What Helps Where – And Why? Semantic Relatedness for Knowledge Transfer

Marcus Rohrbach[1,2,*] Michael Stark[1] György Szarvas[1,†] Iryna Gurevych[1] Bernt Schiele[1,2]
[1]Department of Computer Science, TU Darmstadt [2]MPI Informatics, Saarbrücken

Abstract

Remarkable performance has been reported to recognize single object classes. Scalability to large numbers of classes however remains an important challenge for today's recognition methods. Several authors have promoted knowledge transfer between classes as a key ingredient to address this challenge. However, in previous work the decision which knowledge to transfer has required either manual supervision or at least a few training examples limiting the scalability of these approaches. In this work we explicitly address the question of how to automatically decide which information to transfer between classes without the need of any human intervention. For this we tap into linguistic knowledge bases to provide the semantic link between sources (what) *and targets* (where) *of knowledge transfer. We provide a rigorous experimental evaluation of different knowledge bases and state-of-the-art techniques from Natural Language Processing which goes far beyond the limited use of language in related work. We also give insights into the applicability* (why) *of different knowledge sources and similarity measures for knowledge transfer.*

1. Introduction

Impressive recognition results were reported on a variety of object classes based on robust local features and powerful machine learning techniques. However, these approaches often rely on large amounts of training data limiting their scalability. It is clearly desirable to address the more challenging task of simultaneous recognition of many object classes without the need for large training corpora. Reusing already acquired information by transferring knowledge between object classes has been suggested as a promising way to enable recognition of objects for which training data are scarce. The question what information to re-use in which context has been answered mostly by manual supervision [19, 27] or by providing a few bootstrap training examples [4, 14], limiting the applicability of these approaches.

*Supported by the German National Academic Foundation.

†On leave from the Research Group on Artificial Intelligence of the Hungarian Academy of Sciences.

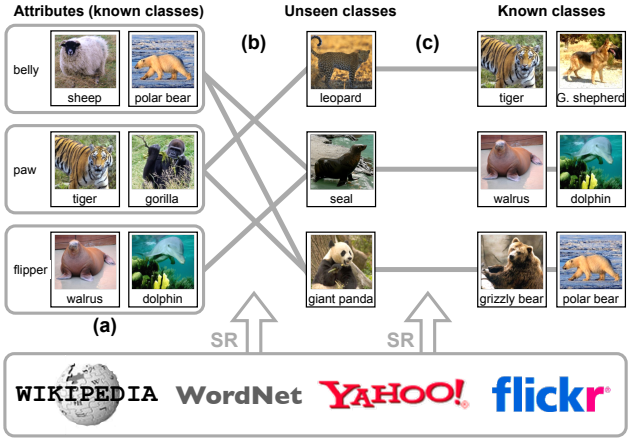


Figure 1. Inter-object class knowledge transfer. (a) Attribute inventory. Semantic Relatedness (SR) used to determine object class-attribute associations (b) and inter-object class similarity (c).

The main objective of our work is to extend such knowledge transfer approaches for object class recognition by significantly reducing the amount of needed manual supervision and training data. We do this by tapping into additional sources of information, namely linguistic knowledge bases, in order to provide the missing semantic link between sources (known object classes) and targets (unseen object classes) of knowledge transfer.

We choose two different models as the starting point of our work. In both models, knowledge transfer is realized by representing unseen object classes relative to known ones. The first model is based on an inventory of descriptive attributes (e.g. "belly", "paw", or "flipper", see Fig. 1(a)). For a given class, each attribute can be either active or inactive, resulting in a characteristic association signature for that class (e.g. "seal" is associated to "belly" and "flipper", see Fig. 1(b)). The second model is based on similarities between an unseen object class and known object classes (e.g. "leopard" is most similar to "tiger" and "G. shepherd", see Fig. 1(c)). For both models we establish the semantic link between known and unseen classes by semantic relatedness (SR), which we measure using linguistic knowledge bases.

Based on these models, our study has two goals. The first goal is to better understand "how far we can get" in

general with replacing manual supervision or seed training data by information acquired automatically from linguistic knowledge bases. The second goal is to evaluate the impact of particular choices of linguistic knowledge bases and semantic relatedness measures and to provide insights into their usefulness for different tasks. In contrast to most related work, we go beyond simple use of tags and image captions, and apply various state-of-the-art Natural Language Processing techniques which have, to our knowledge, not been used in a similar context in computer vision.

The main contributions of our paper are as follows. First, we provide the missing semantic link for inter-object class knowledge transfer by using linguistic knowledge bases, based on two models (attribute-based and direct similarity-based). Second, for the attribute-based model, we explore different levels of automation in the knowledge transfer process. We not only determine the strengths of associations between object classes and attributes automatically using semantic relatedness, but also compile the attribute inventory automatically (see Fig. 1(a)). Third, we provide a rigorous experimental evaluation of different knowledge bases (such as WordNet [12], Wikipedia, or the World Wide Web) and semantic relatedness measures and quantify their usefulness in the context of an object class recognition task. Fourth, we discuss the major differences, together with their possible reasons, between the examined knowledge bases and semantic relatedness measures. We believe that many of these insights are transferable to other vision tasks and may motivate a move-away from using WordNet as the default option for extracting semantic information.

The remainder of this paper is organized as follows. After a review of related work (Sect. 2), we first introduce our model (Sect. 3) and the linguistic knowledge bases / semantic relatedness measures (Sect. 4) used in this study. We then give experimental results (Sect. 5) and conclude with an outlook (Sect. 6).

2. Related work

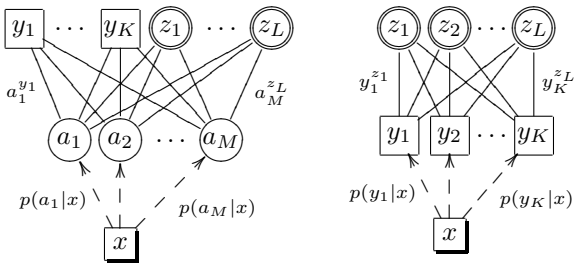
Due to the increasing need for scalable recognition, knowledge transfer between object classes has become an important topic in the vision literature. Amongst other directions, attribute-based representations have gained popularity recently by introducing an interpretable level of indirection between object classes [13, 18, 29]. As attribute activations can characterize object classes without using reference exemplars they lend themselves to *zero-shot* classification of previously unseen object classes. Lampert et al. [19] present zero-shot classification schemes based on attributes, where the associations between attributes and object classes are obtained using manual supervision by human subjects. Farhadi et al. [10] advocate a paradigm shift from "naming" (by object classes) to "describing" (by attributes), distinguishing among common, discriminating, unusual, and unexpected attributes for object classes. In their work, zero-shot classification is phrased as a nearest-neighbor problem: a test image is classified as the most similar object class w.r.t. attribute descriptions. While the first of our object class representations is based on [19], we replace manual supervision with information extracted automatically from linguistic knowledge bases.

Other representations of transferable knowledge focus on more abstract notions of common or discriminating aspects between object classes, as practiced by the hierarchical classification schemes of Marszalek and Schmid [22] and Zweig and Weinshall [32]. Similar in spirit, several works transfer knowledge in the form of learned distance metrics [3, 14, 28] or object class priors [11, 27]. Bart and Ullman [4] encode instances of previously unknown classes as collections of "familiar" classifier responses, i.e., similarities to known classes, and apply a nearest-neighbor scheme for classification. While the second of our object class representations is also based on such direct similarities, we extend this approach to zero-shot classification. That is we do not require the availability of any reference exemplar for unknown classes, but use information obtained from linguistic knowledge bases instead.

Obviously, using vision and language resources in combination promises mutual benefits for both and has been pursued actively in the literature. Approaches range from determining the "visualness" of language entities [2, 6], over the construction of visually grounded ontologies [25, 30] to joint models of images, image tags, and image caption text [1, 20]. [9] adds search engine hit counts for learning object-background co-occurrence statistics. While these approaches clearly demonstrate the benefit of using visual and language information together, most are still limited in the knowledge sources used (mostly WordNet or image tags) and the similarity measures applied (mostly path-length based measures). To our knowledge, our work is the first to give an in-depth exploration of both the possible knowledge bases and semantic relatedness measures for computer vision in general, and more specifically for knowledge transfer between object classes.

3. Two models for knowledge transfer

The main objective of our work is to automatically decide which knowledge to transfer between object classes by tapping into different language resources. We therefore extend two previous lines of work to enable zero-shot object class recognition, namely attribute-based recognition [19] and recognition based on direct similarities [4] between object classes. Both approaches model the relationship between known classes $y_1, \ldots, y_K$ and unseen classes $z_1, \ldots, z_L$. In attribute-based classification, an intermediate layer of descriptive attributes $a_1, \ldots, a_M$ serves as a level of indirection between known and unseen classes (see



(a) Attribute-based [19] (b) Direct similarity-based

Figure 2. Two models for zero-shot object classification. See Sect. 3 for discussions.

Fig. 2(a)). In recognition based on direct similarities, the known classes $y_1, \ldots, y_K$ serve directly as mediators for the unseen classes $z_1, \ldots, z_L$ (see Fig. 2(b)). The following subsections describes these models and our extensions to enable zero-shot learning in more detail.

3.1. Attribute-based classification

Attribute-based classification models object classes relative to an inventory of descriptive attributes. For a given class, each attribute can be either active or inactive, resulting in a characteristic association signature for that class. Fig. 2(a) gives a schematic overview of the Direct Attribute Prediction model (DAP) suggested by [19], which has been shown to yield better classification performance than Indirect Attribute Prediction (IAP).

Following the probabilistic formulation of the DAP model in [19], let $a^y = (a_1^y, \ldots, a_M^y)$ be a vector of binary associations $a_m^y \in \{0, 1\}$ between attributes a_m and training object classes y. A classifier for attribute a_m, trained by labeling all images of all classes for which $a_m^y = 1$ as positive and the rest as negative training examples, can provide an estimate of the posterior probability $p(a_m|x)$ of that attribute being present in image x. Mutual independence yields $p(a|x) = \prod_{m=1}^{M} p(a_m|x)$ for multiple attributes.

In order to transfer attribute knowledge to an unknown class z, we again assume a binary vector a^z for which $p(a|z) = 1$ for $a = a^z$ and 0 otherwise. The posterior probability of class z being present in image x is then obtained by marginalizing over all possible attribute associations a, using Bayes' rule $p(z|a^z) = \frac{p(a^z|z)p(z)}{p(a^z)} = \frac{p(z)}{p(a^z)}$:

$$p(z|x) = \sum_{a \in \{0,1\}^M} p(z|a)p(a|x) = \frac{p(z)}{p(a^z)} \prod_{m=1}^{M} p(a_m|x)^{a_m^z} \quad (1)$$

Assuming identical class priors $p(z)$ and a factorial distribution for $p(a) = \prod_{m=1}^{M} p(a_m)$, we obtain

$$p(z|x) \propto \prod_{m=1}^{M} \left(\frac{p(a_m|x)}{p(a_m)} \right)^{a_m^z} \quad (2)$$

Attribute priors can be approximated by empirical means over the training classes $p(a_m) = \frac{1}{K} \sum_{k=1}^{K} a_m^{y_k}$ or set to $\frac{1}{2}$ [19]. Classifying an image x according to test classes z_L uses MAP prediction $\text{argmax}_{l=1,\ldots,L} p(z_l|x)$.

This leaves us with estimating the class-attribute associations both for the known classes a_m^y as well as for the unknown classes a_m^z. In [19] human judgments of ten subjects [16, 23] are used as the basis of these associations. While this has led to promising recognition results for unseen object classes, the main drawback of the approach is that it requires labor-intensive manual labeling to be applicable to a new domain (new sets of classes and attributes). As the main objective of our work is to reduce this dependency on human labeling effort, in the following subsections we discuss three extensions of the attribute-based model that alleviate this limitation: 1) mining the strengths of associations between classes and attributes by measuring their semantic relatedness using linguistic knowledge bases, 2) finding these attributes automatically, and 3) using the object classes themselves as *objectness attributes* (similarities of classes to each other), thereby eliminating the need for attribute-finding.

Mining object class-attribute associations. Our first extension of the attribute-based classification scheme taps into various language resources in order to automatically mine object class-attribute associations (see Fig. 1(b)). For this purpose, Sect. 4 introduces both different linguistic knowledge bases as well as several text-based semantic relatedness (SR) measures, which quantify the strength of relatedness between pairs of concepts (here class-attribute pairs). Note that we have to (manually) map full-text attribute descriptions designed for human comprehension [16, 23] to concise terms which we can use as input to SR measures, which is inherently prone to drop information and is also susceptible to introducing noise (see Sect. 5).

Mining attributes. While mining class-attribute associations reduces the amount of manual supervision needed significantly, it is still relying on the definition of an appropriate inventory of descriptive attributes to start from. Ideally, these attributes should be able to discriminate between object classes (being associated to some but not all of them), provide sufficient coverage (all classes have at least a single attribute association), and be correlated to visual object class properties that can be observed in images. The creation of an appropriate attribute inventory is clearly a non-trivial task and has undoubtedly required careful engineering by [16, 23]. Our second extension thus aims at avoiding this manual intervention by mining attributes from language resources (see Fig. 1(a)).

In our experiments, and in line with part-based modeling in the vision literature, we found *part attributes* (e.g. flipper for animals, wheel for vehicles) to meet the above described desired characteristics. Part attributes can be mined in var-

ious ways from language resources, most obviously by using the explicit part relations encoded in WordNet. Here, we collect parts of all object class concepts of interest, including the parts of sub- and super-concepts recursively, resulting in 74 mined attributes[1] (compared to 85 manually defined ones).

Objectness as attributes. Our third extension uses object class names as attribute signifiers, characterizing the extent to which object classes are alike, which we casually denote *objectness*. "Giraffness", for example, then characterizes a group of object classes that are sufficiently similar to the giraffe class. As for attributes in general, several objectness attributes can be combined to yield a more precise description of an object class. Similarity is again determined from semantic relatedness measures (see Fig. 1(c)).

Interestingly, objectness attributes tend to encode complementary information to more generic ones, such as part attributes, by nature: while part attributes are often shared across diverse object classes, objectness attributes tend to form groups of classes that are highly related. As an example, consider the "giant panda" (bear) class in Fig. 1: it shares the "paw" attribute with diverse classes, such as "tiger" and "gorilla". On the other hand, similarity on the level of object classes yields "grizzly bear" and "polar bear" as the most similar classes.

3.2. Direct similarity-based classification

Similar in spirit to using objectness as attributes (see Sect. 3.1), but bypassing the level of indirection introduced by the attribute layer, we can use existing classifiers for known object classes $y_1, \ldots, y_K$ directly for zero-shot classification of unseen classes $z_1, \ldots, z_L$ (see Fig. 1(c), 2(b)). The particular choice of using a trained classifier for class y_k for classifying test class z_l depends on the similarity between the two, which can again be determined by using semantic relatedness measures.

This direct similarity model can be interpreted as a DAP model with $M = K$ attributes, where each attribute corresponds to exactly one training class y_k. We thus train classifiers for each class y_k to provide estimates of $p(y_k|x)$ for a test image x. In analogy to attribute-based classification (see Eq. (2)), the posterior of test image x is given as

$$p(z|x) \propto \prod_{k=1}^{K} \left(\frac{p(y_k|x)}{p(y_k)} \right)^{y_k^z}, \qquad (3)$$

where y_k^z can be a binary association variable between the known class y_k and an unknown class z as for attribute-based classification (see Eq. (2)). We found empirically that using continuous weights is beneficial for performance and thus report results consistently for $y_k^z = w_{y_k}^z / \sum_{i=1}^{K} w_{y_i}^z$, assuming continuous weights w_y^z between z and y_k. Similarly, we restrict the considered classifiers to the $K = 5$ most similar ones in all experiments.

[1] All software for computing object class-attribute associations from linguistic knowledge bases and obtained intermediate results (lists of mined attributes, object class-attribute associations) are publicly available on our web pages.

4. Text-based semantic relatedness

In this section, we describe the various linguistic knowledge bases and semantic relatedness (SR) measures that exploit natural language resources to gather information on the (visual) similarity of object classes or object class-attribute associations.[1] The most widely used resources in Natural Language Processing to calculate SR of concepts are without doubt WordNet (as the largest machine readable expert-created language ontology), Wikipedia (as the largest online encyclopedia), and the World Wide Web (as the largest public text collection one can use). For each of these resources we select a single, representative semantic relatedness measure designed to operate on that particular linguistic resource, given that the measure 1) has been widely used in previous studies, and 2) is generally regarded as competitive in terms of performance compared to other measures operating on the same resource.

WordNet and path length-based SR measures. Language ontologies and wordnets in particular are the most popular sources of machine readable information about a language, representing lexicalized concepts, synonymy, concept definitions, and various semantic relations. WordNet [12] is a large scale lexical database of the English language, originally intended as model of human lexical memory. English words are organized into concepts (synonym sets or synsets) according to synonymy and various lexical and semantic relations are provided between these concepts. Due to its impressive size (over 100,000 concepts) and richness in encoded semantic relations, WordNet became the most important expert-created source of language information.

SR measures on WordNet [7] mostly use its graph structure (i.e., the encoded relations) to determine the path length between concepts or the shared information content of concepts. We use the similarity measure proposed by Lin [21] that defines the similarity of two concepts c_1 and c_2 as $sim_{Lin}(c_1, c_2) = \frac{2 * IC(lcs)}{IC(c_1)+IC(c_2)}$, where lcs denotes the lowest common subsumer of the two concepts in the WordNet hierarchy (i.e., the lowest common hypernym) and IC denotes the information content of a concept. IC is computed as $IC(c) = \log p(c)$ where $p(c)$ is the probability of encountering an instance of c in a corpus. The probability $p(c)$ can be estimated from the relative corpus frequency of c and the probabilities of all concepts that c subsumes [26].

Wikipedia and vector-based SR measures. Web based co-occurrence measures for SR often suffer from the noisy nature of web content. Wikipedia has been proposed as a

source of background knowledge for calculating the semantic relatedness of words [15] and argued to provide a stable and noise-free resource w.r.t. this task.

Wikipedia is the largest online collaboratively built Encyclopedia, with more than 3 million articles for the English version. Wikipedia contains pages for concepts and each page provides a detailed and human edited description of the corresponding concept. In the past few years Wikipedia has been increasingly used as a source of world knowledge in Artificial Intelligence and Natural Language Processing in general and in text-based SR calculation in particular. Wikipedia-based SR measures are currently considered to be the state-of-the-art [31].

The Explicit Semantic Analysis (ESA) measure of Gabrilovich and Markovitch used here represents each term as a vector of Wikipedia concepts (according to their use in the corresponding articles) and measures semantic similarity as the cosine of the corresponding concept vectors (thus capturing distributional similarity of the terms over Wikipedia): $sim_{ESA} = \frac{\overrightarrow{c_1} * \overrightarrow{c_2}}{\left|\overrightarrow{c_1}\right| * \left|\overrightarrow{c_2}\right|}$.

World Wide Web and hit-count based SR measures. Apparently the largest source of (textual) information is the World Wide Web itself. Because of this, search results of web search engines (i.e., search hit counts (HC) or text snippets) have been extensively used in many Natural Language Processing applications to model lexical semantic knowledge. With respect to SR, various measures have been proposed that use hit counts to measure term co-occurrence information as an indicator of term relatedness [17]. In our study, we used Yahoo to gather hit count information from the Web. From the many variants proposed in the literature we used the Dice coefficient to measure the similarity of word pairs $sim_{DICE}(t_1, t_2) = \frac{HC(t_1,t_2)}{HC(t_1)+HC(t_2)}$, where HC represents the hit counts for a given term (or the term pair).

In connection with part attributes (see Sect. 3.1), we can refine web-based SR by making explicit use of part-whole (*holonym*) relations [5]. This is achieved by formulating web queries including *holonym patterns*, such as "elephant's tusks" or "patches of leopards". In particular, we use nine holonym patterns[2] suggested by [5] excluding "in" patterns, since these often denote non-visible object class properties or parts.

Web image search and hit-count based SR measures. The World Wide Web provides a natural opportunity to derive more visually oriented SR measures: using the same methods as described above, i.e., web search and Dice coefficient to calculate SR, we can restrict our search to image-related texts (captions, anchor texts, etc.) by using an image search engine like Yahoo Image Search or to human-assigned image tags and description using a collaborative photo management and sharing application like Flickr's search functionality. Performing image-related searches to approximate relatedness of concepts, we expect to get a more visually relevant relatedness measure.

Binarization and normalization. Since attribute-based zero-shot classification (see Sect. 3.1) requires binary associations between object classes and attributes, we binarize the continuous similarity values returned by SR measures as in [19], by a global threshold on the continuous-valued association matrix. The threshold is set to the mean of all matrix entries (we exclude the diagonal for objectness as it contains the similarity of a term with itself). This thresholding obviously requires the similarity values to be comparable across classes and attributes, which we achieve by normalization. We normalize matrix values by dividing by column and row sums prior to binarization.

5. Experiments

In this section, we apply the various zero-shot classification schemes presented in Sect. 3 to a publicly available dataset, namely, the Animals with Attributes (AwA) dataset introduced by [19]. In particular, we reproduce the previously reported results on attribute-based zero-shot classification relying on manual supervision, providing the basis for our evaluation. We then compare the performance of the different knowledge bases introduced in Sect. 4 in place of manual supervision and interpret the differences.

The dataset consists of 50 mammal object classes, each containing at least 92 images, paired with a human provided attribute inventory and corresponding object class-attribute associations [16, 23]. We follow the experimental protocol of [19], using the provided split into 40 training and 10 test classes (24,295 training, 6,180 test images). We also use the provided pre-computed feature descriptors, namely, RGB color histograms, SIFT, rgSIFT, PHOG, SURF, and local self-similarity histograms. In contrast to [19] we concatenate all features to a single vector instead of training independent SVMs.

For computational reasons, we depart slightly from the protocol of [19] in our main experiments. First, we downsample all training images to the minimum of 92 available images per class. Second, we use histogram intersection kernel SVMs [8] instead of χ^2 kernel SVMs. For better comparison to [19], we reproduce their results using all training images and χ^2 kernel SVMs as a reference and report the differences to our reduced setting.

Results. Tab. 1 gives zero-shot classification results in the form of area under ROC curve (AUC) scores for the ten individual test classes (first ten columns) and their average (last but one column). The last column gives the corresponding average multi-class classification accuracies. Each row of the table corresponds to a single classification experiment. The row-wise sections of the table mark the

[2] Nine holonym patterns: (1-2) **whole**'*s* **part**[*s*], (3-4) **whole***s'* **part**[*s*], (5-6) **part**[*s*] *of a* **whole**, (7-8) **part**[*s*] *of the* **whole**, (9) **part***s of* **whole***s*

Source (Measure)	Area under ROC curve (AUC) in %										mean AUC (in %)	mean accuracy (in %)
	chimpanzee	giant panda	leopard	persian cat	pig	hippopatamus	humpback whale	raccoon	rat	seal		
1. Reproduction of the results in [19])												
all images, χ^2	86	65	88	84	73	77	99	78	76	78	80.3	40.3
92 images, hist. int.	87	64	86	84	71	75	98	72	71	77	78.5	34.7
2. Mined object class-attribute associations												
WordNet (Path)	53	52	50	**79**	60	48	86	57	70	51	60.5	15.5
Wikipedia (Vector)	58	65	74	78	62	**70**	88	**73**	63	65	69.7	**27.0**
Yahoo Web (HC)	39	65	63	49	**73**	52	91	39	**76**	56	60.4	22.2
Yahoo Img (HC)	74	**77**	**81**	61	72	57	**97**	63	53	**76**	**71.0**	26.3
Flickr Img (HC)	**79**	72	**81**	76	68	56	94	64	47	63	70.1	20.0
Yahoo Img & Flickr	78	77	83	69	72	57	97	64	49	70	71.6	27.8
3. Mined attributes (and associations)												
WordNet (Path)	49	62	59	**70**	55	43	86	63	58	52	59.8	17.8
Wikipedia (Vector)	65	61	**69**	68	61	70	93	59	58	56	66.0	19.7
Yahoo Web (HC)	39	71	45	47	70	53	95	39	**65**	49	57.4	19.5
Yahoo Img (HC)	66	**79**	66	63	47	58	**98**	59	62	54	65.2	**23.6**
Flickr Img (HC)	60	70	67	60	49	63	**98**	**67**	60	51	64.6	22.9
Yahoo Holonyms (HC)	**78**	61	68	59	**71**	**77**	**98**	66	60	**59**	**69.9**	21.5
Wikipedia & Yahoo Hol.	77	65	74	68	70	76	98	65	63	61	71.6	26.9
4. Objectness as attributes												
WordNet (Path)	**79**	**67**	71	72	**70**	70	97	61	63	62	71.2	25.6
Wikipedia (Vector)	67	48	65	69	52	67	94	68	62	72	66.4	23.3
Yahoo Web (HC)	51	65	69	83	66	**76**	98	52	57	51	66.7	25.5
Yahoo Img (HC)	68	63	76	**86**	67	65	**99**	**77**	**71**	71	**74.1**	**33.0**
Flickr Img (HC)	57	59	**78**	–	63	66	98	70	**71**	**74**	*68.5*	*11.2*
WordNet & Yahoo Img	81	70	80	80	73	70	98	68	71	70	76.1	33.5
5. Direct similarity												
WordNet (Path)	**88**	73	82	59	60	68	**98**	67	66	73	73.4	29.7
Wikipedia (Vector)	79	**77**	84	82	68	60	**98**	74	77	67	76.6	33.2
Yahoo Web (HC)	84	72	**88**	82	**77**	**70**	**98**	**76**	71	60	77.7	34.7
Yahoo Img (HC)	85	70	78	**85**	**77**	64	**98**	73	77	**81**	**78.8**	**35.7**
Flickr Img (HC)	84	72	86	78	**77**	63	**98**	72	**78**	73	77.8	32.5
Yahoo Img & Flickr	84	71	82	84	77	63	98	73	78	78	78.9	33.9
all	84	75	84	84	77	65	98	74	78	79	79.7	34.1

Table 1. Zero-shot classification results on the AwA data set [19]. The best results per table section are given in bold font. "–" denotes unavailable results due to all-inactive attributes, set to chance level = 50% for mean calculation. See Sect. 5 for discussions.

different variants of knowledge transfer (Sect. 1: reproduction of the results of [19], Sect. 2: attribute-based classification using mined object class-attribute associations, Sect. 3: attribute-based classification using mined attributes, Sect. 4: attribute-based classification using objectness as attributes, Sect. 5: direct similarity-based classification). Each row-wise, numbered section (except Sect. 1) gives results for various knowledge bases in the same consistent ordering.

1. Reproduction of the results in [19]. Our implementation, closely following the settings in [19], using all available training images and χ^2 kernel SVM, achieves an average AUC of 80.3% and corresponding multi-class classification accuracy of 40.3% (Tab. 1, Sect. 1). This is very close to 80.7% and 40.5% reported in [19], respectively. Downsampling the training set to 92 images per class and using histogram intersection kernel SVM decreases performance slightly, but not significantly, to 78.5% and 34.7%, respectively (Tab. 1, Sect. 1). All results that follow are based on this computationally more manageable setting.

2. Mined object class-attribute associations. We start by comparing the performance of the various knowledge bases using the original set of attributes proposed by [19], using semantic relatedness to determine class-attribute associations (Tab. 1, Sect. 2). We observe that Yahoo Img performs best on average (71.0% AUC), closely followed by Flickr Img (70.1%). This is expected since both are based on image-related texts and inherently capture important correlations between terms and visual attributes, which can be beneficial for recognition. The difference between Yahoo Img and Flickr Img is minor and may be in consequence of the generally smaller coverage of Flickr compared to the full web used by Yahoo Img. Similar performance is achieved by Wikipedia (69.7%). We attribute this to Wikipedia's encyclopedic nature which provides concise and noise-free explanations of concepts.

Last are WordNet (60.5%) and Yahoo Web (60.4%). They show a significant drop in performance ($\approx 10\%$) compared to the first three knowledge bases. For Yahoo Web this drop is due to an increased level of noise compared to image search or Wikipedia. In particular, we observed incidental co-occurrences on web pages and polysemous expressions (e.g. attribute term "pad", class term "seal") to have a negative effect on performance. Although incidental co-occurrences do not exist in WordNet, it does suffer from (non-disambiguated) polysemous terms. However, more important is the fact that path lengths are a poor indicator of semantic relatedness between object class and attribute concepts, as they are computed from hypernym relations: since object classes and attributes are inherently different in nature, they are likely to lie in entirely different subtrees of the hypernym hierarchy. Consequently, path length is no longer representative for their semantic relatedness.

In an attempt to benefit from potentially complementary information from two different knowledge bases, we report additional results for the fusion of the two best competitors, Yahoo Img and Flickr Img. Fusion is performed by multiplying the respective class probabilities from both knowledge bases. In fact, this fusion performs slightly better than either individual knowledge base (71.6%).

In general, using knowledge bases for acquiring object class-attribute associations (at best 71.6%) performs consistently worse ($\approx 7\%$) than using manually defined associations from [19] (78.5%). We acknowledge that this performance drop is significant, but stress that the information provided to the human judges is far more descriptive than the simplified terms used to query the knowledge bases. E.g. "nest" abbreviates the attribute "keeping their young in a designated, enclosed area". We consider the obtained results, in connection with the significantly reduced amount of manual supervision, highly encouraging and an important contribution towards scalable recognition.

Above results are also reflected in the mined class-

attribute associations. Although it is not always easy to judge if the associations are meaningful, we provide an example for the visual attribute "striped" to give an impression of the quality of mined similarities. In the following we list the four top ranked mammal classes for "striped" in decreasing order: Manual: *zebra, tiger, skunk, raccoon*; WordNet: *elephant, seal, mouse, bat*; Wikipedia: *zebra, skunk, tiger, Chihuahua*; Yahoo Web: *zebra, collie, Dalmatian, polar bear*; Yahoo Img: *zebra, skunk, tiger, Persian cat*; Flickr Img: *skunk, tiger, zebra, leopard*.

3. Mined attributes (and associations). Sect. 3 of Tab. 1 gives results for using automatically mined (part) attributes instead of manually defined ones, as presented in Sect. 2 of Tab. 1, fully avoiding any kind of manual supervision. Disregarding Yahoo Holonyms, we found Wikipedia, Yahoo Img, and Flickr Img again to perform best (66.0%, 65.2%, 64.6%) with a significant margin to the next best knowledge bases (WordNet 59.8%, Yahoo Web 57.4%). This is consistent with the results for manually defined attributes and underpins the differences between the knowledge bases highlighted above.

Moving from manually defined to automatically mined attributes results in a general drop in performance for the measures discussed above. We attribute this to the reduced number of attributes (85 to 74) and the reduced diversity of attributes (colors, context, parts etc. versus parts only).

For automatically mined attributes we give additional results for a specific flavor of Yahoo Web which allows to formulate queries specifically tailored towards part attributes (Yahoo Holonyms, see end of Sect. 4). The resulting semantic relatedness estimates clearly benefit from this specificity, obtaining the best performance of 69.9%, which is comparable to the previously reported results for manually defined attributes. The fusion of the best two, Wikipedia and Yahoo Holonyms, is again beneficial (71.6%) and on par with manually defined attributes.

4. Objectness as attributes. Sect. 4 of Tab. 1 gives results for using objectness as attributes for attribute-based classification. Specifically, we use all 50 class names as attribute signifiers. Yahoo Img is again best (74.1%). In contrast to sections 2 and 3, WordNet (71.2%) gains relative performance and performs second best: using objects (objectness) as attributes renders the corresponding concepts perfectly comparable to object classes with respect to the hypernym hierarchy. Path length becomes a valid measure of semantic relatedness. Next are Yahoo Web (66.7%) and Wikipedia (66.4%). Although Flickr Img performs on average slightly better (68.5%), it could not yield results for all test classes due to a lack in coverage: the user provided text content does often not provide sufficient statistics for co-occurring object classes. The fusion of the best two, WordNet and Yahoo Img is again beneficial (76.1%).

5. Direct similarity. Sect. 5 of Tab. 1 gives results for zero-shot classification based on direct similarity between object classes. We observe that most knowledge bases show very similar and sometimes even slightly better performance compared to using manually defined object class-attribute associations (Yahoo Img 78.8%, Flickr Img 77.8%, Yahoo Web 77.7%, Wikipedia 76.6%, fusion Yahoo Img & Flickr 78.9%, fusion of all 79.7%). Only WordNet performs worse (73.4%). This is in line with the observation that for a given test class, the 5 chosen most similar classifiers used for zero-shot classification are quite reliable on average and generally similar among the different knowledge bases. Even more importantly, the direct similarity model circumvents the need for an intermediate attribute layer, effectively eliminating one potential source of noise from the process: while a specific classifier used in direct similarity-based classification is guaranteed to be trained from appropriate training data, the training set of an attribute classifier is determined using (noisy) semantic relatedness measures.

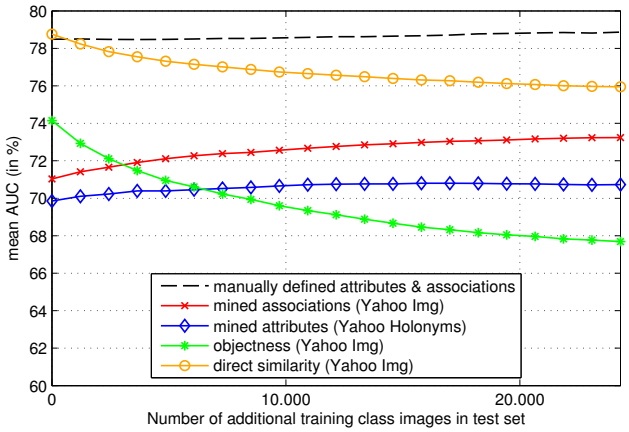


Figure 3. Adding training class images to the test set.

Training class images in the test set. Although the data set and testing protocol suggested by [19] provides a valuable resource for zero-shot classification experiments, it exhibits a distinct property that may lead to a biased view on the results. Namely, the set of object classes used for training and test is disjoint: a zero-shot classifier under test is never challenged to distinguish a previously unseen class from one it actually knows. We expect this distinction to be difficult, since it asks for classifying as negatives those classes that have been used as positive examples during training.

We visualize this effect in Fig. 3, where we plot AUC for the best performing methods per section of Tab. 1, varying the number of training class images which we inject as negatives to the test set. The horizontal axis denotes the number of injected training class images not actually used for training (these images exist due to the down-sampled training set). As expected, we observe performance drops for objectness (Yahoo Img) and direct similarity (Yahoo Img) for added negatives. In contrast to this, all classi-

fiers based on generic attributes (mined associations (Yahoo Img), mined Attributes (Yahoo Holonyms), and manually defined attributes & associations) seemingly generalize better over added negatives, which we explain by their complementary nature (see Sect. 3.1). They tend to form groups of more diverse object classes than objectness attributes, and thus limit the influence of specific (positive) training classes appearing as negatives in the test set.

Summary. In summary, we conclude that Yahoo Img is unparalleled with respect to coverage and precision. It outperforms most other knowledge bases in attribute-, objectness-, and direct similarity-based classification approaches, reaching a level of performance comparable to manually defined object class-attribute associations for direct similarity-based classification. Flickr Img is always inferior to Yahoo Img due to smaller coverage. Wikipedia achieves similar performance as web image search. Yahoo Web and WordNet provide competitive results only for inter-object similarity. Fusion of complementary knowledge bases always helps. Considering the different zero-shot classification approaches, objectness as attributes and direct similarity-based classification are superior to both manually defined and automatically mined attributes and can even level manually provided object class-attribute associations.

6. Conclusions

Having recognized knowledge transfer as a promising route to scalable recognition, we believe that further reducing the needed amount of manual supervision is a vital ingredient for making it a widely applicable tool for computer vision. In this paper, we demonstrate that manual supervision can in principle be fully replaced by tapping into linguistic knowledge bases[1], by the example of zero-shot object class recognition. Although this leads to a decrease in classification accuracy for attribute-based classification, direct similarity-based classifiers achieve performance on par with manual supervision. In our evaluation, we observe that the different characteristics of knowledge bases often result in largely different results. In particular, Yahoo image search and Wikipedia are always among the best choices while Yahoo web search and WordNet are especially inferior for attribute-based associations.

Acknowledgments. The authors would like to thank Torsten Zesch for providing the ESA measure index for Wikipedia [31], Giuseppe Pirrò for the Java WordNet Similarity Library [24], and Yahoo for the BOSS API.

References

[1] K. Barnard, P. Duygulu, D. A. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 2003. 2

[2] K. Barnard and K. Yanai. Mutual information of words and pictures. In *Information Theory and Applications*, 2006. 2

[3] E. Bart and S. Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *CVPR*, 2005. 2

[4] E. Bart and S. Ullman. Single-example learning of novel classes using representation by similarity. In *BMVC*, 2005. 1, 2

[5] M. Berland and E. Charniak. Finding parts in very large corpora. In *Annual Meeting of the ACL*, 1999. 5

[6] E. Boiy, K. Deschacht, and M.-F. Moens. Learning visual entities and their visual attributes from text corpora. In *DEXA*, 2008. 2

[7] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 2006. 4

[8] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. 5

[9] B. Delezoide, G. Pitel, H. L. Borgne, G. Greffenstette, P.-A. Mollic, and C. Millet. Object/background scene classification in photographs using linguistic statistics from the web. In *OntoImage*, 2008. 2

[10] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785, 2009. 2

[11] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, 2006. 2

[12] C. Fellbaum. *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA, 1998. 2, 4

[13] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007. 2

[14] M. Fink. Object classification from a single example utilizing class relevance pseudo-metrics. In *NIPS*, 2004. 1, 2

[15] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *IJCAI'07*. 5

[16] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, 2006. 3, 5

[17] A. Kilgarriff and G. Grefenstette. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29:333–347, 2003. 5

[18] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 2

[19] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2, 3, 5, 6, 7

[20] L. J. Li, R. Socher, and L. F. Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, pages 2036–2043, 2009. 2

[21] D. Lin. An information-theoretic definition of similarity. In J. W. Shavlik and J. W. Shavlik, editors, *ICML*, 1998. 4

[22] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *CVPR*, 2007. 2

[23] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith. Default probability. *Cognitive Science*, 15(2):251–269, 1991. 3, 5

[24] G. Pirrò and N. Seco. Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In *ODBASE*, 2008. 8

[25] A. Popescu, C. Millet, and P.-A. Moëllic. Ontology driven content based image retrieval. In *CIVR*, 2007. 2

[26] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453, 1995. 4

[27] M. Stark, M. Goesele, and B. Schiele. A shape-based object class model for knowledge transfer. In *ICCV*, 2009. 1, 2

[28] S. Thrun. Is learning the n-th thing any easier than learning the first. In *NIPS*, 2006. 2

[29] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, 2009. 2

[30] H. Wang, X. Jiang, L.-T. Chia, and A.-H. Tan. Ontology enhanced web image retrieval: aided by wikipedia & spreading activation theory. In *MIR*, 2008. 2

[31] T. Zesch and I. Gurevych. Wisdom of crowds versus wisdom of linguists - measuring the semantic relatedness of words. *Journal of Natural Language Engineering*, 16, 2010. 5, 8

[32] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *ICCV*, 2007. 2