

Informative Object Annotations: *Tell Me Something I Don't Know*

Lior Bracha

Bar-Ilan University

lior.bracha@live.biu.ac.il

Gal Chechik

Bar-Ilan University, NVIDIA Research

gal.chechik@biu.ac.il

Abstract

Capturing the interesting components of an image is a key aspect of image understanding. When a speaker annotates an image, selecting labels that are informative greatly depends on the prior knowledge of a prospective listener. Motivated by cognitive theories of categorization and communication, we present a new unsupervised approach to model this prior knowledge and quantify the informativeness of a description. Specifically, we compute how knowledge of a label reduces uncertainty over the space of labels and use this uncertainty reduction to rank candidate labels for describing an image. While the full estimation problem is intractable, we describe an efficient algorithm to approximate entropy reduction using a tree-structured graphical model. We evaluate our approach on the open-images dataset using a new evaluation set of 10K ground-truth ratings and find that it achieves $\sim 65\%$ agreement with human raters, close to the upper bound of inter-rater agreement and largely outperforming other unsupervised baselines.

1. Introduction

How would you label the photo in Figure 1? If you answered “a dog”, your response agrees with what most people would answer. Indeed, people are surprisingly consistent when asked to describe what an image is “about” [16]. They intuitively manage to focus on what is “informative” or “relevant” and select terms that reflect this information. In contrast, automated classifiers can produce a large number of labels that are perhaps technically correct, but are often non-interesting (Fig. 1 top right).

A natural approach to ascertain importance lies in the context of the specific task. For instance, classifiers can be efficiently trained to identify dog breeds or animal species. More generally, each task defines importance through a supervision signal provided to the classifier [1, 20, 14]. Here we are interested in a more generic setup, where no downstream task dictates the scene interpretation. This represents the challenge that people face when describing a scene to another person, without any specific task at hand.

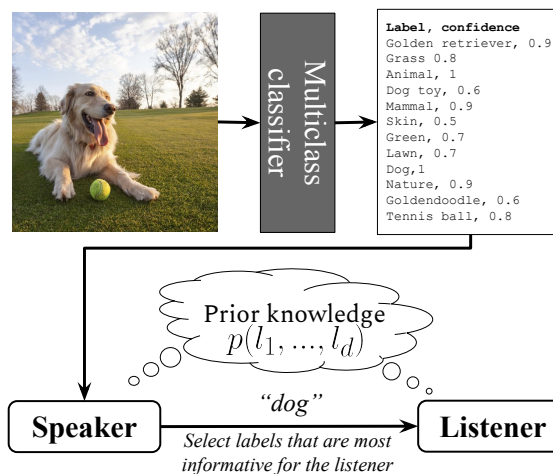


Figure 1. **The problem of informative labeling.** An image is automatically annotated with multiple labels. A “speaker” is then given these labels and their confidence scores and has to select k labels to transmit to a listener, such that the listener finds them informative given his prior knowledge. The prior knowledge is assumed to be common to both the speaker and the listener.

The principles that govern informative communication have long been a subject of research in various fields from philosophy of language and linguistics to computer science. In the discipline of pragmatics, Grice’s maxims state that “one tries to be as informative as one possibly can.” [9]. But the question remains, “Informative about what?” How can we build a practical theory of informative communication that can be applied to concrete problems with real-world data?

In this paper, we address the following concrete learning setup (Figure 1). A speaker receives a set of labels predicted automatically from an image by a multiclass classifier. It also receives the confidence that the classifier assigns to each prediction. Then, it aims to select a few labels (say, one label) to be transmitted to a listener, such that the listener will find those labels informative. The speaker and listener also share the same prior knowledge in the form of the distribution of labels in the image dataset.

We put forward a quantitative theory of how speakers select terms to describe an image. The key idea is that communicated terms are aimed to reduce the uncertainty that a listener has about the semantic space. We show how this “theory-of-mind” can be quantitatively computed using information-theoretic measures. In contrast with previous approaches that focused on visual aspects and their importance [8, 2, 18, 13, 3], our measures focus on information about the semantics of labels.

To compute information content of a label, we build a probabilistic model of the full label space and use it to quantify how transmitting a label reduces uncertainty. Specifically, we compute the entropy of the label distribution as a measure of uncertainty, and also quantify how much this entropy is reduced when a label is set to be true.

Importantly, computing these measures over the full distribution of labels is not feasible because it requires to aggregate an exponentially-large set of label combinations. We show how the entropy and other information theoretic measures can be computed efficiently by approximating the full joint distribution with a tree-structured graphical model (a Chow-Liu tree). We then treat entropy-reduction as a scoring function that allows us to rank all labels of an image, and select those that reduce the entropy most. We name this approach IOTA, for *Informative Object Annotations*.

We test this approach on a new evaluation dataset: 10K images from the open-images dataset [11] were annotated with informative labels by three raters each. We find that human annotations are in strong agreement ($\sim 65\%$) with the uncertainty-reduction measures, just shy of inter-rater agreement and superior to 4 other unsupervised baselines.

Our main contributions are: (1) We describe a novel learning setup of selecting important labels without direct supervision about importance. (2) We develop an information-theoretic framework to address this task, and design scoring functions that can be used to solve it. (3) We further describe an efficient algorithm for computing these scoring functions, by approximating the label distribution using a tree-structured graphical model. (4) We provide a new evaluation set of ground-truth importance ratings based on 10K images from the open-images dataset. (5) We show that IOTA achieves high agreement with human judgment on this dataset.

Learning a measure of importance over the space of visual labels could have wide implications as it allows us to automate what until now required costly human evaluation. By focusing on labels that matter to people, one could, for example, design more relevant loss functions and evaluation metrics for object recognition or steer image captioning toward meaningful descriptions.

2. Related work

Image importance and object saliency. The problem of deciding which components in an image are important has been studied intensively. The main approaches involved identifying characteristics of objects and images that could contribute to importance, and use labeled data for predicting object importance. Elazary and Itti [8] considered the order of object naming in the LabelMe dataset [17] as a measure of the interest of an object and compare that to salient locations predicted by computational models of bottom-up attention. The elegant work of Spain and Perona [18] examined which factors can predict the order in which objects will be mentioned given an image. Berg et al. [2] characterized factors related to semantics, to composition and to the likelihood of attribute-object, and investigated how these affected the measures of importance. [15] focused on predicting entry-level classes using a supervised approach. These studies also make it clear that the object saliency is strongly correlated with its perceived importance [13, 3].

These studies differ from the current work in two significant ways. First, they largely focus on visual properties of objects in images, while our current approach focuses on modeling the labels structure, and only uses image-based information in the form of label confidence as predicted by a classifier. Second, they largely take a supervised approach using measures of importance in a training set to build predictive models of label importance. In contrast, our approach is unsupervised, because our model is not directly exposed to labeled information about object importance.

Information theory and measures of relevance The problem of extracting informative components from a complex signal was studied from an information-theoretic perspective through the information bottleneck principle (IB) [19, 4, 23]. In contrast to the current work, in IB, a signal, X , is compressed into T such that it maximizes information about another variable Y , that can be viewed as a supervisory variable. In [12], information gain was used to select questions in a goal-oriented dialog setup.

Pragmatics, Relevance theory. In pragmatics, effective communication has been characterized by the *cooperative principle* [9], which views communication as a cooperative interaction between a speaker and a listener. These principles were phrased in Grice’s maxims, stating that “one tries to be as informative as one possibly can” and “does not give information that is false or that is not supported by evidence”. Our approach provides a concrete quantitative realization to these principles. Inspired by Grice’s work, Sperber and Wilson proposed a framework called *relevance theory* [21, 22]. They highlighted that a speaker provides cues to a listener, who then interprets them in the context of what she already knows and what the speaker may intended to transmit.

3. Our approach

The key idea of our approach is to quantify the relevant-information content of a message, by modelling what the listener does not know, and find labels that reduce this uncertainty. To illustrate the idea, consider a label that appears in most of the images in a dataset (e.g., *nature*). If the speaker selects to transmit that label, it provides very little information to the listener, because they can already assume that a given image is annotated with that label. In contrast, if the speaker transmits a label that is less common, appearing in only half of the images, more of the listener’s uncertainty would be removed.

A more important property of multi-label uncertainty is that labels are *interdependent*: transmitting one label can reduce the uncertainty of others. This property is evident when considering label hierarchy, for example, *golden-retriever* = true implies that *dog* = true. As a result transmitting a fine-grained label removes more entropy than a more general label. Very importantly however, this effect is not limited to hierarchical relations. For instance, because the label *street* tends to co-occur with *car* and other vehicles, transmitting *street* would reduce the overall uncertainty by reducing uncertainty in correlated co-occurring terms.

Going beyond these examples, we aim to calculate how a revealed label affects the listener’s uncertainty. For this purpose, the Shannon entropy is a natural choice to quantify uncertainty, pending that we can estimate the prior joint distribution of labels. Clearly, modelling the entire prior knowledge about the visual world of a listener is beyond our current reach. Instead, we show how we can approximate the entire joint distribution by building a compact graphical model with a tree structure. This allows us to efficiently compute properties of the joint distribution over labels and more specifically, estimate listener uncertainty and label-conditioned uncertainty.

We start by describing an information-theoretic approach for selecting informative labels by estimating uncertainty and label-conditioned uncertainty. We then describe an algorithm to effectively compute these quantities in practice.

3.1. The problem setup

Assume that we are given a corpus of images, each annotated with multiple labels from a vocabulary of d terms $\mathcal{L} = (l_1, \dots, l_d)$. Since we operate in a noisy labeling setup, we treat the labels as binary random variables $l_i \in \{true, false\}$. We also assume that for each image I , labels are accompanied with a score reflecting the classifier’s confidence in that label, which we denote by $q(l_i|I)$. Such confidence scores can be obtained from classifier predictions, assuming that these confidence scores are calibrated, namely, reflect the true fraction of correct labels. In practice,

many large-scale models indeed calibrate their scores, as we discuss in the experimental section. The goal of the speaker is to select k labels to be transmitted to the listener, such that they are most “useful” or informative.

3.2. Information-theoretic measure of importance

Let us first assume that we can estimate the distribution over labels that a listener has in mind. Clearly, this is a major assumption, and we discuss below how we relax this assumption and approximate this distribution. Given this distribution, we wish to measure the uncertainty it reflects, as well as how much this uncertainty is reduced when the speaker reveals a specific label. A principled measure of the uncertainty about random variables is the Shannon entropy of their joint distribution $H(L_1, \dots, L_d)$ [6]. We use a notation that makes it explicit that the entropy depends on the distribution, where the entropy is defined as

$$H[p(l_1, \dots, l_d)] = - \sum_{l_1, \dots, l_d} p(l_1, \dots, l_d) \log p(l_1, \dots, l_d). \quad (1)$$

Here, summation is over all possible assignments of the d labels, an exponential number of terms that cannot be computed in practice. We show below how to approximate it.

The amount of entropy that is reduced when the speaker transmits a subset of the labels $\mathcal{L}' = \{l_i, l_j, l_k, \dots\}$, is

$$\Delta H(\mathcal{L}') = H[p(l_1, \dots, l_d)] - H[p(l_1, \dots, l_d | \mathcal{L}' = true)] ,$$

where $\mathcal{L}' = true$ means that all labels in \mathcal{L}' are assigned a true value. For simplicity, we focus here on the case of transmitting a single label l_i (see also [7]), and define the *per-label entropy-reduction*

$$\Delta H(i) = H[p(l_1, \dots, l_d)] - H[p(l_1, \dots, l_d | l_i = true)]. \quad (2)$$

This measure has several interesting properties. It has a similar form to the Shannon mutual information, $MI(X; Y) = H(X) - H(X|Y)$, which is always positive. However, the condition on the second term is only over a single value of the label ($l_i = true$). As a result, Eq. (2) can obtain both negative and positive values. When the random variables are independent, $\Delta H(i)$ is always positive, because the entropy can be factored using the chain rule, and obeys $H(L_1, \dots, L_d) - H(L_1, \dots, L_d | L_i) = \sum_{j \neq i} H(L_j) > 0$ (Sec 2.5 [6]). However, when the variables are not independent, collapsing one variable to a True value can actually increase the entropy of other co-dependant variables. As an intuitive example, the base probability of observing a lion in a city is very low, and has low entropy. However, once you see a sign “zoo”, the entropy of facing a lion rises.

The second important property of $\Delta H(i)$ is that it is completely agnostic to the image and only depends on the label distribution. To capture image-specific label relevance, we note that the accuracy of annotating an image

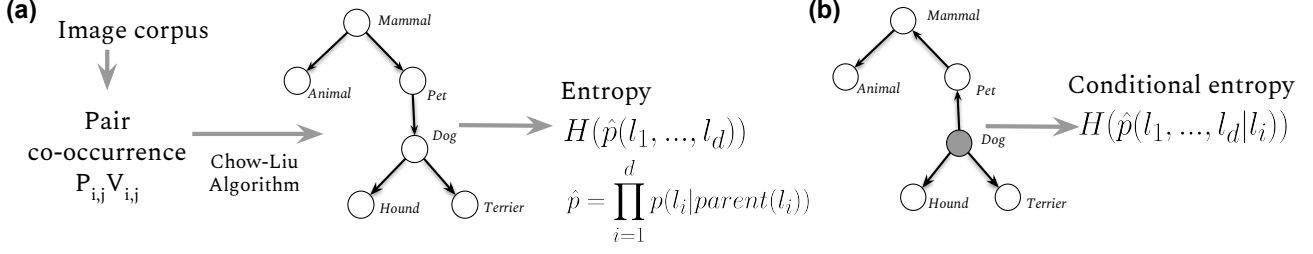


Figure 2. **Uncertainty over labels can be estimated through measuring the entropy of its joint distribution, and computed efficiently using a tree-structured probabilistic graphical model (PGM).** (a) An image corpus is used for collecting pairwise label co-occurrence. Then, a tree-structured graphical model is learned using the Chow-Liu algorithm. Computing the entropy of the approximated distribution \hat{p} has a run-time that is linear in the number of labels. (b) To compute the entropy conditioned on a label $l_{dog} = \text{true}$, the marginal of that node is set to $[0,1]$. Then, the graph edges are redirected and rest of the distribution is updated using the conditional probability tables represented on the edges. Finally, we compute the entropy of the resulting distribution.

with a label may strongly depend on the image. For example, some images may have key aspects of the object occluded. We therefore wish to compute the expected reduction in entropy based on the likelihood that a label is correct $q(l_i|I)$. When an incorrect label is transmitted, we assume here that no information is passed to the listener (there is an interesting research question about negative information value in this case, which is outside the scope of this paper). The **expected entropy-reduction** is therefore

$$E(\Delta H) = q(l_i|I)\Delta H + (1 - q(l_i|I)) \cdot 0$$

this expectation is equivalent to the **confidence-weighted entropy reduction** measure:

$$cw-\Delta H(i) = q(l_i|I) [H(L) - H[L|l_i = \text{true}]] \quad , \quad (3)$$

where $q(l_i|I)$ is the probability that l_i is correct and L is a random variable that holds the distribution of all labels. We propose that this is a good measure of label information in the context of a corpus.

3.3. Other measures of informative labels

Confidence-weighted entropy reduction, Eq. (3), is an intuitive quantification of label informativeness, but other properties of the label distribution may capture aspects of label importance. We now discuss two such measures: information about images, and probabilistic surprise.

Information about images. Informative labels were studied in the context of an *image reference game*. In this setup, a speaker provides labels about an image, and a listener needs to identify the target image among a set of distractor images. Recent versions used natural language captioning for the same purpose [1, 20].

It is natural to define entropy-reduction for that setup. Similar to Eq. (2), compute the difference between the full entropy over images, and the entropy after transmitting a label. When the distribution over images is uniform the entropy reduction is simply $\log(\text{num. images}) -$

$\log(\text{num. matching images})$, where the second term is the number of images annotated by a label. Considering the confidence of a label we obtain

$$cw\text{-Image}\Delta H(i) = q(l_i|I) [\log(q(l_i))] \quad , \quad (4)$$

where $q(l_i|I)$ is again the probability that l_i is correct and $q(l_i)$ is the fraction of images with the label i . This measure is fundamentally different from Eq. (3) in that it focuses on the distribution of labels over *images*, not their on joint distribution.

Probabilistic surprise. Transmitting a label changes the label distribution, the “belief” of the listener. This change can be quantified by the Kullback-Liebler divergence of the label distribution with and without transmission:

$$cw\text{-}D_{KL}(i) = q(l_i|I) D_{KL}(p(l_1, \dots, l_d | l_i = \text{true}) || p(l_1, \dots, l_d)) \quad . \quad (5)$$

We can use this measure as a scoring function to rank labels by how strongly they affect the distribution. As in the entropy reduction approach (Eq. 3), here we exploit the cross-label relationships but provide a different information-theoretic measure for how transmitting a label affects the distribution.

Entropy reduction in a singleton model. An interesting approximation to the joint distribution used in Eq. (1) is provided by the *singleton* model, which models the joint distribution as the product of the marginals $p(l_1, \dots, l_d) = \prod_i p(l_i)$. Here, the joint entropy is simply the sum of per-label entropies. The entropy reduced when transmitting a label is simply its entropy $cw\text{-Singleton}(i) = q(l_i|I)H(l_i)$. Importantly, entropy reduction in this model ignores inter-label relationships.

The entropy is a function that grows monotonically with p for $(p < 0.5)$. This means that if all labels are rare ($p < 0.5$), then ranking labels by their empirical frequency, yields the same ranking as by their singleton entropy reduction.

3.4. Entropy reduction in large label spaces

Given a corpus of images, we wish to compute the joint distribution of label co-occurrence in an image $p(l_1, \dots, l_d)$. The scoring functions described above assume that we can estimate and represent the joint distribution over labels. Unfortunately, even for a modest vocabulary size d , the distribution cannot be estimated in practice since it has 2^d parameters. Instead, we approximate the label distribution using a probabilistic graphical model called a **Chow-Liu tree** [5]. We first describe the graphical model, and then how it is learned from data.

As any probabilistic graphical model, A Chow-Liu tree has two components: First, a tree $G(V, E)$ with d nodes and $d - 1$ edges, where the nodes V correspond to the d labels, and the edges E connect the nodes to form a fully-connected tree. The tree is *directed*, and each node l_i , except a single root node, has a single parent node l_j .

As a second component, every edge in the graph, connecting nodes i and j is accompanied by a conditional distribution, $p(l_i | \text{parent}(l_i))$. Note that this conditional distribution involves only two binary variables, namely a total of four parameters. The full model therefore has only $O(d)$ parameters and can be estimated efficiently from data. With these two components, the Chow-Liu model can be used to represent a joint distribution over all labels, which factorizes over the graph

$$\log p(l_1, \dots, l_d) = \sum_{i=1}^d \log p(l_i | l_{\text{parent}(i)}). \quad (6)$$

While any tree structure can be used to represent a factored distribution as in Eq. (6), the Chow-Liu algorithm finds one specific tree structure: The distribution that is closest to the original full distribution terms of the Kullback-Liebler divergence $D_{KL}(p(\hat{L}) || p(L))$. That tree is found in two steps: First, for every pair of labels i, j , compute their 2×2 joint distribution in the image corpus, then compute the mutual information of that distribution.

$$MI_{ij} = \sum_{l_i=T,F} \sum_{l_j=T,F} p_{ij}(l_i, l_j) \frac{p_{ij}(l_i, l_j)}{p_i(l_i)p_j(l_j)} \quad (7)$$

where the summation is over all combination of True and False value for the two variables, p_{ij} is the joint distribution over label co-occurrence, and p_i and p_j are the marginals of that distribution.

As a second step, assign MI_{ij} as the weight of the edge connecting the nodes of labels i and j and find the maximum spanning tree on the weighted graph. Importantly, the particular directions of the edges of the model are not important. Any set of directions that forms a consistent tree (having at most one parent per node), defines the same distribution over the graph [5]. In practice, since committing

to a single tree may be sensitive to small perturbations in the data, we model the distribution as a mixture of k trees, which are created by a bootstrap procedure.

Representing the joint distribution of labels using a tree provides great computational benefits, since many properties of the distribution can be computed very efficiently. Importantly, when the joint distribution factorizes over a tree, the entropy can be computed exactly using the entropy chain rule:

$$H[p(l_1, \dots, l_d)] = H\left[\prod_{i=1}^d p(l_i | A(l_i))\right] = \sum_{i=1}^d H[p(l_i | A(l_i))], \quad (8)$$

where $A(l_i)$ is the parent of the label l_i . We abused the notation slightly, the root node does not have a parent hence its entropy is not conditioned on a parent but should be $H[p(l_{\text{root}})]$.

Furthermore, in a tree-structured probabilistic model, one can redirect the edges by selecting any node to be a root, and conditioning all other nodes accordingly [10]. This allows us to compute the labeled-conditioned entropy using the following steps. First, given a new root label l_i , iteratively redirect all edges in the tree to make all nodes its descendents. Update the conditional density tables on the edges. Second, assign a marginal distribution of $[0, 1]$ to the node l_i , reflecting the fact that the label is assigned to be true. Third, propagate the distribution throughout the graph using the conditional probability functions on the edges. Finally, compute the entropy of the new distribution using the chain rule as in Eq. (8).

3.5. Selecting labels for transmission

Given the above model, we can compute the expected entropy reduction for each label for a given image. We then take an information-retrieval perspective, rank the labels by their scores and emit the highest rank label.

This process can be repeated for transmitting multiple labels. For example, given that label l_i was transmitted first, we compute how much each of the remaining labels reduces the entropy further. Formally, to decide about a second label to transmit, we compute for every label $l_j \neq l_i$:

$$\Delta H_i(j) = H[p(l_1, \dots, l_d | l_i = \text{true})] - H[p(l_1, \dots, l_d | l_i = \text{true}, l_j = \text{true})] \quad (9)$$

Intuitively, selecting a second label that maximizes this score tends to select labels that are semantically remote from the first emitted labels. If a second label (say, $l_j = \text{pet}$) is semantically similar to the first label (say, $l_i = \text{dog}$), the residual entropy of pet after observing the label dog is low, hence the speaker will prefer other labels.


|  | confidence | $cw-\Delta H$ | $cw-D_{KL}$ | $cw-Image\Delta H$ | $cw-p(l)$ | $cw-Singleton$ |
|---|-------------------|---------------|--------------|--------------------|--------------|----------------|
| | vehicle, airplane | airplane | airliner | airliner | vehicle | vehicle |
| airplane | 1.0 | 52.18 | 56.65 | 5.71 | 0.019 | 0.14 |
| airline | 0.9 | 47.53 | 57.4 | 5.94 | 0.009 | 0.07 |
| airliner | 0.9 | 46.69 | 58.36 | 6.29 | 0.007 | 0.06 |
| aircraft | 0.9 | 46.54 | 46.67 | 4.83 | 0.022 | 0.15 |
| vehicle | 1.0 | 41.02 | 14.34 | 2.33 | 0.199 | 0.72 |
| propeller-aircraft | 0.8 | 41.01 | 49.97 | 5.85 | 0.005 | 0.04 |
| aviation | 0.8 | 40.97 | 40.01 | 4.30 | 0.019 | 0.13 |
| narrow-body aircraft | 0.8 | 40.73 | 55.06 | 6.17 | 0.004 | 0.03 |
| air force | 0.6 | 29.61 | 29.34 | 3.71 | 0.008 | 0.06 |
| aircraft engine | 0.6 | 28.14 | 23.51 | 3.82 | 0.007 | 0.06 |

Table 1. **Ranking image annotations by the compared approaches.** Labels are ranked based on the score functions. Then, the position (namely, k) of the ground-truth label (in bold) is used to compute precision and recall. Later, precision and recall are averaged across images.

4. Experiments

4.1. Data

We tested IOTA on the open-images dataset (OID) [11]. In OID, each image is annotated with a list of labels, together with a confidence score. We approximate the joint label distribution over the validation set (41,620 images annotated with 512,093 labels) and also over the test set (125,436 images annotated with 1,545,835 labels).

Ground-truth data (OID-IOTA-10K). We collected a new dataset of ground-truth “informative” labels for 10K images: 2500 from OID-validation and 7500 from OID-test, 3 raters per image. Raters were instructed to focus on the object or scene that is dominant in the image and to avoid overly generic terms that are not particularly descriptive (“a picture”). Labels were entered as free text, and when possible, matched in real time to the predefined OID knowledge graph (64% of samples) so raters can verify label meaning. The remaining 36% of annotations were matched as a post-process, which included stemming, resolving ambiguities (*e.g.* deciding if a *bat* meant the animal or the sport equipment) and resolving synonyms (*e.g.* pants and trousers). Overall, in many cases raters used exactly the same term to describe an image. In 68% of the images *at least* two raters described the image with the same label, and in 27% all three raters agreed. The data is publicly available at <https://chechiklab.biu.ac.il/brachalior/IOTA/>.

Label co-occurrence. OID lists labels whose confidence is above 0.5. All labels with at least 300 appearances were considered when collecting the label distribution, ignoring their confidence. This yielded a vocabulary of 772 labels. See supp. material for additional experiments.

4.2. Evaluation Protocol

For each importance scoring functions derived above (Sec 3.2) we ranked all labels predicted to each image. Given this label ranking we compared top labels with the ground-truth labels collected from raters, and computed the precision and recall for the top- k ranked labels. Precision

and recall are usually used with more than one ground-truth item. In our case however, for each image, there was only one ground-truth label: the majority vote across the three raters. As a result, the precision@1 is identical to recall@1. We excluded images that had no majority vote (3 unique ratings, 27.6% of images). OID provides confidence values in coarse resolution (1 significant digit), hence multiple labels in an image often share the same confidence values. When ranking by confidence only, we broke ties at random.

We also tested a evaluation setup where instead of a majority label, every label provided by the three raters was considered as ground truth. Precision and recall was computed in the same way.

4.2.1 Clean and noisy evaluation

We evaluated our approach in two setups. In the first, *clean evaluation*, we only considered image labels that were verified to be correct by OID raters. Incorrect labels were excluded from the analysis and not ranked by the scoring functions. We also excluded images whose ground truth label was not in the model’s vocabulary.

In the second setup, *noisy evaluation* we did not force any of these requirements. The analysis included incorrect labels as well as images whose ground truth labels were not in the vocabulary; and thus could not be predicted by our model. As expected, the precision and recall in this setting were significantly lower.

4.3. Compared scoring functions and baselines

We compared the following information-theoretic scoring functions, all weighted by classifier confidence. All CLT-based methods were computed over a mixture of 10 trees, see supplementary material for more details.

(1) **Entropy-reduction** $cw-\Delta H$: See Eq. (3). (2) **Probabilistic surprise** $cw-D_{KL}$: See Eq. (5). (3) **Image entropy reduction** $cw-Image\Delta H$: See Eq. (4). (4) $cw-Singleton$, $q(l_i|I)H(l_i)$: See section (3.3).

We also evaluated three simpler baselines: (5) **Random** A random ranking of labels within each image. (6) **Confidence**, $q(l_i|I)$, which reflects the likelihood that a label is

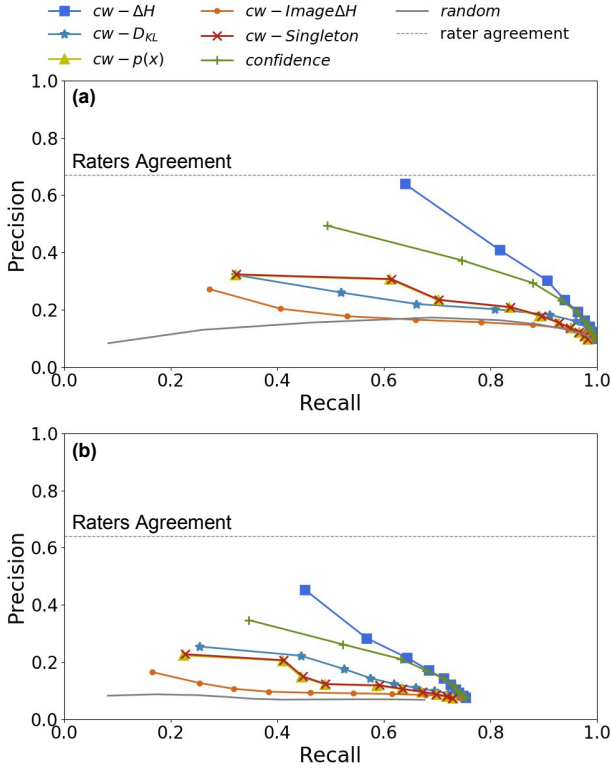


Figure 3. **Precision and recall @k in the clean setup (top) and the noisy setup (bottom)**, computed over the OID-test set. (a) In the *clean* setup, results are an averaged over 2877 images. $cw - \Delta H$ (blue curve) achieves p@1 of 64% and largely outperforms other scoring functions. Rater agreement (dashed line) is at 66%, only slightly higher than $cw - \Delta H$. (b) In the *noisy* setup, results are an average over 3942 images. As with the clean set, $cw - \Delta H$ outperforms other scoring functions, but only achieves p@1 of 45% were the inter-rater agreement is 64%.

correct for an image and is provided by the classifier. Labels with highest confidence were ranked first; ties broken randomly. (7) **Term frequency**, The empirical $p(l_i)$ captures how often a label is observed in corpus, ranked in a descending order. Note that in our data, the term frequency produces the same ranking as singletons, because all labels have a marginal frequency below 0.5, hence their entropy monotonically increases with $p(l_i)$.

5. Results

We first illustrate label ranking by showing the scores for one image. Table 1 the annotations are ordered by $cw - \Delta H$, and the best label per column (scoring function) is highlighted. Singleton and term frequency, $p(l)$, yield the same ranking (but with different values) because the entropy grows monotonically with p . $cw - D_{KL}$ prefers fine-grained classes.

We next present the precision and recall of IOTA and compared methods over the full OID-test in the *clean* setup

| | Single label | | Multiple labels | |
|--------------------------|--------------|-------------|-----------------|-------------|
| | P@1 R@1 | R@5 | P@1 | R@1 |
| Scoring functions | | | | |
| $cw - \Delta H$ | 0.64 | 0.96 | 0.63 | 0.57 |
| $cw - D_{KL}$ | 0.43 | 0.96 | 0.42 | 0.38 |
| $cw - Image\Delta H$ | 0.28 | 0.78 | 0.33 | 0.30 |
| $cw - Singleton$ | 0.33 | 0.89 | 0.34 | 0.31 |
| $cw - p(l)$ | 0.33 | 0.89 | 0.34 | 0.31 |
| Baselines | | | | |
| <i>confidence</i> | 0.49 | 0.96 | 0.50 | 0.46 |
| <i>random</i> | 0.12 | 0.89 | 0.21 | 0.18 |
| Non-weighted | | | | |
| ΔH | 0.29 | 0.86 | 0.34 | 0.31 |
| D_{KL} | 0.22 | 0.87 | 0.29 | 0.26 |
| $Image\Delta H$ | 0.14 | 0.64 | 0.21 | 0.18 |
| $Singleton$ | 0.26 | 0.88 | 0.29 | 0.26 |
| $p(l)$ | 0.26 | 0.88 | 0.29 | 0.26 |

Table 2. **Precision and recall of compared approaches**. Scores are averaged over 10 trees. $cw - \Delta H$ reach an accuracy of 64% for predicting a single label and 63% in a multi-label setup.

(Sec. 4.2.1). Figure 3.a. reveals that IOTA achieves high precision, including a p@1 of 64%. This precision is only slightly lower than the agreement rate of human raters (66%). See details in Table 2 for comparison.

Next, we show similar curves for the *noisy* setup. Here we also considered images where the ground-truth label is not included in the vocabulary, treating model predictions for these images as false. Figure 3.b. shows that in this case too, $cw - \Delta H$ achieves the highest precision and recall compare with the other approaches. As expected, the precision and recall in this setting are lower, reaching p@1=45%.

We further tested all scoring functions using a multi-label evaluation protocol. Here, instead of taking the majority label over three rater annotations, we used all three labels (non-weighted) and computed the precision and recall of the scoring functions against that ground truth set. Results are given in Table 2, showing a similar behavior where $cw - \Delta H$ outperforms the other scoring functions.

Ablation and comparisons. Several comparisons are worth mentioning. First, confidence-weighted approaches (image-dependent) are consistently superior to non-weighted approaches. This suggests that it is not enough to select “interesting” labels if they are not highly confident for the image. Second, the singleton model performs poorly compared to the full CLT, $cw - \Delta H$. This agrees with our observation that a key factor of label importance is how much it affects uncertainty on other labels. Finally, $cw - Image\Delta H$, is substantially worse, which is again consistent with the observation that structure in label space is critical.

We also repeated the analysis while limiting the CL tree to the labels present in each image. This significantly hurt




| | | | |
|--------------------|---|---|---|
| |  |  |  |
| Confidence $q(l)$ | Shoe , footwear, purple | Leaf , plant, tree, nature, yellow, green | land vehicle |
| $cw-\Delta H$ | Shoe | Leaf | Car |
| $cw-D_{KL}$ | Shoe | Autumn | Mercedes-benz |
| $cw-Image\Delta H$ | Violet | Season | Mercedes-benz |
| $cw-p(l)$ | Purple | Plant | Vehicle |
| $cw-Singleton$ | Purple | Plant | Vehicle |

Table 3. **Qualitative example of top-ranked labels by the various scoring functions.** While all annotation are correct, *shoe* (left), *leaf* (middle) and *car* (right) are consistent with human annotations. In the car example, *cw-p(l)* and singleton select an overly abstract label, while *cw-D_{KL}* and *cw-Image Δ H* select more fine grained labels. This effect was pervasive in our dataset.

the precision ($p@1 = 0.48$), suggesting that information about out-of-image labels is important. More broadly, this work takes a view that separates the prediction problem into two factors: One that models the listener prior knowledge, hence is image independent, and a second that is image dependent. In the OID dataset, the blind-listener approximation proves very effective.

Originally, we expected that labels relations can be modelled well using a known semantic hierarchy. We tested a hierarchy provided with OID (600 labels), but we found it far less effective than CLT ($p@1=0.34$ for $cw-\Delta H$), presumably because semantic relatedness differs substantially from visual co-occurrence. E.g., “*dog*” and “*park*” are remote in a semantic hierarchy, but interdependent and hence closer in a co-occurrence-based tree. Thus, using an existing ontology of the labels does not necessarily model the visual co-occurrences that can be learned from data.

Qualitative results. Table 3 lists top-ranked labels by various scoring functions for three images. $cw-\Delta H$ consistently agrees with human annotations (in bold), capturing an intermediate, more informative category compared with other scoring functions. Ranking based on confidence only for the left column described the image as either *shoe*, *footwear* or *purple*. While all three are technically correct, *shoe* is the most natural, informative title for that image. For the middle column (leaf) there were 20 highly-confident predicted annotations (only 6 shown); all approaches other than $cw-\Delta H$ failed to return "leaf". Finally, the *car* example (bottom) demonstrates a common phenomena where $cw-p(l)$ and $cw-Singleton$ prefer to more abstract categories whereas $cw-D_{KL}$ and $cw-Image\Delta H$ prefer fine-grained labels.

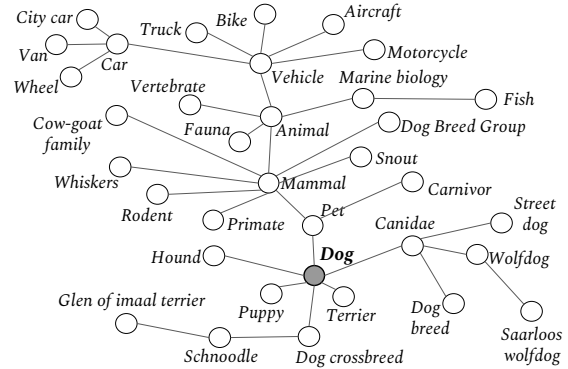


Figure 4. **Part of Chow-Liu tree around the label “dog”** learned from the OID validation set with 765 labels. The model clearly captures semantic relations, even-though they are not explicitly enforced. For instance the label “*per*” is connected directly to “*dog*”, and “*truck*” and “*bike*” connected to “*vehicle*”.

These results are all built on a Chow-Liu graphical model. To test if its label-dependency structure reflects sensible label semantics, Figure 4 illustrates parts of the tree that was formed around the label *dog* (38 of 765 labels). Semantic concepts are grouped in a way that agrees with their meaning (mostly). Note that this tree structure, not a hierarchical model, but only captures the pairwise dependencies among label co-occurrence in the open-images dataset.

Robustness to hyper parameters. We tested the robustness of IOTA to the two hyper parameters of the model. (1) The number of trees in the mixture model; and (2) The size of the vocabulary analyzed. The model was largely robust to these parameters. Detailed results are given in the suppl.

6. Conclusion

We present an unsupervised approach to select informative annotation for a visual scene. We model the prior knowledge about visual experience using the joint distribution of labels, and use it to rank labels per-image by how much entropy they can remove over the label distribution. The top ranked labels capture labels that are “intuitive”, showing high agreement with human raters. This is surprising, since the model does not use any external source of semantic information besides label concurrence.

Several questions remain open. First, while our current experiments captures common context, the approach can be extended to any context. It would be interesting to apply this method to expert annotators with the aim of retrieving listener-specific context. Second, easy-to-learn quantifiers of label importance can be used to improve loss functions in multi-class training, assigning more weight to more important labels.

References

- [1] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *Empirical Methods in Natural Language Processing*, 2016. 1, 4
- [2] Alexander C Berg, Tamara L Berg, Hal Daume, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. Understanding and predicting importance in images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3562–3569. IEEE, 2012. 2
- [3] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722, 2015. 2
- [4] Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. *Journal of machine learning research*, 6(Jan):165–188, 2005. 2
- [5] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, May 1968. 5
- [6] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012. 3
- [7] Michael R DeWeese and Markus Meister. How to measure the information gained from one symbol. *Network: Computation in Neural Systems*, 10(4):325–340, 1999. 3
- [8] Lior Elazary and Laurent Itti. Interesting objects are visually salient. *Journal of vision*, 8(3):3–3, 2008. 2
- [9] H Paul Grice. Logic and conversation. 1975, pages 41–58, 1975. 1, 2
- [10] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 5
- [11] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyaneesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from <https://github.com/openimages>*, 2017. 2, 6
- [12] Sang-Woo Lee, Yu-Jung Heo, and Byoung-Tak Zhang. Answerer in questioners mind: Information theoretic approach to goal-oriented visual dialog. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2584–2594. Curran Associates, Inc., 2018. 2
- [13] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2011. 2
- [14] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6964–6974, 2018. 1
- [15] Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. From large scale image categorization to entry-level categories. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013. 2
- [16] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439, 1976. 1
- [17] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008. 2
- [18] Merrielle Spain and Pietro Perona. Measuring and predicting object importance. *International Journal of Computer Vision*, 91(1):59–76, 2011. 2
- [19] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*, page 368377, 1999. 2
- [20] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2017. 1, 4
- [21] Deirdre Wilson and Dan Sperber. Relevance theory. In *Handbook of pragmatics*. Blackwell, 2002. 2
- [22] Deirdre Wilson and Dan Sperber. *Meaning and relevance*. Cambridge University Press, 2012. 2
- [23] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient human-like semantic representations via the information bottleneck principle. *arXiv preprint arXiv:1808.03353*, 2018. 2