Studying Object Naming with Online Photos and Caption

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 explore what names people use to describe visual concepts and why these names are chosen. Choosing object names has been a topic of interest in cognitive psychology, but a systematic, data-driven approach for naming at the scale of thousands of objects does not yet exist. First, we find that visual context has interpretable effects on visual naming, by analysing the MSCOCO dataset that has manually annotated objects and captions containing the natural language names for the object. We show that taking into account other objects as context helps improve the prediction of object names. We then analyse the naming patterns on a large dataset from Flickr, using automatically detected concepts. Preliminary results indicate that naming patterns can be identified on a large scale, but contrary to the conventional wisdom in cognitive psychology, are not dominated by *genus* for animals. We further validate the automatic method with a pilot Amazon Mechanical Turk naming experiment, and explore the impact of automatic concept detectors with t-SNE visualizations.

Category and Subject Descriptors I.2.6 Artificial Intelligence Learning — Knowledge acquisition

Keywords multimedia; learning; naming.

1. INTRODUCTION

Categorisation and naming is central to how we describe and interpret the physical world. A particular concept can often be named in many different ways, though humans are generally consistent in the names they use under a given context. For example *Ursus arctos horribilis* could be called a *brown bear*, *Ursus arctos*, *bear* or *mammal*; while typically the term *brown bear* is used. The psychology literature calls this commonly used name the basic-level name [16] for a concept. Under this model each concept is associated with a single name. Clearly this is only a first order approximation, which works well in general, but does not capture any of the subtleties of human categorisation and naming. It is not clear that a single basic-level name can be chosen for all categories, moreover the context of the naming choice is known to play a significant role [15].

MMCommons'15, October 30, 2015, Brisbane, Australia.

978-1-4503-3744-1/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2814815.2814817

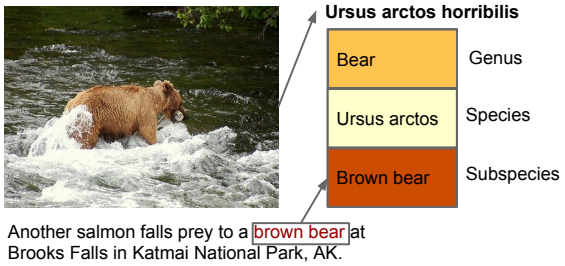

Figure 1: An image-caption where the concept *Ursus arctos horibilis* was described at the level of subspecies.

The primary issue with realising the context-dependent naming seems to be one of scale, as also recognised by proponents for the basic-level name model [15]. Typical experiments involve paying human test subjects to come into the lab and name objects [3, 9]. As a result it is unrealistic to unroll the contextual effects on naming over a large number of categories. With large scale image-caption datasets and improved image classification techniques, we can finally start to explore some of these subtleties.

Understanding how people name visual concepts could be usefull for tasks such caption generation, tag suggestion and image search. For caption generation, choosing the right words to describe the contents of the image can have a big impact on how natural the caption is. Basic-level names are specific enough to get the point across but common enough that the target audience understands. Suggesting tags is a similar problem, we want to provide potential tags that actually match how people describe the image. From Flickr tag statistics, for instance, when a *red fox* is detected a likely tag could be *fox*, whereas when a *brown bear* is identified the tag *brown bear* would be better. Similarly search could be improved by allowing images matched to relevant ancestor terms to be included in results. For example the query *cat* is more likely to match a *domestic cat* rather than a *lion*, whereas the query *felidae* has no such bias.

In this work we present several preliminary experiments on object naming. We first examine naming patterns in the MSCOCO dataset, where object ground truth and human-generated natural language descriptions are available (Sections 3). We then describe a method to determine the distribution of basic-level category names with automatic visual concept detectors on a large Flickr dataset (Sections 4). We observe that many concepts do not have a single basic-level name. Moreover we observe that across a few

hundred classes of mammals, reptiles and birds, the level of specificity for naming differs from concept to concept. We also note that the quality of concept detectors has a large influence on the estimated names. In future we plan to explore naming patterns on a much larger scale than has previously been possible.

2. RELATED WORK

A model for how people name objects was introduced to the psychology literature by Rosch [16]; this model used the idea of basic-level categories. The basic-level category represents the ideal trade-off between low in-category visual variability and high between-class visual variability. Work by Rosch [15] notes that there are contextual effects on both the level of abstraction used to name an object and even if it will be named. Chaigneau et al. [3] demonstrate, using adult subjects, that situational information changes the way subjects categorise unfamiliar objects into familiar categories.

In the case naming plants and animals Lakoff [9] explains that the genus is typically thought to be the level most commonly used. The reason that genus is important stems from how it was originally defined in the Linnaean system, as a level where each category can be easily identified.

Automatically assigning labels to images is a very active topic within computer vision. In the standard image classification problem a classifier is trained to recognise a collection of visual concepts with each concept given a single label, which may or may not be representative of how that concept is normally described. Convolutional Neural Network (CNN) [8, 17] models are the current state-of-the-art for this task. Building on the features produced by CNNs some authors have used joint vector space embeddings to define mappings from images to words mined from tags or captions [7]. We differ from these state-of-the-art image to text methods, in that our focus is modelling and interpreting the psychological processes that drive naming.

Recently there has been interest in developing machine learning techniques to choose the most appropriate name to give to a visual object. One approach taken by Deng et al. [5] is to optimise the accuracy-specificity trade off by using a semantic hierarchy to select the appropriate name. This technique does not take into account how people actually describe objects. Ordonez et al. [13] present a model predicting the labels people will actually use to name objects. Their model uses a text based component which trades off name frequency with linguistic proximity and a visual component that assesses the visual saliency of names to the image. Later work extends this model to a larger number of visual concepts and re-casts the problem to one of discovering semantic *Refer-to-as* relationships [6]. Neither of these models capture the important effects of visual context, nor do they attempt to understand and interpret the patterns in the basic-level names.

Our recent work [11] applied visual context to naming concepts, the goal was to use visual features to choose names rather than interpret how context affects naming. This work builds upon [11] to both understand how context affects naming and uncover patterns in naming across a broad range of visual categories.

3. OBJECT NAMING WITH CONTEXT

We first study the effect of image and language context on object naming by analysing the words used in image captions. To this end, we use the Microsoft Common Objects in Context (MSCOCO) [10] dataset to show that context has a measurable and interpretable effect on how objects are named in captions. Our method considers the caption word statistics for naming each object and naming prediction accuracy with and without additional image/caption context. We use decision tree classifiers to show concrete and interpretable cases where context has an effect on naming. An example of how context can effect concept naming is shown in Figure 2.

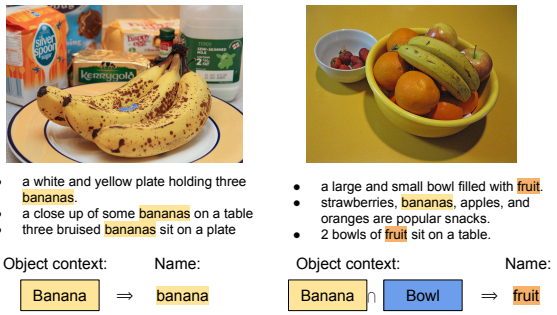

Figure 2: An example where context changes the name used to describe a concept. Top: Two example images containing *bananas* from MSCOCO data set. Middle: Three captions written by mturk workers to describe each image, with valid names for *banana* highlighted. Bottom: Relevant visual context for the names used. Presence of other fruits in the right image led to a collective name – this is one of the mechanisms of context-dependent naming.

3.1 Object naming in MSCOCO

The MSCOCO training set has over 80000 images each with five captions collected from Amazon Mechanical Turk [4]. Also available are manual annotations identifying which of 80 concepts are present in each image. These annotations were collected independently of the captions and with techniques in place to ensure that if a concept is in the image it will be annotated.

We define objects in terms of the WordNet [12] synsets, which are groups of words which have the same meaning. Synsets are arranged hierarchically, where a synset's ancestors (called the hypernyms) are more general terms, while a synsets children (called hyponyms) are more specific terms. For example the word *cat* in the sense of a feline mammal has a direct hypernym *feline* and a direct hyponym *domestic cat*.

We first manually matched each of the 80 concepts to a unique node in WordNet. We then take the manually annotated concepts in each image as *concept* ground truth. The *naming* ground truth comes from parsing the five captions for each image as follows. The nouns in each caption are identified using a parts of speech tagger, then uni-grams and bi-grams are formed from the words surrounding each noun. Each of these n-grams is then matched to the WordNet hierarchy and filtered to keep only those words which are ancestors/parents of the ground truth concepts for that image. Overlapping n-grams are then removed by keeping only the most specific; defined as the one that matches to the deepest node in WordNet. This ensures that the bi-gram *tennis racquet* is matched to the *tennis racquet* synset, rather than the *racquet* synset. While the uni-gram *racquet*, occurring alone, is matched to the *racquet* synset.

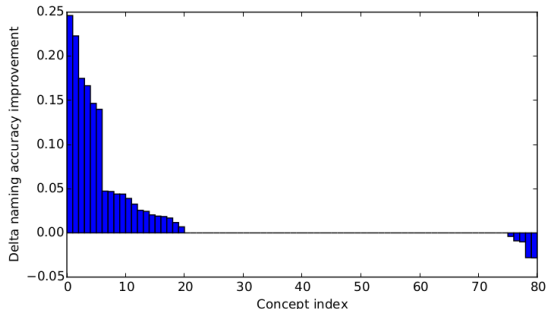

Figure 3: Improvement in name prediction accuracy when context is given compared to no context. All 80 MSCOCO concepts are shown ordered by improvement. See Sec 3.3 for details.

3.2 Measuring context effects on naming

We first identify the number of objects with more than one common name in the corresponding captions. For these concepts we show that the other concepts present in the scene will have an effect on how the concept is named. We train classifiers to predict object names with contextual information, i.e., the ground truth of other objects in this dataset. This *context-name* model learn random forest classifiers to predict which name will be used for each concept, given the full set of object ground-truths in each image. The input to *context-name* is a boolean vector with 80 elements, indicating which concepts are in the image. The output is one of several possible names for the concept. Random forest is preferred over other classifiers as it helps interpretation. We use the Gini importance metric [2] to understand which contextual concepts are the most relevant to making a naming choice. This metric computes the decrease in the Gini impurity for each of the features, averaged across all decision trees in the forest. This model is compared against a *frequent-name* baseline, which always assigns the most common name for each concept. For example if images marked as *bicycle* in the *concept* ground are most frequently named as *bike* in the *naming* ground truth then the name *bike* is always predicted.

3.3 Results: naming with context

On the MSCOCO dataset, of our 80 concepts, 48 have a second common name used at least 10% as frequently as the most common name. This suggests that a single basic-level name is not an appropriate simplification for a large number of concepts.

To learn the *context-name* model, we use 80% of the MSCOCO training set for collecting name counts and training our name classifiers. A further 10% is used for selecting hyper-parameters. With the final 10% used for testing. Each random forest name classifier has 100 estimators, the minimum number of samples for splitting an internal node is 4 and the minimum number of samples per leaf is 2.

Out of the 48 concepts, 9 showed an improvement in naming accuracy of greater than 5% when all ground truth concepts were given as context; this is in comparison to always choosing the most common name. Results are shown in Figure 3. The most-improved concepts are *car*, *ball*, *orange* and *backpack*. In the case of *orange* the most common names are *fruit*, *oranges*, *food*. While the most important object context as measured by the Gini metric are *apple*, *dining table* and *bowl*. Intuitively, when people name an orange they are more likely to use the collective term *fruit* in the presence of other fruit such as oranges; the concept *bowl* likely indicates that there is a fruit bowl in the image. In the case of *ball* the most common names are *tennis ball*, *baseball* and *ball* while the most important concepts are *tennis racket*, *baseball bat* and *baseball glove*. This is a case where the concept has multiple sub-concepts each with their own basic-level name. The context allows us to differentiate between the sub-concepts and select the most appropriate name. There are five concepts where the *frequent-name* baseline outperforms the *context-name* method. These concepts are characterised by relatively small testing and training sets. Classifier over-fitting is the likely cause of the performance difference.

4. OBJECT DETECTORS TO NAMES

We extend the analysis of naming patterns in Section 3 to a large number of objects. To this end, we automatically detect visual objects in a large image-caption dataset. We then use the object names present in the caption as the *naming* ground truth for the corresponding concept. In this analysis we restrict ourselves to objects from the animal kingdom, allowing a clean definition of specificity under the rigid taxonomical structure of the Linnaean system.

4.1 Large-scale visual object naming

We use the SBU 1-Million image-caption dataset [14], sourced from Flickr. This dataset consists of 1-Million images with captions that are likely to be visually relevant.

Any sub-string of the caption is a candidate for the description of the visual concept. In our case, however, we have chosen to only use uni-grams or bi-grams which match to nodes in the taxonomy. The matching method is similar to the text to concept matching described in Section 3, though we match to the ITIS animal taxonomy rather than WordNet. An n-gram is matched to a node in the taxonomy using exact string matching to vernacular names or scientific names. Word concatenation, lemmatization and punctuation removal are used to improve the recall.

Defining objects. We adopt ITIS to define the objects in the animal kindom. ITIS [1] is a collection of taxonomic information for plants, animals, fungi and microbes around the world. ITIS was developed and continues to be supported by a collection of federal agencies in the United States. The system has over 690000 scientific names and 124000 common names arranged hierarchically by their classifications e.g. kingdom, class, genus and species. The ITIS taxonomy is used in preference to WordNet because the depth of a synset from the root does not have a natural interpretation, whereas in ITIS the depth corresponds to groups such as class, genus or species.

Given an image in the SBU dataset, we automatically detect the object label based on pre-trained visual classifiers. Specifically, we first identify the synset label using the Oxford VGG 16-layer network [17], which was pre-trained on ImageNet. This network is trained to classify 1000 different visual synsets in WordNet. We map these synsets to nodes in the animal taxonomy using a string matching from synset lemmas to taxonomy vernaculars and taxonomy entry names. The resulting mapping is many-to-many, though typically taxonomy entries only have one synset mapped to them. We only consider the most confident, i.e., top one, visual prediction of the VGG network for each image. If an

image has a top one visual concept that is not an animal the image is ignored.

Matching objects to names. We select a sub-tree of the taxonomy such as *Mammalia* (Mammals) or *Aves* (Birds). For each image we match the highest confidence visual concept to names in the caption. We require that both the visual concept and the possible name map to taxonomy entries in the sub-tree of interest, and that the name and the visual concept have to have a descendant/ancestor relationship. If this condition is met then the name is counted as a way of describing the visual concept. The reason for using such a strong condition is that we can be far more confident that a name is actually being used to describe a visual concept if both the classifier and caption agree with respect to the taxonomy.

4.2 Analyzing large-scale naming patterns

Using both automatically detected concepts and the names matched to the ITIS taxonomy we explore how different classes of animals such as birds, mammals, reptiles, are described. We count both the frequency of concept/name pairs and concept/taxonomy level pairs. The frequency of each name for a concept gives a fine grained look at how a concept is described. Normalising the taxonomy level counts independently for each concept gives us a broad overview of the level of specificity used.

Using the SBU 1-Million image-caption dataset we calculate the level at which each animal concept is named. There are over 59000 images in the SBU dataset which both trigger an animal classifier and have a descendant/ancestor animal name in the caption. Figure 4 shows the results for the *Mammalia* class. We can see that, for the subset of mammals which we can detect, many are commonly described at only one level. This is consistent with the idea that many concepts have a basic-level at which they are described. There are also a number of cases where this does not hold, and multiple names are used with similar frequency to describe the same concept. For example *black bear* and *bear* are used with very similar frequency to describe the animal *Ursus americanus*. This supports the idea that a single basic-level is not applicable in all cases.

Using Figure 4 we see that animals in the *Mammalia* class are commonly described at the level of species, genus or family. Similar figures for both *Aves* (birds) and *Reptilia* (Reptiles) are provided online[1]. *Aves* are typically described at the level of class by the name *bird* or occasional at the level of family or genus, while *Reptilia* are typically described at the level of order or genus. This leads to the observation that animals in the class *Mammalia* are, in general, described more specifically. We suspect this is because mammals often have shape differences which are obvious at human scales. Further confirming this is the observation that the classes of birds and reptiles which are described at more specific levels tend to be large and have distinctive shapes such as *ostrich, black swan, alligator and iguana*. This is consistent with the claim that distinctive shapes are important in categorisation and naming [9].

4.3 Mechanical Turk naming analysis

We conduct a small scale animal naming experiment on Amazon Mechanical Turk (AMT) to act as a pilot validation of the large scale naming results in Section 4.2. We ask AMT workers to label 30 images for each of 10 animal categories with the name they would use to describe the animal. Three different AMT workers are assigned to each image, giving a total of 900 judgements. These judgements are then matched to the animal taxonomy and filtered as in Section 4.1.

Animal	MTurk Names	SBU Names
Dasyurus		
Bos	ox	cattle
Canis lupus	wolf, dog	wolf, dog
Ursus arctos horribilis	bear	brown bear, bear
Cebus capucinus		capuchin monkey
Marmota	squirrel	
Ailurus fulgens	red panda	red panda
Elephas maximus	elephant	elephant
Vulpes lagopus	arctic fox, fox	
Panthera leo	lion	lion

Table 1: Common names selected for each animal. Names are in order from most frequent to least. Only showing names that occur in at least 10% of cases with a matching name, and with the total count greater than 20.

The names chosen by turkers, shown in Table 1, demonstrate that for some animal categories multiple names are commonly used (eg *arctic fox* and *fox*). Moreover, we see that the results are similar to those obtained automatically from the SBU dataset. For example *Canis lupus*, *Panthera leo*, *Elephas maximus* and *Ailurus fulgens* all have the same most common name across both datasets. In the case of *Ursus arctos horribilis* (brown bear) we see that turkers tended to use more general names than those used in the image-caption corpus. We suspect that this is because people have greater contextual information about the photos they upload themselves, which encourages the use of more specific names. Note that mechanical turk names for *Cebus capucinus* (white-headed capuchin monkey) and *Dasyurus* did not match the taxonomy. *Dasyurus* is a nocturnal marsupial native to Australia and New Guinea, so it is reasonable to assume that the annotators could not identify it correctly. *Cebus capucinus* was overwhelmingly described as a *monkey*, though it is technically a *new world monkey*, this disconnect between the ITIS vernaculars and the names being used by turkers is the reason *Cebus capucinus* was ignored.

These pilot results support the ideas that some animals are commonly named in different ways with similar frequency and that the specificity with which animals are named may vary.

4.4 Naming and concept detector performance

The concept detector we used was a state-of-the-art CNN trained on ImageNet, the top-1 performance of which is approximately 70% [17]. To qualitatively observe the effect of visual detection on naming, we visualise the images matched to different names of the same object using a t-Distributed Stochastic Neighbor Embedding (t-SNE) [18]. t-SNE is an unsupervised approach for embedding feature vectors into a low dimensional space; it has been shown to produce good visualisations for high-dimensional data. Our original feature space is 4096 dimensional and extracted from the second last layer of the VGG 16-layer CNN.

The t-SNE embedding for *Ursus maritimus* (polar bear) shown in Figure 5 divides the images into at least three dis-

[1] http://users.cecs.anu.edu.au/~u4534172/animal_naming_heat_maps/

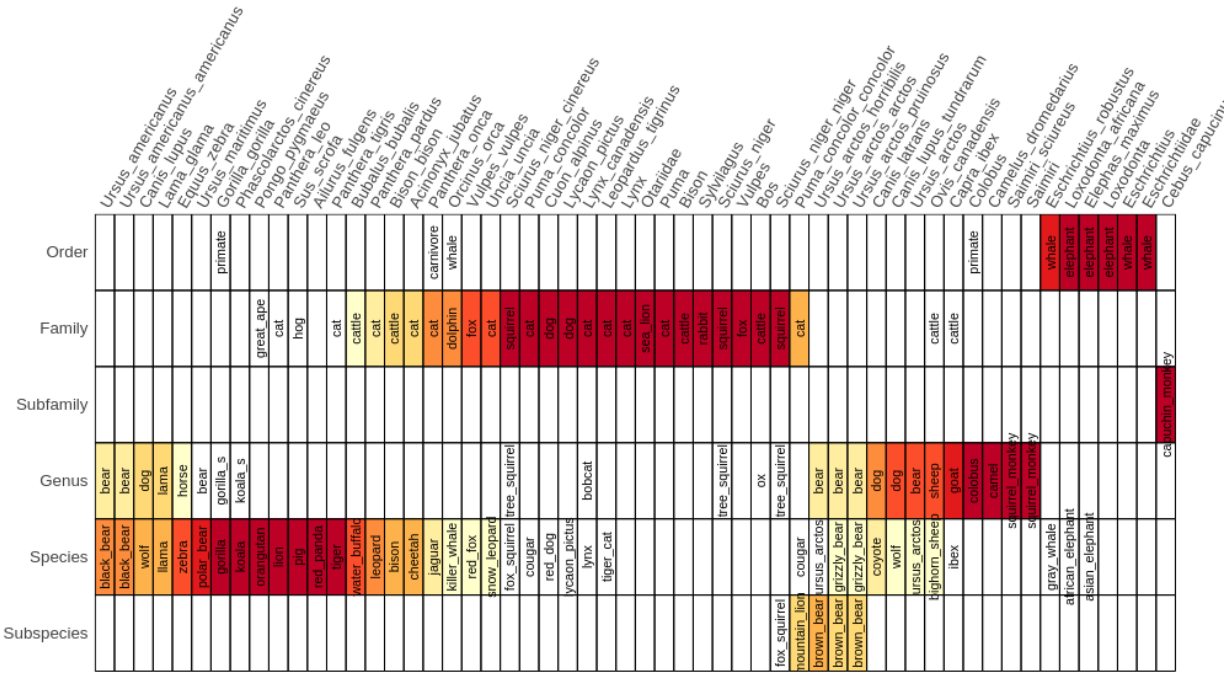

Figure 4: The level of names used to describe different mammals. The first row is the most general level, and the last is the most specific. Each column is a different animal corresponding to a visual classifier. Darker colours indicate larger counts; with columns normalised. Each column must have at least 20 detections to be included.

tinct regions. In the upper right are have polar bears in icy environments, in the lower left are polar bears swimming in the water and in the middle are polar bears in enclosures or other environments. The name *polar bear* is used relatively uniformly throughout the space, the name *bear* is primarily used in the middle section and generally not in the upper right hand corner where the polar bears are in icy environments. This indicates that people are less likely to name *Ursus maritimus* as *bears* when they are shown in a visually icy context.

The t-SNE embedding for *Cygnus atratus* (black swan), Figure 6, shows a number of classifier failures. All the images in this figure were classified as *Cygnus atratus*. The upper right of the figure shows white swans, the lower right shows ducks, while the left of the figure is mostly black swans. It is clear from this that the *Cygnus atratus* is typically described as a *black swan* and that the other names *duck* and *swan* are mostly spurious detections. The names *duck* and *swan* slipped past the caption matching procedure because they are different possible names for *black swan*.

It is interesting to note that the t-SNE embedding uses the same CNN features which are used for classification by the fully connected output layer, and that these features generally show separation between clusters of errors in the classifier. We should be able to remove a number of classifier errors by training more classifiers on animal classes and then setting a higher detection cut-off. This would reduce the recall, which would necessitate using a much larger dataset.

5. CONCLUSION AND DISCUSSIONS

In this work we have identified that assigning a single basic-level name to all visual concept is in-sufficient to capture the complexities of naming and categorisation. We show this on both, a dataset with manual ground-truth concepts (MSCOCO), and a much larger image-caption dataset (SBU) through automatic concept detection. Preliminary experiments on the MSCOCO dataset show that in some cases, such as collective naming and sub-concept naming, the effect of context on the name chosen can be interpreted and identified automatically. On the SBU dataset we demonstrate that an entirely automatic method is a feasible way to identify interpretable naming patterns on a large scale. Preliminary results indicate that mammals, reptiles and birds are typically described at different levels of specificity.

There are a large number of contextual factors that are thought to affect naming, many of which have remained relatively unexplored on a large scale. We propose to explore the effects of an individuals context on the names they choose, for example geographical region, tagging vs captioning and specialised knowledge. In doing so we hope to better understand how and why people name concepts they way they do. This should enable us to improve text generation systems, by tailoring their outputs to an individuals context.

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. This work was supported in part by the ARC under project DP140102185.

6. REFERENCES

[1] Integrated taxonomic information system (ITIS). `http://www.itis.gov`. Accessed: 2015-06-04.

[2] K. J. Archer and R. V. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 2008.

[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.

[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[5] J. Deng, J. Krause, a. C. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. *ICCV*, 2012.

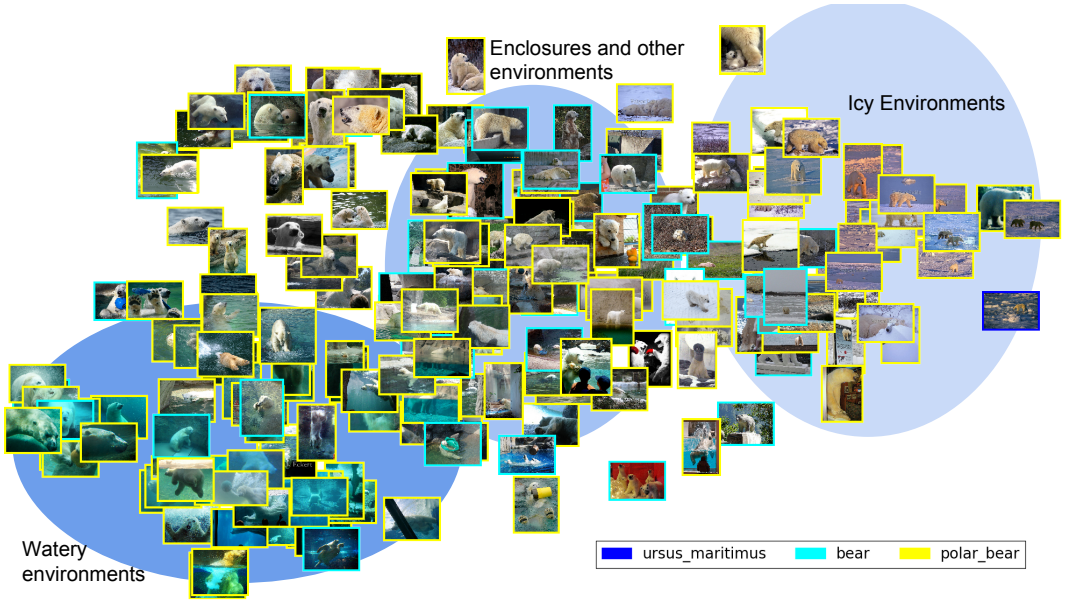

Figure 5: Embedding of *Ursus maritimus* (polar bear) images with CNN features using t-SNE. The colour of an image's border shows the name matched to that image.

Figure 6: Embedding of *Cygnus atratus* (black swan) images with CNN features using t-SNE. The colour of an image's border shows the name matched to that image.

[6] S. Feng, S. Ravi, R. Kumar, P. Kuznetsova, W. Liu, A. C. Berg, T. L. Berg, and Y. Choi. Refer-to-as Relations as Semantic Knowledge. *AAAI'15*, 2015.

[7] A. Frome, G. Corrado, J. Shlens, and S. Bengio. DeViSE: A Deep Visual-Semantic Embedding Model. 2013.

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

[9] G. Lako. Women, fire, and dangerous things. *U-niversity of Chicago Press, Chicago*, 1987.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. pages 740–755, 2014.

[11] A. Mathews, L. Xie, and X. He. Choosing basic-level concept names using visual and language context. *WACV*, 2015.

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

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

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

[15] E. Rosch. *Principles of categorization*. MIT Press, 1999.

[16] 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.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[18] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 2008.