A Poodle or a Dog? Evaluating Automatic Image Annotation Using Human Descriptions at Different Levels of Granularity

Josiah K. Wang[1] **Fei Yan**[2] **Ahmet Aker**[1] **Robert Gaizauskas**[1]

[1] Department of Computer Science, University of Sheffield, UK
[2] Centre for Vision, Speech and Signal Processing, University of Surrey, UK

{j.k.wang, ahmet.aker, r.gaizauskas}@sheffield.ac.uk f.yan@surrey.ac.uk

Abstract

Different people may describe the same object in different ways, and at varied levels of granularity (“poodle”, “dog”, “pet” or “animal”?) In this paper, we propose the idea of ‘granularity-aware’ groupings where semantically related concepts are grouped *across* different levels of granularity to capture the variation in how different people describe the same image content. The idea is demonstrated in the task of automatic image annotation, where these semantic groupings are used to alter the results of image annotation in a manner that affords different insights from its initial, category-independent rankings. The semantic groupings are also incorporated during evaluation against image descriptions written by humans. Our experiments show that semantic groupings result in image annotations that are more informative and flexible than without groupings, although being too flexible may result in image annotations that are less informative.

1 Introduction

Describing the content of an image is essential for various tasks such as image indexing and retrieval, and the organization and browsing of large image collections. Recent years have seen substantial progress in the field of visual object recognition, allowing systems to automatically annotate an image with a list of terms representing concepts depicted in the image. Fueled by advances in recognition algorithms and the availability of large scale datasets such as ImageNet (Deng et al., 2009), current systems are able to recognize thousands of object categories with reasonable accuracy, for example achieving an error rate of 0.11 in classifying $1,000$ categories in the ImageNet Large Scale Visual Recognition Challenge 2013 (ILSVRC13) (Russakovsky et al., 2013).

However, the ILSVRC13 classification challenge assumes each image is annotated with only *one* correct label, although systems are allowed up to five guesses per image to make the correct prediction (or rather, to match the ground truth label). The problem with this is that it becomes difficult to guess what the ‘correct’ label is, especially when many other categories can equally be considered correct. For instance, should a system label an image containing an instance of a dog (and possibly some other objects like a ball and a couch) as “dog”, “poodle”, “puppy”, “pet”, “domestic dog”, “canine” or even “animal” (in addition to “ball”, “tennis ball”, “toy”, “couch”, “sofa”, *etc.*)? The problem becomes even harder when the number of possible ways to refer to the same object instance increases, but the number of prediction slots to fill remains limited. With so many options from which to choose, how do we know what the ‘correct’ annotation is supposed to be?

In this paper, we take a *human-centric* view of the problem, motivated by the observation that humans are likely to be the end-users or consumers of such linguistic image annotations. In particular, we investigate the effects of grouping semantically related concepts that may refer to the same object instance in an image. Our work is related to the idea of *basic-level* categories (Biederman, 1995) in Linguistics, where most people have a natural preference to classify certain object categories at a particular level of granularity, *e.g.* “bird” instead of “sparrow” or “animal”. However, we argue that what one person considers

Proceedings of the 25th International Conference on Computational Linguistics, pages 38–45,
Dublin, Ireland, August 23-29 2014.

‘basic-level’ may not necessarily be ‘basic-level’ to another, depending on the person’s knowledge, expertise, interest, or the context of the task at hand. For example, Rorissa (2008) shows that users label groups of images and describe individual images differently with regards to the level of abstraction. The key idea behind our proposed ‘*granularity-aware*’ approach is to group semantically related categories *across* different levels of granularity to account for how different people would describe content in an image differently.

We demonstrate the benefits of the ‘granularity-aware’ approach by producing a re-ranking of visual classifier outputs for groups of concept nodes, *e.g.* WordNet synsets. The concept nodes are grouped across different levels of specificity within a semantic hierarchy (Section 3.1). This models better the richness of the vocabulary and lexical semantic relations in natural language. In this sense these groupings are used to alter the results of image annotation in a manner that affords different insights from its initial, category-independent rankings. For example, if the annotation mentions only “dog” but not “poodle”, a system ranking “poodle” at 1 and “dog” at 20 will have a lower overall score than a system ranking “dog” at 1, although both are equally correct. Grouping (“poodle” or “dog”) however will allow a fairer evaluation and comparison where both systems are now considered equally good. The ‘granularity-aware’ groupings will also be used in evaluating these re-rankings using textual descriptions written by humans, rather than a keyword-based gold-standard annotation. The hypothesis is that by modeling the variation in granularity levels for different concepts, we can gain a more informative insight as to how the output of image annotation systems can relate to how a person describes what he or she perceives in an image, and consequently produce image annotation systems that are more human-centric.

Overview. The remainder of the paper is organized as follows: Section 2 discusses related work. Section 3 describes our proposed ‘granularity-aware’ approach to group related concepts across different levels of granularity. It also discusses how to apply the idea both in automatic image annotation, by re-ranking noisy visual classifier outputs in a ‘granularity-aware’ manner, and in evaluation of classifier outputs against human descriptions of images. The results of the proposed method are reported in Section 4. Finally, Section 5 offers conclusions and proposes possible future work.

2 Related work

Work on automatic image annotation traditionally relies heavily on image datasets annotated with a fixed set of labels as training data. For example, Duygulu et al. (2002) investigated learning from images annotated with a set of keywords, posing the problem as a machine translation task between image regions and textual labels. Gupta and Davis (2008) includes some semantic information by incorporating prepositions and comparative adjectives, which also requires manual annotation as no such data is readily available. Recent work has moved beyond learning image annotation from constrained text labels to learning from real world texts, for example from news captions (Feng and Lapata, 2008) and sports articles (Socher and Fei-Fei, 2010).

There is also recent interest in treating texts as richer sources of information than just simple bags of keywords, for example with the use of semantic hierarchies for object recognition (Marszałek and Schmid, 2008; Deng et al., 2012b) and the inclusion of attributes for a richer representation (Lampert et al., 2009; Farhadi et al., 2009). Another line of recent work uses textual descriptions of images for various vision tasks, for example for recognizing butterfly species from butterfly descriptions (Wang et al., 2009) and discovering attributes from item descriptions on fashion shopping websites (Berg et al., 2010). There has also been interest in recent years in producing systems that annotate images with full sentences rather than just a list of terms (Kulkarni et al., 2011; Yang et al., 2011). We consider our work to complement the work of generating full sentences, as it is important to filter and select the most suitable object instances from noisy visual output. The shift from treating texts as mere labels to utilizing them as human-centric, richer forms of annotations is important to gain a better understanding of the processes underlying image and text understanding or interpretation.

Deng et al. (2012b) address the issue of granularity in a large number of object categories by allowing classifiers to output decisions at the optimum level in terms of being accurate and being informative, for example outputting “mammal” rather than “animal” while still being correct. Their work differs from ours in that the semantic hierarchy is used from *within* the visual classifier to make a decision about its output, rather than for evaluating existing outputs. More directly related to our work is recent work by Ordonez et al. (2013), which incorporates the notion of basic-level categories by modeling word 'naturalness' from text corpora on the web. While their focus is on obtaining the most 'natural' basic-level categories for different encyclopedic concepts as well as for image annotation, our emphasis is on accommodating different levels of naturalness, not just a single basic level. We adapt their model directly to our work, details of which will be discussed in Section 3.1.

3 Granularity-aware approach to image annotation

The proposed 'granularity-aware' approach to image annotation consists of several components. We first define semantic groupings of concepts by considering hypernym/hyponym relations in WordNet (Fellbaum, 1998) and also how people describe image content (Section 3.1). The groupings are then used to re-rank the output of a set of category-specific visual classifiers (Section 3.2), and also used to produce a grouped 'gold standard' from image captions (Section 3.3). The re-ranked output is then evaluated against the 'gold standard', and the initial rankings and 'granularity-aware' re-rankings are compared to gain a different insight into the visual classifiers' performance as human-centric image annotation systems.

3.1 Semantic grouping across different granularity levels

The goal of semantic grouping is to aggregate related concepts such that all members of the group refer to the same instance of an object, even across different specificity levels. In particular, we exploit the hypernym/hyponym hierarchy of WordNet (Fellbaum, 1998) for this task. WordNet is also the natural choice as it pairs well with our visual classifiers which are trained on ImageNet (Deng et al., 2009) categories, or *synsets*.

The WordNet hypernym hierarchy alone is insufficient for semantic grouping as we still need a way to determine what constitutes a reasonable group, *e.g.* putting all categories into a single "entity" group is technically correct but uninformative. For this, we draw inspiration from previous work by Ordonez et al. (2013), where a 'word naturalness' measure is proposed to reflect how people typically describe image content. More specifically, we adapt for our purposes their proposed approach of mapping encyclopedic concepts to basic-level concepts (mapping "*Grampus griseus*" to the more 'natural' "dolphin"). In this approach, the task is defined as learning a translation function $\tau(v, \lambda) : V \mapsto W$ that best maps a node v to a hypernym node w which optimizes a trade-off between the 'naturalness' of w (how likely a person is to use w to describe something) and the distance between v and w (to constrain the translation from being too general, *e.g.* "entity"), with the parameter λ controlling this trade-off between naturalness and specificity. Formally, $\tau(v, \lambda)$ is defined as:

$$\tau(v, \lambda) = \underset{w \,\in\, \Pi(v)}{\arg\max} \left[\, \lambda\, \phi(w) - (1 - \lambda)\, \psi(w, v) \,\right] \tag{1}$$

where $\Pi(v)$ is the set of hypernyms for v (including v), $\phi(w)$ is naturalness measure for node w, and $\psi(w, v)$ is the number of edges separating nodes w and v in the hypernym structure of WordNet.

For our work, all synsets that map to a common hypernym w are clustered as a single semantic group G_w^λ:

$$G_w^\lambda = \{v : \forall_v\ \tau(v, \lambda) = w\} \tag{2}$$

In this sense, the parameter $\lambda \in [0, 1]$ essentially also controls the average size of the groups: $\lambda = 0$ results in no groupings, while $\lambda = 1$ results in synsets being grouped with their most 'natural' hypernym, giving the largest possible difference in the levels of granularity within each group.

Estimating the naturalness function using Flickr. Ordonez et al. (2013) use n-gram counts of the Google IT corpus (Brants and Franz, 2006) as an estimate for term naturalness $\phi(w)$. Although large, the corpus might not be optimal as it is a general corpus and may not necessarily mirror how people describe image content. Thus, we explore a different corpus that (i) better reflects how humans describe image content; (ii) is sufficiently large for a reasonable estimate of $\phi(w)$. The Yahoo! Webscope Yahoo Flickr Creative Commons 100M (YFCC-100M) dataset (Yahoo! Webscope, 2014) fits these criteria with 100 million images containing image captions written by users. Hence, we compute term occurrence statistics from the title, description, and user tags of images from this dataset. Following Ordonez et al., we measure $\phi(w)$ as the maximum log count of term occurrences for all terms appearing in synset w.

Internal nodes. Unlike Ordonez et al. (2013), we do not constrain v to be a leaf node, but instead also allow for internal nodes to be translated to one of their hypernyms. We could choose to limit visual recognition to leaf nodes and estimate the visual content of internal nodes by aggregating the outputs from all its leaf nodes, as done by Ordonez et al. (2013). However, since the example images in ImageNet are obtained for internal nodes pretty much in the same way as leaf nodes (by querying "dog" rather than by combining images from "poodle", "terrier" and "border collie") (Deng et al., 2009), the visual models learnt from images at internal nodes may capture different kinds of patterns than from their hyponyms. For example, a model trained with ImageNet examples of "dog" might capture some higher-level information that may otherwise not be captured by merely accumulating the outputs of the leaf nodes under it, and *vice versa*.

3.2 Re-ranking of visual classifier output

The visual classifier used in our experiments (Section 4.2) outputs a Platt-scaled (Platt, 2000) confidence value for each synset estimating the probability of the synset being depicted in a given image. The classifier outputs are then ranked in descending order of these probability values, and are treated as image annotation labels.

As mentioned, these rankings do not take into consideration that some of these synsets are semantically related. Thus, we aggregate classifier outputs within our semantic groupings (Section 3.1), and then re-rank the scores of each *grouped* classifier. Formally, the new score of a classifier c, $\rho_c(G_w^\lambda)$, for a semantic group G_w^λ is defined as:

$$\rho_c(G_w^\lambda) = \max_{v \in G_w^\lambda} \; p_c(v) \tag{3}$$

where v is a synset from the semantic group G_w^λ, and $p_c(v)$ is the original probability estimate of classifier c for synset v. I.e., the probability of the most probable synset in the group is taken as the probability of the group.

To enable comparison of the rankings against a gold standard keyword annotation, a word label is also generated for each semantic group. We assign as the semantic group's label $\ell(G_w^\lambda)$ the first term of synset w, the common hypernym node to which members of the group best translates. Note that the term merely acts a label for evaluation purposes and should not be treated as a word in a traditional sense. We also merge semantic groups with the same label to account for polysemes/homonyms, again taking the maximum of ρ_c among the semantic groups as the new score.

The semantic grouping of synsets is performed independently of visual classifier output. As such, we only need to train each visual classifier *once* for each synset, without requiring re-training for different groupings since we only aggregate the *output* of the visual classifiers. This allows for more flexibility since the output for each semantic group is only aggregated at *evaluation time*.

3.3 Evaluation using human descriptions

The image dataset used in our experiments (Section 4.1) is annotated with five full-sentence captions per image but *not* keyword labels. Although an option would be to obtain keyword annotations via crowdsourcing, it is time consuming and expensive and also requires validating the annotation quality. Instead, we exploit the existing full-sentence captions from the dataset to automatically generate a gold standard keyword annotation for evaluating our ranked classifier outputs. The use of such captions is also in line with our goal of making the evaluation of image annotation systems more human-centric. For each caption, we extract nouns using the open source tool FreeLing (Padrö and Stanilovsky, 2012).

						λ					
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Semantic Grouping	0.3450	0.3450	0.3548	0.3735	0.4025	0.4417	0.4562	0.4702	0.4834	0.5059	0.5395
Random Grouping	0.3450	0.3450	0.3493	0.3529	0.3585	0.3689	0.3823	0.4067	0.4241	0.4359	0.4467
Number of groups	1294	1294	1237	1105	949	817	693	570	474	419	368

Table 1: Results of re-ranking with semantic groupings. The first two rows show the average NDCG scores for the proposed groupings and the random baseline groupings, for different groupings formed by varying λ. The bottom row shows the number of semantic groups formed for different values of λ.

For each image, each noun is assigned an individual relevance score, which is the number of captions that mentions the noun. This upweights important objects while downweighting less important objects (or errors from the annotator or the parser). The result is a list of nouns that humans use to describe objects present in the image, each weighted by its relevance score. We assume nouns that appear in the same WordNet synset ("bicycle" and "bike") are synonyms and that they refer to the same object instance in the image. Hence, we group them as a single label-group, with the relevance score taken to be the maximum relevance score among the nouns in the group.

Since there are only five captions per image, the proposed approach will result in a *sparse* set of keywords. This mirrors the problem described in Section 1 where systems have to 'guess' the so-called 'correct' labels, thus allowing us to demonstrate the effectiveness of our 'granularity-aware' re-rankings.

In order to compare the annotations against the re-rankings, we will need to map the keywords to the semantic groupings. This is done by matching the nouns to any of the terms in a semantic group, with a corresponding label $\ell(G_w^\lambda)$ for each group (Section 3.2). Nouns assigned the same label are merged, with the new relevance score being the maximum relevance score among the nouns. If a noun matches more than one semantic group (polyseme/homonym), we treat all groups as relevant and divide the relevance score uniformly among the groups. Evaluation is then performed by matching the semantic group labels against the image annotation output.

4 Experimental evaluation

Our proposed method is evaluated on the dataset and categories as will be described in Section 4.1, by re-ranking the output of the visual classifiers in Section 4.2. The effects of semantic groupings are explored using different settings of λ (see Section 3.1).

Baseline. To ensure any improvements in scores are not purely as a result of having a shorter list of concepts to rank, we compare the results to a set of baseline groupings where synsets are grouped in a random manner. For a fair comparison the baselines contain the same number of groups and cluster size distributions as our semantic groupings.

4.1 Dataset and Object Categories

The Flickr8k dataset (Hodosh et al., 2013) is used in our image annotation experiments. The dataset contains 8,091 images, each annotated with five textual descriptions. To demonstrate the notion of granularity in large-scale object hierarchies, we use as object categories synset nodes from WordNet (Fellbaum, 1998). Ideally, we would like to be able to train visual classifiers for all synset categories in ImageNet (Deng et al., 2009). However, we limit the categories to only synsets with terms occurring in the textual descriptions of the Flickr8k dataset to reduce computational complexity, and regard the use of more categories as future work. This results in a total of 1,372 synsets to be used in our experiments. The synsets include both *leaf nodes* as well as *internal nodes* in the WordNet hierarchy.

4.2 Visual classifier

Deep learning (LeCun et al., 1989; Hinton and Salakhutdinov, 2006) based approaches have become popular in visual recognition following the success of deep convolutional neural networks

(CNN) (Krizhevsky et al., 2012) in the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC12) (Deng et al., 2012a). Donahue et al. (2013) report that features extracted from the activation of a deep CNN trained in a fully supervised fashion can also be re-purposed to novel generic tasks that differ significantly from the original task. Inspired by Donahue et al. (2013), we extract such activation as feature for ImageNet images that correspond to the 1,372 synsets, and train binary classifiers to detect the presence of the synsets in the images of Flickr8k. More specifically, we use as our training set the 1,571,576 ImageNet images in the 1,372 synsets, where a random sample of 5,000 images serves as negative examples, and as our test set the 8,091 images in Flickr8k. For each image in both sets, we extracted activation of a pre-trained CNN model as its feature. The model is a reference implementation of the structure proposed in Krizhevsky et al. (2012) with minor modifications, and is made publicly available through the Caffe project (Jia, 2013). It is shown in Donahue et al. (2013) that the activation of layer 6 of the CNN performs the best for novel tasks. Our study on a toy example with 10 ImageNet synsets however suggests that the activation of layer 7 has a small edge. Once the 4,096 dimensional activation of layer 7 is extracted for both training and test sets, 1,372 binary classifiers are trained and applied using LIBSVM (Chang and Lin, 2011), which give probability estimates for the test images. For each image, the 1,372 classifiers are then ranked in order of their probability estimates.

4.3 Evaluation measure

The systems are evaluated using the Normalized Discounted Cumulative Gain (NDCG) (Wang et al., 2013) measure. This measure is commonly used in Information Retrieval (IR) to evaluate ranked retrieval results where each document is assigned a relevance score. This measure favours rankings where the most relevant items are ranked ahead of less relevant items, and does not penalize irrelevant items.

The NDCG at position k, $NDCG_k$, for a set of test images $\mathcal{I}$ is defined as:

$$NDCG_k(\mathcal{I}) = \frac{1}{|\mathcal{I}|} \sum_{i=1}^{|\mathcal{I}|} \frac{1}{IDCG_k(i)} \sum_{p=1}^{k} \frac{2^{R_p} - 1}{\log_2(1+p)} \tag{4}$$

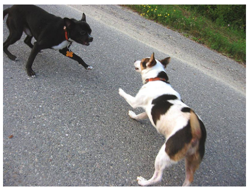
where R_p is the relevance score of the concept at position p, and $IDCG_k(i)$ is the ideal discounted cumulative gain for a perfect ranking algorithm at position k, which normalizes the overall measure to be between 0.0 to 1.0. This makes the scores comparable across rankings regardless of the number of synset groups involved. For each grouping, we report the results of $NDCG_k$ for the largest possible k (*i.e.* the number of synset groups), which gives the overall performance of the rankings.

4.4 Results

Table 1 shows the results of re-ranking the output of the visual classifiers (Section 4.2), with different semantic groupings formed by varying λ. The effects of the proposed groupings is apparent when compared to the random baseline groupings. As we increase the value of λ (allowing groups to have a larger range of granularity), the NDCG scores also consistently increase. However, higher NDCG scores do not necessarily equate to better groupings, as semantic groups with too much flexibility in granularity levels may end up being less informative, for example by annotating a “being” in an image. The informativeness of the groupings is a subjective issue depending on the context, and makes an interesting open question. To provide insight into the effects of our groupings, Figure 1 shows an example where at low levels of λ (rigid flexibility), the various dog species are highly ranked but none of them is considered relevant by the evaluation system. However, at $\lambda = 0.5$ most dog species are grouped as a “dog” semantic group resulting in a highly relevant prediction, while at the same time allowing the “sidewalk” group to rise higher in the rankings. At higher levels of λ, however, the semantic groupings become less informative when superordinate groups like “being”, “artifact” and “equipment” are formed, suggesting that higher flexibility with granularity levels may not always be more informative.

5 Conclusions and future work

We presented a ‘granularity-aware’ approach to grouping semantically related concepts across different levels of granularity, taking into consideration that different people describe the same thing in different



dog (5)
road (2)
pavement (1)
street (1)

$\lambda = \mathbf{0.0}$	$\lambda = \mathbf{0.3}$	$\lambda = \mathbf{0.5}$	$\lambda = \mathbf{0.8}$	$\lambda = \mathbf{1.0}$
score: 0.241	score: 0.480	score: 0.941	score: 0.943	score: 0.944
beagle	beagle	***dog***	***dog***	***dog***
boston_terrier	boston_terrier	animal	animal	animal
corgi	***dog***	leash	leash	leash
basset	basset	kitten	being	being
hound	hound	pet	bird	bird
spaniel	spaniel	penguin	***sidewalk***	***sidewalk***
border_collie	border_collie	***sidewalk***	cat	cat
terrier	terrier	cat	artifact	artifact
dachshund	dachshund	artifact	student	sport
pup	pup	person	baseball	chair
st_bernard	bulldog	student	chair	child
bulldog	springer_spaniel	goat	child	device
springer_spaniel	leash	livestock	frisbee	ball
leash	kitten	rabbit	ball	slope
kitten	pet	duck	slope	equipment
pet	penguin	baseball	equipment	fabric
dog	***sidewalk***	chair	fabric	rug
sheepdog	doberman	child	rug	seat
penguin	collie	frisbee	seat	support
sidewalk	cat	spectator	support	stick

Figure 1: Example re-ranking of our visual classifier by semantic groupings, for selected values of λ. Words directly below the image indicate the 'gold standard' nouns extracted automatically from its corresponding five captions. The number next to each noun indicate its relevance score. For each re-ranking, we show the labels representing the semantic groupings. ***Italicized labels*** indicate a match with the (grouped) 'gold standard' nouns (see Section 3.3).

ways, and at varied levels of specificity. To gain insight into the effects of our semantic groupings on human-centric applications, the proposed idea was investigated in the context of re-ranking the output of visual classifiers, and was also incorporated during evaluation against human descriptions. We found that although the groupings help provide a more human-centric and flexible image annotation system, too much flexibility may result in an uninformative image annotation system. Future work could include (i) exploring different ways of grouping concepts; (ii) incorporating the output of visual classifiers to improve both groupings and rankings; (iii) using information from more textual sources to improve image annotation; (iv) taking the approach further to generate full sentence annotations. We believe that these steps are important to bridge the semantic gap between computer vision and natural language.

Acknowledgements

The authors would like to acknowledge funding from the EU CHIST-ERA D2K programme, EPSRC grant reference: EP/K01904X/1.

References

Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *Proceedings of ECCV*, volume 1, pages 663–676.

Irving Biederman. 1995. Visual object recognition. In S. F. Kosslyn and D. N. Osherson, editors, *An Invitation to Cognitive Science, 2nd edition, Volume 2, Visual Cognition*, pages 121–165. MIT Press.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1. In *Linguistic Data Consortium*.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of CVPR*.

Jia Deng, Alexander C. Berg, Sanjeev Satheesh, Hao Su, Aditya Khosla, and Li Fei-Fei. 2012a. ImageNet large scale visual recognition challenge (ILSVRC) 2012. `http://image-net.org/challenges/LSVRC/2012/`.

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

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2013. DeCAF: A deep convolutional activation feature for generic visual recognition. arXiv:1310.1531 [cs.CV].

Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David A. Forsyth. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of ECCV*, pages 97–112.

Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. 2009. Describing objects by their attributes. In *Proceedings of CVPR*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Yansong Feng and Mirella Lapata. 2008. Automatic image annotation using auxiliary text information. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 272–280. Association for Computational Linguistics.

Abhinav Gupta and Larry S. Davis. 2008. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *Proceedings of ECCV*, pages 16–29.

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of CVPR*.

Chris H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of CVPR*.

Y. LeCun, B. Boser, J. Denker, D. Henerson, R. Howard, W. Hubbard, and L. Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.

Marcin Marszałek and Cordelia Schmid. 2008. Constructing category hierarchies for visual recognition. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Proceedings of ECCV*, volume 5305 of *Lecture Notes in Computer Science*, pages 479–491. Springer Berlin Heidelberg.

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

Lluïs Padrö and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference*, LREC '12, Istanbul, Turkey, May. ELRA.

John C. Platt. 2000. Probabilities for SV machines. *Advances in Large-Margin Classifiers*, pages 61–74.

Abebe Rorissa. 2008. User-generated descriptions of individual images versus labels of groups of images: A comparison using basic level theory. *Information Processing and Management*, 44(5):1741–1753.

Olga Russakovsky, Jia Deng, Jonathan Krause, Alexander C. Berg, and Li Fei-Fei. 2013. ImageNet large scale visual recognition challenge (ILSVRC) 2013. `http://image-net.org/challenges/LSVRC/2013/results.php`.

Richard Socher and Li Fei-Fei. 2010. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Proceedings of CVPR*, pages 966–973.

Josiah Wang, Katja Markert, and Mark Everingham. 2009. Learning models for object recognition from natural language descriptions. In *Proceedings of BMVC*.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*.

Yahoo! Webscope. 2014. Yahoo! Webscope dataset YFCC-100M. `http://labs.yahoo.com/Academic_Relations`.

Yezhou Yang, Ching Lik Teo, Hal Daumé, III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of EMNLP*, pages 444–454.