From Large Scale Image Categorization to Entry-Level Categories

Vicente Ordonez[1], Jia Deng[2], Yejin Choi[3], Alexander C. Berg[1], Tamara L. Berg[1]
[1]University of North Carolina at Chapel Hill, [2]Stanford University, [3]Stony Brook University
[vicente,aberg,tlberg]@cs.unc.edu, jdeng@stanford.edu, ychoi@cs.stonybrook.edu

Abstract

Entry level categories – the labels people will use to name an object – were originally defined and studied by psychologists in the 1980s. In this paper we study entry-level categories at a large scale and learn the first models for predicting entry-level categories for images. Our models combine visual recognition predictions with proxies for word "naturalness" mined from the enormous amounts of text on the web. We demonstrate the usefulness of our models for predicting nouns (entry-level words) associated with images by people. We also learn mappings between concepts predicted by existing visual recognition systems and entry-level concepts that could be useful for improving human-focused applications such as natural language image description or retrieval.

1. Introduction

Computational visual recognition is beginning to work. Although far from solved, algorithms have now advanced to the point where they can recognize or localize thousands of object categories with reasonable accuracy [17, 4, 3, 12]. 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 *grampus griseus*, while most people are more likely to simply say *dolphin*.

This is closely related to ideas of *basic and entry level categories* formulated by psychologists such as Eleanor Rosch [18] and Stephan Kosslyn [11]. 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). More generally such questions are very relevant to recent work on the connection between computer vision outputs and (generating) natural language descriptions of images [8, 13, 16, 14].

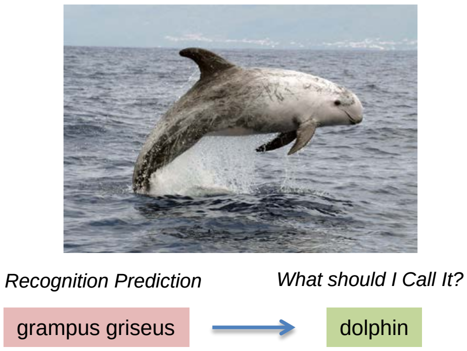


Figure 1. Example translation between a WordNet based object category prediction and what people might call the depicted object.

In this paper we consider two related problems 1) learning a mapping from specific categories – *e.g.*, leaf nodes in WordNet [9] – to what people are likely to call them 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 the image.

Our proposed methods take into account several sources of structure and information: the structure of WordNet, frequencies of word use from Google n-grams, outputs of a large-scale visual recognition system, and large amounts of paired image and text data. In particular, we make use of the SBU Captioned Photo Dataset [16], which consists of 1 million images with natural language captions, as a source of natural image naming patterns. Taken together, we are able to study patterns for choice of basic level categories at a much larger scale than previous psychology experiments.

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 hierarchy. One key difference is that our approach allows a reward function over the WordNet hierarchy that is not monotonic along paths from the root to leaves. This allows reward based on factors including frequency of word use that are not monotonic along such paths in WordNet. This 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.* [6] because although optimal in some abstract 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 always "starting from scratch" in developing new applications. It is wasteful not to take advantage of the CPU weeks [10, 12], months [3, 6], or even millennia [15] invested in developing recognition models for increasingly large labeled datasets [7, 19, 22, 5, 20]. 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).

Evaluations show that our models can effectively emulate the naming schemes 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 the raw noisy vision estimates or the results of using the state of the art *hedging* system from Deng *et al.* [6].

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 — has been discussed in *Prototype Theory* [18].
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.

1.2. Paper Overview

Our paper is divided as follows. In section 2 we run experiments to gather entry-level category labels directly from people. In section 3 we learn translations between leaf node concepts and entry-level concepts. In section 4 we propose two models and a joint model that can take an image as input and predict entry-level concepts. Finally, in section 5 we provide experimental evaluations.

2. Obtaining Natural Categories from Humans

We use Amazon Mechanical Turk to crowd source translations of ImageNet synsets into entry-level categories $D = \{x_i, y_i \mid x_i \text{ is a leaf node, } y_i \text{ is a word}\}$. Our experiments present users with a 2x5 array of images sampled from an ImageNet synset, x_i, and users are 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 preference for particular entry-level categories.

This experiment expands on previous studies in psychology [18, 11]. Cheap and easy online crowdsourcing enables us to gather these labels for a much larger set of (500) concepts than previous experiments. Furthermore, we use the results of our experiments to automatically learn generalizations to a substantially larger set of ImageNet synsets in section 3.

3. Translating Encyclopedic Concepts to Entry-Level Concepts

Our objective in this section is to discover mappings between encyclopedic concepts (ImageNet leaf categories, e.g. Chlorophyllum molybdites) to output concepts that are more *natural* (e.g. mushroom). In section 3.1 we present an approach that relies on the wordnet hierarchy and frequency of words in a web scale corpus. In section 3.2 we follow an approach that uses visual recognition models learned on a paired image-caption dataset.

3.1. Language-Only Translation

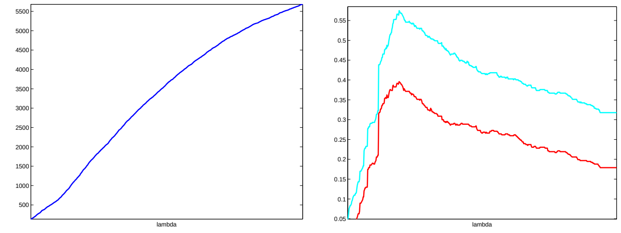
For comparison purposes, we first consider a translation approach that relies only on language-based information. We hypothesize that the frequency of terms computed from massive amounts of text on the web reflects the "naturalness" of concepts. We use the n-gram counts of the Google 1T corpus [2] as a proxy for term "naturalness". Specifically, for a synset w, we quantify "naturalness" as the maximum log count $\phi(w)$ of all of the terms in the synset.

To control the degree of naturalness, we constrain the translation using the hyponym/hypernym structure of WordNet. More specifically, we define $\psi(w, v)$ as a function that measures the distance between leaf node v and node w in the hypernym structure. Then the translation function $\tau(v, \lambda) : V \mapsto W$ maps a leaf node v to a target node w by maximizing a trade-off between naturalness and semantic proximity.

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

$\Pi(v)$ is the set of (inherited) hypernyms including v. We find the optimal λ based on a sub-set of translation pairs $D = (x_i, y_i)$ collected using MTurk (section 2).

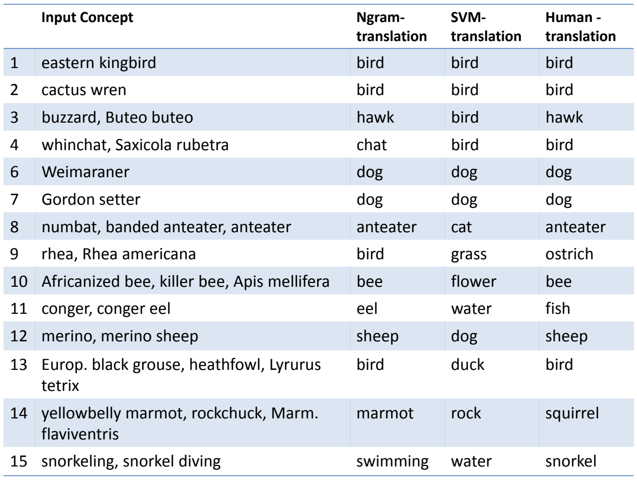
$$\Phi(D, \lambda) = \sum_{i} 1[word(\tau(x_i, \lambda)) = y_i] \qquad (2)$$



(a) vocabulary size vs. λ (b) naturalness vs. λ

Figure 2. **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).

We show the relationship between λ and the size of the output vocabulary $|W|$ on the left side of Fig. 2 and the relationship between λ and $\Phi(D, \lambda)$ on the right side. The size of the output vocabulary increases monotonically with λ. At a high level, increasing λ serves to encourage mappings to be close to the input node in the WordNet hierarchy, thereby increasing the vocabulary size and limiting the generalization of concepts. Conversely, "naturalness", $\Phi(D, \lambda)$, increases initially and then decreases as too much specificity or generalization hurts the naturalness of the outputs. For example, generalizing from "grampus griseus" to "dolphin" is good for "naturalness", but generalizing all the way to "entity" decreases "naturalness". In Figure 2 the red line shows accuracy for predicting the most agreed upon word for a synset, while the cyan line shows the accuracy for predicting any word collected from any user.



	Input Concept	Ngram-translation	SVM-translation	Human - translation
1	eastern kingbird	bird	bird	bird
2	cactus wren	bird	bird	bird
3	buzzard, Buteo buteo	hawk	bird	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	cat	anteater
9	rhea, Rhea americana	bird	grass	ostrich
10	Africanized bee, killer bee, Apis mellifera	bee	flower	bee
11	conger, conger eel	eel	water	fish
12	merino, merino sheep	sheep	dog	sheep
13	Europ. black grouse, heathfowl, Lyrurus tetrix	bird	duck	bird
14	yellowbelly marmot, rockchuck, Marm. flaviventris	marmot	rock	squirrel
15	snorkeling, snorkel diving	swimming	water	snorkel

Figure 3. Translations from ImageNet leaf node synset categories to *entry level categories* using our automatic approaches from sections 3.1 (Ngram-) and 3.2 (SVM-) and crowd-sourced human annotations from section 2 (Human-).

3.2. Visually-Informed Translation

In this approach, for a given leaf synset v we sample a set of $n = 100$ images $s = \{I_1, I_2, ..., I_n\}$ and each image is automatically annotated with nouns $N_i = \{n_{i1}, n_{i2}, ..., n_{im}\}$ using the models learned in section 4.2. We use the 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 where we consider each category v in our experimental setting as a document for the *inverse document frequency* term. We pick the most relevant keyword for each node v as the entry-level categorical translation.

4. Predicting Entry-Level Concepts for Images

Our objective in this section is to explore approaches that can take an image as input and predict its entry-level labels. The models we propose are: 1) a method that combines "naturalness" measures computed from the web with direct estimates of visual content computed at leaf nodes and inferred for internal nodes (section 4.1), 2) a method that learns models for entry-level recognition from a large collection of images with associated captions (section 4.2), and 3) a joint method combining the two approaches (section 4.3).

4.1. Prediction using Propagated Visual Estimates

As our first method for predicting entry level categories for an image, we present a variation on the hedging approach [6]. In the hedging work, the output is the node with the maximum expected reward, where the reward is monotonic in the hierarchy and has been smoothed by adding a carefully chosen constant to the reward for all nodes. In our modification, we construct a non-monotonic reward γ based on naturalness and a smoothing offset that is scaled by the position in the hierarchy.

The image content for an image, I, is estimated using trained models from [6]. These models predict presence or absence of 7404 leaf node concepts in the ImageNet hierarchy. Following the approach of [6], we compute estimates of visual content for internal nodes by hierarchically accumulating all predictions below a node:[1]

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

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 for the category corresponding to input leaf node v. Each linear SVM is trained on sift features with locally-constrained linear coding and spatial pooling on a regular 3x3 grid. Following our approach from section 3.1, we define for every node in the ImageNet hierarchy a trade-off function between "naturalness" (ngram counts) and specificity (relative position in the wordnet hierarchy):

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

Where $\tilde{\psi}(v) = \max_{w \in Z(v)} \psi(v, w)$ measures the max height over $Z(v)$, the set of leaf nodes under v. We parameterize this trade-off by $\hat{\lambda}$.

For entry-level category prediction on images, we would like to maximize both "naturalness" and content estimates. 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 5 Wordnet synsets according to:

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

$$f_{nat}(v, I, \hat{\lambda}) = f(v, I)[\phi(v) - \hat{\lambda}\tilde{\psi}(v)] \tag{6}$$

As we change $\hat{\lambda}$ we expect similar behavior to our web based concept translations (section 3.1). Again, 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 hedging [6] for different settings of $\hat{\lambda}$. For a side by side comparison we modify hedging to output the top 5 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 4) that under different parameter settings we *consistently* obtain much higher levels of precision for predicting entry-level categories than hedging [6].

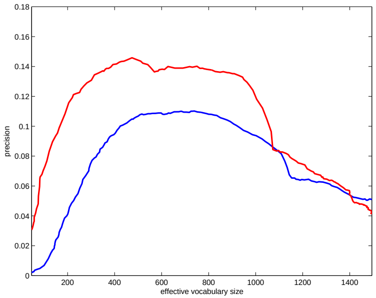


Figure 4. Relationship between average precision agreement and working vocabulary size (on a set of 1000 images) for the hedging method [6] (blue) and our direct translation method (red).

4.2. Prediction using Supervised Learning

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 7404 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 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 large scale dataset of [16], containing *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 an input space of estimates of visual content for $S = 7404$ ImageNet leaf node categories and $\mathbf{Y} = [0, 1]^D$ is a set of binary output labels for D target categories.

For estimating the presence of objects from our set of 7404 ImageNet leaf node categories we use the same models as the previous section with one additional consideration. We run the classifiers on a set of bounding boxes $B = \{b_k\}$ for each training image using the window selection method of [21]. We then aggregate the results across multiple bounding boxes by max pooling of visual concepts scores. So the feature descriptor for an image $I^{(i)}$ is:

$$X^{(i)} = \{x_j^{(i)} \mid x_j^{(i)} = max(\hat{f}(v_j, I^{(i)}, b_k))\} \tag{7}$$

Where $\hat{f}(v_j, I^{(i)}, b_k)$ is the output score for the presence

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

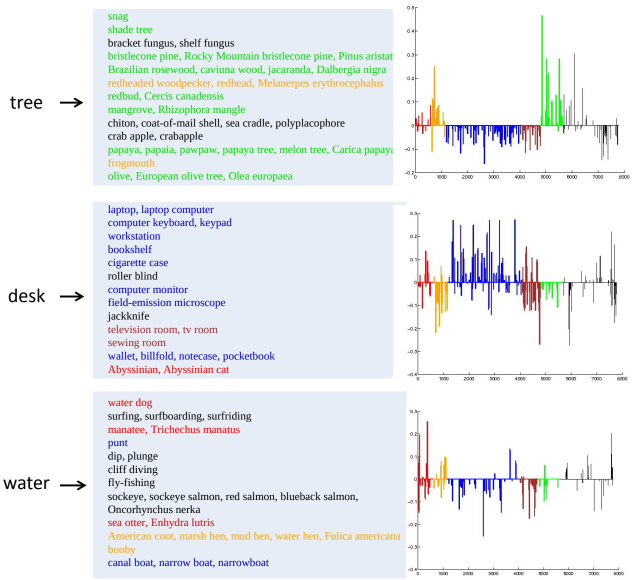


Figure 5. Entry-level categories with their corresponding top weighted leaf node features after training an SVM on our noisy data and a visualization of weights grouped by an arbitrary categorization of leaf nodes. vegetation(green), birds(orange), instruments(blue), structures(brown), mammals(red), others(black).

of the visual concept represented by the leaf node v_j and bounding box b_k on image $I^{(i)}$.

For training our D target categories, we obtain labels Y from the million captions by running a POS-tagger [1] and defining $Y_i = [y_j \mid \text{image } i \text{ has noun } j]$. The POS-tagger helps cleans up some word sense ambiguity due to polysemy. $|D|$ is determined experimentally from data by learning models for the most frequent words 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 10000 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 subject to platt scaling for each of our target concepts. We keep 800 of the best performing models. Our combined scoring prediction function is then (note that the operations here are pointwise operators):

$$F_{svm}(I, \Theta) = [f_{svm}(v_i, I, \theta_i)] \tag{8}$$

$$F_{svm}(I, \Theta) = \frac{1}{1 - exp(a\Theta^\top X + b)} \tag{9}$$

$$R(\theta_i) = \frac{1}{2}\|\theta_i\| + c\sum_{j=1}^{|D|} max(0, 1 - y_i^{(j)}\theta_i^\top X^{(j)})^2 \tag{10}$$

We minimize the squared hinge-loss with ℓ_1 regularization (eqn 10). The latter provides a natural way of modeling the relationships between the input and output label spaces that encourages sparseness[2]. See examples in Figure 5. Since we learn each linear SVM independently, θ_i represents a row in the joint matrix Θ. We fit Platt scaling parameters $a = [a_i]$ and $b = [b_i]$ for each target label i on a held out validation set.

One of the drawbacks of using the ImageNet hierarchy to aggregate estimates of visual concepts (section 3) 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 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 5 for some example learned relationships.

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). We show some results of images with predicted output labels for a group of images in Figure 6.

4.3. Joint Prediction

Finally, we explore methods to combine our two approaches from section 4.1 and section 4.2. We start by associating the SVM based scores f_{svm} (section 4.2) to synsets in the ImageNet hierarchy. Here we map words from our

[2] We find $c = 0.01$ to yield good results for our problem and use this value for training all individual models.

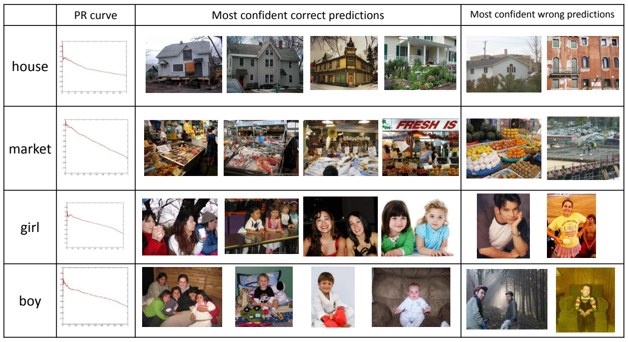

PR curve
Most confident correct predictions
Most confident wrong predictions
house
market
girl
boy
FRESH IS

Figure 6. 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.

Method	Dataset A Precision	 Recall	 N+	Dataset B Precision	 Recall	 N+
Flat classifier	1.85 ± 0.45	0.92 ± 0.24	1635	2.63 ± 0.58	1.41 ± 0.32	1652
Hedging [6]	10.21 ± 1.10	5.44 ± 0.67	705	13.26 ± 1.46	7.55 ± 0.73	823
Ngram-biased Mapping	14.20 ± 1.28	7.60 ± 0.81	447	17.59 ± 1.36	10.11 ± 1.01	576
SVM Mapping	19.13 ± 1.91	9.95 ± 1.04	207	24.17 ± 2.63	14.27 ± 1.48	244
Ngram-biased + SVM	19.87 ± 1.21	10.44 ± 0.69	336	25.08 ± 2.37	14.42 ± 1.35	415

Table 1. Performance at predicting the union of labels provided by 3 Turkers on dataset A (random images) and Dataset B (images with high confidence scores). Precision/Recall are computed per image and averaged across each dataset, computed over 10 splits.

Method	Dataset A Precision	 Recall	 N+	Dataset B Precision	 Recall	 N+
Flat classifier	0.95 ± 0.40	1.67 ± 0.89	1635	1.42 ± 0.43	2.37 ± 0.96	1652
Hedging [6]	6.28 ± 1.01	10.92 ± 1.86	705	8.96 ± 0.96	16.96 ± 2.44	823
Ngram-biased Mapping	9.06 ± 1.47	16.35 ± 2.96	447	11.66 ± 1.18	22.01 ± 2.79	576
SVM Mapping	11.85 ± 1.55	20.23 ± 2.24	207	15.93 ± 2.05	30.25 ± 3.91	244
Ngram-biased + SVM	12.68 ± 1.49	21.96 ± 2.77	336	16.95 ± 1.83	31.52 ± 3.76	415

Table 2. Performance at predicting the labels agreed upon by 2 (of 3) Turkers on dataset A (random images) and Dataset B (images with high confidence scores). Precision/Recall are computed per image and averaged across each dataset, computed over 10 splits.

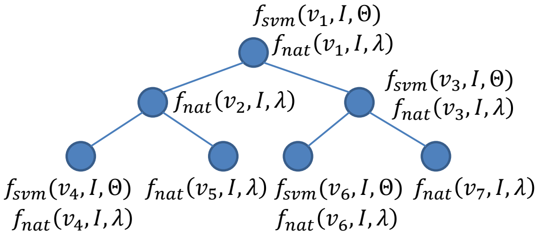

$f_{svm}(v_1, I, \Theta)$
$f_{nat}(v_1, I, \lambda)$
$f_{nat}(v_2, I, \lambda)$
$f_{svm}(v_3, I, \Theta)$
$f_{nat}(v_3, I, \lambda)$
$f_{svm}(v_4, I, \Theta)$
$f_{nat}(v_4, I, \lambda)$
$f_{nat}(v_5, I, \lambda)$
$f_{svm}(v_6, I, \Theta)$
$f_{nat}(v_6, I, \lambda)$
$f_{nat}(v_7, I, \lambda)$

Figure 7. For every node in the tree we have estimates of visual content coming from two sources a) naturalness and hierarchical aggregation f_{nat} and b) supervised learning f_{svm}.

target nouns D to the best matching synset concept. For each synset, v, we also associate its direct translation score, $f_{nat}(v, I, \hat{\lambda})$ (section 4.1), illustrated in Fig 7. This means that for all WordNet synsets we have a direct translation score, and for some synsets we have a mapped SVM score $f_{svm}(v, I, \theta_v)$ (for nodes not appearing in D we set this score to be zero). Likewise the SVM scoring function introduces some new concepts not present in the WordNet hierarchy that have a value of zero for $f_{nat}(v, I, \hat{\lambda})$. We redefine the domain of our scoring function (eqns 11 and 12) and use a parameter α to control for tradeoff between the two models (13).

$$\widetilde{f}_{nat}(v) = \begin{cases} f_{nat}(v, I, \hat{\lambda}), & \text{if } v \in dom(f_{nat}) \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

$$\widetilde{f}_{svm}(v) = \begin{cases} f_{svm}(v, I, \theta_v), & \text{if } v \in dom(f_{svm}) \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

$$f_{joint}(v, \alpha) = \alpha \widetilde{f}_{nat}(v) + \widetilde{f}_{svm}(v) \quad (13)$$

We define our prediction function that associates a set of ImageNet nodes $v_1, v_2, ..., v_n$ to an input image based on the joint scores (13) as follows:

$$\hat{v_1}, \hat{v_2}, ..., \hat{v_n} = \underset{\hat{v_1}, \hat{v_2}, ..., \hat{v_n}}{\operatorname{argmax}} \sum_i f_{joint}(v_i, \alpha) \quad (14)$$

This means we can select the set of n nodes that have the highest scores. We use $n = 5$ and find α that minimizes the error on the average annotation F1 score per image on a training set of 1000 images with human labels.

5. Experimental Evaluation

We evaluate learning general translations from encyclopedic to entry-level concepts (section 5.1) and predicting entry-level concepts for images (section 5.2).

5.1. Evaluating Translations

We show sample results from each of our methods to learn concept translations in Figure 3 (more are included in the supplemental material). In some cases Ngram-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”. SVM 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 IDF. 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, ngram translation agrees 37% of the time with human supplied translations and the SVM translation agrees 21% of the time, indicating that translation learning is non-trivial.

5.2. 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 MTurk. We select 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 predic-

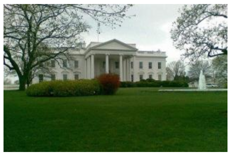
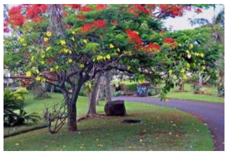
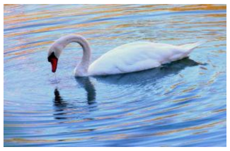
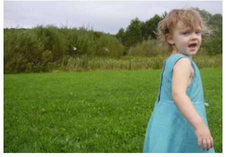
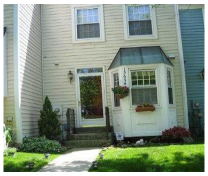
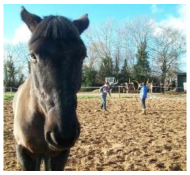
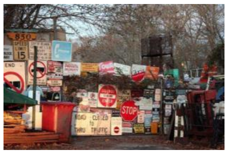
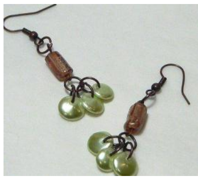
	Images	Labels	Flat Classifier	Hedging [6]	Ngram-based	SVM-based	Joint
Results in the top 25%		building bush, field fountain grass, home house, window manor, sky tree, yard white house	farmhouse stately ranch courthouse **manor**	**house** **home** **building** housing residence	**building** **home** **house** structure housing	neighborhood street **tree** **house** bridge	**building** **house** **home** structure **tree**
		bush driveway field, flower grass road, rock street, tree	umbrella flamboyant titus grape gleditsium	woody **tree** plant vascular flowering	**tree** plant oak structure framework	**grass** **field** **road** mountain forest	**tree** plant **grass** **field** **road**
		beak, bird feather, ripple lake, neck pond, pool swan, water	hooded cygnet whooper drake bottlenose	aquatic anseriform waterfowl **swan** duck	**bird** duck **swan** tree material	duck **water** **lake** beach sand	duck **swan** **water** **lake** boat
		blue dress bush, dress girl, child grass, plant sky, tree	Hyla large wind Honduras Salix	woody **tree** **plant** vascular conifer	**tree** **plant** material flower wear	**dress** **girl** field beach boy	**dress** **girl** field **tree** beach
		front yard grass, window house, lawn potted plant sidewalk stair, tree	camper stoop chicken dacha detach	camper trailer stoop porch structure	structure trailer porch stoop camper	neighborhood **house** **window** bedroom door	neighborhood **house** building **window** bedroom
		farm, fence field horse, mule kite, dirt people tree, zoo	gelding yearling shire yearling draft	**horse** equine perissodactyl ungulate male	**horse** **tree** equine male gelding	**horse** pasture **field** cow **fence**	**horse** pasture **field** cow **fence**
Results in the bottom 25%		fence, junk sign stop sign street sign trash can tree	feeder Hyla cleaner box large	woody **tree** structure plant vascular	**tree** structure building plant area	logo street neighborhood building office building	logo street neighborhood building office
		circle earring hook jewel jewelry make up stone	clasp fob enamel chain gold	clasp fix constraint device chain	clasp fix constraint device chain	bead pearl bracelet silver sterling	clasp fix constraint device bead

Figure 8. Example translations. 1^{st} col shows images. 2^{nd} col shows MTurk associated nouns. These represent the ground truth annotations (entry-level categories) we would like to predict (colored in blue). 3^{rd} col shows predicted nouns using a standard multiclass flat-classifier. 4^{th} col shows nouns predicted by the method of [6]. 5^{th} col shows our n-gram based method predictions. 6^{th} col shows our SVM mapping predictions and finally the 7^{th} column shows the labels predicted by our joint model. Matches are colored in green. Tables 1,2 show the measured improvements in recall and precision. We provide more examples in supplemental material.

tions. 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 – 3610 distinct nouns total in our evaluation sets. This makes noun prediction extremely challenging!

We evaluate prediction of all nouns associated with an image by Turkers (Table 1) and prediction of nouns assigned by at least 2 of 3 Turkers (Table 2). Here N+ refers to the working vocabulary of the method – the total number of unique words output by the method for the given test set. 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 [6], and our proposed entry-level category predictors (ngram-biased mapping, SVM mapping, and a joint model). Performance at this task on Set B is in general better than performance on Dataset A, because 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.

On all datasets and tasks we find the joint model to perform the best (section 4.3), followed by supervised prediction (section 4.2), and propagated prediction (section 4.1). In addition, we greatly outperform both leaf node classification and the hedging technique [6] (approximately doubling their performance on this task).

6. 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. These methods could apply to many different end-user applications that require recognition outputs that are useful for human consumption, including tasks related to description generation and retrieval.

Acknowledgments: This work was supported by NSF Career Award #1054133 and NSF Award #1161876.

References

[1] S. Bird. Nltk: the natural language toolkit. In *COLING/ACL*, 2006. 5

[2] T. Brants and A. Franz. Web 1t 5-gram version 1. In *Linguistic Data Consortium*, 2006. 3

[3] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010. 1, 2

[4] J. Deng, A. C. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei. Large scale visual recognition challenge. In *www.image-net.org/challenges/LSVRC/2012*, 2012. 1

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 2

[6] J. Deng, J. Krause, A. C. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *CVPR*, 2012. 1, 2, 3, 4, 6, 7, 8

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, June 2010. 2

[8] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: generating sentences for images. In *ECCV*, 2010. 1

[9] C. Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998. 1

[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *tPAMI*, 2009. 2

[11] P. Jolicoeur, M. A. Gluck, and S. M. Kosslyn. Pictures and names: making the connection. cognitive psychology. *Cognitive Psychology*, 16:243–275, 1984. 1, 2

[12] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 2

[13] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. In *CVPR*, 2011. 1

[14] P. Kuznetsova, V. Ordonez, A. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In *ACL*, 2012. 1

[15] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, and A. Ng. Building high-level features using large scale unsupervised learning. In *ICML*, 2012. 2

[16] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 1, 4

[17] F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid. Towards good practice in large-scale learning for image classification. In *CVPR*, 2012. 1

[18] E. Rosch. Principles of categorization. *Cognition and Categorization*, page 2748, 1978. 1, 2

[19] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77:157–173, 2008. 2

[20] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *PAMI*, 30:1958–1970, 2008. 2

[21] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2001. 4

[22] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large scale scene recognition from abbey to zoo. In *CVPR*, 2010. 2