Predicting Entry-Level Categories

**Vicente Ordonez · Wei Liu · Jia Deng · Yejin Choi ·
Alexander C. Berg · Tamara L. Berg**

Abstract Entry-level categories — the labels people use to
name an object — were originally defined and studied by
psychologists in the 1970s and 80s. In this paper we extend
these ideas to study entry-level categories at a larger scale
and to learn models that can automatically predict entry-
level categories for images. Our models combine visual recog-
nition predictions with linguistic resources like WordNet and
proxies for word "naturalness" mined from the enormous
amount of text on the web. We demonstrate the usefulness
of our models for predicting nouns (entry-level words) asso-
ciated with images by people, and for learning mappings be-
tween concepts predicted by existing visual recognition sys-
tems and entry-level concepts. In this work we make use of
recent successful efforts on convolutional network models
for visual recognition by training classifiers for 7,404 object
categories on *ConvNet* activation features. Results for cate-
gory mapping and entry-level category prediction for images
show promise for producing more natural human-like labels.
We also demonstrate the potential applicability of our results
to the task of image description generation.

V. Ordonez (✉) · W. Liu · A.C. Berg · T.L. Berg
Department of Computer Science
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
E-mail: vicente@cs.unc.edu

J. Deng
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI, USA

Y. Choi
Computer Science and Engineering
University of Washington, Seattle, WA, USA

Recognition Prediction   *What should I Call It?*

grampus griseus ⟶ dolphin

Fig. 1 Example translation between a WordNet based object category
prediction and what people might call the depicted object.

1 Introduction

Computational visual recognition is beginning to work. Al-
though far from solved, algorithms have now advanced to
the point where they can recognize or localize thousands
of object categories with reasonable accuracy (Deng et al.,
2010; Perronnin et al., 2012; Krizhevsky et al., 2012; Dean
et al., 2013; Simonyan and Zisserman, 2014; Szegedy et al.,
2014). Russakovsky et al. (2014) present an overview of
recent advances in classification and localization for up to
1000 object categories. While one 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 *grampus griseus*, while most people would simply
call this object a *dolphin*. We propose to develop categoriza-
tion systems that are aware of these kinds of human naming
choices.

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

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

Fig. 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.

This notion is closely related to ideas of *basic and entry-level categories* formulated by psychologists such as Eleanor Rosch (Rosch, 1978) and Stephen Kosslyn (Jolicoeur et al., 1984). Rosch defines *basic-level categories* as roughly those categories at the highest level of generality that 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. Super-ordinate, more general, categories such as *animal* will share fewer attributes and demonstrate more variability. Subordinate, more specific categories, such as *American Robin* will share even more attributes like shape, color, and size. Rosch studied basic level categories through human experiments, e.g. asking people to enumerate common attributes for a given category. The work of Jolicoeur et al. (1984) further studied the way people identify categories, defining the concept of *entry-level categories*. Entry level categories are essentially the categories that people 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 the entry-level category for a *penguin* would be at a lower level of abstraction (see Figure 2).

So, while objects are members of many categories – e.g. 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 cat-

egories – *e.g.*, leaf nodes in WordNet (Fellbaum, 1998) – 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 a depicted object.

Evaluations show that our models can effectively emulate the naming choices of human observers. Furthermore, we show that using noisy vision estimates for image content, our system can output words that are significantly closer to human annotations than either raw visual classifier predictions or the results of using a state of the art hierarchical classification system (Deng et al., 2012) that can output object labels at varying levels of abstraction from very specific terms to very general categories.

1.1 Insights into 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* (Rosch, 1978).
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 methods to predict entry-level categories for input images guided by semantic word knowledge or by using a large-scale corpus of images with text.

1.2 Paper Overview

Our paper is divided as follows. Section 2 presents a summary of related work. Section 3 introduces a large-scale image categorization system based on convolutional network activations. In Section 4 we learn translations from subordinate concepts to entry-level concepts. In Section 5 we propose two models that can take an image as input and predict entry-level concepts. Finally, in Section 6 we provide experimental evaluations.

The major additions in this journal version compared to our previous publication (Ordonez et al., 2013) include an expanded discussion about related work in Section 2. We have also replaced the large scale image categorization system based on hand-crafted SIFT + LLC features used in our previous work with a system based on state-of-the-art convolutional network activations obtained using the Caffe framework (Jia, 2013) (Section 3). Sections 4 and 5 have been updated accordingly and include additional qualitative examples. Section 6 contains a more thorough evaluation studying the effect of the number of predicted outputs on precision and recall, and an additional extrinsic evaluation of the system in a sentence retrieval application.

2 Related work

Questions about *entry-level categories* are directly relevant to recent work on the connection between computer vision outputs and (generating) natural language descriptions of images (Farhadi et al., 2010; Ordonez et al., 2011; Kuznetsova et al., 2012; Mitchell et al., 2012; Yang et al., 2011; Gupta et al., 2012; Kulkarni et al., 2013; Hodosh et al., 2013; Ramnath et al., 2014; Mason and Charniak, 2014; Kuznetsova et al., 2014). Previous works have not directly addressed naming preference choices for entry-level categories when generating sentences. Often the computer vision label predictions are used directly during surface realization (Mitchell et al., 2012; Kulkarni et al., 2013), resulting in choosing non-human like namings for constructing sentences even when handling a relatively small number of categories (i.e. Pascal VOC categories like potted-plant, tv-monitor or person). For these methods, our entry-level category predictions could be used to generate more natural names for objects. Other methods handle naming choices indirectly in a data-driven fashion by borrowing human references from other visually similar objects (Kuznetsova et al., 2012, 2014; Mason and Charniak, 2014).

Our work is also related to previous works that aim to discover visual categories from large-scale data. The works of Yanai and Barnard (2005) and Barnard and Yanai (2006) learn models for a set of categories by exploring images with loosely associated text from the web. We learn our set of categories directly as a subset of the WordNet (Fellbaum,

1998) hierarchy, or from the nouns used in a large set of carefully selected image captions that directly refer to images. The more recent works of Chen et al. (2013) and Divvala et al. (2014) present systems capable of learning any type of visual concept from images on the web, including efforts to learn simple common sense relationships between visual concepts (Chen et al., 2013). We provide a related output in our work, learning mappings between *entry-level categories* and subordinate/leaf-node categories. The recent work of Feng et al. (2015) proposes that entry-level categorization can be viewed as lexical semantic knowledge, and presents a global inference formulation to map all encyclopedic categories to their entry-level categories collectively.

On a technical level, our work is related to (Deng et al., 2012) that tries to "hedge" predictions of visual content by *optimally* backing off in the WordNet hierarchy. 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 the leaves. Another difference is that we have replaced the underlying leaf node classifiers from Deng et al. (2012) with recent convolutional network activation features. 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 those of (Deng et al., 2012) because although optimal in some sense, they are 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 "starting from scratch" whenever developing new applications. It is wasteful not to take advantage of the CPU weeks (Felzenszwalb et al., 2010; Krizhevsky et al., 2012), months (Deng et al., 2010, 2012), or even millennia (Le et al., 2012) invested in developing recognition models for increasingly large labeled datasets (Everingham et al., 2010; Russell et al., 2008; Xiao et al., 2010; Deng et al., 2009; Torralba et al., 2008). 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 in large amounts of web text, outputs of a large-scale visual recognition system, and large amounts of paired image and text data. In particular, we use the SBU Captioned Photo Dataset (Ordonez et al., 2011), which consists of 1 million images with natural language descriptions, and Google n-gram frequencies collected for all words on the web. Taking all of these resources together, we are able

to study patterns for choice of entry-level categories at a much larger scale than previous psychology experiments.

3 A Large-Scale Image Categorization System

Large-scale image categorization has improved drastically in recent years. The computer vision community has moved from handling 101 categories (Fei-Fei et al., 2007) to 100,000 categories (Dean et al., 2013) in a few years. Large-scale datasets like ImageNet (Deng et al., 2009) and recent progress in training deep layered architectures (Krizhevsky et al., 2012) have significantly improved the state-of-the-art. We leverage a system based on these as the starting point for our work.

For features, we use activations from an internal layer of a convolutional network, following the approach of (Donahue et al., 2013). In particular, we use the pre-trained reference model from the Caffe framework (Jia, 2013) which is in turn based on the model from Krizhevsky et al. (2012). This model was trained on the 1,000 ImageNet categories from the ImageNet Large Scale Visual Recognition Challenge 2012. We compute the 4,096 activations in the 7th layer of this network for images in 7,404 leaf node categories from ImageNet and use them as features to train a linear SVM for each category. We further use a validation set to calibrate the output scores of each SVM with Platt scaling (Platt, 1999).

4 Translating Encyclopedic Concepts to Entry-Level Concepts

Our objective in this section is to discover mappings between subordinate encyclopedic concepts (ImageNet leaf categories, e.g. Chlorophyllum molybdites) to output concepts that are more *natural* (e.g. mushroom). In Section 4.1 we present an approach that relies on the WordNet hierarchy and frequency 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

We first consider a translation approach that relies only on language-based information: the hierarchical semantic structure from WordNet (Fellbaum, 1998) and text statistics from the Google Web 1T corpus (Brants and Franz., 2006). 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 (Brants and Franz., 2006) 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
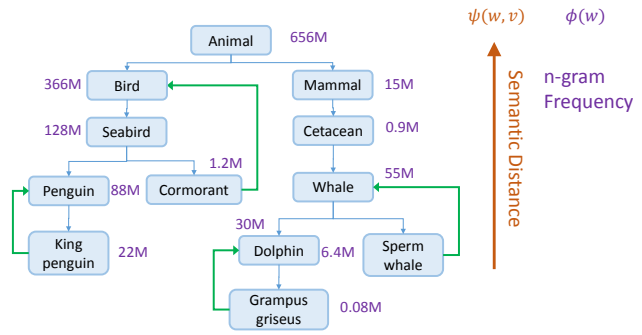
Fig. 3 Our first categorical translation model uses the WordNet hierarchy to find an hypernym that is close to the leaf node concept (*semantic distance*) and has a large naturalness score based on its n-gram frequency. The green arrows indicate the ideal category that would correspond to the entry-level category for each leaf-node in this sample semantic hierarchy.

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, $\tau(v, \lambda)$, that maximizes a trade-off between naturalness, $\phi(w)$, and semantic proximity, $\psi(w, v)$, measuring the distance between leaf 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, e.g. *penguin, seabird, bird, animal* (see Figure 3). An ideal prediction for this concept would be *penguin*. To control how the overall system trades off naturalness vs semantic proximity, we perform line search to set λ. For this purpose we use a held out set of subordinate-category, entry-level category pairs (x_i, y_i) collected using Amazon Mechanical Turk (MTurk) (for details refer to Section 6.1). Our objective is 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 vocabulary size in Figure 4(a), and between λ and overall translation accuracy, $\Phi(D, \lambda)$, in Figure 4(b). As we increase λ, $\Phi(D, \lambda)$ 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". In Figure 4(b) the red line shows accuracy for predicting the most agreed upon word for a synset, while

(a) vocabulary size vs. λ (b) naturalness vs. λ
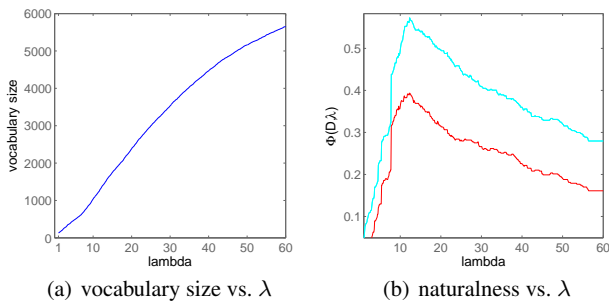
Fig. 4 Left: shows the relationship between parameter λ and the target vocabulary size. **Right:** shows the relationship between parameter λ and agreement accuracy with human labeled synsets evaluated against the most agreed human label (red) and any human label (cyan).

the cyan line shows the accuracy for predicting any word collected from any user. Our experiment also supports that *entry-level categories* seem to lie at a certain level of abstraction where there is a discontinuity. Going beyond this level of abstraction suddenly makes our predictions considerably worse (see Figure 4(b)). Rosch (1978) 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 describe later in Section 5.2. We use the union of this set of labels $N = N_1 \cup N_2 ... \cup N_n$ as keyword annotations for synset v and rank them using a TFIDF information retrieval measure. We consider each category v as a document for computing the *inverse document frequency* (IDF) term. We pick the most highly ranked noun for each node, v, as its entry-level categorical translation (see an example in Figure 5).

5 Predicting Entry-Level Concepts for Images

In Section 4 we proposed models to translate between one linguistic concept, e.g. *grampus griseus*, to a more natural concept, e.g. *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
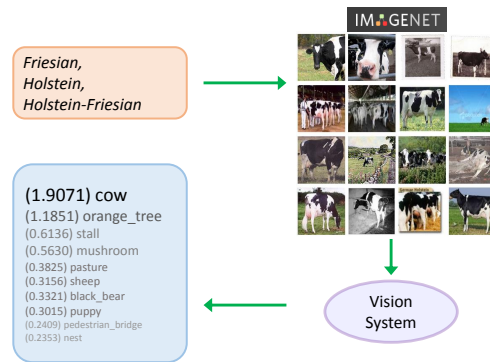
Fig. 5 We show the system instances of the category *Friesian, Holstein, Holstein-Friesian* and the vision system pre-trained with candidate entry-level categories ranks a set of candidate keywords and outputs the most relevant, in this case *cow*.

(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

We estimate image content for an image, I, using the pre-trained models from Section 3. These models predict presence or absence of 7,404 leaf node concepts in ImageNet (WordNet). Following the approach of Deng et al. (2012), we compute estimates of visual content for internal nodes by hierarchically accumulating all predictions below a node:[1]

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

where $Z(v)$ is the set of all leaf nodes under node v and $\hat{f}(v, I)$ is a score predicting the presence of leaf node category v from our large scale image categorization system introduced in Section 3. Similar to our approach in Section 4.1, we define for every node in the ImageNet hierarchy a trade-off function between "naturalness" ϕ (ngram counts) and specificity $\tilde{\psi}$ (relative position in the WordNet hierarchy):

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

where $\phi(w)$ is computed as the log counts of the nouns and compound nouns in the text corpus from the *SBU Captioned Dataset* (Ordonez et al., 2011), and $\tilde{\psi}(w)$ is an upper bound on $\psi(w, v)$ from equation (1) equal to the maximum path in the WordNet structure from node v to node w. We parameterize this trade-off by $\hat{\lambda}$.

---

[1] This function might bias decisions toward internal nodes. Other alternatives could be explored to estimate internal node scores.

|	Input Concept	Language-based Translation	Visual-based Translation	Human Translation
1	eastern kingbird	bird	bird	bird
2	cactus wren	bird	bird	bird
3	buzzard, Buteo buteo	hawk	hawk	hawk
4	whinchat, Saxicola rubetra	chat	bird	bird
6	Weimaraner	dog	dog	dog
7	Gordon setter	dog	dog	dog
8	numbat, banded anteater, anteater	anteater	dog	anteater
9	rhea, Rhea americana	bird	grass	ostrich
10	Africanized bee, killer bee, Apis mellifera	bee	bee	bee
11	conger, conger eel	eel	fish	fish
12	merino, merino sheep	sheep	sheep	sheep
13	Europ. black grouse, heathfowl, Lyrurus tetrix	bird	bird	bird
14	yellowbelly marmot, rockchuck, Marm. flaviventris	marmot	male	squirrel
15	snorkeling, snorkel diving	swimming	sea turtle	snorkel
16	cologne, cologne water, eau de cologne	essence	bottle	perfume

Fig. 6 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 from Section 6.1 (right).

For entry-level category prediction in images, we would like to maximize both "naturalness" and estimates of image content. For example, text based "naturalness" will tell us that both *cat* and *dog* are good entry-level categories, but a confident visual prediction for *German shepherd* for an image tells us that *dog* 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 a similar behavior as in our language-based concept translations (Section 4.1). We can tune $\hat{\lambda}$ to control the degree of specificity while trying to preserve "naturalness" using n-gram counts. We compare our framework to the "hedging" technique of Deng et al. (2012) 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 7) that under different parameter settings we *consistently* obtain much higher levels of precision for predicting entry-level categories than hedging (Deng et al., 2012). We also obtain an additional gain in performance than in our previous work (Ordonez et al., 2013) by relying on the dataset-specific text-statistics of the
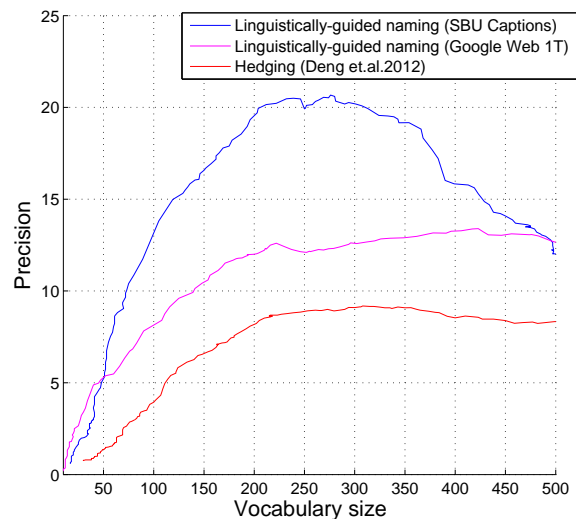
Fig. 7 Relationship between average precision agreement and working vocabulary size (on a set of 1000 images) for the hedging method (Deng et al., 2012) (red) and our linguistically-guided naming method that uses text statistics from the generic Google Web 1T dataset (magenta) and from the SBU Caption Dataset (Sec. 5.1). We use $K = 5$ to generate this plot and a random set of 1000 images from the SBU Captioned Dataset.

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 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 7,404 ImageNet leaf nodes we are missing models for many potentially important entry-level categories that are not at the leaf level.

As an 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 (Ordonez et al., 2011), which contains *1 million* captioned images. We transform this dataset into a set $D = \{X^{(j)}, Y^{(j)} \mid X^{(j)} \in \mathbf{X}, Y^{(j)} \in \mathbf{Y}\}$, where $\mathbf{X} = [0\text{–}1]^s$ is a vector of estimates of visual content for $s = 7,404$ 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 SVM content predictors based on convolutional network activation features described in Section 3. We run these SVM predictors over the whole image as opposed to the max-pooling approach over bounding boxes from our previous paper (Ordonez et al., 2013) so that we have a more uniform comparison to our linguistically-guided naming approach (Section 5.1) which does the same. There was some minor drop in performance when running our models exclusively on the whole image. Compared to our previous work, our visually-guided naming approach still has a significant gain from using the *ConvNet* features introduced in section 3.

For training our d target categories, we obtain labels Y from the million captions by running a POS-tagger (Bird, 2006) and defining $Y^{(j)} = \{y_{ij}\}$ such that:

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

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. 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 $d = 1,169$ of the best performing models. Our scoring function f_{svm} for a target concept v_i is then:

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

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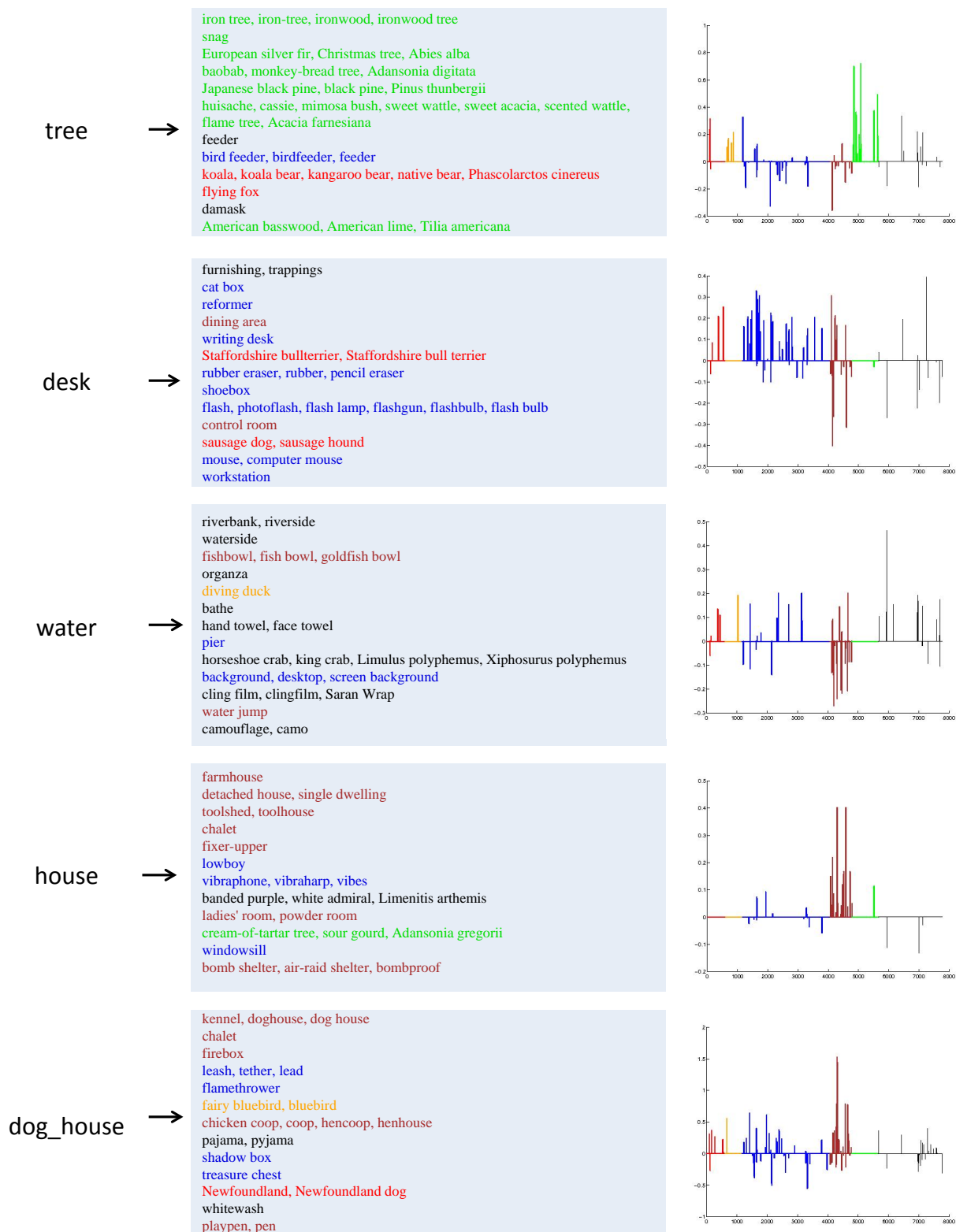
$$R(\theta_i) = \frac{1}{2}\|\theta_i\| + c \sum_{j=1}^{|D|} max(0, 1 - y_{ij}\theta_i^\top X^{(j)})^2. \quad (8)$$

We learn the parameters θ_i by minimizing the squared hinge-loss with ℓ_1 regularization (eqn 8). The latter provides a natural way of modeling the relationships between the input and output label spaces that encourages sparseness (examples in Figure 8). 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 a co-occurrence relationship with *bird* that may be useful for
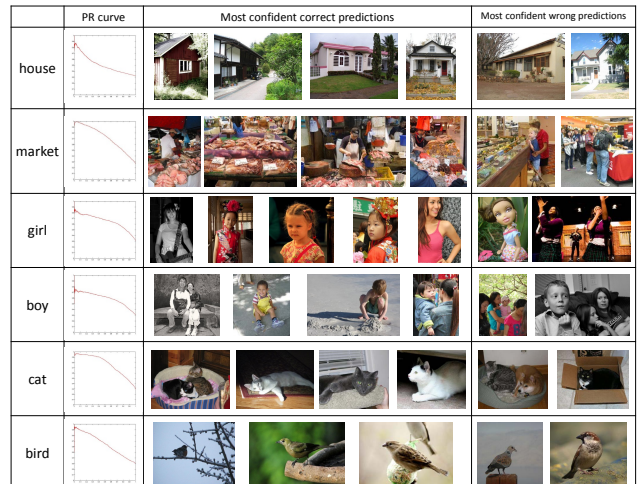
Fig. 9 Sample predictions from our experiments on a test set for each type of category. Note that image labels come from caption nouns, so some images marked as correct predictions might not depict the target concept whereas some images marked as wrong predictions might actually depict the target category.

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

desk →

furnishing, trappings
cat box
reformer
dining area
writing desk
Staffordshire bullterrier, Staffordshire bull terrier
rubber eraser, rubber, pencil eraser
shoebox
flash, photoflash, flash lamp, flashgun, flashbulb, flash bulb
control room
sausage dog, sausage hound
mouse, computer mouse
workstation

water →

riverbank, riverside
waterside
fishbowl, fish bowl, goldfish bowl
organza
diving duck
bathe
hand towel, face towel
pier
horseshoe crab, king crab, Limulus polyphemus, Xiphosurus polyphemus
background, desktop, screen background
cling film, clingfilm, Saran Wrap
water jump
camouflage, camo

house →

farmhouse
detached house, single dwelling
toolshed, toolhouse
chalet
fixer-upper
lowboy
vibraphone, vibraharp, vibes
banded purple, white admiral, Limenitis arthemis
ladies' room, powder room
cream-of-tartar tree, sour gourd, Adansonia gregorii
windowsill
bomb shelter, air-raid shelter, bombproof

dog_house →

kennel, doghouse, dog house
chalet
firebox
leash, tether, lead
flamethrower
fairy bluebird, bluebird
chicken coop, coop, hencoop, henhouse
pajama, pyjama
shadow box
treasure chest
Newfoundland, Newfoundland dog
whitewash
playpen, pen

Fig. 8 Entry-level categories with their 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), others(black).

prediction. A chair is often occluded by the objects sitting on the chair, but evidence of those types of objects, e.g. *people* or *cat* or co-occurring objects, e.g. *table* can help us predict the presence of a chair. See Figure 8 for some example learned relationships.

Given this large dataset of images with noisy visual predictions and text labels, we manage to learn quite good estimators of high-level content, even for categories with relatively high intra-class variation (e.g. girl, boy, market, house). We show some results of images with predicted output labels for a group of images in Figure 9.

6 Experimental Evaluation

We evaluate two results from our paper – models that learn general translations from encyclopedic concepts to entry-level concepts (Section 6.1) and models that predict entry-level concepts for images (Section 6.2). We additionally provide an extrinsic evaluation of our naming prediction methods by using them for a sentence retrieval application (Section 6.3).

6.1 Evaluating Translations

We obtain translations from ImageNet synsets to entry-level categories using Amazon Mechanical Turk (MTurk). In our experiments, users are presented with a 2x5 array of images sampled from an ImageNet synset, x_i, and asked to label 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 in agreement) 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 categories. We denote our resulting set of reference translations as: $D = \{(x_i, y_i)\}$, where each element pair corresponds to a translation from a leaf node x_i to an entry-level word y_i.

We show sample results from each of our methods to learn concept translations in Figure 6. In some cases language-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". Visual-based translation fails when it learns to weight context words highly, for example "snorkeling" $\rightarrow$ "water", or "African bee" $\rightarrow$ "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. Our visual-based translation benefits significantly from using *ConvNet* features (Section 3) compared to the 21% agreement that we previously reported in Ordonez et al. (2013). Note that our visual-based translation unlike our language-based translation does not use the WordNet semantic hierarchy to constrain the output categories to the set of inherited hypernyms of the input category.

This experiment expands on previous studies in psychology (Rosch, 1978; Jolicoeur et al., 1984). Readily available and inexpensive 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 prediction methods by evaluating how well we can predict nouns freely associated with images by users on Amazon Mechanical Turk. We initially selected two evaluation image sets. **Dataset A:** contains 1000 images selected at random from the million image dataset. **Dataset B:** contains 1000 images selected from images displaying high confidence in concept predictions. We additionally collected annotations for another 2000 images so that we can tune tradeoff parameters in our models. Both sets are completely disjoint from the sets of images used for learning. For each image, we instruct 3 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 of associated nouns – 3,610 distinct nouns total in our evaluation sets. This makes noun prediction extremely challenging!

For evaluation, we measure how well we can predict all nouns associated with an image by Turkers (Figure 10) and how well we can predict the nouns commonly associated by Turkers (assigned by at least 2 of 3 Turkers, Figure 11). For reference we compute the precision of one human annotator against the other two and found that on Dataset A humans were able to predict what the previous annotators labeled with 0.35 precision and with 0.45 precision for Dataset B.

Results show precision and recall for prediction on each of our Datasets, comparing: leaf node classification performance (flat classifier), the outputs of hedging (Deng et al., 2012), and our proposed entry-level category predictors (linguistically guided naming (Section 5.1) and visually guided naming (Section 5.2)). Qualitative examples for Dataset A are shown in Figure 13 and for Dataset B in Figure 14. Performance at this task on Dataset B is in general better than performance on Dataset A. This is unsurprising since
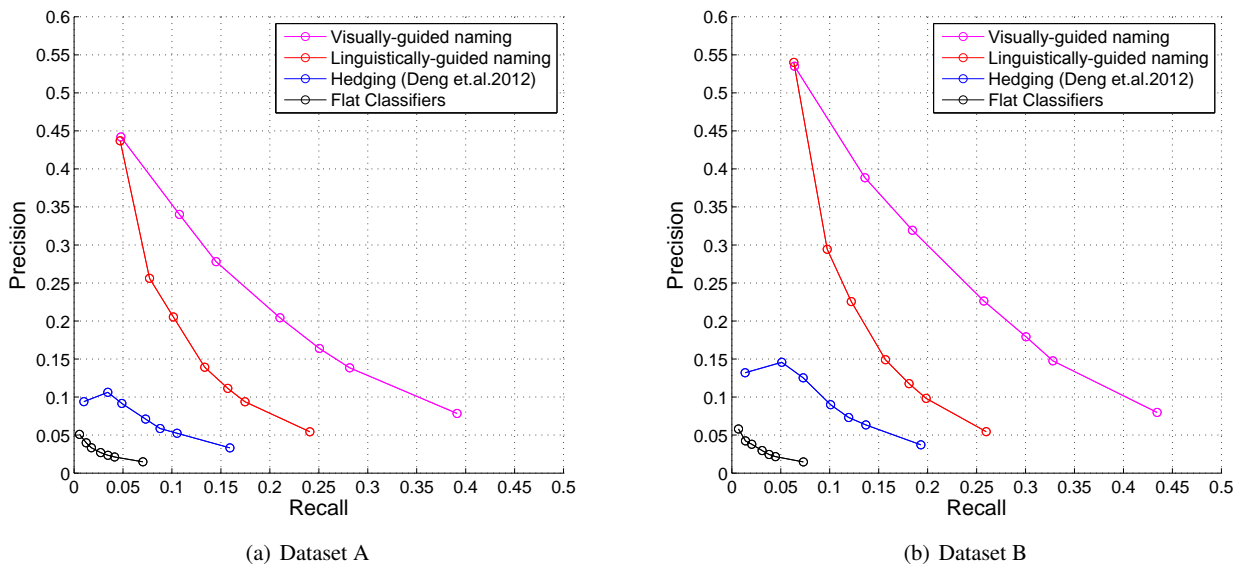
Fig. 10 Precision-recall curves for different entry-level prediction methods when using the top K categorical predictions for $K = 1, 3, 5, 10, 15, 20, 50$. The ground truth is the union of labels from all users for each image.
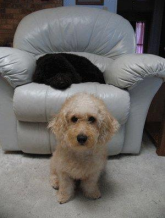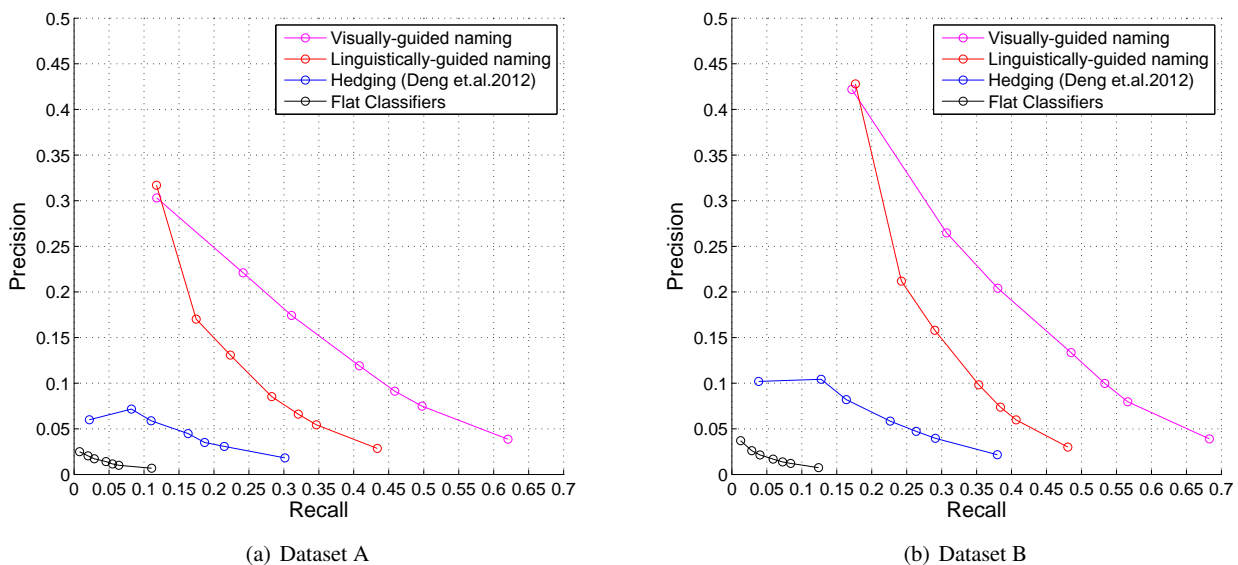

Fig. 11 Precision-recall curves for different entry-level prediction methods when using the top K categorical predictions for $K = 1, 3, 5, 10, 15, 20, 50$. The ground truth is the set of labels where at least two users agreed.

Dataset B contains images which have confident classifier scores. Surprisingly their difference in performance is not extreme and performance on both sets is admirable for this challenging task. When compared to our previous work (Ordonez et al., 2013) that relies on SIFT + LLC features, we found that the inclusion of *ConvNet* features provided a significant improvement in the performance for the visually-guided naming predictions but it did not improve the results using the WordNet semantic hierarchy for both Hedg-

ing (Deng et al., 2012) and our linguistically-guided naming method.

On the two datasets we find the visually-guided naming model to perform better (Section 5.2) than the linguistically-guided naming prediction (Section 5.1). In addition, we outperform both leaf node classification and the hedging technique (Deng et al., 2012).

We additionally collected a third test set Dataset C consisting of random ImageNet images belonging to the 7,404

Method	Images	Original Caption	Top 5 Retrieved Sentences
Visually-guided Naming		(808) "dining area in great room open to kitchen opens to seat 8 people"	(1) [table area beside kitchen] (2) [work table sitting area in separate room bathroom kitchen area sleeping area] (3) [dining table in kitchen area] (4) [by the kitchen table area] (5) [dining room table in kitchen]
Visually-guided Naming		(1105) "fresh snow on pine trees in yosemite national park"	(1) [pine trees forest under snow] (2) [pine tree in snow] (3) [pine tree in snow] (4) [snow in pine tree] (5) [pine tree in snow]
Visually-guided Naming		(60747) "theres no room in the chair for me so i am sitting in daddys spot on the floor"	(1) [dog and cat in chair] (2) [dog and cat in chair] (3) [bear in a chair poor chair bear] (4) [dog in cat] (5) [cat in chair]
Linguistically-guided Naming		(519) "cat in the box"	(1) [cat in box cat on box] (2) [cat in the cat box] (3) [obligatory cat in box picture] (4) [cat in cats] (5) [cat in box upside down cat]
Linguistically-guided Naming		(37153) "we were wondering where you could sail a boat in colorado we passed this boat about 4 times"	(1) [car under boat] (2) [car in truck] (3) [car in car mirror] (4) [portable car toy box in cars and trucks] (5) [car in car mirror bw]

Fig. 12 Good examples of retrieved sentences describing image content. We show the original sentence for each image with its corresponding rank in parenthesis. We also show the top 5 retrieved sentences for each image. We are showing here only images that ranked highly the original caption (within the top 10%) .

Method	Precision $K = 1, 2, 3$	Recall $K = 1, 2, 3$
Flat classifier	4.40, 4.00, 3.43	2.10, 3.82, 4.87
Hedging	9.00, 9.55, 10.25	4.90, 11.72, 19.64
Linguist.-guided	26.70, 16.15, 12.90	17.59, 19.52, 22.25
Visually-guided	25.80, 17.95, 13.73	17.50, 22.76, 25.73

Table 1 Here we show results on Dataset C which consists of images from ImageNet. The human labels for each image are the union of the labels collected from different Mechanical Turk users.

categories represented in our leaf node classifiers. We make sure not to include those images in the training of our leaf node classifiers. These images are more object-centric, often displaying a single object. This resulted in a smaller number

of unique labels provided by users for each image with an average of 2 unique labels per image. We report the precision and recall at $K = 1, 2, 3$ for all of our methods in this dataset in Table 1. We observe that at $K = 1$ there is a small advantage of our linguistically-guided naming method compared to the visually-guided naming approach. Both methods surpass the flat mapping classifiers and the Hedging approach. In this different dataset the entry-level category predictors using our visually-guided naming approach still offer better performance than the linguistically-guided naming approach at $K = 2, 3$. Note that our linguistically-guided naming does not require expensive retraining of visual models like our visually-guided naming. Also, the gap between

Method	Dataset A		Dataset B	
	Top 1%	Top 10%	Top 1%	Top 10%
Flat classifier	40	80	48	93
Hedging	62	172	92	266
Linguistically-guided	71	310	104	416
Visually-guided	162	516	210	617

Table 2 Here we show the number of images (for each dataset and method) for which we could retrieve its original image description within the top 1% and the top 10%. Note that each dataset has 1000 images in total.

our two naming approaches is smaller than in the previous experiments on Datasets A and B.

6.3 Evaluating Image Entry-Level Predictions for Sentence Retrieval

Entry-level categories are also the natural categories that people use in casual language. We evaluate our produced naming predictions indirectly by using them to retrieve image descriptions. Our sentence retrieval approach works as follows: We predict entry-level categories with $K = 5$ and use them as keywords to retrieve a ranked list of sentences from the entire 1 million image descriptions in the *SBU Captioned Dataset*. We use cosine similarity on a bag-of-words model for representation and ranking.

The images in our test Dataset A and Dataset B in the previous section come from the *SBU Captioned Dataset* and therefore already have one image description associated with each of them. This image description was written by the owner of each picture. Note that these "ground truth" image descriptions for each of our test images are included in the pool of 1 million captions. We use the rank of the ground truth image description for each image as a measure of performance in this task. We report on Table 2 the number of images for which its "ground truth" description was ranked within the top 1% and the top 10% for the various methods compared in our paper and for each test set. Although our evaluation uses a rough metric of performance, we observed that the top 5 sentences retrieved for images that had its original sentence ranked within the top 1% were also often very good descriptions for the query image. We show some qualitative examples in Figure 12.

7 Conclusion

Results indicate that our inferred concept translations are meaningful and that our models are able to predict entry-level categories—the words people use to describe image content—for images. Our models managed to leverage a large scale visual categorization system to make new types of predictions. These methods could apply to a wide range of end-user applications that require recognition outputs to be use-

ful for human consumption, including tasks related to description generation and retrieval. We presented an initial experiment on this direction for image description using a sentence retrieval approach.

References

Kobus Barnard and Keiji Yanai. Mutual information of words and pictures. *Information Theory and Applications*, 2006.

Steven Bird. Nltk: the natural language toolkit. In *COLING/ACL*, 2006.

Thorsten Brants and Alex Franz. Web 1t 5-gram version 1. In *Linguistic Data Consortium*, 2006.

Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. NEIL: Extracting Visual Knowledge from Web Data. In *ICCV*, 2013.

Thomas Dean, Mark A Ruzon, Mark Segal, Jonathon Shlens, Sudheendra Vijayanarasimhan, and Jay Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. In *CVPR*, 2013.

Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

Jia Deng, Alexander C. Berg, Kai Li, and Fei-Fei Li. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010.

Jia Deng, Jonathan Krause, Alexander C. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *CVPR*, 2012.

Santosh Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: generating sentences for images. In *ECCV*, 2010.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 2007.

Images	Labels	Flat Classifier	Hedging [Deng et.al.2012]	Linguistically-guided Naming	Visually-guided Naming
	bird, duck feather fin, fur goose, lake pond, swan water, wing	pen cob whooper cygnet Cygnus	**swan** aquatic **bird** anseriform waterfowl	**swan** **bird** snow **duck** pen	**swan** **duck** **pond** **bird** **water**
	boat, crowd flag, harbor lake, ocean people, wave race, sail ship, warf	drill container **harbor** clipper seaside	vessel craft transport vehicle **ship**	**ship** tree coast shore **boat**	**boat** **ship** beach **sail** **harbor**
	car, home house, land power line road, sky street, tree truck, wire	secondhand chair golf power **car**	transport wheel structure self-propelled motor	**tree** building **car** **house** tower	bus **street** **car** cable car **road**
	bag, basket bike, boy, hat man, person sidewalk, jacket stone, street tire, tree	pannier dirt ice skateboard push-bike	wheel container vehicle transport cover	**bike** **bag** **basket** dog building	**bike** mountain bike seat **boy** girl
	big ben building clock, roof sky, street light tower wall	jigsaw integrate chatelaine turret masjid	structure **building** **tower** circuit house	**building** **tower** house home tree	clock tower **building** **tower** castle church
	bathroom cabinet doorway faucet mirror, sink towel, vanity	console credence armoire Murphy **vanity**	furniture furnish room area box	box room area table **cabinet**	**bathroom** **sink** **cabinet** room floor
	fence, junk sign stop sign street sign trash can tree	jigsaw gift trophy display comic	outlet place establishment store structure	store place building box window	market shop bar street book
	circle earring hook jewel jewelry make up stone	skeleton clasp toggle pull corkscrew	constraint fix implement device chain	chain bottle bit tree flower	bead silver chain sterling glass

Results in the top 25% (left margin, upper section)

Results in the bottom 25% (left margin, lower section)

Fig. 13 Example translations on Dataset A (random images). 1st col shows images. 2nd col shows MTurk associated nouns. These represent the ground truth annotations (entry-level categories) we would like to predict (colored in blue). 3rd col shows predicted nouns using a standard multi-class flat-classifier. 4th col shows nouns predicted by the method of (Deng et al., 2012). 5th col shows our n-gram based method predictions. 6th col shows our SVM mapping predictions and finally the 7th column shows the labels predicted by our joint model. Matches are colored in green. Figures 10,11 show the measured improvements in recall and precision.

Images	Labels	Flat Classifier	Hedging [Deng et.al.2012]	Linguistically-guided Naming	Visually-guided Naming

Results in the top 25%

building, bush, field, fountain, grass, home, house, window, manor, sky, tree, yard, white house

summer	**home**	**building**	**house**
farmhouse	**building**	**house**	barn
background	**house**	**home**	**field**
detach	housing	**tree**	hill
tombstone	structure	country	**home**

dirt, flower, grass, leaf, petal, plant, pot, rain, rise, rose, stem, white

cauliflower	vegetable	**flower**	**flower**
terrarium	solid	dog	**plant**
gypsophilum	food	tree	**rose**
West	produce	fruit	**grass**
mash	matter	**white**	**pot**

beach, beach sand, bridge, cloud, coast, grass, water, man, ocean, weed, sand, shirt, shorts, structure

seaside	formation	**bridge**	**beach**
oceanfront	shore	shore	**sand**
strand	elevation	**water**	boat
sand	psychological	**coast**	**bridge**
waterside	event	side	**water**

blue dress, bush, dress, girl, child, grass, plant, sky, tree

frame	wear	**dress**	**grass**
tudung	good	woman	shirt
Frisbee	consumer	**tree**	**dress**
raglan	cover	**dress**	**girl**
skirt	garment	shirt	field

animal, barn, brown, building, cabin, dirt, dog, farm, field, grass, meadow, shack, shed, tree, turkey

barnyard	housing	**building**	**barn**
corncrib	home	house	**field**
farmhouse	structure	home	horse
sod	**building**	**tree**	**farm**
frame	house	area	truck

brick, building, door, flower, market, product, sign, table, window

window	outlet	**flower**	**market**
jigsaw	place	store	**flower**
florist	establishment	place	fruit
gift	store	**market**	pot
afghan	structure	tree	street

Results in the bottom 25%

architecture, bench, dome, fence, field, grass, sky, stage, structure, tent, tree

geodesic	**structure**	building	water tower
planetarium	building	roof	tower
mosque	protection	bridge	bridge
dome	**dome**	tower	building
observation	roof	**dome**	background

beam, chair, chandelier, gathering, wine, glass, indoor, light, napkin, party, people, silverware, suit, table

control	structure	room	bar
conference	room	building	pizza
game	area	area	shirt
war	building	restaurant	**table**
conference	restaurant	store	office

Fig. 14 Example translations on Dataset B (images with high response to visual models). 1st col shows images. 2nd col shows MTurk associated nouns. These represent the ground truth annotations (entry-level categories) we would like to predict (colored in blue). 3rd col shows predicted nouns using a standard multi-class flat-classifier. 4th col shows nouns predicted by the method of (Deng et al., 2012). 5th col shows our n-gram based method predictions. 6th col shows our SVM mapping predictions and finally the 7th column shows the labels predicted by our joint model. Matches are colored in green. Figures 10,11 show the measured improvements in recall and precision.

C. Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010.

Song Feng, Sujith Ravi, Ravi Kumar, Polina Kuznetsova, Wei Liu, Alexander C Berg, Tamara L Berg, and Yejin Choi. Refer-to-as relations as semantic knowledge. In *AAAI*, 2015.

Ankush Gupta, Yashaswi Verma, and CV Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, 2012.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899, May 2013.

Yangqing Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.

Pierre Jolicoeur, Mark A Gluck, and Stephen M Kosslyn. Pictures and names: making the connection. *Cognitive Psychology*, 16:243–275, 1984.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *TPAMI*, 2013.

Polina Kuznetsova, Vicente Ordonez, Alex Berg, Tamara L Berg, and Yejin Choi. Collective generation of natural image descriptions. In *ACL*, 2012.

Polina Kuznetsova, Vicente Ordonez, Tamara Berg, and Yejin Choi. Treetalk: Composition and compression of trees for image descriptions. *TACL*, 2014.

Quoc V Le, Rajat Monga, Matthieu Devin, Kai Chen, Greg S Corrado, Jeff Dean, and Andrew Y Ng. Building high-level features using large scale unsupervised learning. In *ICML*, 2012.

Rebecca Mason and Eugene Charniak. Nonparametric method for data-driven image captioning. ACL, 2014.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *EACL*, 2012.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara L Berg. From large scale image categorization to entry-level categories. In *ICCV*, 2013.

Florent Perronnin, Zeynep Akata, Zaid Harchaoui, and Cordelia Schmid. Towards good practice in large-scale learning for image classification. In *CVPR*, 2012.

John C Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, 1999.

K. Ramnath, S. Baker, L. Vanderwende, M. El-Saban, S.N. Sinha, A. Kannan, N. Hassan, M. Galley, Yi Yang, D. Ramanan, A. Bergamo, and L. Torresani. Autocaption: Automatic caption generation for personal photos. In *WACV*, 2014.

Eleanor Rosch. Principles of categorization. *Cognition and Categorization*, pages 27–48, 1978.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2014.

Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 2008.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv e-prints*, September 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. *ArXiv e-prints*, September 2014.

Antonio Torralba, Robert Fergus, and William T Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *TPAMI*, 2008.

Jianxiong Xiao, James Hays, Krista Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large scale scene recognition from abbey to zoo. In *CVPR*, 2010.

Keiji Yanai and Kobus Barnard. Probabilistic web image gathering. In *MIR*. ACM, 2005.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011.