to the samples for which the value of that element is above 80%, then assign the value **0** for the rest. If the feature is negatively skewed, we assign the value **1** to the samples for which the value of that element is below 80%, then assign the value **0** for the rest.

Naive Prototype Model -[2] We implemented the naive prototype model of representativeness following the procedure of [6]. Given a dataset $\mathcal{D}$, we select its prototype sample by:

$$\mathbf{x}_{\text{proto}} = \text{argmin}_{\mathbf{x}} \sum_{i=1}^{N} |\mathbf{x}_i - \mathbf{x}|_2, \mathbf{x}, \mathbf{x}_i \in \mathcal{D}, \tag{9}$$

which aims to minimize the mean square error within the dataset. The representativeness score is then defined as the similarity between the input and the prototype:

$$\mathbf{NPT}(\mathbf{x}^*, \mathcal{D}_s) = \exp\{-\lambda |\mathbf{x}^* - \mathbf{x}_{\text{proto}}|_2\}, \tag{10}$$

where $\mathbf{x}_{\text{proto}}$, represented as a BoW vector, is the prototype of $\mathcal{D}_s$, and λ is a scaling constant which actually does not affect the ranking results.

4.1. Ranking Images on ImageNet

ImageNet [16] is a large-scale image ontology dataset which provides us the essential knowledge of the visual world including not only semantic hierarchies but also the relevant image instances. Although all images in ImageNet are manually verified to contain the relevant concepts, their quality and representativeness are still left un-labeled. In this section, we apply our model to rank the images of ImageNet in order to give representative images higher priority in potential applications.

We show quantitative evaluation results for two groups of data from ImageNet: *big cat* and *pome*. Three models including Bayesian Model(BM[6]), Naive Prototype (NPT) and our model are adopted to re-rank the images of the 11

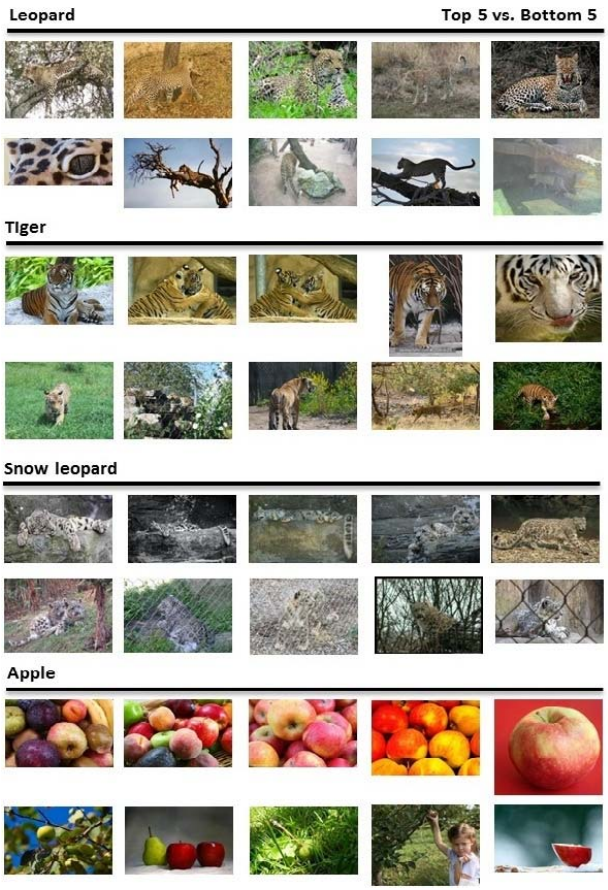

Figure 4. Re-ranking images on ImageNet. Compared to the unorganized data, the top-ranked images always contain conspicuous instances for the underlying synset and have much less occlusions and distracters, e.g. human bodies and man-made objects.

children synsets[1]. Table 1 shows the web-click score S_w of the ranking results for the tested models as well as the original rank. Relatively, our model achieved the largest improvement compared with the other two approaches. Figure 4 shows some typical ranking results of our method.

Table 1. Quantitative Evaluation S_W on ImageNet.

ID	Origin	BM	NPT	Ours
1	6.6822E-1	1.5054E+0	7.2690E-1	**6.4989E+0**
2	4.5906E+0	1.0599E+1	1.9066E+1	**4.0684E+1**
3	5.7921E+0	3.8826E+0	6.1455E+0	**8.0257E+0**
4	1.1633E-1	**2.9740E-1**	9.0174E-2	9.1124E-2
5	5.0671E+1	5.7508E+1	**1.2151E+2**	2.9409E+1
6	1.1764E+0	8.1034E-1	1.6402E+0	**2.7653E+0**
7	**4.3777E+1**	1.7893E+0	2.0931E+0	1.9262E+0
8	**1.0728E+1**	1.0211E+0	7.2240E-1	1.8984E+0
9	3.0949E-2	**1.9927E-1**	6.6682E-2	2.6842E-2
10	8.1893E-3	1.4129E-2	9.0000E-3	**2.4793E-2**
11	1.1230E+0	4.7773E-1	1.5035E+0	**1.4657E+1**

[1]ID 1-11 in Table 1 denote WordNet synsets: n02129991, n02128925, n02128757, n02130086, n02128385, n02129604, n02130308, n07739125, n07767847, n07769584

4.2. General Representative Image Mining

In this section we consider a more general case: given a concept, how to automatically select the most representative images without human interaction. Practically, given a keyword, we search for the k most related words using a public available semantic ontology named NeedleSeek [18]. Then, we crawl images by querying Google with all the related keywords to build a customized image ontology. Based on the auto-built image ontology, we set up the representativeness model following the procedure in Section 3. Representative Images can then be obtained by ranking the concept-related images with our model. Note that this procedure can be applied to refine the results of commercial image search engines since it is fully automatic, semantic-aware, and psychological plausible. We test the representativeness models with three concepts: *Wolf*(animal), *Paris*(city) and *Rose* (flower). The related keywords obtained from NeedleSeek[2] are shown in Table 2. For each keyword, we crawled 200 images from Google Image Search[3] to build the customized image ontology. We use S_W and S_S as quantitative evaluation metrics for the mining performance. The results are shown and discussed in Figure 5.

Table 2. Related keywords of the tested concepts

concept	related keywords at NeedleSeek [18]
Wolf	**wolf**, moose, elk, deer, coyote, bear, beaver, black bear, bobcat, lynx
Paris	**Paris**, Berlin, London, Tokyo, Beijing, Moscow, Rome, Sydney, Washington, New Delhi
Rose	**rose**, lily, chrysanthemum, daisy, carnation, orchid, tulip, daffodil, violet, gladiolus

5. Discussion

Relation to Saliency People might be confused when being asked to point out both the "representative" and the "salient" parts of an image. By taking the logarithm of Eq 2, we could explicitly show the connection between these two concepts, which might lead to a reasonable explanation for the cause of such confusions.

$$\log(\mathrm{R}(\mathbf{x}^*, \mathcal{D}_s)) = \underbrace{\log(p(\mathbf{x}^*|\mathcal{D}_s))}_{\text{Log likelihood}} \underbrace{- \log(p(\mathbf{x}^*|\mathcal{D}_n))}_{\text{Self Information}} \quad (11)$$

The first term in Eq 11 is simply the log likelihood which has also been used as a measure of representativeness in previous works [28]. The second term is the Self-Information of $\mathbf{x}^*$ conditioned on the observation of $\mathcal{D}_n$. Note that, Self-Information is a well accepted measure of bottom-up saliency in early vision modeling research [29, 30, 31]. Thus, our representativeness can be explained as "Likelihood +

[2]http://www.needleseek.msra.cn/
[3]http://images.google.com

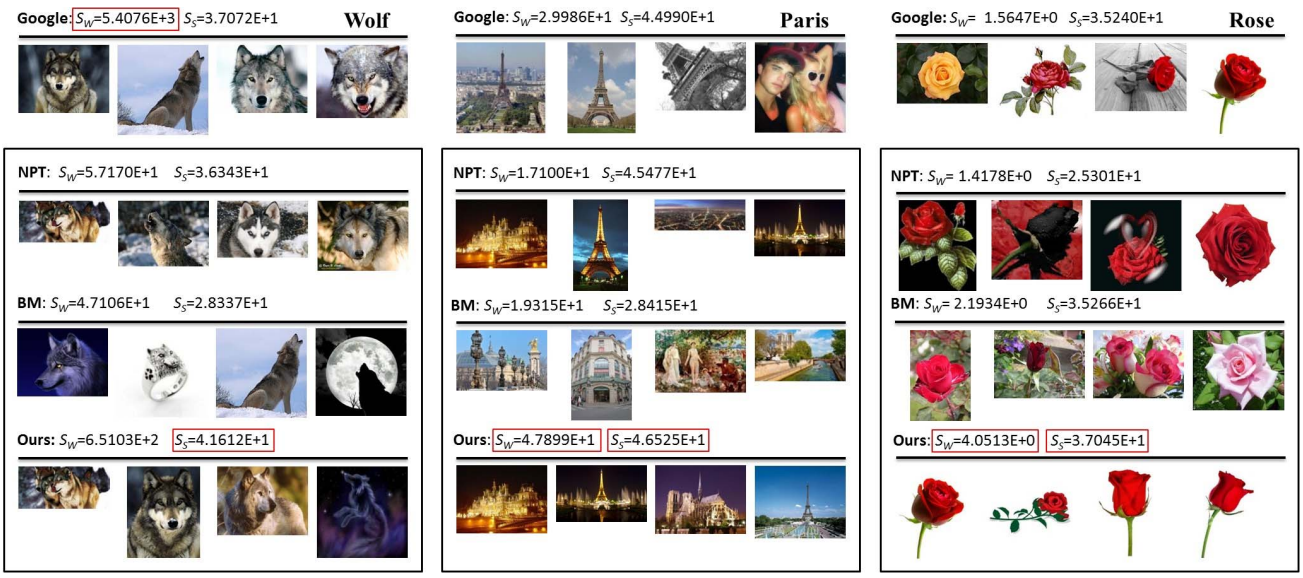

Figure 5. Representative image mining for "Wolf", "Paris" and "Rose". For each concept, we show both intuitive (images) and quantitative (S_W & S_S) comparisons between Google, Bayesian Model [6] (BM), Naive Prototype Model (NPT), and our method. Intuitively, our model is less sensitive to outliers compared to Google and BM, and it also generates comprehensive results that have better diversity than NPT.

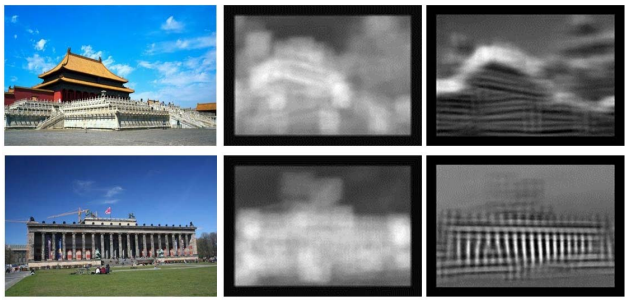

Figure 6. Comparisons between bottom-up saliency and the proposed representativeness. For each row, we show the input image, the AIM saliency map [31], and the visualized response map of our representativeness. Unlike the saliency method, our model locates those regions which contain both salient and discriminative contents such as the golden roof of the Chinese Palace Museum and the huge pillars of the German Berlin Museum.

Saliency", which favors the items that are not only well fitted into the observed concept but also remarkably salient to other related, confusable concepts. Figure 6 shows some comparisons between our representativeness model and AIM saliency (Attention by Information Maximization [31]) on natural images. We run the AIM code package[4] with its default parameter settings. To show the real differences, we use the same features to compute the response map of our representativeness model. The input image is first decomposed by 25 $21 \times 21 \times 3$ color bases. The coefficient vector of each location is treated as a single item whereas all items compose the reference dataset ($\mathcal{D}_r$). In each test, we use the input image in the other test to generate the negative reference ($\mathcal{D}_n$). The two input images are collected from the Internet, which capture the classic view of the Palace Museum of China and the Berlin Museum in Germany. As shown in the results, AIM saliency map highlights the unique objects, e.g. the building corners and human bodies, whereas our model favors the representative components such as the golden roof and huge pillars which are indeed the most recognizable elements for eastern and western architectural styles.

Evaluation Bias In the experiment, our second evaluation metric S_W explicitly characterizes the representativeness of a given image by the number of user-clicks the image has received. This measure benefits from the large number of web users, and mostly it is more reliable and objective compared to human evaluation because it is not restricted by subject numbers and variety. The potential problem with this metric is that users might click an image according to their personal interests instead of the real semantic relevance. Such kind of clicks will bias the evaluation to favor attractive images rather than the representative ones. However, such imperfection does not affect the validity of this evaluation metric for comparing the relative performance of different representativeness models.

Disambiguation Polysemy is a practical problem for large scale implementation of our framework. Intuitively, we could rely on the strength of NeedleSeek [18] to tackle this problem. For example, in NeedleSeek, the concept "apple" corresponds to three categories including "fruit", "brand" and "tree". Based on such knowledge, images can be collected and organized into different semantic branches by which we can effectively eliminate the polysemous problem for the subsequent processing.

[4] http://www-sop.inria.fr/members/Neil.Bruce/

6. Conclusion

In this paper, we have introduced a novel computational model for visual representativeness based on ontological semantic embedding and dynamic prototype discovery. Unlike previous works that focus on the visible data, we exploit the possibility of integrating implicit information to support semantic-aware visual analysis. The embedded image ontology provides additional image statistics helping the model to identify true outliers. Meanwhile, the intermediate prototype representation enhances the cognitive plausibility of our model and ensures the accuracy and effectiveness of the probabilistic inference. Experimental results demonstrate the superior performance of the proposed approach against the state-of-the-art representativeness models as well as commercial image search engines. The computational cost of our method is approximately $O(n^2)$, where n is the number of the retrieved/indexed images of a concept. Since n has an expectable upper bound (e.g.$<$1000 for online retrieval), the cost can be regarded as constant, so that our approach is highly scalable and thus can be applied in billion-level applications.

7. Acknowledgement

Xiaoshuai Sun and Hongxun Yao were partially supported by National Natural Science Foundation of China (Grant No. 61071180) and Key Program (Grant No. 61133003).

References

[1] D. Kahneman and A. Tversky. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972.

[2] S.K. Reed. Pattern recognition and categorization. *Cognitive psychology*, 3(3):382–407, 1972.

[3] D.L. Medin and M.M. Schaffer. Context theory of classification learning. *Psychological Review*, 85(3):207, 1978.

[4] R.M. Nosofsky. Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(1):87, 1987.

[5] J.B. Tenenbaum, T.L. Griffiths, et al. The rational basis of representativeness. *Proceedings of the 23rd annual conference of the Cognitive Science Society*, pp.1036–1041, 2001.

[6] J.T. Abbott, K.A. Heller, Z. Ghahramani, and T.L. Griffiths. Testing a bayesian measure of representativeness using a large image database. *NIPS*, 2011.

[7] I. Simon, N. Snavely, and S.M. Seitz. Scene summarization for online image collections. *ICCV*, 2007.

[8] L.S. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. *WWW*, 2008.

[9] Xu Z. Zhang L. Liu C. Wang, X.J. and Y. Rui. Towards indexing representative images on the web. *ACM MM*, 2012.

[10] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A.A. Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4):101, 2012.

[11] Z. Ghahramani and K. Heller. Bayesian sets. *NIPS*, 2005.

[12] S. Singh, A. Gupta, and A.A. Efros. Unsupervised discovery of mid-level discriminative patches. *ECCV*, 2012.

[13] R. Rahul and S. Lazebnik Computing iconic summaries of general visual concepts. *CVPR Workshops*,2008.

[14] T.L. Berg and A.C. Berg. Finding Iconic Images *CVPR Workshops*, 2009.

[15] K.A. Heller and Z. Ghahramani. A simple bayesian framework for content-based image retrieval. *CVPR*, 2006.

[16] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009.

[17] C. Fellbaum. Wordnet. *Theory and Applications of Ontology: Computer Applications*, pages 231–243, 2010.

[18] S. Shi, H. Zhang, X. Yuan, and J.R. Wen. Corpus-based semantic class mining: distributional vs. pattern-based approaches. *ICCL*, pp.993–1001, 2010.

[19] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. *CVPR*, 2011.

[20] A. Torralba, R. Fergus, and W.T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *TPAMI*, 30(11):1958–1970, 2008.

[21] E. Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192, 1975.

[22] E. Rosch. Principles of categorization. *Concepts: core readings*, pp.189–206, 1999.

[23] G. Hamerly and C. Elkan. Learning the k in k-means. *NIPS*, 2004.

[24] Y.F.G. Hamerly. Pg-means: learning the number of clusters in data. *NIPS*, 2007.

[25] J.A. Hartigan and PM Hartigan. The dip test of unimodality. *The Annals of Statistics*, 13(1):70–84, 1985.

[26] A. Kalogeratos and A. Likas. Dip-means: an incremental clustering method for estimating the number of clusters. *NIPS*, 2012.

[27] J. Deng, A. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? *ECCV*, 2010.

[28] G. Gigerenzer and U. Hoffrage. How to improve bayesian reasoning without instruction: Frequency formats. *Psychological review*, 102(4):684, 1995.

[29] N. Bruce and J. Tsotsos. Saliency based on information maximization. *NIPS*, 2006.

[30] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G.W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.

[31] N.D.B. Bruce and J.K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3), 2009.