Deep Classifiers from Image Tags in the Wild

Hamid Izadinia[*]
University of Washington
Seattle, WA
izadinia@cs.uw.edu

Bryan C. Russell
Adobe Research
San Francisco, CA
brussell@adobe.com

Ali Farhadi
University of Washington
Seattle, WA
ali@cs.uw.edu

Matthew D. Hoffman
Adobe Research
San Francisco, CA
mathoffm@adobe.com

Aaron Hertzmann
Adobe Research
San Francisco, CA
hertzman@adobe.com

http://deep-tagging.cs.washington.edu

ABSTRACT

This paper proposes direct learning of image classification from image tags in the wild, without filtering. Each *wild tag* is supplied by the user who shared the image online. Enormous numbers of these tags are freely available, and they give insight about the image categories important to users and to image classification. Our main contribution is an analysis of the Flickr 100 Million Image dataset, including several useful observations about the statistics of these tags. We introduce a large-scale robust classification algorithm, in order to handle the inherent noise in these tags, and a calibration procedure to better predict objective annotations. We show that freely available, wild tags can obtain similar or superior results to large databases of costly manual annotations.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning*; I.5.1 [**Pattern Recognition**]: Models—*Neural nets*

Keywords

Deep Learning; Tags in the Wild; Large-scale Robust Classification; Image Tag Suggestion; Image Retrieval

1. INTRODUCTION

Image classification has made dramatic strides in the past few years. This progress is partly due to the creation of large-scale, hand-labeled datasets. Collecting these datasets involves listing object categories, searching for images of each category, pruning irrelevant images and providing detailed labels for each image. There are several major issues with this approach. First, gathering high-quality annotations for large datasets requires substantial effort and expense. Second, it remains unclear how best to determine the list of categories. Existing datasets comprise only a fraction of recognizable visual concepts, and often miss concepts that are important to end-users. These datasets draw rigid distinctions between different types of concepts (e.g., scenes, attributes, objects) that exclude many important concepts.

This paper introduces an approach to learning about visual concepts by employing *wild tags*. That is, *we directly use the tags provided by the users that uploaded the images to photo-sharing services*, without any subsequent manual filtering or curation. While previous vision datasets have used data from photo-sharing sites [25] or other, smaller-scale sources [5, 15, 29], the scale and scope of these datasets is tiny in comparison to the uncurated wild tags available on sharing sites. Enormous number of images and tags are freely provided by users worldwide, representing a vast, untapped source of data for computer vision. Moreover, tags in the photosharing services give insight into the image categories that are important to users and include scenes (beach), objects (car), attributes (rustic), activities (wedding), and visual styles (portrait), as well as concepts that are harder to categorize (family). Furthermore, tags are shared on many types of popular sharing sites, such as Behance, Shapeways, Imgur, and Lookbook. Learning to harness these data sources could benefit both computer vision, as well as consumer interfaces, such as tag suggestion.

While wild tags offer tremendous potential for computer vision, they also present significant challenges. These tags are entirely uncurated, so users provide different numbers of tags for their images, and choose different subsets of tags [1, 20]. Each tag may have multiple meanings, and, conversely, multiple terms may be used for the same concept. To paraphrase a famous saying, an image is worth thousands of visual concepts, but, yet, the average Flickr image has only 4.07 tags, and the precise sense of each is ambiguous. Is it even possible to learn good models from such messy data?

The main contribution of this work is to study the statistics of the tags that users provide for their uploaded images. A crucial first step in any data modeling is to understand the dataset, and these online datasets exhibit many types of structure and bias. We make many observations about the structure of this data that will inspire future research. Furthermore, we show that, perhaps surprisingly, good deep classifiers can be trained from this data. As a first step toward exploiting the structure of the data, we describe a stochastic EM approach to robust logistic regression, for large-scale training with randomly-omitted positive labels. Since tag noise is different for different tags, the tag outlier probabilities are learned simultaneously with the classifier weights. Furthermore, we describe calibration of the trained model probabilities from a small annotation set.

We demonstrate results for several useful tasks: predicting the tags that a user would give to an image, predicting objective annotations for an image, and retrieving images for a tag query. For the latter two tasks, which require objective annotations, we cali-

[*] This work was done while the author was an intern at Adobe Research.

.

brate and test on the manually-annotated NUS-WIDE [4] dataset. We show that training on a large collection of freely available, wild tags alone obtains comparable performance to using a smaller, manually-annotated training set. That is, we can learn to predict thousands of tags *without any curated annotations at all.* Moreover, if we calibrate the model with a small annotated dataset, we can obtain superior performance to conventional annotations at a tiny fraction (1/200) of the labeling cost. Our methods could support several annotation applications, such as auto-suggesting tags to users, clustering user photos by activity or event, and photo database search. We also demonstrate that using robust classification substantially improves image retrieval performance with multi-tag queries.

2. RELATED WORK

Current computer vision research is driven, in part, by datasets. These datasets are built through a combination of webscraping and crowd-sourcing, with the aim of labeling the data as cleanly as possible. Important early tagging datasets such as Corel 5k [5] and IAPR TC 12 [15] comprise only a few thousand images each, and at most a few hundred possible tags. ImageNet [25] is now the most prominent whole-image classification dataset, but other significant recent datasets include NUS-WIDE [4], SUN scene attribute database [22, 30], and PLACES [31]. The curation process has a number of drawbacks, such as the cost of gathering clean labels and the difficulty in determining a useful space of labels. It is unclear that this procedure alone will scale to the space of all important visual concepts [25]. We take a complementary approach of using a massive database of freely available images with noisy, unfiltered tags.

Previous work in automatic tagging from tagged examples uses nearest-neighbor-like approaches, notably TagProp [8], or by model-based regression (e.g., [13]). This work has used small-scale datasets, which are quite different in scale and statistics than wild tag datasets. Our work is also the first to explore deep feature learning for tagging. Li et al. [14] give a detailed survey of the related work in this area, and perform comparisons on large collections of Flickr tags. Gong et al. [7] use raw Flickr tags as side-information for associating images with descriptive text. Gong et al. [6] train deep features to predict NUS-WIDE tags.

Classification with noisy labels is a well-studied learning problem, e.g., [19, 23, 24, 28]. We extend robust logistic regression [23] to large-scale learning with Stochastic EM. The model is similar to concurrent research on deep learning with noisy labels [24, 28] but with a simpler interpretation and better stability guarantees.

3. ANALYSIS OF WILD TAGS

When can wild tags be useful, and when can they be trusted? In this section, we analyze the tags provided on Flickr, and compare them to two datasets with ground truth labels. Some of these observations motivate our algorithm in Section 4, and others provide fodder for future research. Our main dataset is the Yahoo/Flickr Creative Commons 100M dataset[1], which we refer to as YFCC100M. This dataset comprises 99.3 million images, each of which includes a list of the tags supplied by the user that uploaded the image.

3.1 Types of Tags

The YFCC100M dataset provides an enormous number of images and tags (Figure 1) that could be used for learning. There are 5400 tags that occur in at least 1000 images. The set of tags provides a window into the image concepts that are important to users. Many of these represent types of image label that are not represented in previous datasets. Some of the most important tag types are as follows: **events and activities** such as travel, music, party, festival, football, school; **specific locations** such as california and italy; **scene types** such as nature, park, urban, sunset, etc.; **the seasons** (fall, winter, summer, spring); **image style** such as portrait, macro, vintage, hdr; and **art and culture** such as painting, drawing, graffiti, fashion, punk. Many frequent tags also represent categories that do not seem learnable from image data alone, which we call **non-image tags**, including years (2011, 2012, ...), and specific camera and imaging platforms (nikon, iphone, slr).

Flickr tag	# Flickr	synset	# node	# subtree
travel	1221148	*travel.n.01*	0	0
wedding	734438	*wedding.n.03*	1257	1257
flower	907773	*flower.n.01*	1924	339376
art	902043	*art.n.01*	0	11353
music	826692	*music.n.01*	0	0
party	669065	*party.n.01*	0*	0
nature	872029	*nature.n.01*	0	0
beach	768752	*beach.n.01*	1713	1773
city	701823	*city.n.01*	1224	1224
tree	697009	*tree.n.01*	1181	563038
vacation	694523	*vacation.n.01*	0	0
park	686458	*park.n.01*	0	0
people	641571	*people.n.01*	1431	1431
water	640259	*water.n.06*	759	7585
architecture	616299	*architecture.n.01*	1298	1298
car	610114	*car.n.01*	1307	40970
festival	609638	*festival.n.01*	0	0
concert	605163	*concert.n.01*	1322	1322
summer	601816	*summer.n.01*	0	0
sport	564703	*sport.n.01*	1888	200402

Table 1: The 20 most frequent tags in YFCC100M, after merging plurals and omitting location and non-image tags. Corresponding ImageNet synsets are given, along with synset node and subtree counts. These statistics are typical: we estimate that nearly half of popular Flickr tags are absent from ImageNet. Moreover, even when there is correspondence, some ImageNet tags do not capture all meanings of a term (Section 3.2). (***There are 66 party images in ImageNet, in the wrong synset *party.n.04*.)**

3.2 ImageNet and The Dataset Gap

We hypothesized that YFCC100M contains information missing from existing, curated datasets. Does it? We compare YFCC100M to the ImageNet image classification dataset [25], which comprises 14 million images gathered from several image search engines, and labeled according to the WordNet hierarchy [16] through a carefully-designed crowdsourcing procedure.

Missing concepts. In order to quantify the dataset gap, we studied the 100 most frequent tags in YFCC100M (after omitting the non-image and location tags described above). For each tag, we manually determined a correspondence to WordNet, as follows. In WordNet, each concept is represented by a synonym set, or *synset*. WordNet synsets are ordered, and most tags (78%) correspond to the first WordNet noun synset for that word. For example, the tag beach corresponds to the synset *beach.n.01*. In other cases, we corrected the match manually. The most-frequent examples are shown in Table 1. Based on this analysis and some simple calculations, we estimate that about half of the common Flickr image tags are absent from ImageNet. Some of these missing tags are covered by scene [22, 30, 31] and style databases [11, 18]. Some common tags in Flickr do not even exist in the WordNet hierarchy, such as cosplay (costume play), macro (macro photography), and vintage (in the sense of "retro" or "old-style"). We also observed a few large set of images assigned to the wrong ImageNet synset, including "party," "landscape," and "tree/tree diagram."

Poorly-represented concepts. Even when there is a corresponding tag in ImageNet, the tag may be poorly represented. There

[1] http://yahoolabs.tumblr.com/post/89783581601

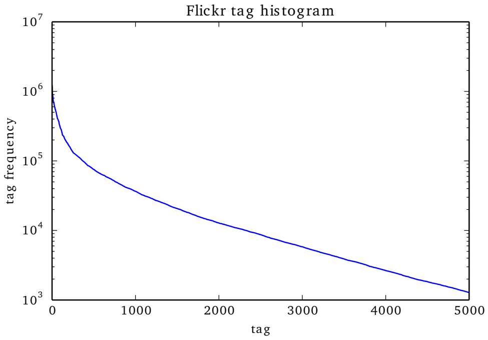

Figure 1: Tag histogram for the most popular tags, excluding non-image tags. The distribution is heavy-tailed, and there are 5400 tags with more than 1000 images each.

are 11k images in the ImageNet *art.n.01* hierarchy, but there are only 8 subtrees of *art.n.01* with at least 1000 images; the biggest ones are “olympian zeus,” “cinquefoil,” and “finger-painting;” and there are no subtrees for “painting,” “drawing,” or “illustration.” The ImageNet synset for “band” includes only images for “marching bands” and not, say, “rock bands.” Many image categories that are significant to users—for example, in analyzing personal photo collections—are not well represented in the ImageNet categories. Examples include family, travel, festival, and summer. Based on the above observations, we conclude that ImageNet falls far short of the full Flickr database for representing the set of visual concepts important to users. This is not in any way meant to disparage the substantial, important efforts of the ImageNet team, but to emphasize the enormous difficulty in trying to precisely curate a dataset including all important visual concepts.

3.3 Label Noise and Ambiguities

A fundamental challenge in dealing with wild tags is that the mapping from observed tags to true concepts is ambiguous. Here we discuss some of the ambiguities that we have observed.

Many terms have multiple or overlapping meanings. The simplest case is plurals, e.g., car and cars, which have different meanings but which seem to be more or less interchangeable tags on Flickr. Some tags have multiple distinct meanings [26], e.g., rock can mean both “rock-and-roll music,” and “rocky landscapes.” Trickier cases include terms like music, concert, and performance, which often overlap, but often do not. Some words are used nearly interchangeably, such as cat and kitten, even though their meanings are not the same. It seems that nearly all common tags exhibit some multiple meanings, though often one sense dominates the others. Synonyms are also common, as well as misspellings.

Multi-word tags often occur split up, e.g., images in New York are frequently tagged as New and York rather than New York. For this reason, tags like New and San are largely meaningless on their own. Merging these split tags (especially using cues from the other image metadata) is a problem for future research.

3.4 Analysis with Ground Truth

In this section, we perform analysis using the annotated subset of the **NUS-WIDE** dataset [4]. This is a set of 269,642 Flickr images with both wild tags, and “ground truth” annotations by undergraduate and high school students according to 81 concepts. There are a number of potential sources of noise with this dataset. Since the dataset was constructed by keyword searches, it is not an unbiased sample of Flickr, e.g., only one image in the dataset has zero keywords. Annotators were not asked to judge every image for every concept; a query expansion strategy was used to reduce annotator effort. Annotators were also asked to judge whether concepts were present in images in ways that may differ from how the images were originally tagged.

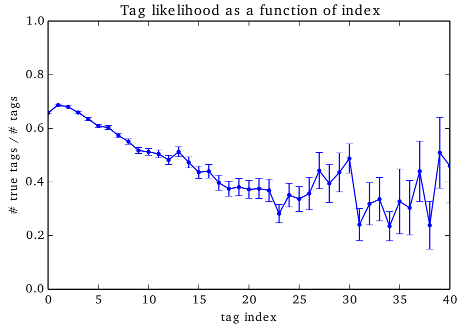

Figure 2: Tag likelihood as a function of index. (Error bars show standard error.)

Tagging likelihoods. We now quantify the accuracy of Flickr tags. We consider the Flickr images in NUS-WIDE that contain manual annotations, and we treat these 81 labels as ground truth. We assume an identity mapping between tags and annotations, i.e., the Flickr tag cat corresponds to the NUS-WIDE annotation cat. Overall, given that a tag correctly applies to an image, there is empirically a 38% chance that the uploader will actually supply it. This probability varies considerably for different tags, ranging from 2% for person to 94% for cat. Frequently-omitted tags are often non-entry-level categories [21] (e.g., person) or they are not an important subject in the scene [2] (e.g., clouds, buildings). Given that a tag does not apply, there is a 1% chance that the uploader supplies it anyway. Across the NUS-WIDE tags, this probability ranges from 2% (for street) to 0.04% (for toy).

Despite these percentages, false tags and true tags are almost equally likely, since only a few of the 81 tags correctly apply to each image. Each image has an average of 1.3 tags (of the 81), and an observed tag has only a 62% chance of being true. This percentage varies across different tags. None of these numbers should be taken as exact, because the NUS annotations are far from perfect. Additionally, many “false" tags are due to differences in word senses between Flickr and NUS-WIDE. For example, many earthquake images are clearly the result of earthquakes, but are labeled as negatives in NUS-WIDE. Many cat images that are annotated as non-cat are images of tigers, lions, and cat costumes. Many nighttime images were probably taken at night but indoors.

Tag index effects on accuracy. Flickr tags are provided in an ordered list. We observed that tags earlier in the list are often more accurate than later tags, and we again treat the NUS-WIDE annotations as ground truth in order to quantify this.

We find that the effect is substantial, as shown in Figure 2. A tag that appears first or second in the list of tags has about 65% chance of being accurate. A tag that occurs in position 20 or later has about a 35% chance of being accurate. The scales and shape of these plots also vary considerably across different tags.

Effect of total number of tags. We hypothesized that tag reliability depends on the total number of tags provided for an image. This was motivated by our observation of commercially-oriented sharing sites, where uploaders are incentivized to include extraneous tags in order to boost search results. However, we did not find substnatial evidence of this in Flickr.

4. WILD TAG CLASSIFICATION VIA ROBUST LOGISTIC REGRESSION

We now describe a Robust Logistic Regression (RLR) algorithm, designed to address the following observations from the previous section: wild tags often omit relevant tags, and the rate of omission is different for each tag. A conventional robust loss (e.g., Huber,

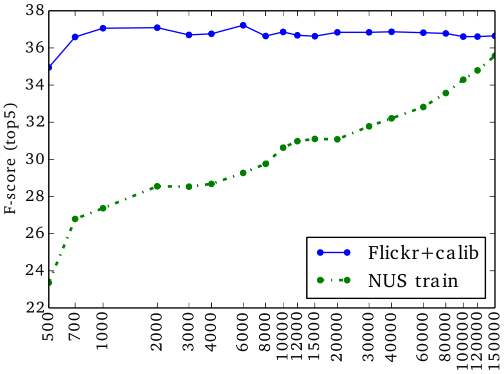

Figure 3: Effect of calibration set size on image annotation score. By first training on the annotation cost can be reduced by a factor of 200, while obtaining the same results.

Geman-McClure) would not be appropriate because of the need to set the loss function's parameters individually for each tag. The method is based on previous robust logistic regression methods [23], and we adapt these methods to the large-scale setting using Stochastic EM [3].

The classifier takes as input image features $\mathbf{x}$ and predicts class labels $y \in \{0, 1\}$. We perform prediction for each possible tag independently (i.e., mulitilabel classification), and so we consider simple binary classification in this section. We use as image features $\mathbf{x}$ the output of the last fully-connected layer of a Convolutional Neural Network [12] and assume a prediction model $z = \sigma(\mathbf{w}^T\mathbf{x})$, where $\sigma(t) = (1 + \exp(-t))^{-1}$ is the sigmoid function.

As discussed in Section 3, wild tags are often noisy. However, the logistic regression model assumes that the observed labels $\{y_i\}$ are mostly reliable—that is, it assumes that $y_i = 1$ almost always when $\mathbf{w}^T\mathbf{x}_i$ is large. To cope with this issue, we add a variable π defined as the probability that a user will supply a tag, conditioned on the tag being true for the image. Lower values of π dampens the influence of confident predictions, which allows the model to be robust to outliers. The model parameters are then $\theta = \{\mathbf{w}, \pi\}$, and the loss function for training is the negative log-likelihood of the data:

$$L(\mathbf{w}, \pi) = -\sum_i y_i \ln(\pi\sigma(\mathbf{w}^T\mathbf{x}_i)) + (1 - y_i)\ln(1 - \pi\sigma(\mathbf{w}^T\mathbf{x}_i)) \tag{1}$$

Optimization via stochastic EM algorithm. Learning the model for a given tag entails minimization of the loss with respect to $\mathbf{w}$ and π. Stochastic gradient descent could be used for all parameters [28]. However, we use Stochastic Expectation-Maximization (EM) [3], since the steps are simpler to interpret and implement, and the updates to π are numerically stable by design. Our stochastic EM algorithm applies the following steps to each minibatch:

1. Define the sufficient statistics for the minibatch as:
$$S_\alpha^{\text{mb}} \equiv \sum_i \alpha_i / N; \quad S_{y\alpha}^{\text{mb}} \equiv \sum_i y_i\alpha_i / N, \tag{2}$$
where N is the number of data points in the minibatch and:
$$\alpha_i = \begin{cases} 1 & y_i = 1 \\ \frac{(1-\pi)\sigma(\mathbf{w}^T\mathbf{x}_i)}{1-\pi\sigma(\mathbf{w}^T\mathbf{x}_i)} & y_i = 0 \end{cases} \tag{3}$$
Estimates of the average sufficient statistics for the full dataset are updated with a step size η, $S^{\text{ds}} \leftarrow (1 - \eta)S^{\text{ds}} + \eta S^{\text{mb}}$. In our experiments, we initialized S_α^{ds} and $S_{y\alpha}^{\text{ds}}$ to 1 and used a fixed step size of $\eta = 0.01$.

2. Compute π from the current estimate of the sufficient statistics, so that π is an estimate of the percentage of true labels that were actually supplied as tags, $\pi \leftarrow S_{y\alpha}^{\text{ds}} / S_\alpha^{\text{ds}}$.

	Supervision	Recall	Precision	F-score
Visual feature+kNN [6]	Clean(Train)	32.1	22.6	26.5
Visual feature+SVM [6]	Clean(Train)	34.2	18.8	24.3
CNN+Softmax [6]	Clean(Train)	48.2	22	30.2
CNN+Ranking [6]	Clean(Train)	42.5	22.8	29.7
CNN+WARP [6]	Clean(Train)	52	22.3	31.2
NUS-Wide, LR	Clean(Train)	58.2	26.1	36
NUS-Wide, LR, ft	Clean(Train)	**58.9**	**27.7**	**37.7**
YFCC, LR	Wild(Train)	61.4	22.3	32.7
YFCC, RLR	Wild(Train)	62.1	**22.6**	**33.1**
YFCC, LR, ft	Wild(Train)	63.4	21.9	32.6
YFCC, RLR, ft	Wild(Train)	**66.9**	21.1	32.1
YFCC, LR, Calib.	Wild(Train)+Clean(Calib.)	40.4	**39.3**	39.9
YFCC, RLR, Calib.	Wild(Train)+Clean(Calib.)	43.3	35.9	39.3
YFCC, LR, ft, Calib.	Wild(Train)+Clean(Calib.)	42.6	37.8	**40**
YFCC, RLR, ft, Calib.	Wild(Train)+Clean(Calib.)	**44.7**	36	39.9

Table 2: Image annotation scores, illustrating how the freely-available wild tags can augment or supplant costly manual annotations. Testing is performed on the NUS-WIDE test set. The first set of rows show training only on the NUS-WIDE training set with logistic regression, and the previously-reported state-of-the-art [6]. Each of the second set of rows is trained on YFCC100M with either LR or Robust LR, with or without fine-tuning (ft). The third section are also calibrated on the NUS train set. All scores are predictions-at-5.

3. The weights $\mathbf{w}$ are updated using stochastic gradient on L. It is straightforward to show that the gradient is computed by $\frac{dL}{d\mathbf{w}} = \sum_i (\sigma(\mathbf{w}^T\mathbf{x}_i) - \alpha_i)\,\mathbf{x}_i$. Neural network parameters may also be fine-tuned with this gradient.

Calibration. To cope with differences across datasets, we apply a calibration step to adapt the learned weights for a tag to a new dataset. We tested the calibration method from [27], but found it degraded performance. Instead we recover a bias β, given learned weights $\mathbf{w}$ on a large dataset such that the prediction model is $z = \sigma(\mathbf{w}^T\mathbf{x} + \beta)$. Very little curated data is necessary for this process, since only one new parameter is being estimated per tag. In our experiments, we train on the YFCC100M data and calibrate on a subset of NUS-WIDE annotations. More general domain adaptation methods (e.g., [9]) could also be used.

5. EXPERIMENTS

We now describe experiments for testing models learned from YFCC100M on several tasks, including tag prediction, image annotation, and image retrieval with one or more tags.

We use the architecture of Krizhevsky's network [12] and initialize our weights with pre-trained network for the large-scale object recognition task (ILSVRC2012) which is hosted in Caffe website [10]. Training is performed for 20,000 minibatches, with minibatch size of 500 images. Each minibatch takes 2 seconds, and a complete run takes 11 hours on a GeForce GTX780 GPU. Based on the observations in Section 3.4, we only keep the first 20 tags in all Flickr images in our experiments. We use a subset of 4,768,700 images from YFCC100M as training set and hold out another 200,000 for testing. The sets are split by user ID in order to ensure that images from the same user do not occur in both sets. Plural and singular tags are combined using WordNet's lemmatization.

5.1 Tag Prediction

We first test the following prediction task: given a new image, what tags would a user be likely to apply to this image? This task could be useful for consumer applications, for example, auto-suggesting tags for users when sharing their images. Note that this task is different from ground-truth prediction; we want to suggest tags that are both objectively accurate and likely to be applied by a user.

We trained a logistic regression baseline and a robust logistic regression model on our 4.7M-image YFCC100M training set, and

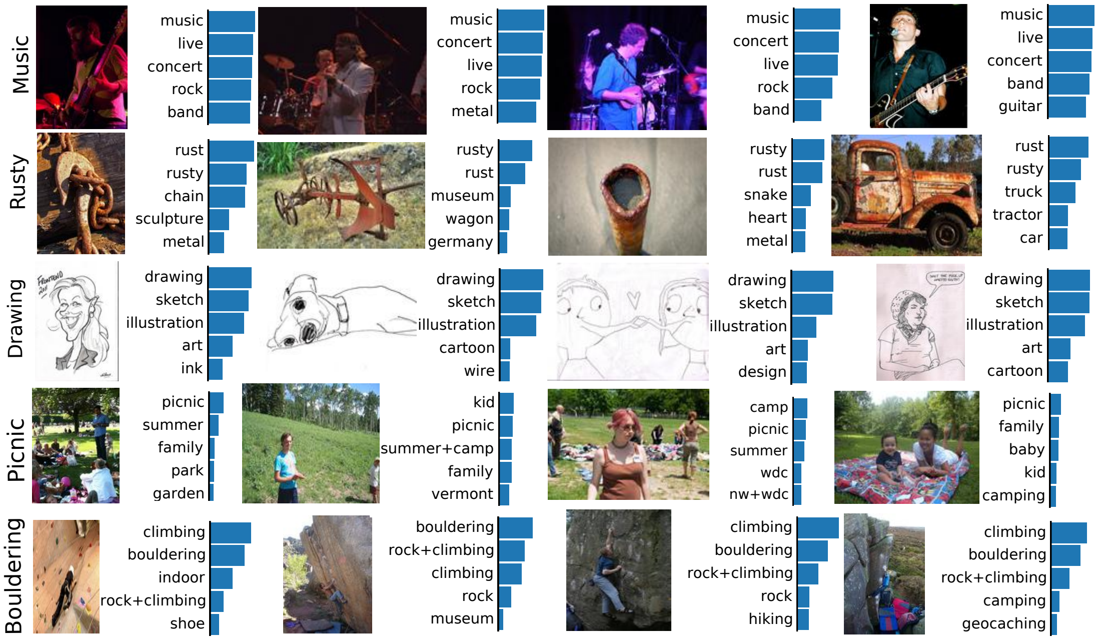

Figure 4: Single-tag retrieval results, and automatically-generated annotations. None of the query tags are in NUS-WIDE, and most (music, rusty, drawing, bouldering) are also absent from ImageNet. Many of the annotations are also absent from the other datasets.

	Recall	Precision	F-score
LR	9.7	7.9	8.7
RLR	11.7	8.0	9.5

Table 3: Tag prediction results on YFCC100M dataset. Robust logistic regression improves over logistic regression's ability to predict which tags a user is likely to apply to an image.

evaluated the models' ability to annotate images in the 200K-image YFCC100M test set. For each test image, the model predicts the probability of each tag occurring: $P(y = 1|\mathbf{x}, \mathbf{w}, \pi) = \pi\sigma(\mathbf{w}^T\mathbf{x})$. The final annotations were produced by selecting the top 5 most likely tags for each image. We evaluate overall precision, recall at 5 for each image, averaged over all images, as well as F-score. RLR achieves higher recall without sacrificing precision (Table 3). Figure 4 shows qualitative results using RLR on the Flickr test set.

5.2 Image Annotation

We next test the task: given an image, which labels objectively apply to it? We use the same YFCC100M training set as above, but evaluate on the manually-annotated 81 labels, treating them as ground truth. We also compare to models trained on NUS-WIDE. We evaluate per-tag precision, recall, and F-score, which are computed for each tag separately, and then averaged across tags (Table 2). We define precision for a tag that is never predicted to be 0. To predict annotations with RLR, we compute $\sigma(\mathbf{w}^T\mathbf{x})$, adding β when calibrated. Testing and training LR on NUS produces better scores than training on YFCC100M alone; it also produces better scores than the reported state-of-the-art on this dataset [6]. We get the best scores by training on YFCC100M and then calibrating on NUS (Section 4).

It is important to consider the cost of annotated labels. The wild tags in YFCC100M are basically free, whereas obtaining manual annotations is a very costly process. We compare training on a subset of NUS training annotations, versus YFCC100M training plus calibration with the same NUS subset. As shown in Figure 3, the calibration process can yield scores superior to training on the full annotation set, but with a 200x reduction in annotation cost.

We also compared our method to nearest-neighbor tagging methods, using published scores on IAPR TC12 (Table 4). Our method produces higher scores, except for one precision score. Applying nearest neighbor methods to YFCC100M is impractical in practice, due to the need to search in the entire dataset at test-time, even with nearest-neighbor acceleration.

Method	Recall	Precision	N+
JEC [15]	29	28	250
TagProp σML [8]	35	46	266
SKL-CRM [17]	32	51	274
RLR (ours)	41	46	277

Table 4: Results of comparison with nearest neighbor methods on IAPR TC12 [15] dataset. We follow the experimental setting of [8]. N+ shows the number of tags with non-zero recall.

	1 Tag	2 Tags	3 Tags
NUS-Wide, LR	*81*	*17.9*	*9.1*
YFCC, LR	70.1	8.5	2.3
YFCC, RLR	**71.9**	9.2	2.7
YFCC, LR, Calib	70.1	10.3	3.6
YFCC, RLR, Calib	**71.9**	**11**	**3.9**

Table 5: Image retrieval results, showing precision at 5 for multi-tag retrieval. Testing is performed on the NUS-WIDE test set. Columns show performance for each method for the number of tags that need to be matched. See the caption to Table 2 for an explanation of the rows. Robust LR consistently outperforms LR, and calibration consistently improves results. These trends are clearer for longer (and therefore more difficult) queries.

5.3 Image Retrieval

Finally, we consider the tag-based image retrieval task: given a set of query tags, find images that match all the tags. We measure performance using normalized precision at 5; each system returns a set of 5 images, and its score for a given query is the number of those images that are characterized by all tags divided by the smaller of 5 and the total number of relevant images in the test set. We use the NUS-WIDE annotations as ground truth. We tested the same models from the previous section. We tested each method with queries consisting of every combination of one, two, and three tags that had at least one relevant image in the test set. All models

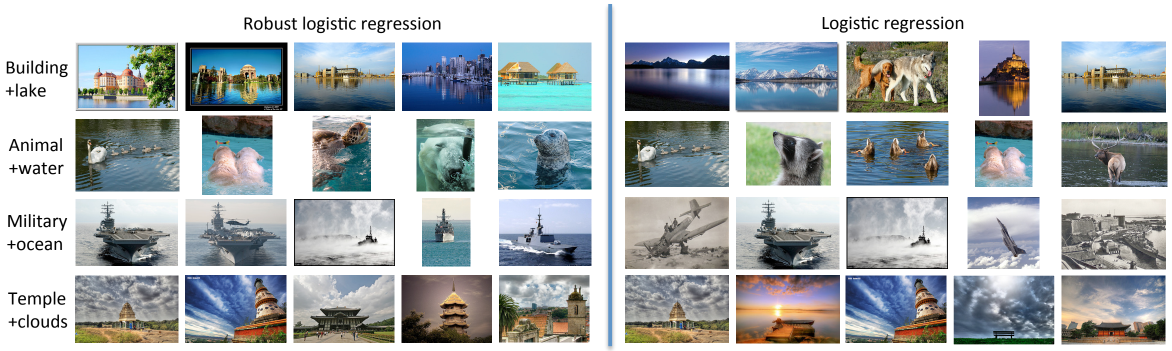

Figure 5: Multi-tag retrieval queries where Robust LR gives notably superior results to LR. Retrieval results are sorted from left-to-right.

perform well on single-tag queries (Table 5), but the differences in precision grow rapidly as the number of tags that the retrieved images must match increases. RLR consistently outperforms LR, and calibration significantly improves the models trained on Flickr. Figure 5 shows some queries for which RLR outperforms LR.

The model trained on NUS-WIDE achieves the best score. However, there are many thousands of tags for which no annotations are available, and these results show that good results can be obtained on these tags as well.

6. DISCUSSION AND FUTURE WORK

Online wild tags represent a great, untapped natural resource. We show that, despite their noise, these tags can be useful, either alone or together with a small amount of calibration. Though we have tested the Flickr dataset, there are numerous other online datasets with different kinds of wild tags that can also be leveraged and explored for different applications. As noted in Section 3, there is a great deal of structure in these tags that could be exploited in future work. These tags could also provide mid-level features for other classification tasks and consumer applications, such as tag suggestion and organizing personal photo collections.

7. REFERENCES

[1] M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *Proc. CHI*, 2007.

[2] A. C. Berg, T. L. Berg, H. Daumé III, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and predicting importance in images. In *Proc. CVPR*, 2012.

[3] O. Cappé and E. Moulines. On-line expectation-maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

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

[5] P. Duygulu, K. Barnard, N. D. Freitas, and D. a. Forsyth. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In *Proc. ECCV*, 2002.

[6] Y. Gong, Y. Jia, T. K. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. In *Proc. ICLR*, 2014.

[7] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *Proc. ECCV*, 2014.

[8] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proc. ICCV*, 2009.

[9] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell. Efficient learning of domain-invariant image representations. In *Proc. ICLR*, 2013.

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding, 2014. http://arxiv.org/abs/1408.5093.

[11] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. Recognizing image style. In *Proc. BMVC*, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012.

[13] X. Li and C. G. M. Snoek. Classifying tag relevance with relevant positive and negative examples. In *Proc. MM*, 2013.

[14] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. M. Snoek, and A. del Bimbo. Socializing the semantic gap: A comparative study on image tag assignment, refinement and retreival, 2015. http://arxiv.org/abs/1503.08248.

[15] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *Proc. ECCV*, 2008.

[16] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11), 1995.

[17] S. Moran and V. Lavrenko. A sparse kernel relevance model for automatic image annotation. *Int J Multimed. Info Retr*, 3(4), 2014.

[18] N. Murray, D. Barcelona, L. Marchesotti, and F. Perronnin. AVA: A Large-Scale Database for Aesthetic Visual Analysis. In *CVPR*, 2012.

[19] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari. Learning with noisy labels. In *Proc. NIPS*, 2013.

[20] O. Nov, M. Naaman, and C. Ye. What drives content tagging: The case of photos on flickr. In *Proc. CHI*, 2008.

[21] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg. From large scale image categorization to entry-level categories. In *Proc. ICCV*, 2013.

[22] G. Patterson, C. Xu, H. Su, and J. Hays. The sun attribute database: Beyond categories for deeper scene understanding. *IJCV*, 108:59–81, 2014.

[23] V. C. Raykar, S. Y, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *JMLR*, 11:1297–1322, 2010.

[24] S. E. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *Proc. ICLR*, 2015.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2014. http://arxiv.org/abs/1409.0575.

[26] K. Saenko and T. Darrell. Unsupervised learning of visual sense models for polysemous words. In *Proc. NIPS*, 2008.

[27] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[28] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolution networks with noisy labels. In *Proc. ICLR*, 2015.

[29] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. CHI*, 2004.

[30] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proc. CVPR*, 2010.

[31] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. In *Proc. NIPS*, 2014.