Social Negative Bootstrapping for Visual Categorization

Xirong Li, Cees G.M. Snoek, Marcel Worring, and Arnold W.M. Smeulders
Intelligent Systems Lab Amsterdam, University of Amsterdam
Science Park 904, 1098XH, Amsterdam, The Netherlands
{x.li, cgmsnoek, m.worring, a.w.m.smeulders}@uva.nl

ABSTRACT

To learn classifiers for many visual categories, obtaining labeled training examples in an efficient way is crucial. Since a classifier tends to misclassify *negative* examples which are visually similar to positive examples, inclusion of such informative negatives should be stressed in the learning process. However, they are unlikely to be hit by random sampling, the de facto standard in literature. In this paper, we go beyond random sampling by introducing a novel *social negative bootstrapping* approach. Given a visual category and a few positive examples, the proposed approach adaptively and iteratively harvests informative negatives from a large amount of social-tagged images. To label negative examples without human interaction, we design an effective virtual labeling procedure based on simple tag reasoning. Virtual labeling, in combination with adaptive sampling, enables us to select the *most misclassified* negatives as the informative samples. Learning from the positive set and the informative negative sets results in visual classifiers with higher accuracy. Experiments on two present-day image benchmarks employing 650K virtually labeled negative examples show the viability of the proposed approach. On a popular visual categorization benchmark our precision at 20 increases by 34%, compared to baselines trained on randomly sampled negatives. We achieve more accurate visual categorization without the need of manually labeling any negatives.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning*

General Terms

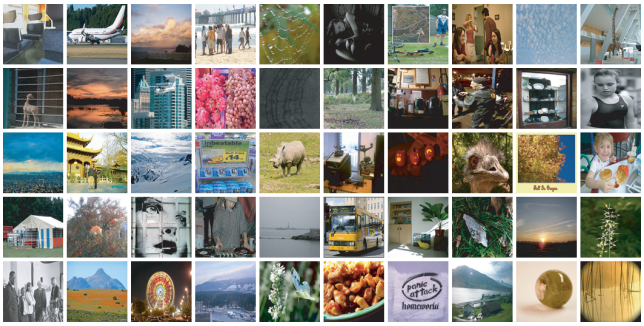
Algorithms, Measurement, Experimentation

Keywords

Social-tagged examples, negative bootstrapping

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMR '11, April 17-20, Trento, Italy
Copyright ©2011 ACM 978-1-4503-0336-1/11/04 ...$10.00.

(a) Positive examples of 'aeroplane'



(b) Randomly sampled negative examples of 'aeroplane'

(c) Automatically generated negative examples (this paper)

Figure 1: A positive set and two negative sets of the visual category 'aeroplane'. The negative set (b) is obtained by random sampling, while the negative set (c) is automatically generated by our approach. Note that, compared to (b), our negatives are visually more similar to the positive set (a). Hence, they are more informative, yielding more accurate visual classifiers. Such negatives are found *without the need of actually labeling them.*

1. INTRODUCTION

Labeled examples are crucial to learn classifiers for visual categorization. To be precise, we need positive and negative examples with respect to a specific visual category. When the number of categories is large, obtaining labeled examples in an efficient way is essential. To that end, current research focuses on obtaining positive examples [6, 19].

In [6], for instance, the authors investigate online collaborative annotation. The authors in [19] gather positive examples by re-ranking web image search results. While intensive effort has been devoted to the positive examples, random sampling is the de facto standard for achieving the negatives [6, 10, 11, 17, 19, 25, 32].

In practice, a classifier tends to misclassify negative examples which are visually similar to positive examples. As Fig. 1 shows, to derive an accurate classifier for a category, say 'aeroplane', confusing negatives such as images of birds or sky should be included when training classifiers. However, such informative negatives are unlikely to be hit by random sampling.

In this paper we go beyond random sampling by proposing a novel, yet technically simple, *social negative bootstrapping* approach. Our approach conceptually bears some resemblance to active learning [22] and AdaBoost [8], as all of them seek informative examples for learning a new classifier. However, there are two notable differences between our approach and traditional active learning. First, in contrast to active learning which requires human interaction to label examples selected in each round, our approach selects informative negatives without human interaction. Second, our definition of informative examples differs from its counterpart in a typical active learning setting. There, one assumes that the input data consists of both positive and negative examples. Thus, examples the system is most uncertain about, namely closest to the decision boundary [22], are considered informative. Our approach, by contrast, selects negatives falling on the positive side and far away from the boundary. Compared to AdaBoost which works on fully labeled data, our approach is grounded on social-tagged data, without the need of manually labeling any negative examples. Since most annotation efforts to create fully labeled data are consumed by annotating the negatives [11], social negative bootstrapping is suited for exploiting large-scale datasets.

2. RELATED WORK

This study is about sampling negative examples for visual categorization. But, it cannot stand alone without positive examples. So we first review recent progress in obtaining positive examples, and then discuss work on the negatives.

2.1 Obtaining Positive Examples

Much research has been conducted towards devicing efficient solutions to acquire positive examples. E.g., by data-driven learning from web image search results [19, 23, 28, 30] or social-tagged data [12, 14, 20, 25, 31, 32], or by online collaborative annotation [6, 16, 18]. In [19], for instance, the authors train a visual classifier on web image search results of a given category, and re-rank the search results by the classifier. By estimating the relevance of user tags to image content [12], social-tagged data can be cleaned up. The authors in [18] develop an online annotation tool, letting web users label images as volunteers. Though the automated approaches are not comparable to human annotation [10, 25], their output already gives a good starting point for manual labeling. Therefore, a recent trend is to combine data-driven learning and online annotation. For instance, the authors in [6] build an ImageNet wherein positive examples of a WordNet category [15] are obtained by labeling web image search results of the category using the Amazon Mechanical Turk service. In this service, web annotators are paid by micro payments. In a recent release of ImageNet, for almost 9,000 categories, there are at least five hundred positive examples per category. Compared to traditional expert labeling, the new labeling mechanism yields positive examples for many categories with lower cost. In this paper we assume that positive examples are obtained by (one of) the approaches described above, and focus on obtaining negative examples.

2.2 Obtaining Negative Examples

Despite the achievement of gathering positive examples, the problem of how to effectively obtain the negatives remains unclear and its importance underestimated. One might consider bypassing the negative labeling problem by one-class learning, which creates classifiers using positive examples only [21]. However, as in principle learning from more information will lead to better results, visual classifiers trained by one-class learning are inferior to classifiers trained by two-class learning with randomly sampled negatives [11].

To automatically create a negative training set for a given category, the mainstream approach is to randomly sample a relatively small subset from a large pool of (social-tagged) examples [6, 10, 11, 17, 19, 25, 32]. Apart from the obvious fact that random sampling is simple and easy to use, we attribute its popularity to the following two reasons. First, as the possible negatives significantly outnumber the positive training set, down-sampling the negatives bypasses class imbalance which is known to affect classifier learning [9]. Second, except for some over-frequent categories such as 'sky' and 'person', the chance of finding genuine positive examples in a random fraction of the pool is low. If the pool is sufficiently large, one might end with a set of reliable negatives, but not necessarily the most informative ones.

Since negative examples are selected at random, the performance of individual classifiers may vary. According to the bootstrap aggregation theory [2], such variance can be reduced by model averaging. Hence, the authors in [17] perform random sampling multiple times to create multiple classifiers, and combine them uniformly. Although the robustness of the final classifier might be improved by classifier aggregation, such a "random+aggregation" approach seems not strategically better than random sampling.

Negative bootstrapping has also been studied in the context of text categorization, e.g., [13]. There, unlabeled examples are inserted into the negative set, if they are most dissimilar to the positives, or predicted as negatives with high confidence by the current classifier. A similar idea is reported in [29] for video retrieval, where negatives are selected at the bottom when ranking unlabeled examples by their scores of being positives in descending order. Though sampling at the bottom probably yields reliable negatives, an intrinsic drawback is that those negatives are already correctly classified, adding them to the training process is not so useful by definition. Indeed, empirical evidence from [17] indicates that such conservative sampling is inferior to random sampling.

In this paper, we strive to reveal the true value of social-tagged images as negative training examples. As a reward, we obtain visual classifiers which are more accurate than classifiers trained on randomly sampled negatives or their aggregated version.

3. SOCIAL NEGATIVE BOOTSTRAPPING

3.1 Problem Statement

Let x be a target image which we want to categorize, and V a large set of visual categories. Let S be a large set of images, where each image is labeled with at least one category from V by social tagging. Given a specific category $w \in V$, let B_{w+} be a positive training set, which are obtained, for instance, by the approaches described in Section 2.1. In a classical two-class learning setting, one has access to B_{w+} and a set of manually labeled negative examples. In random negative bootstrapping, manually labeled negatives are replaced by randomly sampled pseudo-negatives. Social negative bootstrapping derives a visual classifier $G(x, w)$ from B_{w+} and from B_{w-} which contains negative examples obtained from S. The output of $G(x, w)$ is a likelihood score of the image x being positive with respect to the category w. We aim for negative examples most informative to train classifiers, but without actually labeling any negatives.

3.2 The Algorithm

For a given category w and a positive set B_{w+}, we adaptively and iteratively select informative negatives B_{w-} from S. In particular, we select the informative examples from those negatives having the highest probability of being misclassified. We detail our proposal as follows.

3.2.1 Virtual Labeling

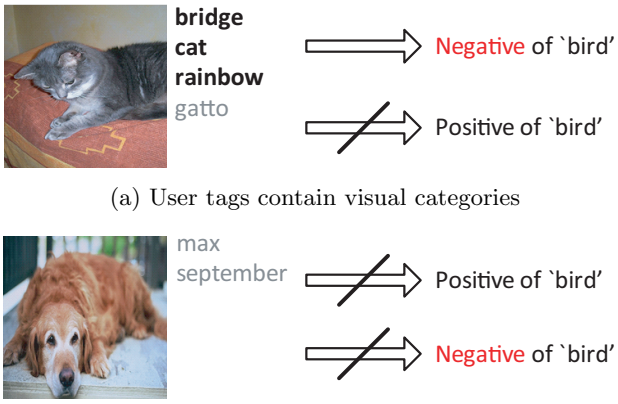
For social-tagged data, even though user tags are often unreliable for identifying positive examples, we argue that they are reliable for determining negative examples, allowing us to construct an effective virtual labeling procedure exploiting tag statistics and semantics.

We base the virtual labeling procedure on our observation about social image tagging. User tags of an image may contain some visual categories, or they may contain no visual categories, as shown in Fig. 2. In both cases, determining the positiveness of the image to a given category w is difficult, due to the subjectiveness of social tagging. However, on a set of randomly sampled images we observe that if an image is labeled with visual categories, but not labeled with w or its semantically related tags, the image is likely to be a negative example of w. We illustrate this observation in Fig. 2(a).

To obtain a set of reliable negatives, we need to determine V_w, a tagging vocabulary the average user uses to depict the category w, where $V_w \subset V$. By simply excluding images labeled with tags from V_w, we will obtain a set of reliable negative examples. We use S_{w-} to represent the virtually labeled negative set,

$$S_{w-} \leftarrow virtual\ labeling(S, w). \quad (1)$$

Concerning general criteria for creating V_w, to cope with the diversity of user tags, we construct V_w as a set of tags semantically correlated to the category. Semantic correlation between tags can be measured, say, by tag co-occurrence in a large corpus [5] or by human knowledge [15]. Note that our virtual labeling is conducted in the tag space, rather than in the visual feature space wherein visual categorization is performed. As a consequence, we obtain reliable negatives, among which we expect sufficient samples which are informative for training classifiers.



(a) User tags contain visual categories

(b) User tags do not contain visual categories

Figure 2: Inferring negative examples of a specific visual category by tag reasoning. Tags corresponding to visual categories are marked by a bold font. If an image is labeled with visual categories, but not with a target category, say 'bird', or its semantically related tags, the image is likely to be a negative example of the category. Images in all other cases are not taken as positives, nor as negatives.

3.2.2 Adaptive Sampling

The virtually labeled set S_{w-} is large, say having millions of images. Directly training classifiers on B_{w+} and S_{w-} is not only computationally challenging, but also suffering from extreme class imbalance. To make learning feasible, we iteratively exploit S_{w-} by performing multi-round learning. We use T to denote the number of learning rounds, and $t = 1, \ldots, T$ to index the rounds. In each round t, we adaptively select the most informative negative examples based on classifiers trained in previous rounds. To that end, we propose a two-stage adaptive sampling strategy. Suppose we have a classifier $G_t(x, w)$ obtained in the round t. In the first stage, we randomly sample n_u samples from S_{w-} to form a candidate set U_t,

$$U_t \leftarrow random\ sampling(S_{w-}, n_u). \quad (2)$$

To scale down the computational cost of selecting informative negatives from U_t and to reduce the chance of having genuine positives in U_t, we make $n_u \ll |S_{w-}|$. In the second stage, we use $G_{t-1}(x, w)$ to predict labels for each example in U_t, and obtain $\widetilde{U}_t$ in which each example is associated with a likelihood score of being positive to w,

$$\widetilde{U}_t \leftarrow predction(U_t, G_{t-1}(x, w)). \quad (3)$$

We consider examples which are *most misclassified*, i.e., predicted as positive with the largest scores, the *most informative* negatives. We rank examples in $\widetilde{U}$ by their scores in descending order and select the top ranked examples as the informative negative set found in the round t. We denote this negative set as $B^{(t)}_{w-}$. To bypass class imbalance, we enforce the number of the selected negatives to be equal to $|B_{w+}|$, namely

$$B^{(t)}_{w-} \leftarrow selection(\widetilde{U}_t, |B_{w+}|), \quad (4)$$

where $|\cdot|$ denotes set cardinality. By repeating the adaptive

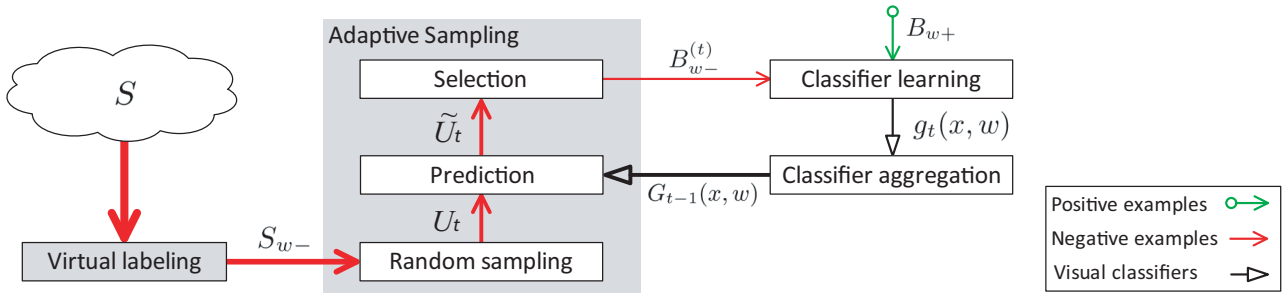


Figure 3: The proposed *social negative bootstrapping* approach. Given a specific visual category w and a positive set B_{w+}, we obtain a series of informative negative sets $\{B_{w-}^{(t)}\}$ from a large set of virtually labeled negative examples S_{w-} by multi-round adaptive sampling. In round t, we use $G_{t-1}(x,w)$ to classify a candidate set U_t, and select the most misclassified negatives to form $B_{w-}^{(t)}$. To initialize the bootstrapping process, $B_{w-}^{(1)}$ is randomly sampled from S_{w-}. By iteratively exploiting the informative negatives, we obtain visual classifiers with better discrimination ability, but without the cost of manually labeling any negatives.

sampling procedure, we iteratively select informative negatives from S_{w-} in an adaptive manner.

3.2.3 *Classifier Learning and Aggregation*

In each round t, we learn a new classifier $g_t(x,w)$ from B_{w+} and $B_{w-}^{(t)}$. As $B_{w-}^{(t)}$ is composed of negatives which are most misclassified by previous classifiers, we suppose that the new classifier is complementary to its ancestors. Therefore, we choose classifier aggregation to obtain the final classifier. Let $G_t(x,w)$ be an aggregated classifier which uniformly combines $g_t(x,w)$ and the previous t-1 classifiers:

$$G_t(x,w) = \frac{t-1}{t} G_{t-1}(x,w) + \frac{1}{t} g_t(x,w). \quad (5)$$

To trigger the bootstrapping process, we train an initial classifier $g_1(x,w)$ on B_{w+} and $B_{w-}^{(1)}$, which consists of examples randomly sampled from S_{w-}, with $|B_{w-}^{(1)}| = |B_{w+}|$.

We illustrate the entire framework in Fig. 3, with the algorithm given in Table 1. By adaptively selecting informative negative sets, social negative bootstrapping enables us to derive visual classifiers with better discrimination ability.

Table 1: The proposed social negative bootstrapping algorithm.

INPUT: visual concept w, expert-labeled positive examples B_{w+}, social-tagged examples S, and the number of learning rounds T.
OUTPUT: visual classifier $G_T(x,w)$.

1. **Creating negative example pool**:
   $S_{w-} \leftarrow virtual\ labeling(S,w)$.
2. **Creating an initial classifier**:
   (a) $B_{w-}^{(1)} \leftarrow random\ sampling(S_{w-}, |B_{w+}|)$.
   (b) $g_1(x,w) \leftarrow classifier\ learning(B_{w+}, B_{w-}^{(1)})$.
   (c) $G_1(x,w) = g_1(x,w)$.
3. **For** $t = 2,\ldots,T$ **do**
   3.1 **Adaptive sampling**:
   (a) $U_t \leftarrow random\ sampling(S_{w-}, n_u)$.
   (b) $\widetilde{U}_t \leftarrow prediction(U_t, G_{t-1}(x,w))$.
   (c) $B_{w-}^{(t)} \leftarrow selection(\widetilde{U}_t, |B_{w+}|)$.
   3.2 **Classifier learning**:
   $g_t(x,w) \leftarrow classifier\ learning(B_{w+}, B_{w-}^{(t)})$.
   3.3 **Classifier aggregation**:
   $G_t(x,w) = \frac{t-1}{t} G_{t-1}(x,w) + \frac{1}{t} g_t(x,w)$.

4. EXPERIMENTAL SETUP

We compare the proposed approach with the following two types of baselines, both of which rely on random sampling to obtain negative training data: 1) "random sampling" [10, 11, 19, 25, 32], and 2) "random+aggregation" [17]. For a fair comparison, whenever applicable, we will make our approach and the baselines share the same input and parameters.

4.1 Data sets

Positive training set B_{w+}. We choose the PASCAL VOC 2008 training set [7], collected from Flickr, with expert-labeled ground truth for 20 visual categories. For each category, we randomly sample 50 positive examples as B_{w+}.

Social-tagged image set S. We construct S as follows. We create the visual category vocabulary V by taking the intersection between the ImageNet vocabulary [6] and a social tagging vocabulary in which each tag is used by at least 100 distinct users in a set of 10 million Flickr images. The size of V is 5,009. Next, we go through 3.5 million Flickr images[1] created in our previous work [12], and remove images batch-tagged or having no tags from V. We end with S consisting of 650K images.

Two test sets. To evaluate classifiers derived from the same training set but by different approaches, we adopt the following two test sets, which were created independently by manually labeling different subsets of Flickr images. For within-dataset visual categorization, we adopt the VOC2008 validation set [7]. To test the robustness of the proposed approach in a cross-dataset setting, we choose the NUS-OBJECT test set [4]. We present in Table 2 data statistics of the training and test sets.

4.2 Implementation

Image representation. Since vector-quantized keypoint descriptors are effective features for visual categorization, we follow this convention. In particular, we adopt dense sampling for keypoint localization and SURF [1] for keypoint description, using a fast implementation of dense-SURF [24].

[1]Data available at `http://staff.science.uva.nl/~xirong`

Table 2: Statistics of the training and test sets used in our experiments. For each category w, we train classifiers on a small number of positive set B_{w+} and a large amount of social-tagged negative set S_{w-}.

	Training set		Positives in test sets (%)	
Category w	$\lvert B_{w+}\rvert$	$\lvert S_{w-}\rvert$	VOC08-val	NUS-OBJECT
aeroplane	50	521,010	5.3	4.8
bicycle	50	484,144	4.5	–
bird	50	395,079	6.3	5.8
boat	50	438,637	4.3	7.1
bottle	50	390,601	5.2	–
bus	48	511,708	2.3	–
car	50	383,319	10.1	3.5
cat	50	482,091	7.6	3.5
chair	50	327,967	8.0	–
cow	37	521,429	1.7	1.3
diningtable	50	484,960	2.4	–
dog	50	489,730	9.1	4.0
horse	50	525,110	4.6	2.6
motorbike	50	513,191	4.6	–
person	50	190,541	48.9	–
pottedplant	50	520,920	4.3	–
sheep	32	508,885	1.4	–
sofa	50	401,056	3.0	–
train	50	515,572	3.3	2.1
tv/monitor	50	228,876	4.9	–

With the SURF descriptors quantized by a codebook of 4000 bins, an image is represented by a 4000-dimensional feature which describes dominant structural patterns of that image.

Base classifiers. The proposed approach does not rely on specific classification models. Here we instantiate $g_t(x, w)$ using Support Vector Machine (SVM) for its good performance [26]. Since we do not aim for the best possible performance, but rather focus on the performance gain, we train two-class SVM classifiers using LIBSVM's default cost parameter [3], and the χ^2 kernel.

Parameters of social negative bootstrapping. To create the social-tagged negative pool for a given category w, we compute the Normalized Google Distance (NGD) [5] between tags and w on the 10 million set. Tags whose distance to w is smaller than 1 are considered as semantically correlated to w. Notice that the tag 'face' is strongly correlated to 'person' related concepts, but it tends to be underused in social tagging. Thus, if an image has faces detected by the Viola-Jones detector [27], we add 'face' to the tags of that image. We combine the correlated tags and childnodes of w in WordNet to form the correlated tag set,

$$V_w = \{w' \in V | NGD(w, w') < 1 \text{ or } w' \text{ is a WordNet childnode of } w\}. \quad (6)$$

For the size of the candidate set in each learning round, namely n_u in Eq. 2, we strike a balance between effectiveness and efficiency, with $n_u = 1000$ as our choice. We observe that the overall performance becomes stable after 50 learning rounds, therefore we set $T = 50$.

Parameters of the two negative sampling baselines. We use the same negative pool S_{w-} as used in the proposed approach for the two baseline approaches. In each learning round t, "random sampling" randomly selects $|B_{w+}|$ negative examples from S_{w-} to train a classifier, while "random+aggregation" uniformly aggregates this classifier and the previous t-1 classifiers.

Given the parameter setting above, we train up to $20 \times 50 \times 2$=2,000 base classifiers in total.

Evaluation criteria. For each visual category, we predict the presence of that category in a test image with a real-valued confidence score. Images in a test set are ranked according to their scores in descending order. To evaluate the performance, we adopt Precision at 20 (P20) to compare the top ranked results, and Average Precision (AP) for the whole ranked list.

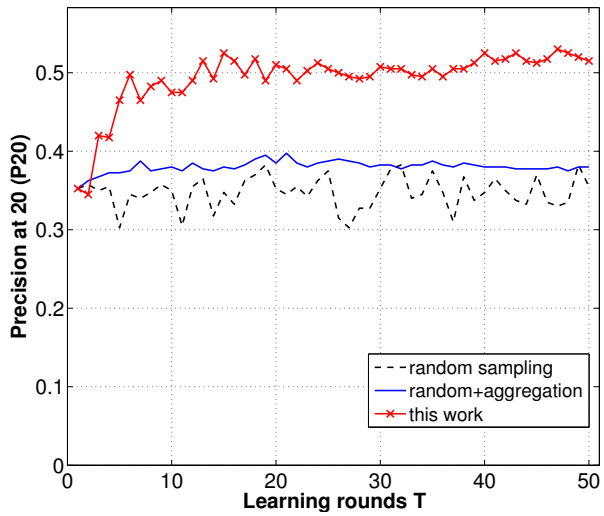
5. RESULTS

5.1 Comparing Different Approaches

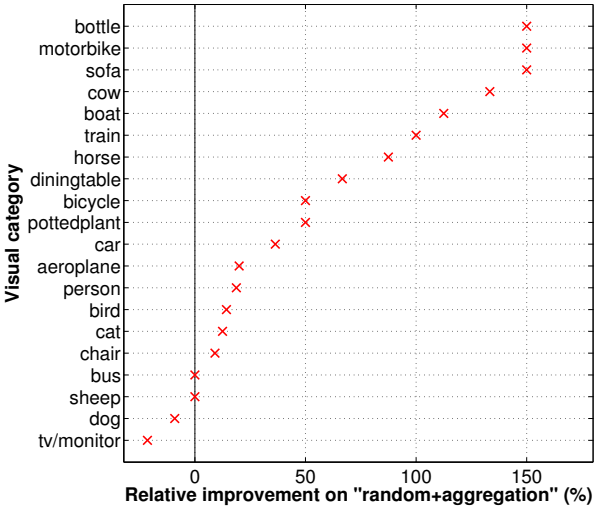
As shown in Fig. 4(a), the proposed approach compares favorably to the baselines for ranking positive results at the top. In the first round, as no classifier is available, all approaches start with the same negative set $B^{(1)}_{w-}$, and consequently produce the same classifier $G_1(x, w)$. Afterwards, while the baseline approaches keep selecting random negatives, our approach starts seeking the most informative negatives. The "random sampling" approach is affected by the random factor in sampling, so its performance varies. The "random+aggregation" approach reduces such variance by combining classifiers. However, as the performance curves show, "random+aggregation", with a P20 score of 0.380 at T=50, can hardly go beyond the best performance of "random sampling", which is 0.383. The results clearly show the limitation of obtaining negatives by random sampling. In contrast, our approach reaches a P20 score of 0.513, which is 34.1% better than the best performance of "random sampling", and a 35.0% relative improvement over "random+aggregation". By adaptively and iteratively sampling the most informative negatives, we obtain visual classifiers with higher accuracy.

As shown in Fig. 4(b), for 10 out of the 20 categories, we obtain a relative improvement of at least 50% on "random+aggregation". Note that for four categories, i.e., 'bus', 'sheep', 'dog', and 'tv/monitor', our approach does not improve the baseline. For the category 'dog', we find that closeups of flowers are frequently selected as the most informative negatives, and consequently, examples of other good negative classes, e.g., 'horse', are outnumbered. Probably because of such bias, the performance degenerates. For the category 'tv/monitor', images of rectangular objects are continuously selected. Classifiers trained on such negatives seem to be less powerful to separate the category from 'person', the most frequent category in the VOC08-val set. These results suggest that for certain categories, the diversity of negative training examples might be reduced to some extent. Nevertheless, our approach indeed correctly ranks the first result of the four categories, while the baseline fails. In general, when compared to random sampling, adaptive sampling results in more accurate visual classifiers.

The proposed approach is also effective for ranking an entire set, as shown in Fig. 5. Notice that our performance curve dips at T=2. This is because $g_2(x, w)$ is derived from $B^{(1)}_{w-}$, which are the most misclassified negatives by $g_1(x, w)$, and thus much distinct from generic negatives. Consequently, $g_2(x, w)$ is less effective for classifying generic negatives. Nevertheless, as subsequent classifiers are designed to be complementary to their ancestors, such ineffectiveness is tentative and will be resolved by adaptive sam-

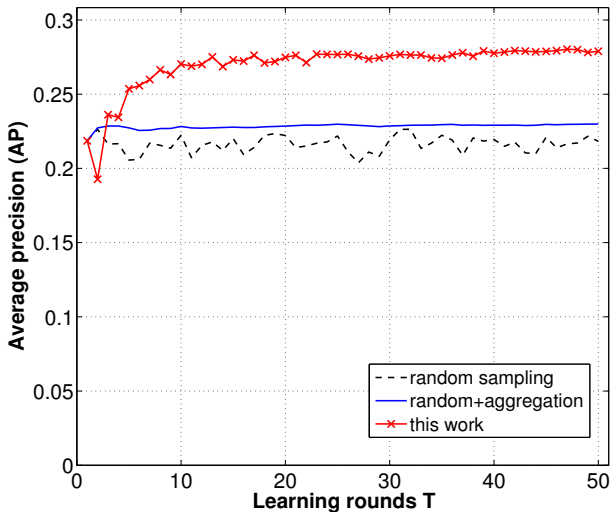


(a) Overall comparison

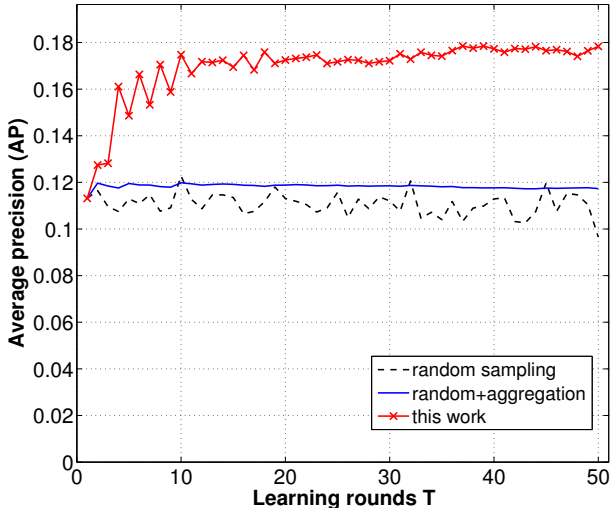


(b) Per-category comparison (T=50)

Figure 4: Within-dataset visual categorization. Test set VOC08-val. For 10 out of the 20 categories, classifiers trained by the proposed approach is at least 50% better in terms of Precision at 20. The considerable improvement is achieved without manually labeling any negative examples.



(a) Test set VOC08-val



(b) Test set NUS-OBJECT

Figure 5: Cross-dataset visual categorization. The proposed approach is not only superior to the baselines for ranking an entire set, but also effective in the cross-dataset setting.

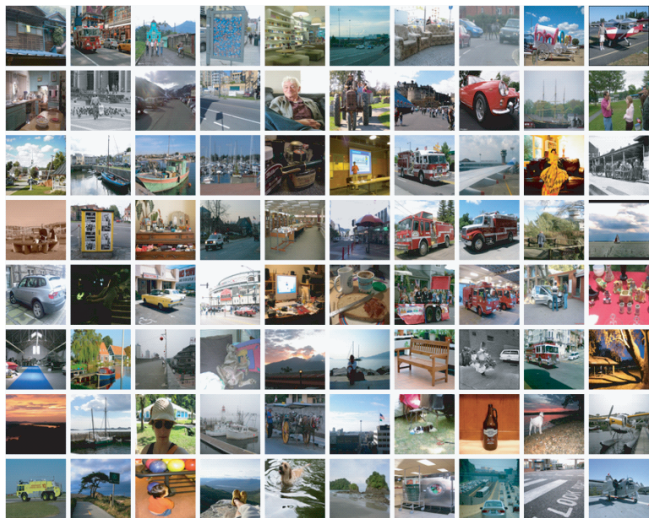
pling. When compared to "random+aggregation" with an AP score of 0.117, our approach reaches an AP score of 0.178 on NUS-OBJECT. The cross-dataset experiment shows the robustness of the proposed approach.

5.2 Examples

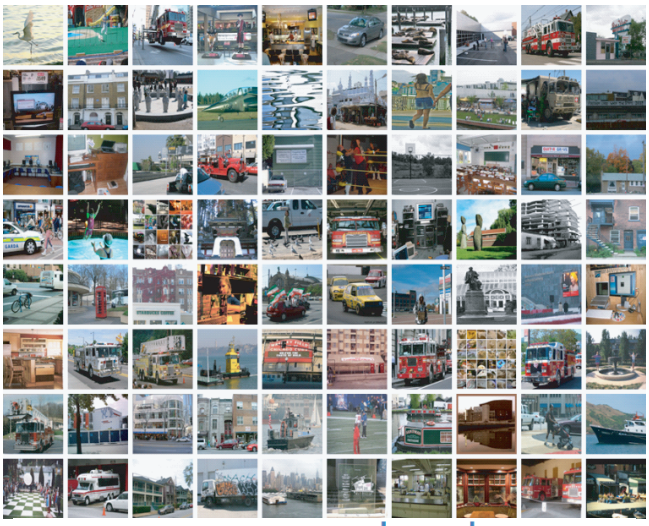
We show in Fig. 6 the most informative negative examples found by our approach. As we use the dense-SURF feature, negative examples visually close to the positives in terms of their structural patterns are predicted as informative for classifier training. See the categories 'cow', 'train', and 'bus' for instance. Because the examples are selected without manual verification, genuine positives may be included occasionally, see Fig. 6(c). Nevertheless, as they are in the minority, their impact on the bootstrapping process is minimal. Further, by visualizing the distribution of user tags in the selected negatives with a tag cloud, we see which negative classes are most informative to a certain category. We conjecture that such a relationship is feature-dependent, i.e., different features result in different informative negatives for the same category. Moreover, we observe that the informativeness relationship between categories seems to be asymmetric. For instance, while 'bus' and 'car' are most informative to 'train', the most informative negative class for 'bus' and 'car' is 'firetruck', rather than 'train'. A plausible explanation is that 'firetruck' bears more resemblance to the former two categories in terms of properties of the object, e.g., rectangular curves, and visual context, in casu, street. In sum, the qualitative results in Fig. 6 further illustrate the effectiveness of the proposed approach in finding negative examples informative for learning visual classifiers.

(a) Negatives of 'cow'

(b) Negatives of 'train'



(c) Negatives of 'car'



(d) Negatives of 'bus'

Figure 6: The 80 most informative negative examples, found by the proposed approach, for specific visual categories. By visualizing the distribution of user tags in the selected negatives as a tag cloud, we see which negative classes are most informative to a given category.

6. CONCLUSIONS

In this paper we study how to sample informative negative examples from widely available social-tagged images for visual categorization. To that end, we propose the *social negative bootstrapping* approach. Our major findings are as follows. Negative examples can be obtained, with no human interaction, by the designed virtual labeling procedure which exploits tag statistics and semantics. Virtual labeling, in combination with adaptive sampling, allows us to harvest informative negatives from those negatives having the highest probability of being misclassified. When compared to classifiers trained on randomly sampled negatives, classifiers derived from such informative negatives have better discrimination ability. The proposed approach is thus strategically better than random negative bootstrapping.

Experiments on two image benchmarks and 650K virtually labeled negative examples verify our proposal. For the majority of visual categories, we obtain a relative improvement of at least 50%, in terms of precision at 20. Moreover, cross-dataset visual categorization shows the robustness of the proposed approach. Notice that the substantial progress is achieved without the need of labeling any negative examples. As the promising results suggest, social negative bootstrapping opens up interesting avenues for future research.

7. REFERENCES

[1] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359, 2008.

[2] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. NUS-WIDE: A real-world web image database from National University of Singapore. In *CIVR*, 2009.

[5] R. Cilibrasi and P. Vitanyi. The Google similarity distance. In *IEEE Trans. on Knowl. and Data Eng.*, 2004.

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

[7] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[8] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55:119–139, 1997.

[9] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6(5):429–449, 2002.

[10] L. Kennedy, S.-F. Chang, and I. Kozintsev. To search or to label?: Predicting the performance of search-based automatic image classifiers. In *ACM MIR*, 2006.

[11] X. Li and C. Snoek. Visual categorization with negative examples for free. In *ACM MM*, 2009.

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

[13] B. Liu, Y. Dai, X. Li, W. Lee, and P. Yu. Building text classifiers using positive and unlabeled examples. In *ICDM*, 2003.

[14] D. Liu, X.-S. Hua, and H.-J. Zhang. Content-based tag processing for internet social images. *Multimedia Tools Appl.*, 51:723–738, 2011.

[15] G. Miller. WordNet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[16] M. Naphade, J. Smith, J. Tešić, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MM*, 13(3):86–91, 2006.

[17] A. Natsev, M. Naphade, and J. Tešić. Learning the semantics of multimedia queries and concepts from a small number of examples. In *ACM MM*, pages 598–607, 2005.

[18] B. Russell, A. Torralba, K. Murphy, and W. Freeman. LabelMe: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3):157–173, 2008.

[19] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010. in press.

[20] A. Sun and S. Bhowmick. Quantifying tag representativeness of visual content of social images. In *ACM MM*, 2010.

[21] D. Tax. *One-class classification*. PhD thesis, Delft University of Technology, 2001.

[22] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *ACM MM*, 2001.

[23] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, 2008.

[24] J. Uijlings, A. Smeulders, and R. Scha. Real-time visual concept classification. *IEEE Trans. MM*, 12(7):665–681, 2010.

[25] A. Ulges, C. Schulze, M. Koch, and T. Breuel. Learning automatic concept detectors from online video. *Comput. Vis. Image Underst.*, 114(4):429–438, 2010.

[26] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 2000.

[27] P. Viola and M. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57:137–154, 2004.

[28] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma. Annotating images by mining image search results. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1919–1932, 2008.

[29] R. Yan, A. Hauptmann, and R. Jin. Negative pseudo-relevance feedback in content-based video retrieval. In *ACM MM*, 2003.

[30] K. Yanai and K. Barnard. Probabilistic web image gathering. In *ACM MIR*, 2005.

[31] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ACM MM*, 2010.

[32] S. Zhu, G. Wang, C.-W. Ngo, and Y.-G. Jiang. On the sampling of web images for learning visual concept classifiers. In *CIVR*, 2010.