Choosing Basic-Level Concept Names using Visual and Language Context

Alexander Mathews[*], Lexing Xie[*†], Xuming He[†*]
[*]The Australian National University, [†]NICTA
alex.mathews@anu.edu.au, lexing.xie@anu.edu.au, xuming.he@nicta.com.au

Abstract

We study basic-level categories for describing visual concepts, and empirically observe context-dependant basic-level names across thousands of concepts. We propose methods for predicting basic-level names using a series of classification and ranking tasks, producing the first large-scale catalogue of basic-level names for hundreds of thousands of images depicting thousands of visual concepts. We also demonstrate the usefulness of our method with a picture-to-word task, showing strong improvement over recent work by Ordonez et al, by modeling of both visual and language context. Our study suggests that a model for naming visual concepts is an important part of any automatic image/video captioning and visual story-telling system.

1. Introduction

Automatically describing the objects, people and the scene in an image is one of the most ambitious tasks of computerised image understanding. Progress on this task has significant practical implications, since there are billions of pictures on the web, as well as in personal and professional collections. There are two aspects to this picture-to-words problem. The first is computational visual recognition – to recognize or localize thousands of visual semantic categories. Systems that solve this problem are beginning to work [13, 6]. The second aspect is to mimic the human description of categories – recent work by Ordonez *et al.* [17] addressed this aspect by identifying basic-level categories. We note, however, that there are two key limitations of the recent literature on pictures to words. The first is in assuming that the basic-level name for a visual category is unique, whereas recent work cognitive psychology found that object names are context-dependent [1, 14, 3] and are affected by attributes such as typicality [11]. The second is that the mapping from categories to basic-level names relies on crowd-sourced explicit labeling. While crowdsourcing is an efficient way to gather one name-per category for tens thousands of categories, it is not scalable to cases where different instances of an object have different basic-level names that vary from one task and context to another. In this work, we study the interplay between basic-level names and the visual and language context of each image, and scale image-to-words systems to millions of images. To this end, we make use of millions of online images with human-supplied descriptions [4, 22], and large-scale visual recognition systems [8, 13].

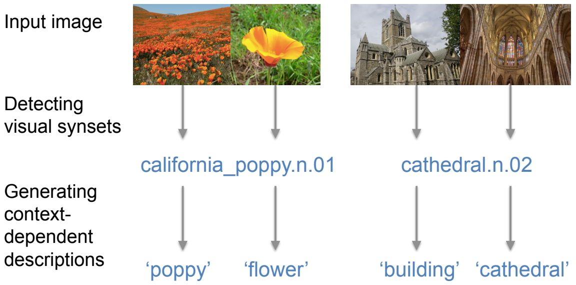


Figure 1. Examples of basic-level name variations due to object appearance (right) and image context (left). Given an input image, this paper proposes novel methods for assigning context-dependent basic-level names using object appearance with visual and language context.

1.1. Variations in basic-level categories

The *basic* level of categorization is "the most inclusive (abstract) level at which the categories can mirror the structure of attributes perceived in the world" [19]. However, a number of factors affect how people name a particular visual concept, or assign *basic-level names*.

Visual appearance plays an important role in naming object instances which belong to the same category. Psychologists have identified these factors as typicality [11], or perceptual variability and kind diversity [14]. In images, people tend to assign different names for different view points. As shown in Figure 1, pictures of a *cathedral* can be called *building* given an exterior view, and *cathedral* given an interior view.

Other objects or the visual setting may also change how semantically identical objects are named. Psychologists have observed a similar effect, namely context-independent and context-dependent properties [1]. For example, an entire field of *california poppy* flowers is described with the word *poppy*, while a single flower is often described with the word *flower* (Figure 1 left); in close-up an *apple* is likely to be named as an *apple*, while in the presence of other fruits it is named as *fruit* (Figure 4 right, second row).

Cognitive psychology experiments are typically conducted with isolated concepts [20], up to a dozen concepts with toys or object drawings [11, 14], and with a few dozen to a hundred subjects due to limitations of experiment condition. Using images with captions generated naturally by hundreds of thousands of users, we can train algorithms to predict the most appropriate basic-level names for thousands of semantic categories in context.

1.2. Solution overview

We propose a three step approach to automatically determine the basic-level names of objects in an image. We first use ImageNet [4] to learn visual concept models from images for more than 2,000 visual synsets (Section 3.1). We then learn a model to choose a basic-level name for a given concept, which explicitly takes into account visual context, object importance and object appearance (Section 3.2.2). Finally, we fine-tune the obtained descriptions for an image using language context – this is done via a model for ranking descriptions from different visual concepts, as described in Section 3.3. This system is evaluated on the SBU 1 Million Flickr image dataset (Section 4-5). Our system achieves a higher precision and recall than Ordonez *et al*. [17] when predicting the 5 most likely basic-level names. Furthermore, on 1,200+ synsets predicting basic-level names using visual context leads to more accurate matches with human descriptions.

The main contributions of this work are as follows.

- We produce the first large-scale catalogue of fine-grained basic-level categories with thousands of visual synsets and hundreds of thousands of images. This is automatically constructed by analysing tagged online datasets with associated natural language descriptions, and can easily scale to an order of magnitude more concepts and images.
- We propose a new method to predict context-dependent basic-level categories taking into account visual and language information. This is achieved by decomposing the problem into a set of classification and ranking tasks.
- We perform experiments and benchmark on a dataset two orders of magnitude larger than that of prior work [17], our system shows significant improvement for the picture-to-word task using context-dependent basic-level naming of visual concepts.

A catalogue of context-dependent basic-level categories, and a word prediction benchmark of 150,000 images are released online.

2. Related Work

Since the late 1970s, psychologists have studied how people typically name objects. Rosch *et al*. [20] noted that people typically describe objects at what is called their *basic-level category*. Thereafter, several groups noted context effects for naming objects, including Barsalou [1] and Rosch herself [19]. More recently, strong experimental evidence of context-dependent categorization is observed in children [14] and adults [3]. The constraints of experimenting with in-lab subjects means that no comprehensive catalogue of basic-level categories is yet available.

Automatically associating images with natural language is a very active topic within computer vision. Most recent systems rely on visual recognition as a component, such as state-of-the-art approaches using convolutional neural networks (CNN) [13, 8]. The approaches for associating images to words and sentences started with visual detection over a small number of object categories, followed by language model [23], caption retrieval [18], and explicitly capturing syntactic and semantic features [7]. A few approaches explicitly relate visual semantics to their expression in words, such as studying how objects, attributes and visual relations correlate with their descriptions [25], and learning visual variations of object categories [5]. In terms of capturing human descriptions of natural images, our work is inspired by the studies of importance [21, 2] and the first work to identify basic-level categories from images [17]. We offer two points of departure from the state-of-the-art. First, we note that the goal of deciding which basic-level names describe an image is different from predicting objects or a visual sentence; the focus here is modeling the psychological processes that affect naming rather than visual detection or linguistic regularities. Second, a number of recent works model context [5, 21, 2] or naming [17], but these two parts have not yet been connected, to the best of our knowledge.

3. Predicting basic-level categories

We propose a method to predict the basic-level names of objects given an image, an overview is shown in Figure 2. We take a probabilistic approach to model the probability of using a word y to describe an image $\mathbf{x}$, denoted by $p(y=1|\mathbf{x})$. However, directly mapping $\mathbf{x}$ to y does not take into account the relationships between visual concept and words, as well as the context among co-existing concepts. We exploit the structure of this problem by approximately modeling this distribution with three main components: first recognizing visual concepts from images, followed by naming *individual* concepts in the second step; and then ranking these name-and-concept pairs from *all* concepts using context for each image.

3.1. Detecting semantic visual concepts

The first step towards deciding on descriptive names for a visual scene is to capture semantics, or meanings. In linguistics, a word *sense* is “an element from a given set of meanings (of the word)” [16]. WordNet, the widely-used lexical database, uses *synsets*, i.e., *synonym sets*, to represent word senses [16]. The well-known ImageNet [4] database illustrates each WordNet synset with a few hun-

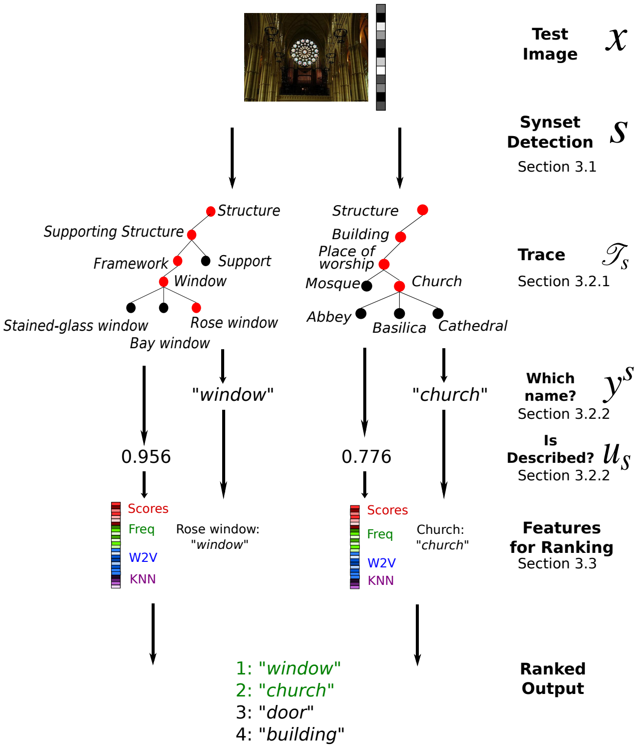


Figure 2. Method overview of context-dependent prediction of basic-level names. See Section 3 for details.

dred images. We capture visual semantics by learning a visual representation of each synset from ImageNet.

Our model uses the output of the last fully-connected layer of a CNN [8] as feature vector $\mathbf{x}$, and logistic regression classifiers. Equations 1, 3, and 4 *adapt* the last supervised layer of CNN features to a number of new classification tasks – new synsets, new set of training data, and fine-grained appearance categories. Such adaptation is known to yield performance competitive to retraining the deep convolution network [12]. This learning scheme is efficient enough to handle a large amount of training data and thousands of target classes.

We first learn a *synset* classifier to estimate the probability that a synset s appears in image x:

$$p(s=1|\mathbf{x}) = \sigma(\mathbf{w}_s^T \mathbf{x}) \tag{1}$$

Here σ is the sigmoid function $\sigma(x) = 1/(1+e^{-x})$, and $\mathbf{w}_s$ is a weight vector for distinguishing synset s from all other synsets.

3.2. Naming individual visual concepts

Given a detected visual concept s in image $\mathbf{x}$, we want to model the probability of mentioning a word y related to s. To this end, we define a generative process of naming each concept as follows. First, given the image and concept, we generate a distribution over a switch variable u_s that indicates whether s would be described at all. If u_s is *on*, we then generate a word y_i^s with a distribution $p(y_i^s = 1|\mathbf{x}, s = 1, u_s = 1)$. Here we assume each individual concept contribute support to only one name. Overall, the probability of generating a word y_i^s given concept and image can be written as,

$$\begin{aligned} p(y_i^s = 1|\mathbf{x}, s) &= \sum_{u_s \in \{0,1\}} p(y_i^s = 1|\mathbf{x}, s, u_s)p(u_s|\mathbf{x}, s) \\ &= p(y_i^s = 1|\mathbf{x}, s, u_s = 1)p(u_s = 1|\mathbf{x}, s). \end{aligned} \tag{2}$$

The word probability allows us to select the most likely name for each individual concept s. We will generate a set of proposals of word-concept pairs based on these concept specific probabilities and globally rank them in Section 3.3.

In the following subsection, we describe how to obtain a set of target nouns y^s for each concept. The modelling of $p(u_s = 1|\mathbf{x}, s)$ and $p(y_i^s = 1|\mathbf{x}, s, u_s = 1)$ are detailed in Section 3.2.2.

3.2.1 Identifying a set of names for a concept

For each synset s, we define a set $\mathscr{T}_s$ that contains all words that can be used as the *name* for the semantic concept s. For example, $\mathscr{T}_{cow}$ would include *cattle*, *bovine*, and *animal* which are the larger categories that it can be classified into, or *moo-cow* as an informal name, and *kine* as an archaic plural for *cow*.

We define an indicator vector $\mathbf{y}^s \in \{0,1\}^{|\mathscr{T}_s|}$, where each element corresponds to one word in $\mathscr{T}_s$. Then the task of choosing an entry-level description for an image x becomes a multi-class classification problem, i.e., estimating the probability of each choice $p(y_i^s|\mathbf{x}, s, u_s)$, for $i = 1, \ldots, |\mathscr{T}_s|$, such that $\sum_i p(y_i^s|\mathbf{x}, s, u_s) = 1$.

In this work, we obtain $\mathscr{T}_s$ by *tracing* the WordNet hierarchy up to 5 hypernym (i.e., parent concept) levels, and obtain a set of lemmas at each level (e.g., *ridding horse* and *mount* are both lemmas of the same synset). The final set $\mathscr{T}_s$ is the union of these lemmas. Such a construction excludes the hyponyms (i.e. children nodes) of a synset, to avoid confusing this task of choosing context-dependent names with fine-grained classification (e.g., distinguishing a *male horse* from a *mare*). Note that this construction of $\mathscr{T}_s$ does not include names that may be sibling or other related nodes (such as *zebra* or *mule* for *horse*). We found this to yield much cleaner candidates for entry-level descriptions (albeit lower recall) than simply treating the most frequent words as candidates. This restriction could be relaxed by looking at nouns within a certain WordNet distance of the synset of interest, but we leave this as future work.

Due to its construction using the WordNet hierarchy, set $\mathscr{T}_s$ is henceforth referred to as the *trace* of s.

3.2.2 Choosing a name for a concept

According to the generative model for concept names in Equation 2, we learn an *is_described* classifier to estimate how likely synset s is to be explicitly described given that it is visually present.

$$p(u_s = 1|\mathbf{x}, s = 1) = \sigma(\mathbf{w}_u^T \mathbf{x}) \tag{3}$$

The intuition behind the *is_described* classifier is similar to that of predicting the importance of an object [2]. We expect CNN feature $\mathbf{x}$ to be able to capture most of the information that relates to concept importance and describability, as it has shown state-of-the-art performance in capturing scene types and contextual factors [13].

We implement the remaining part of Equation 2 by learning a *description* classifier, to infer the most likely noun used to describe synset s.

$$p(y_i^s = 1|\mathbf{x}, s = 1, u_s = 1) \propto \exp(\mathbf{w}_{yi}^T \mathbf{x}) \tag{4}$$
$$i = 1, \ldots, |\mathcal{T}_s|$$

As one can see from the definitions, these *synset*, *is_described*, and *description* classifiers are learned on successively smaller training sets that are increasingly tuned to the differences in description generation, i.e. first distinguishing $s = 0$ versus $s = 1$, then $u_s = 0$ versus $u_s = 1$ only when $s = 1$, and finally choosing among $\mathbf{y}^s$ when $u_s = 1$. For efficiency reasons, a concept s is considered for an image when $p(s = 1|\mathbf{x})$ exceeds a high threshold. According to Equation 2, the *is_described* probability $p(u_s = 1|\mathbf{x}, s = 1)$ is a shared scaling factor for all possible words y_i^s, hence it is not used for selecting names, and only used for ranking names across synsets.

3.3. Ranking names and concepts

Equation 2 describes the probability of generating a name for each synset s, but it does not prescribe ways to rank the descriptions generated by different synsets. One way to impose a ranking is with the confidence of the *is_described* classifier $p(u_s = 1|\mathbf{x}, s = 1)$. This confidence does not, however, take into account additional side information, such as the reliability of each description classifier, the prior likelihood of seeing each concept in an image, and the context of other nouns generated for the same image.

We aim to learn a ranking score r for each triple composed of image $\mathbf{x}_i$, synset s_m, word y_k, referred to by their respective indexes (i, m, k). We use a linear ranking function

$$r_{i,m,k} = \mathbf{w}_r^T h_{i,m,k}.$$

The optimization problem for obtaining the ranking weights $\mathbf{w}_r$ follows the RankSVM formulation [9]. Here the training data consist of pairs of image-synset-word tuples (i, m, k) and (j, q, l), where word k, synset m is associated with image i, while word l , synset q is *not* associated with image j, and $\xi_{i,m,k;j,q,l}$ are non-negative slack variables.

$$\begin{aligned} &\min\ J(\mathbf{w}_r, \xi) = \frac{1}{2}\mathbf{w}_r^T \mathbf{w}_r + C \sum_{i,m,k;j,q,l} \xi_{i,m,k;j,q,l} \qquad (5) \\ &s.t.\ \forall (i, m, k; j, p, l) \\ &\qquad \mathbf{w}_r^T h_{i,m,k} \geq \mathbf{w}_r^T h_{j,q,l} + 1 - \xi_{i,m,k;j,q,l} \\ &\qquad \xi_{i,m,k;j,q,l} \geq 0 \end{aligned}$$

We use four types of features to capture different information that is relevant for ranking words.

SCORES from different classifiers. These include: *is-described-score*, probability that a synset description appears in the list of words as in Eq (3); *direct-to-noun-score*, the probability that a word k is used to describe image $\mathbf{x}$ – $p(y_k|\mathbf{x})$, obtained using logistic regression; *synset-score*: the probability that synset s_m, corresponding to word k is in the image, according to Eq (1).

AUX-iliary information about words and synsets. These include: *in-synset-frequency*, the prior of word k within its corresponding synset m; *global-noun-freq*, the prior probability of word k in all images; *description-accuracy*, the accuracy of the description classifier for this synset based on cross-validation performance; *is-described-accuracy*, the accuracy of the *is_described* classifier from cross-validation; *trace-size*, or $|\mathcal{T}_s|$ in Eq (4).

KNN-rank. We find the k-nearest neighbours in the training set to image $\mathbf{x}$, we then rank the nouns associated with these retrieved images by TF-IDF. *knn-rank* is then $1/rank$ for the word in question. k is chosen to be 500 in this work.

WORD2VEC features are used to capture the word context. We use a modified version of the Word2Vec Continuous Bag of Words (CBOW) model [15] without the hierarchical softmax. Our CBOW training method take a random selection of words in each training iteration, thus making our word-context model order-independent. The CBOW model projects words into a 100 dimensional feature space.

We extract two different types of Word2Vec features from the set of candidate nouns for each image, broadly described as similarity and score features. *word2vec-similarity-max* and *word2vec-similarity-avg* are the maximum and average cosine similarity between word k and the top 6 words according to *is_described* scores. *word2vec-score-max* and *word2vec-score-avg* are the maximum and average probability of predicting the target word k given a random subset of context words under the modified CBOW model, taking max- and average- over 10 random inputs.

We augment these features using non-linear transformations. Specifically we append the products of all pairs of features. Augmenting the feature vector of linear SVM is known to produce competitive performance compared to SVMs with non-linear kernels [24]. For each image, we generate the final set of ranked words by removing duplicate words from the ranked list of synset-word pairs (m, k).

4. Experimental setup

Training and testing datasets. We use the IMAGENET-FLICKR dataset to train synset classifiers and for a preliminary validation of the concept of context-dependent basic-level names in Section 5.1. This dataset is the subset of ImageNet[4] originally from Flickr [22], containing over 5.7 million images, with both WordNet synset labels and Flickr metadata such as caption and tags.

We use the SBU dataset [18] to train our *is_described* and *description* classifiers as well as for evaluation. Originally this dataset consisted of 1 Million Flickr images with associated captions; however at the time of collection only 95%, or 950K images were still publicly available. Another 2000 images are removed because they form the two datasets, of 1000 images each, used by Ordonez *et al.* [17], here referred to as SBU-1KB and SBU-1KB. We used 80% of the remaining images, or 760,000, for training the *is_described* and *description* classifiers; while 40,000, or 4% were used for training the rankSVM. The remaining 148,832 images were used for evaluation, which we refer to as SBU-148K. SBU-1KA contains randomly selected images, while SBU-1KB contains images for which synset detectors produced high confidences. A list of nouns associated with each image is produced by Amazon Mechanical Turk workers. We evaluate against the union of nouns selected by all Turkers, since the level of agreement across different Turk workers was not released.

Generating groundtruth descriptions for the SBU dataset We extract lemmatized nouns from the image captions and filter out nouns that are not part of ImageNet. The ground truth is noisy as captions may not directly refer to the visual content. But they are more realistic for capturing people's language use in natural settings than those from the naming exercises in SBU-1KA and SBU-1KB.

Model learning The image feature $\mathbf{x}$ is a 4096 dimensional vector from the last fully-connected layer of a pre-trained CNN [8]. We learn 2633 synset classifiers (Equation 1) on ImageNet. These synsets have enough training examples in the IMGNET-FLICKR dataset, and have at least 100 positive instances in the SBU dataset. The threshold on $p(s=1|\mathbf{x})$, for considering an image to represent a synset, is chosen to be 0.95. The *is_described* classifier for u_s is trained on images representing synset s, images associated with any word in $\mathcal{T}_s$ are positives examples, the rest are negatives. We learn a set of one-vs-one SVM classifiers as the *description* classifier for each synset. We use SVM^{rank} [10] to learn description ranking over different feature settings. Regularization parameters for all classifiers are tuned with cross validation.

Evaluation metric For each image, the system outputs a list of nouns sorted by confidence. For each image we calculate the precision and recall at each position in the list; we further average these precision and recall points across 10 random testing set partitions – generating the P-R curve and its standard deviation bars. This is the same metric reported by Ordonez *et al.* [17] though they report only the precision and recall at top 5 words per image.

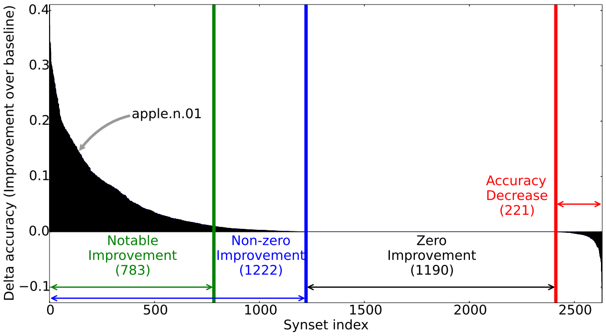


Figure 3. Per-synset accuracy improvement of the basic-level name classification proposed in Section 3.1–3.2.2 over the *Frequency+described* baseline. See Section 5.1 for discussions.

Four baselines are used for comparison.

- *Ngram-biased-SVM*, as presented by Ordonez *et al.* [17]. This baseline only applies to the SBU-1KA and SBU-1KB datasets.
- *Direct-to-noun* , a method consisting of 2,549 separate logistic regressors which directly predict each noun y_i using CNN feature $\mathbf{x}$ as input, calibrated with Platt-scaling using hold-out data from SBU training set.
- *Most-frequent name*, a method that outputs the most-frequent noun y_i in each trace $\mathcal{T}_s$ for synset s, instead of using Equation (3–4). Nouns are ranked by their prior frequency.
- *Frequency+described* is the same as the *Most-frequent name* method except using the *is_described* classifier (Eq. 3) for ranking.

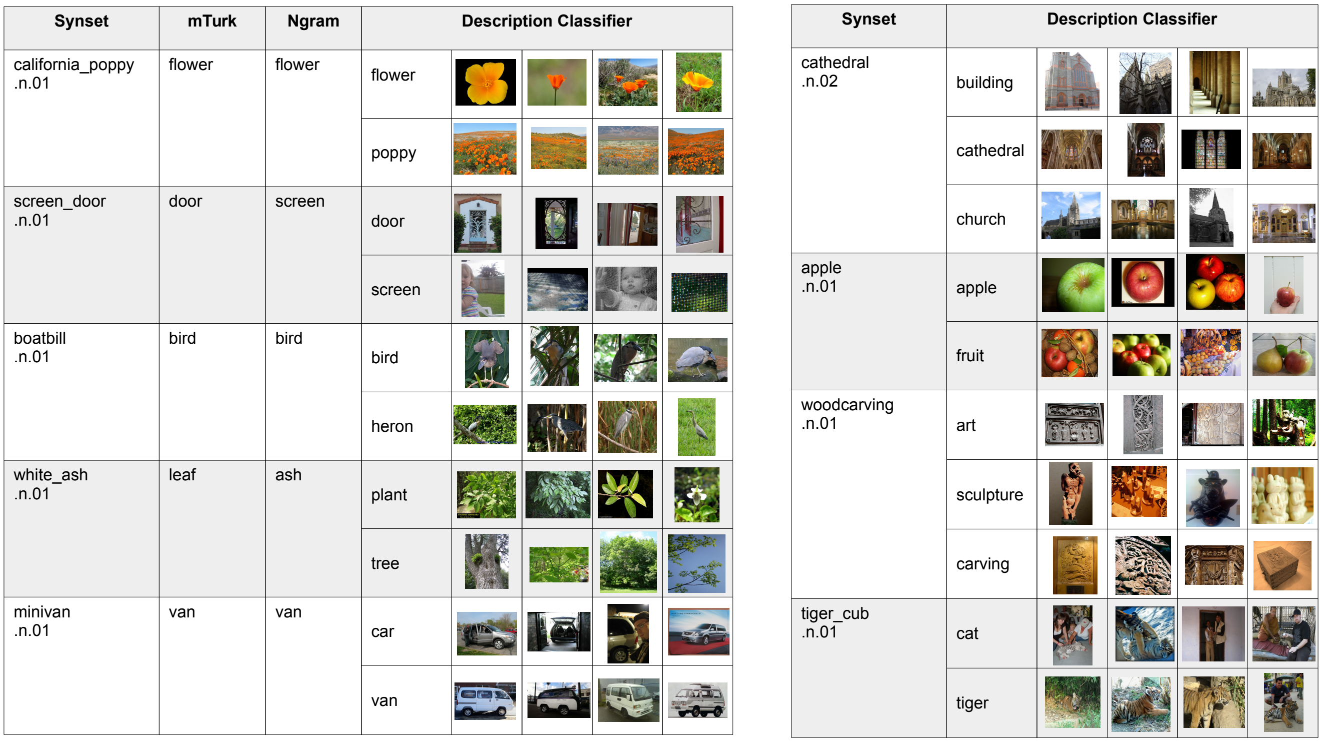
5. Experimental Results

We evaluate the performance of word classification and ranking in three ways: description classification within each synset trace (Section 5.1), picture to word prediction, and word ranking method taking into account language context (Section 5.2).

5.1. Choosing basic-level names with visual context

Preliminary validation. We examine 3,398 synsets from the ImageNet-Flickr dataset with at least 200 captioned images, and examine which words on the trace are used in their descriptions. Among these synsets, the second most common name is used in at least 10% of image captions for 1,026 synsets. This shows that using multiple names for a single semantic concept is a common phenomena.

Automatic description classification. We use methods outlined in Section 3.1–3.2 to detect each synset and choose descriptions on the SBU dataset. Our per-synset evaluation



Synset	mTurk	Ngram	Description Classifier
california_poppy .n.01	flower	flower	flower
			poppy
screen_door .n.01	door	screen	door
			screen
boatbill .n.01	bird	bird	bird
			heron
white_ash .n.01	leaf	ash	plant
			tree
minivan .n.01	van	van	car
			van

Synset	Description Classifier
cathedral .n.02	building
	cathedral
	church
apple .n.01	apple
	fruit
woodcarving .n.01	art
	sculpture
	carving
tiger_cub .n.01	cat
	tiger

Figure 4. Examples of context-dependent basic-level categories. Left: For each synset, we display crowd-sourced one-name-per-synset [17], n-gram based most frequent name [17], and context-dependent descriptions chosen according to Sec 3, along with four image examples. Right: synsets for which no previous results [17] were available.

reports the accuracy of 3-fold cross-validation with an internal 3-fold cross validation to select the hyper-parameters for the description classifier. Figure 3 displays accuracy improvements of our method over the *Frequency+described* baseline. Among the 2,633 synsets in the SBU dataset , our method improves upon the *Most frequent name* baseline in 1,222 synsets, among which 783 improved by more than 1%. No change in accuracy is measured for 1190, and 221 synsets exhibiting a small accuracy decrease. We found that our method provides the most improvement for synsets with ambiguous basic-level names – two or more names used with similar frequency. The fraction of improved synsets is on par with the preliminary validation above. The 221 synsets that exhibit an accuracy decrease are characterised by ambiguous basic-level names and fewer than average training examples. The ambiguous basic level names cause the classifier to choose names other than the most common, while the small training set size reduces the quality of the classifier.

Illustrative examples. Figure 4 shows results of basic-level name classification in comparison to labels from Mechanical Turk workers [17] and N-gram frequency [17] (left table). The right table shows a few synsets for which no results are available from prior work [17]. We can see several aspects of visual context come into play when choosing the basic-level names – including view point variation (*plant* vs *tree*; or *bird* vs *heron*); the presence of other object or part (*apple* vs *fruit*; *door* vs *screen*); and the appearance variations within the category (*art*, *sculpture* vs *carving*).

This is the first large-scale fully automatic classification of basic-level names. Using image collections with objects in their visual context enables us to discern context-dependent basic-level names previously observed in controlled lab environments [1, 14], and alleviates the need for crowd-sourced labels [17].

5.2. Predicting nouns

We evaluate basic-level name classification (Sec 3.1–3.2) in an end-to-end noun prediction task on three SBU test subsets described in Section 4. The proposed approach, *BasicName-Visual*, uses *description* classifiers (Equation 4) to choose names from the traces, and the *is_described* scores (Equation 3) for ranking. We remove duplicate nouns appearing in multiple traces, and generate precision-recall curves by allowing 1, 2, 3, . . . nouns per image.

Figure 5 (left and middle) compares *BasicName-Visual* to four baselines (Section 4) on SBU-1KA and SBU-1KB. When predicting 5 words per image on SBU-1KA, *BasicName-Visual* achieves a precision of 0.26 with 0.13 recall, while *Ngram-biased-SVM* [17] achieves a precision of 0.20 with 0.10 recall. Detailed statistics for SBU-1KB are available in the supplemental material. Furthermore, *BasicName-Visual* is slightly better than *Most-frequent name* and *Frequency+described* while *Ngram-biased-SVM* is on par with *Direct-to-noun*. The same eval-

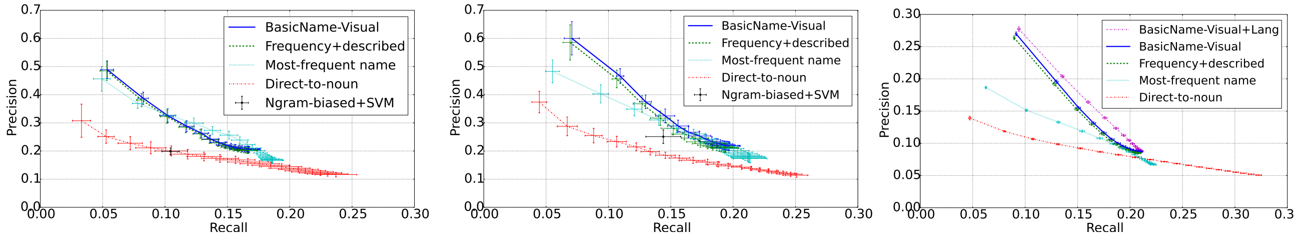


Figure 5. Left+Middle: Precision-recall curves for our method and the four baselines on SBU-1KA and SBU-1KB. Right: Precision-recall curves on SBU-148K. Error bars show standard deviation.

uations are carried out on SBU-148K (Figure 5 right). With more than two orders of magnitude more testing data - hence lower variance for performance estimates - we can see a significant difference between *BasicName-Visual* and *Most-frequent name* and *Frequency+described*. This shows that both Equation 3 and 4 contribute to choosing the right description. Further restricting predictions to the 783 noticeably improved synsets from Section 5.1 leads to a significant and consistent gain of 0.02 precision across the recall range (Figure 1 in supplemental material).

Figure 5 (right) shows the effect of language context and auxiliary information for ranking words in the same image, denoted as *BasicName-Visual+Lang*. This approach significantly out-performs *BasicName-Visual* across the recall range. The average precision of *BasicName-Visual* is 0.336 ± 0.003, improved to 0.341 ± 0.002 with SCORES features, and further improved to 0.347 ± 0.002 with KNN, WORD2VEC and AUX features.

Figure 6 shows several examples of description classification. The first four rows show examples where *BasicName-Visual+Lang* correctly predicts more specific names than the most frequent name, such as preferring *church* over *building* in row 3 and *sunflower* over *flower* in row 4. Row 5 contains an example where synset classifiers break down, here the main objects (bikes and people) are small and subject to poor lighting. Row 6 shows a difficult case where the scene contains several objects that do not usually appear together. The prediction *ball* can be considered correct but is not in the groundtruth.

6. Conclusion

We studied context-dependent basic-level categorization for objects in natural images, and proposed a method to predict basic-level names using visual and language context. We produced the first automatically generated catalogue of basic-level categories for thousands of visual concepts and hundreds of thousands of images. Our approach to basic-level naming of concepts showed superior performance in a few picture-to-word tasks. Future work includes using basic level names for image-to-sentence applications, generalized trace construction by expanding the candidate names beyond direct ancestors, and learning interpretable context using constructs such as visual attributes.

Acknowledgments NICTA is funded by the Australian Government as represented by the Dept. of Communications and the ARC through the ICT Centre of Excellence program.

References

[1] L. W. Barsalou. Context-independent and context-dependent information in concepts. *Memory & Cognition*, 10(1):82–93, 1982. 1, 2, 6

[2] A. Berg, T. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and predicting importance in images. In *CVPR*, 2012. 2, 4

[3] S. E. Chaigneau, L. W. Barsalou, and M. Zamani. Situational information contributes to object categorization and inference. *Acta Psychologica*, 130(1):81–94, 2009. 1, 2

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, June 2009. 1, 2, 5

[5] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning Everything about Anything : Webly-Supervised Visual Concept Learning. In *CVPR*, 2014. 2

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. 1

[7] M. Hodosh, P. Young, and J. Hockenmaier. Framing Image Description as a Ranking Task : Data , Models and Evaluation Metrics. *JAIR*, 47:853–899, 2013. 2

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 1, 2, 3, 5

[9] T. Joachims. Optimizing search engines using clickthrough data. In *ACM KDD*, 2002. 4

[10] T. Joachims. Training linear svms in linear time. In *ACM KDD*, 2006. 5

[11] P. Jolicoeur, M. a. Gluck, and S. M. Kosslyn. Pictures and names: making the connection. *Cognitive psychology*, 16(2), 1984. 1, 2

[12] S. Karayev, A. Hertzmann, H. Winnemoeller, A. Agarwala, and T. Darrell. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013. 3

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 2, 4

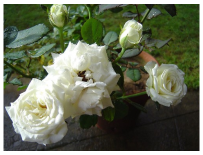
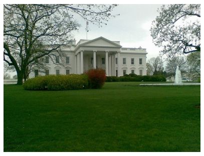
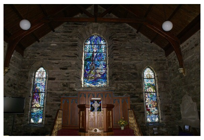
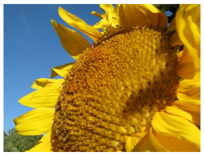
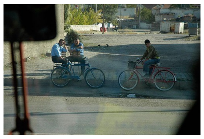
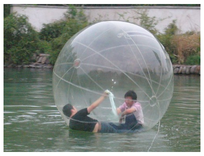
	Images	Labels	Ngram-biased-SVM	Direct-to-noun	Frequency+described	BasicName-Visual+Lang
Successful Examples		dirt, flower, grass, leaf, petal, plant, pot, rain, rise, rose, stem, white	flower tree plant dog face	tree window flower river house	plant rose white flower tree	rose white plant tree flower
		building, bush, field, tree, fountain, sky, grass, home, house, manor, white house, window, yard	building house home structure tree	tree house flower glass road	house grass building tree plant	house grass tree building road
		altar, alter, roof, art, building, flower, church, door, lamp, light, wall, podium, window, stain glass, vase, stain glass window,	street building door tower room	tree window house stained glass glass	window building tower monument column	window tower building church monument
		close-up, flower, petal, sky, tree, stamen, sunflower	bird pink tree plant white	house girl sign dog mountain	flower plant yellow	sunflower flower yellow plant
Failed Examples		adult, house, bicycle, bike, bush, parking lot, people, pole, street, tree, wall, window	neighborhood mountain zoo rock ground	flower bridge girl house car	car chair rider plant place	car seat person chair plant
		bubble, float, lake, man, pants, plastic, pond, shirt, shrub, water	shift dog zoo grass ball	water girl river beach house	ball fish building sail bird	ball fish building way sail

Figure 6. Example images from the SBU-1KA and SBU-1KB datasets with Amazon Mechanical Turk labels. We show the top few nouns predicted by our method, *BasicName-Visual*, and three baselines. Words are printed in green if they are present in the list of labels. The first four images are examples of where our method performs well, the last two images are examples of where our method performs poorly.

[14] D. Mareschal and S. H. Tan. Flexible and context-dependent categorization by eighteen-month-olds. *Child Development*, 78(1):19–37, 2007. 1, 2, 6

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, 2013. 4

[16] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11), Nov. 1995. 2

[17] V. Ordonez, J. Deng, Y. Choi, A. Berg, and T. Berg. From large scale image categorization to entry-level categories. In *ICCV*, 2013. 1, 2, 5, 6

[18] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 2, 5

[19] E. Rosch. *Principles of categorization*. MIT Press, 1999. 1, 2

[20] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439, 1976. 1, 2

[21] M. Spain and P. Perona. Measuring and predicting object importance. *IJCV*, 91:59–76, 2011. 2

[22] L. Xie and X. He. Picture tags and world knowledge: learning tag relations from visual semantic sources. In *ACM Multimedia*, 2013. 1, 5

[23] Y. Yang, C. L. Teo, H. Daumé, III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011. 2

[24] G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Recent advances of large-scale linear classification. *Proceedings of the IEEE*, 100(9):2584–2603, Sept 2012. 4

[25] C. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013. 2