Learning to Name Objects

By Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg

Abstract

We have seen remarkable recent progress in computational visual recognition, producing systems that can classify objects into thousands of different categories with increasing accuracy. However, one question that has received relatively less attention is "what labels should recognition systems output?" This paper looks at the problem of predicting category labels that mimic how human observers would name objects. This goal is related to the concept of entry-level categories first introduced by psychologists in the 1970s and 1980s. We extend these seminal ideas to study human naming at large scale and to learn computational models for predicting entry-level categories. Practical applications of this work include improving human-focused computer vision applications such as automatically generating a natural language description for an image or text-based image search.

1. INTRODUCTION

Computational visual recognition is beginning to work. Although far from solved, algorithms for analyzing images have now advanced to the point where they can recognize or localize thousands of object categories with reasonable accuracy.[3, 14, 24, 25] While we could predict any one of many relevant labels for an object, the question of "What *should* I actually call it?" is becoming important for large-scale visual recognition. For instance, if a classifier were lucky enough to get the example in Figure 1 correct, it might output *Cygnus Colombianus*, while most people would probably simply say *swan*. Our goal is to learn models to map from specific, encyclopedic terms (*Cygnus Colombianus*) to how people might refer to a given object (*swan*).

These learned mappings could add a new type of structure to hand-built linguistic resources, such as WordNet.[9] WordNet enumerates a large set of English nouns augmented by relationships, including hyperonymy (*is-a* connections) linking more general categories, for example, passerine, to more specific categories, for example, firebird (a firebird is a kind of passerine). Our models might learn that an image of a firebird is more likely to be described by the term "bird" instead of a more technical term like "passerine." When combined with a computer vision system that attempts to recognize many very specific types of objects in a particular image, our models allow mapping to the words people are likely to use for describing the depicted objects. For end-user applications, these types of outputs may be more useful than the outputs of very accurate but overly specific visual categorization systems. This is especially relevant for human computer interaction mediated by text—for instance, in text-based image search.

Our work is inspired by previous research on *basic and entry-level categories* formulated by psychologists, including Rosch[23] and Kosslyn.[13] Rosch defines *basic-level categories*

as those categories at the highest level of generality whose members still share many common attributes and have fewer distinctive attributes. An example of a basic level category is *bird* where most instances share attributes like having feathers, wings, and beaks. Subordinate, more specific categories, such as *American Robin* will have members that share even more attributes such as shape, color, and size. Super-ordinate, more general categories, such as *animal* have members that share fewer attributes and demonstrate more variability. Rosch studied basic level categories through human experiments, for example, asking people to enumerate common attributes for a given category.

The work of Jolicoeur et al.[13] further studied the way people identify categories, defining the notion of *entry-level categories*. Entry-level categories are essentially the categories that people will naturally use to identify objects. The more prototypical an object the more likely it will have its entry point at the basic-level category. For less typical objects the entry point might be at a lower level of abstraction. For example an *American robin* or a *penguin* are both members of the same basic-level *bird* category. However, the *American robin* is more prototypical, sharing many features with other birds and thus its entry-level category coincides with its basic-level category of *bird*, while a *penguin* would be identified at a lower level of abstraction (see Figure 2).

Thus, while objects are members of many categories—for

Figure 1. Example translation from a WordNet based object category prediction to what people might call the depicted object.

Recognition Prediction   *What Should I Call It?*

Cygnus Colombianus   →   Swan

Figure 2. An *American Robin* is a more prototypical type of bird hence its *entry-level category* coincides with its *basic-level category* while for penguin which is a less prototypical example of bird, the *entry-level category* is at a lower level of abstraction.

Superordinates: animal, vertebrate
Basic Level: bird
Entry Level: bird
Subordinates: American robin

Superordinates: animal, vertebrate
Basic Level: bird
Entry Level: penguin
Subordinates: Chinstrap penguin

example, Mr. Ed is a palomino, but also a horse, an equine, an odd-toed ungulate, a placental mammal, a mammal, and so on—most people looking at Mr. Ed would tend to call him a *horse*, his entry level category (unless they are fans of the show). Our paper focuses on the problem of object naming in the context of *entry-level categories*. We consider two related tasks: (1) learning a mapping from *fine-grained*/encyclopedic categories—for example, leaf nodes in WordNet[9]—to what people are likely to call them (*entry-level categories*) and (2) learning to map from outputs of thousands of noisy computer vision classifiers/detectors evaluated on an image to what a person is likely to call depicted objects.

Evaluations show that our models can effectively emulate the naming choices of humans. Furthermore, we show that using noisy computer vision estimates for image content, our system can output words that are significantly closer to human annotations than either the raw visual classifier predictions or the results of using a state of the art hierarchical classification system[6] that can output object labels at varying levels of abstraction, from very specific terms to very general categories.

1.1. Outline
The rest of this paper is organized as follows. Section 2 presents a summary of related work. Section 3 introduces a large-scale image categorization system based on deep convolutional neural network (CNN) activations. In Section 4, we learn translations between input linguistic concepts and entry-level concepts. In Section 5, we propose two models that can take an image as input and predict entry-level concepts for the depicted objects. Finally, in Sections 6 and 7 we provide experimental evaluations and conclusions.

2. RELATED WORK
Questions about *entry-level categories* are directly relevant to recent work on generating natural language descriptions for images.[8, 11, 15, 16, 19, 21] In these papers, the goal is to automatically produce natural language that describes the content of an image or video. We attack one specific facet of this problem, how to name objects in images in a human-like manner. Previous approaches that construct image descriptions directly from computer vision predictions often result

in unnatural constructions, for example, "Here we see one TV-monitor and one window."[15] Other methods handle naming choices indirectly by sampling human written text written about other visually similar objects.[16, 17]

On a technical level, our work is related to recent work from Deng et al.[6] that tries to "hedge" predictions of visual content by *optimally* backing off in the WordNet semantic hierarchy. For example, given a picture of a *dog*, a noisy visual predictor might easily mistake this for a *cat*. Therefore, outputting a more general prediction, for example, *animal*, may sometimes be better for overall performance in cases of visual ambiguity. One key difference is that our approach uses a reward function over the WordNet hierarchy that is non-monotonic along paths from the root to leaves, as it is based on word usage patterns, rather than perplexity. Another difference is that we make use of recent convolutional neural network based features for our underlying visual classifiers.[12] Our approach also allows mappings to be learned from a WordNet leaf node, l, to natural word choices that are not along a path from l to the root, "entity." In evaluations, our results significantly outperform the "hedging" technique[6] because although optimal for maximizing classification accuracy, it is not optimal with respect to how people describe image content.

Our work is also related to the growing challenge of harnessing the ever increasing number of pre-trained recognition systems, thus avoiding always "starting from scratch" in developing new applications. With the advent of large labeled datasets of images, including ImageNet[5] with over 15,000,000 labeled images for a subset of the WordNet hierarchy, a large amount of compute effort has been dedicated to building vision based recognition systems. It would be wasteful not to take advantage of the CPU weeks,[10, 14] months,[4, 6] or even millennia[18] invested in developing and training such recognition models. However, for any specific end user application, the categories of objects, scenes, and attributes labeled in a particular dataset may not be the most useful predictions. One benefit of our work can be seen as exploring the problem of translating the outputs of a vision system trained with one vocabulary of labels (WordNet leaf nodes) to labels in a new vocabulary (commonly used visually descriptive nouns).

Our proposed methods take into account several sources of structure and information: the structure of WordNet, frequencies of word use on the web,[2] outputs of a large-scale visual recognition system,[12] and large amounts of paired image and text data. In particular, we make use of the SBU Captioned Photo Dataset[21] which contains 1 million images with natural language captions as a source of natural image naming patterns. By incorporating all of these resources, we are able to study entry-level categories at a much larger scale than in previous settings.

2.1. Challenges of predicting entry-level categories
At first glance, the task of finding the entry-level categories may seem like a linguistic problem of finding a *hypernym* of any given word. Although there is a considerable conceptual connection between entry-level categories and hypernyms, there are two notable differences:

1. Although *"bird"* is a hypernym of both *"penguin,"* and

"sparrow," "bird" may be a good entry-level category for *"sparrow,"* but not for *"penguin."* This phenomenon, that some members of a category are more prototypical than others, is discussed in *Prototype Theory*.[23]

2. Entry-level categories are not confined by (inherited) hypernyms, in part because encyclopedic knowledge is different from common sense knowledge. For example *"rhea"* is not a kind of *"ostrich"* in the strict taxonomical sense. However, due to their visual similarity, people generally refer to a *"rhea"* as an *"ostrich."* Adding to the challenge is that although extensive, WordNet is neither complete nor practically optimal for our purpose. For example, according to WordNet, *"kitten"* is not a kind of *"cat,"* and *"tulip"* is not a kind of *"flower."*

In fact, both of the above points have a connection to visual information of objects, as visually similar objects are more likely to belong to the same entry-level category. In this work, we present the first extensive study that (1) characterizes entry-level categories in the context of translating encyclopedic visual categories to natural names that people commonly use, and (2) provides approaches that infer entry-level categories from a large-scale image corpus, guided by semantic word knowledge.

3. A LARGE-SCALE IMAGE CATEGORIZATION SYSTEM

We take advantage of recent advances in deep learning based visual features for training a large number of visual classifiers for leaf-node object categories. In particular, we use the pre-trained CNN model from the Caffe framework[12] based on the model from Krizhevsky et al.,[14] trained on 1000 imagenet categories from the ImageNet Large Scale Visual Recognition Challenge 2012. This model consists of a feed-forward neural network with multiple layers, each with different levels of connectivity between units in contiguous layers. The last few layers in the network consist of fully connected layers, where all the units in a given layer are connected to all the units in the subsequent layer. The output layer of the network consists of 1000 units corresponding to each category for the classification task. Donahue et al.[7] showed that the activations of some of the intermediate layers, particularly the fully connected layers before the output layer, were a useful generic image representation for a variety of other recognition tasks.

We similarly compute the 4096 activations in the last fully connected layer of this network and use these as features to train a linear SVM for each of 7404 leaf level categories in ImageNet. We also use a validation set to calibrate the output scores of each SVM using Platt scaling.[22] These 7404 visual classifiers will be used to predict image content either directly (Section 5.1) or to train entry-level visual predictors (Sections 4.2 and 5.2).

4. TRANSLATING ENCYCLOPEDIC CONCEPTS TO ENTRY-LEVEL CONCEPTS

Our first goal toward understanding how people name objects, is to learn mappings between encyclopedic concepts (ImageNet leaf categories, e.g., Chlorophyllum molybdites) and concepts that are more *natural* (e.g., mushroom). In Section 4.1, we present an approach that relies on the WordNet

hierarchy and frequencies of words in a web scale corpus. In Section 4.2, we follow an approach that uses visual recognition models learned on a paired image-caption dataset.

4.1. Language-based translation

As a baseline, we first consider a translation approach that relies only on language-based information: the hierarchical semantic tree from WordNet[9] and text statistics from the Google Web 1T corpus.[2] We posit that the frequencies of terms computed from massive amounts of text on the web reflect the "naturalness" of concepts. We use the n-gram counts of the Google Web 1T corpus[2] as a proxy for naturalness. Specifically, for a synset w, we quantify naturalness as, $\phi(w)$, the log of the count for the most commonly used synonym in w. As possible translation concepts for a given category, v, we consider all nodes, w in v's inherited hypernym structure (all of the synsets along the WordNet path from w to the root).

We define a translation function for a category v, $\tau(v, \lambda)$, that maps v to a new node w, such that w maximizes the trade-off between naturalness, $\phi(w)$, and semantic proximity, $\psi(w, v)$, measuring the distance between node v and node w in the WordNet hypernym structure:

$$\tau(v, \lambda) = \arg\max_w [\phi(w) - \lambda \psi(w, v)], \ w \in \Pi(v), \qquad (1)$$

where $\Pi(v)$ is the set of (inherited) hypernyms from v to the root, including v. For instance given an input category $v = King \ penguin$ we consider all categories along its set of inherited hypernyms, for example, *penguin, seabird, bird, animal* (see Figure 3). An ideal prediction for this concept would be *penguin*. We use line search to find the optimal λ, which controls how much we care about naturalness versus semantic proximity, based on a held out set of subordinate-category, entry-level category pairs $D = (x_i, y_i)$ collected using crowdsourcing to maximize the number of correct translations predicted by our model:

$$\Phi(D, \lambda) = \sum_i \mathbb{1}[\tau(x_i, \lambda) = y_i], \qquad (2)$$

where $\mathbb{1}[\cdot]$ is the indicator function. We show the relationship between λ and translation accuracy, $\Phi(D, \lambda)$, in Figure 4, where

Figure 3. Our first categorical translation model uses the WordNet hierarchy to find an hypernym that is close to the leaf node concept (*semantic distance*) and has a large naturalness score based on its n-gram frequency. The green arrows indicate the ideal category that would correspond to the entry-level category for each leaf-node in this sample semantic hierarchy.
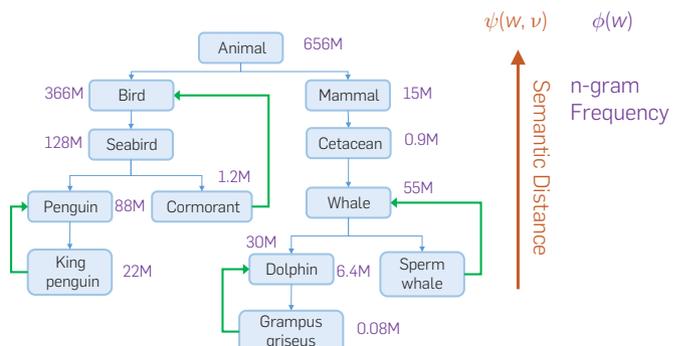
Figure 4. Relationship between parameter λ and translation accuracy, Φ(*D*, λ), evaluated on the most agreed human label (red) or any human label (cyan).
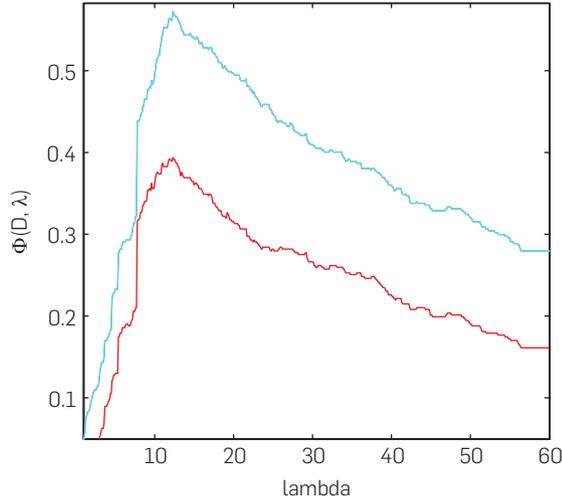
the red line shows accuracy for predicting the word used by the most people for a synset, while the cyan line shows the accuracy for predicting any word used by a labeler for the synset. As we increase λ, Φ(*D*, λ) increases initially and then decreases as too much generalization or specificity reduces the naturalness of the predictions. For example, generalizing from *grampus griseus* to *dolphin* is good for "naturalness," but generalizing all the way to "entity" decreases "naturalness."

Our experiment also supports the idea that *entry-level categories* lie at a level of abstraction where there is a discontinuity. Going beyond this level of abstraction suddenly makes our predictions considerably worse. Rosch[23] indeed argues in the context of basic level categories that basic cuts in categorization happen precisely at these discontinuities where there are bundles of information-rich functional and perceptual attributes.

4.2. Visual-based translation
Next, we try to make use of pre-trained visual classifiers to improve translations between input concepts and entry-level concepts. For a given leaf synset, *v*, we sample a set of *n* = 100 images from ImageNet. For each image, *i*, we predict some potential entry-level nouns, N_i, using pre-trained visual classifiers that we will further describe in Section 5.2. We use the union of this set of labels $N = N_1 \cup N_2 \dots \cup N_n$ as keyword annotations for synset *v* and rank them using a *term frequency-inverse document frequency* (TFIDF) information retrieval measure. This ranking measure promotes labels that are predicted frequently for our set of 100 images, while decreasing the importance of labels that are predicted frequently in all our experiments across different categories. We pick the most highly ranked noun for each node, *v*, as its entry-level categorical translation.

We show a comparison of the output of this approach with our Language-based Translation approach and mappings provided by human annotators in Table 1. We explain the collection of human annotations in the evaluation section (Section 6.1).

Table 1. Translations from ImageNet leaf node synset categories to *entry-level categories* using our automatic approaches from Sections 4.1 (left) and 4.2 (center) and crowd-sourced human annotations (right).

Input concept	Language-based translation	Visual-based translation	Human translation
Cactus wren	Bird	Bird	Bird
Buzzard, Buteo buteo	Hawk	Hawk	Hawk
Whinchat, Saxicola rubetra	Chat	Bird	Bird
Weimaraner	Dog	Dog	Dog
Numbat, banded anteater, anteater	Anteater	Dog	Anteater
Rhea, Rhea americana	Bird	Grass	Ostrich
Conger, conger eel	Eel	Fish	Fish
Merino, merino sheep	Sheep	Sheep	Sheep
Yellowbelly marmot, rockchuck	Marmot	Male	Squirrel
Snorkeling, snorkel diving	Swimming	Sea turtle	Snorkel

5. PREDICTING ENTRY-LEVEL CONCEPTS FOR IMAGES
In Section 4, we proposed models to translate between one linguistic concept, for example, *grampus griseus*, to a more natural object name, for example, *dolphin*. Our objective in this section is to explore methods that can take an image as input and predict entry-level labels for the depicted objects. The models we propose are: (1) a method that combines "naturalness" measures from text statistics with direct estimates of visual content computed at leaf nodes and inferred for internal nodes (Section 5.1) and (2) a method that learns visual models for entry-level category prediction directly from a large collection of images with associated captions (Section 5.2).

5.1. Linguistically guided naming
In our first image prediction method, we estimate image content for an image, *I*, using the pre-trained visual models described in Section 3. These models predict the presence or absence of 7404 leaf level visual categories in the ImageNet (WordNet) hierarchy. Following the "hedging" approach,[6] we compute estimates of visual content for internal nodes in the hierarchy by accumulating all predictions below a node:

$$f(v, I) = \begin{cases} \hat{f}(v, I), & \text{if } v \text{ is an leaf node,} \\ \sum_{v' \in Z(v)} \hat{f}(v', I), & \text{if } v \text{ is an internal node,} \end{cases} \quad (3)$$

where *Z*(*v*) is the set of all leaf nodes under node *v* and $\hat{f}(v, I)$ is the output of a platt-scaled decision value from a linear SVM trained to recognize category *v*. Similar to our approach in Section 4.1, we define for every node in the ImageNet hierarchy

a trade-off function between "naturalness" ϕ (n-gram counts) and specificity $\tilde{\psi}$ (relative position in the WordNet hierarchy):

$$\gamma(v, \hat{\lambda}) = [\phi(w) - \hat{\lambda}\tilde{\psi}(w)], \qquad (4)$$

where $\phi(w)$ is computed as the log counts of the nouns and compound nouns in the text corpus from the *SBU Captioned Dataset*,[21] and $\tilde{\psi}(w)$ is an upper bound on $\psi(w, v)$ from Equation (1) equal to the maximum height in the WordNet hierarchy for node w. We parameterize this trade-off by $\hat{\lambda}$.

For entry-level category prediction for images, we would like to maximize both "naturalness" and visual content estimates. For example, text based "naturalness" will tell us that both *cat* and *swan* are good entry-level categories, but a confident visual prediction for *Cygnus Colombianus* for an image tells us that *swan* is a much better entry-level prediction than *cat* for that image.

Therefore, for an input image, we want to output a set of concepts that have a large prediction for both "naturalness" and content estimate score. For our experiments we output the top K WordNet synsets with the highest f_{nat} scores:

$$f_{nat}(v, I, \hat{\lambda}) = f(v, I)\gamma(v, \hat{\lambda}). \qquad (5)$$

As we change $\hat{\lambda}$ we expect similar behavior to our concept translations (Section 4.1), tuning $\hat{\lambda}$ to control the degree of specificity while trying to preserve "naturalness." We compare our framework to the "hedging" technique[6] for different settings of $\hat{\lambda}$. For a side by side comparison we modify hedging to output the top K synsets based on their scoring function. Here, the working vocabulary is the unique set of predicted labels output for each method on this test set. Results demonstrate (Figure 5) that under different parameter settings we *consistently* obtain much higher levels of precision for predicting entry-level categories than hedging.[6] We also obtain an additional gain in performance over our previous work[20] by incorporating dataset-specific text-statistics from the *SBU Captioned Dataset* rather than the more generic *Google Web 1T* corpus.

5.2. Visually guided naming
In the previous section, we rely on WordNet structure to compute estimates of image content, especially for internal nodes. However, this is not always a good measure of content prediction because: (1) The WordNet hierarchy doesn't encode knowledge about some semantic relationships between objects (i.e., functional or contextual relationships), (2) Even with the vast coverage of 7404 ImageNet leaf nodes we are missing models for many potentially important entry-level categories that are not at the leaf level.

As one alternative, we can directly train models for entry-level categories from data where people have provided entry-level labels—in the form of nouns present in visually descriptive image captions. We postulate that these nouns represent examples of entry-level labels because they have been naturally annotated by people to describe what is present in an image. For this task, we leverage the SBU Captioned Photo Dataset[21] which contains *1 million* captioned images. We transform this dataset into a set $D = \{X^{(i)}, Y^{(i)} \mid X^{(i)} \in \mathbf{X}, Y^{(i)} \in \mathbf{Y}\}$, where $\mathbf{X} = [0\text{--}1]^s$ is a vector of estimates of visual content for

Figure 5. Relationship between average precision agreement and working vocabulary size (on a set of 1000 images) for the hedging method (red) and our Linguistically guided naming method using text statistics from the generic Google Web 1T dataset (magenta) and from the SBU Caption Dataset (Section 5.1). We use $K = 5$ to generate this plot and a random set of 1000 images from the SBU Captioned Dataset.
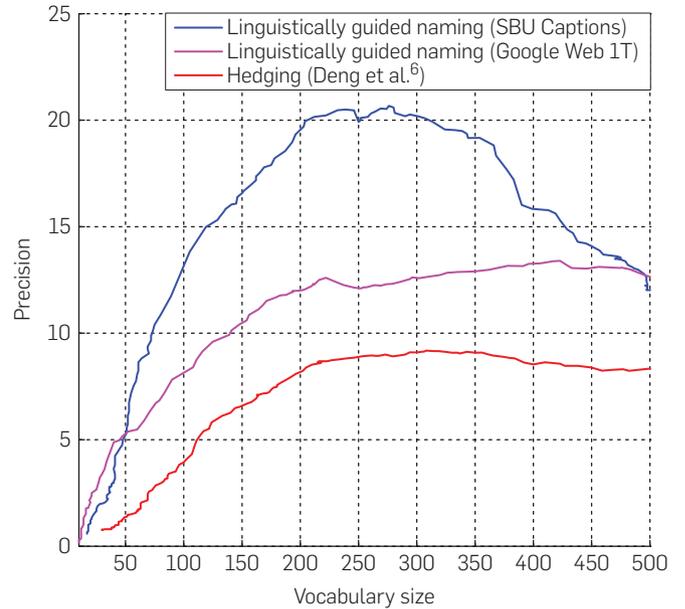
$s = 7404$ ImageNet leaf node categories and $\mathbf{Y} = [0, 1]^d$ is a set of binary output labels for d target categories. Input content estimates are provided by the deep learning based SVM predictions (described in Section 3).

For training our d target categories, we obtain labels Y from the million captions by running a POS-tagger[1] and defining $Y^{(j)} = \{y_{ij}\}$ such that:

$$y_{ij} = \begin{cases} 1, & \text{if caption for image } j \text{ has noun } i, \\ 0, & \text{if otherwise.} \end{cases} \qquad (6)$$

The POS-tagger helps clean up some word sense ambiguity due to polysemy, by only selecting those instances where a word is used as a noun. The number of target categories, d, is determined experimentally from data by learning models for the most frequent nouns in this dataset. This provides us with a target vocabulary that is both likely to contain entry-level categories (because we expect entry-level category nouns to commonly occur in our visual descriptions) and to contain sufficient images for training effective recognition models. We use up to 10,000 images for training each model. Since we are using human labels from real-world data, the frequency of words in our target vocabulary follows a power-law distribution. Hence we only have a very large amount of training data for a few most commonly occurring noun concepts. Specifically, we learn linear SVMs followed by Platt scaling for each of our target concepts. We keep $d = 1169$ of the best performing models. Our scoring function f_{svm} for a target concept v_i is then:

$$f_{svm}(v_i, I, \theta_i) = \frac{1}{1 - \exp(a_i\theta_i^\top X + b_i)}, \qquad (7)$$

where θ_i are the model parameters for predicting concept v_i,

and a_i and b_i are Platt scaling parameters learned for each target concept v_i on a held out validation set.

$$R(\theta_i) = \frac{1}{2}\|\theta_i\| + c\sum_{j=1}^{|D|}\max(0, 1 - y_{ij}\theta_i^\top X^{(j)})^2. \qquad (8)$$

We learn the parameters θ_i by minimizing the squared hinge-loss with ℓ_1 regularization (Equation 8). The latter provides a natural way of modeling the relationships between the input and output label spaces that encourages sparseness (examples in Figure 6). We find $c = 0.01$ to yield good results for our problem and use this value for training all individual models.

One of the drawbacks of using the ImageNet hierarchy to aggregate estimates of visual concepts (Section 5.1) is that it ignores more complex relationships between concepts. Here, our data-driven approach to the problem implicitly discovers these relationships. For instance a concept like *tree* has co-occurrence relationships with various types of birds, and other animals that live on trees (see Figure 6).

Given this large dataset of images with noisy visual predictions and text labels, we manage to learn quite good predictors of high-level content, even for categories with relatively high intra-class variation (e.g., girl, boy, market, house).

6. EXPERIMENTAL EVALUATION
We evaluate results on our two proposed naming tasks—learning translations from encyclopedic concepts to entry-level concepts (Section 6.1), and predicting entry-level concepts for objects in images (Section 6.2).

6.1. Evaluating translations
We use Amazon Mechanical Turk to crowd source translations of ImageNet synsets into entry-level categories $D = \{x_i, y_i \mid x_i$ is a leaf node, y_i is a word$\}$. Our experiments present users with a 2×5 array of images sampled from an ImageNet synset, x_i, and users are asked to provide a label, y_i, for the depicted concept. Results are obtained for 500 ImageNet synsets and aggregated across 8 users per task. We found agreement (measured as at least 3 of 8 users agreeing) among users for 447 of the 500 concepts, indicating that even though there are many potential labels for each synset (e.g., *Sarcophaga carnaria* could conceivably be labeled as fly, dipterous insect, insect, arthropod, etc.) people have a strong preference for particular entry-level categories.

We show sample results from each of our methods to learn concept translations in Table 1. In some cases linguistics-based translation fails. For example, *whinchat* (a type of bird) translates to "chat" most likely because of the inflated counts for the most common use of "chat." Our visual-based translation fails when it learns to weight context words highly, for example "snorkeling" → "water," or "African bee" → "flower" even when we try to account for common context words using TFIDF. Finally, even humans are not always correct, for example "Rhea Americana" looks like an ostrich, but is not taxonomically one. Even for categories like "marmot" most people named it "squirrel." Overall, our language-based translation (Section 4.1) agrees 37% of the time with human supplied translations and the visual-based translation (Section 4.2) agrees 33% of the time, indicating that translation learning is a non-trivial task. This experiment expands on previous studies in psychology.[13, 23] Cheap and easy online crowdsourcing enables us to gather these labels for a much larger set of (500) concepts than previous experiments and to learn generalizations for a substantially larger set of ImageNet synsets.

6.2. Evaluating image entry-level predictions
We measure the accuracy of our proposed entry-level category image prediction methods by evaluating how well we can predict nouns freely associated with images by users on Amazon Mechanical Turk. Results are evaluated on a test set containing 1000 images selected at random from the million image dataset. We additionally collect annotations for another 2000 images so that we can tune trade-off parameters in our models. This test set is completely disjoint from the sets of images used for learning the pre-trained visual models. For each image, we instruct three users on MTurk to write down any nouns that are relevant to the image content. Because these annotations are free associations we observe a large and varied set with 3610 distinct nouns total in our evaluation sets. This makes noun prediction extremely challenging!

Figure 6. Entry-level category *tree* with its corresponding top weighted leaf node features after training an SVM on our noisy data, and a visualization of weights grouped by an arbitrary categorization of leaf nodes. Vegetation (green), birds (orange), instruments (blue), structures (brown), mammals (red), and others (black).

tree →
iron tree, iron-tree, ironwood, ironwood tree
snag
European silver fir, Christmas tree, Abies alba
baobab, monkey-bread tree, Adansonia digitata
Japanese black pine, black pine, Pinus thunbergii
huisache, cassie, mimosa bush, sweet wattle, sweet acacia, scented wattle,
flame tree, Acacia farnesiana
feeder
bird feeder, birdfeeder, feeder
koala, koala bear, kangaroo bear, native bear, Phascolarctos cinereus
flying fox
damask
American basswood, American lime, Tilia americana
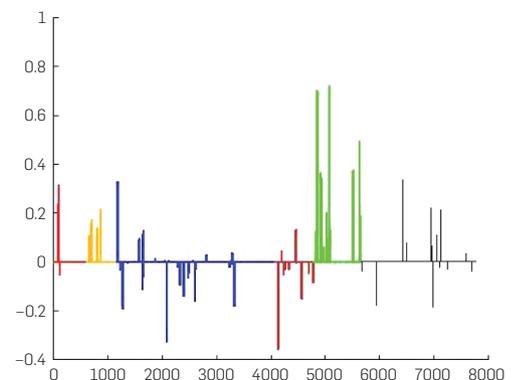
For evaluation, we measure how well we can predict all nouns associated with an image by Turkers (Figure 7a) and how well we can predict the nouns commonly associated by Turkers (assigned by at least two of three Turkers, Figure 7b). For reference we compute the precision of one human annotator against the other two and found that on our test set humans were able to predict what the previous annotators labeled with 0.35 precision when compared to the agreed set of nouns by Turkers.

Results show precision and recall for prediction on our test set, comparing: leaf node classification performance (flat classifier), the outputs of "hedging,"[6] and our proposed entry-level category predictors (Linguistically guided, Section 5.1, and Visually guided, Section 5.2). Performance on the test set is admirable for this challenging task. On the two datasets we find the Visually guided naming model to perform better (Section 5.2) than the Linguistically guided naming (Section 5.1). In addition, we significantly outperform both leaf node classification and the "hedging"

technique.[6] We show an image with sample output from our methods at $K = 5$ in Figure 8.

7. CONCLUSION
We have explored models for mapping encyclopedic concepts to entry-level concepts, and for predicting natural names for objects depicted in images. Results indicate that our inferred concept translations are meaningful and that our models provide a first step toward predicting entry-level categories—the nouns people use to name objects—depicted in images. These methods could be helpful for many different end-user applications that require recognition outputs that are useful for human consumption, including tasks related to description generation and image retrieval from complex text queries.

Acknowledgments

Figure 7. Precision-recall curve for different entry-level prediction methods when using the top K categorical predictions for $K = 1, 3, 5, 10, 15, 20, 50$. (a) An evaluation using the union of all human labels as ground truth and (b) using only the set of labels where at least two users agreed.
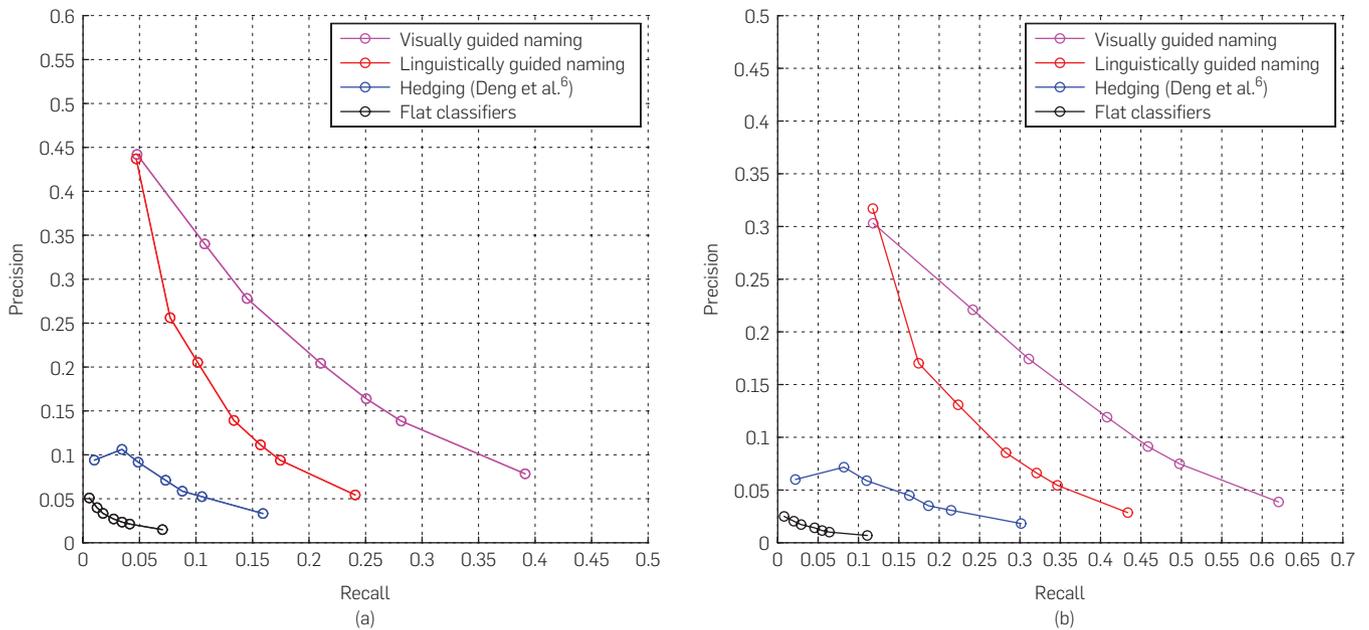
Figure 8. Category predictions for an example input image for a large-scale categorization system and our translated outputs using linguistically and visually guided models. The first column contains nouns associated with the image by people. We highlight in green the predicted nouns that were also mentioned by people. Note that *oast* is a type of farm building for drying hops and a *dacha* is a type of Russian farm building.

References
1. Bird, S. Nltk: The natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions* (July 2006). Association for Computational Linguistics, Sydney, Australia, 69–72.
2. Brants, T., Franz, A. Web 1t 5-gram version 1. In *Linguistic Data Consortium (LDC)* (2006), Linguistic Data Consortium, Philadelphia.
3. Dean, T., Ruzon, M.A., Segal, M., Shlens, J., Vijayanarasimhan, S., Yagnik, J. Fast, accurate detection of 100,000 object classes on a single machine. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2013), 18141–821.
4. Deng, J., Berg, A.C., Li, K., Li, F.-F. What does classifying more than 10,000 image categories tell us? In *European Conference on Computer Vision (ECCV)*, Daniilidis, Kostas and Maragos, Petros and Paragios, Nikos, eds. Volume 6315 of *Lecture Notes in Computer Science* (2010), Springer, Berlin, Heidelberg, 71–84.
5. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009 (June 2009), 248–255.
6. Deng, J., Krause, J., Berg, A.C., Fei-Fei, L. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012 (June 2012), 3450–3457.
7. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition, 2013. arXiv preprint arXiv:1310.1531.
8. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D. Every picture tells a story: Generating sentences for images. In *European Conference on Computer Vision (ECCV)*, Daniilidis, Kostas and Maragos, Petros and Paragios, Nikos, eds. Volume 6314 of *Lecture Notes in Computer Science* (2010), Springer, Berlin, Heidelberg, 15–29.
9. Fellbaum, C., ed. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, Massachusetts, 1998.
10. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell. 32,* 9 (Sept 2010), 1627–1645.
11. Hodosh, M., Young, P., Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res. 47,* 1 (May 2013), 853–899.
12. Jia, Y. Caffe: An open source convolutional architecture for fast feature embedding, 2013. http://caffe.berkeleyvision.org/.
13. Jolicoeur, P., Gluck, M.A., Kosslyn, S.M. Pictures and names: Making the connection. cognitive psychology. *Cogn. Psychol. 16,* (1984), 243–275, 1984.
14. Krizhevsky, A., Sutskever, I., Hinton, G. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25,* F. Pereira and C.J.C. Burges and L. Bottou and K.Q. Weinberger, eds. (2012), Curran Associates, Inc., 1097–1105.
15. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A., Berg, T. Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Machine Intell. 35,* 12 (Dec 2013), 2891–2903.
16. Kuznetsova, P., Ordonez, V., Berg, A., Berg, T.L., Choi, Y. Collective generation of natural image descriptions. In *Association for Computational Linguistics (ACL)*, 2012.
17. Kuznetsova, P., Ordonez, V., Berg, T., Choi, Y. Treetalk: Composition and compression of trees for image descriptions. *Trans. Assoc. Comput. Linguist. 2,* 1 (2014), 351–362.
18. Le, Q., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J., Ng, A. Building high-level features using large scale unsupervised learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML12)*, John Langford and Joelle Pineau, eds. (Edinburgh, Scotland, GB, July 2012), Omnipress, New York, NY, USA, 81–88.
19. Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., Daumé, H. III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (April 2012), Association for Computational Linguistics, Avignon, France, 747–756.
20. Ordonez, V., Deng, J., Choi, Y., Berg, A.C., Berg, T.L. From large scale image categorization to entry-level categories. In *2013 IEEE International Conference on Computer Vision (ICCV)* (Dec 2013), 2768–2775.
21. Ordonez, V., Kulkarni, G., Berg, T.L. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24,* J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, eds. (2011), Curran Associates, Inc., 1143–1151.
22. Platt, J.C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers* (1999), MIT Press, 61–74.
23. Rosch, E. Principles of categorization. In *Cognition and Categorization,* E. Rosch and B.B. Lloyd, eds. (1978), 27–48.
24. Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition, Sept. 2014. arXiv preprint arXiv:1409.1556.
25. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. Going deeper with convolutions, Sept. 2014. arXiv preprint arXiv:1409.4842.

Vicente Ordonez (vicenteor@alienai.org), Allen Institute for Artificial Intelligence, Seattle, WA.

Wei Liu, Alexander C. Berg, and Tamara L. Berg ({wliu, aberg, tlberg}@cs.unc.edu), Department of Computer Science, University of North Carolina at Chapel Hill, NC.

Jia Deng (jiadeng@umich.edu), Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI.

Yejin Choi (yejin@cs.washington.edu), Department of Computer Science and Engineering, University of Washington, Seattle, WA.