Combining Language Sources and Robust Semantic Relatedness for Attribute-Based Knowledge Transfer

Marcus Rohrbach[1,2], Michael Stark[1,2], György Szarvas[1,⋆], and Bernt Schiele[1,2]

[1] Department of Computer Science, TU Darmstadt
[2] Max Planck Institute for Informatics, Saarbrücken, Germany

Abstract. Knowledge transfer between object classes has been identified as an important tool for scalable recognition. However, determining which knowledge to transfer where remains a key challenge. While most approaches employ varying levels of human supervision, we follow the idea of mining linguistic knowledge bases to automatically infer transferable knowledge. In contrast to previous work, we explicitly aim to design robust semantic relatedness measures and to combine different language sources for attribute-based knowledge transfer. On the challenging Animals with Attributes (AwA) data set, we report largely improved attribute-based zero-shot object class recognition performance that matches the performance of human supervision.

1 Introduction

While remarkable recognition performance has been reported on a wide variety of object classes, scaling recognition to large numbers of classes remains a key challenge, mostly because of the prohibitive amount of required training data. Knowledge transfer between object classes has been advocated to reduce the amount of required training data by re-using acquired information in the context of related, but previously unknown recognition tasks (zero-shot recognition). Knowledge transfer on the level of attribute-based object class models has received particular attention [1–3]. In [4] we proposed to combine attribute-based object class models with information mined automatically from linguistic knowledge bases, thereby avoiding any kind of human supervision. While we could show first promising results, only standard semantic relatedness measures were employed thereby limiting their robustness for visual object class recognition. At the same time, we suggested an alternative model for knowledge transfer [4], bypassing the intermediate layer of attributes. While this direct similarity-based model [5] exhibited superior performance for zero-shot recognition compared to the attribute-based model, it generalized significantly worse for a more realistic testing scenario in which training and test classes cannot be assumed disjoint.

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

K.N. Kutulakos (Ed.): ECCV 2010 Workshops, Part I, LNCS 6553, pp. 15–28, 2012.

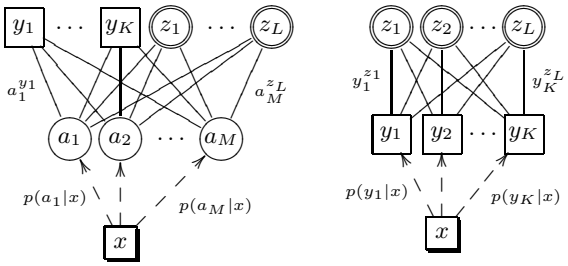
The main objective of our work is therefore to explicitly adapt semantic relatedness to the specific task of attribute-based object class recognition, to improve the robustness and reliability of inter-class knowledge transfer. The first important tool for this task is the combination of different semantic relatedness measures and language sources, where we can benefit from their complementary strengths, compensating their weaknesses. The second important tool is to expand a given attribute inventory by additional attributes, in order to solidify the basis upon which class-level decisions are taken. Both tools aim at replacing individual semantic relatedness estimates taken between a pair of concepts by several measurements, to increase robustness against errors.

The main contributions of our paper are as follows. First, we explore novel semantic relatedness measures which we show to be more appropriate for attribute-based object class recognition than the ones used before [4] (Sec. 5). Second, we suggest to combine individual semantic relatedness measures to yield more robust composite measures explicitly combining different language sources (Sec. 6). Third, we show how to expand a given attribute inventory with the help of semantic relatedness and demonstrate superior performance of the expanded inventory over the original one (Sec. 7). Fourth, we show that classifier level fusion further improves performance thereby attaining performance of human supervision (Sec. 8).

2 Related Work

Transferring knowledge between object classes has become an important direction towards scalable recognition. A prerequisite for knowledge transfer is an appropriate representation of transferable knowledge. Different representations have been proposed, ranging from discriminating aspects [6, 7] to distance metrics [5, 8, 9] and class priors [10, 11]. Descriptive attributes offer an intuitive characterization of transferable knowledge [1, 3, 12, 13]. The second prerequisite for knowledge transfer is to specify which knowledge can be transferred where. [2] introduced an attribute-based object class model for zero-shot recognition, based on human-provided associations between object classes and attributes.

Recently, we demonstrated the successful combination of this object class model and semantic relatedness, replacing human supervision by information automatically mined from linguistic knowledge bases [4]. In a similar zero-shot setting, [14] compare the performance of a linguistic knowledge base (Google Trillion-Word-Corpus) to manual labels. However, the model is applied in the context of a completely different domain, namely, neural decoding of novel thoughts. [15] classify unseen butterfly categories according to text descriptions. While encouraging results using standard linguistic knowledge bases and semantic relatedness measures have been reported [4, 14], we believe there is significant room for improvement in the design of these measures towards their use in object class recognition. E.g. we found important differences among individual knowledge bases and semantic relatedness measures [4] that one should exploit to improve robustness of the approach. The first goal of our work is therefore to build upon



(a) Attribute-based (b) Direct similarity-based

Fig. 1. Two models for zero-shot object classification. See Sec. 3 for discussions. Reproduced from [4] with permission.

our previous work and to carefully design a customized inventory of semantic relatedness measures for zero-shot object class recognition.

We also investigate a second object class model for knowledge transfer, the so called direct similarity model [4]. This model is also based on representing previously unknown object classes relative to known ones, characterizing unknown classes by their semantic relatedness to known classes [5, 16]. Interestingly, both models exhibit quite different behavior [4]. While at first glance, direct similarity shows better absolute performance in zero-shot recognition, the attribute-based model seemingly generalizes better when leaving the rather artificial experimental setup of the Animals with Attributes data set [2], which assumes disjoint sets of object classes appearing in training and test. The second main goal of our work is therefore to leverage this essential advantage of the attribute-based model and push its performance to match that of direct similarity and human supervision.

As concerns linguistic knowledge bases and individual semantic relatedness measures, we go beyond the ones considered in [4], e.g., by adding Yahoo Snippets [17] and Yahoo Near [18] (see Sec. 5).

3 Object Class Models for Knowledge Transfer

We briefly review the attribute-based models (see Fig. 1(a)) for knowledge transfer at the core of our approach, as introduced by [2] (direct attribute prediction model, DAP). Additionally we shortly introduce the direct-similarity based model [4] (see Fig. 1(b)) which we compare to. For a more detailed derivation, we refer the reader to [2] and [4], respectively.

3.1 Attribute-Based Classification

In the attribute-based model, the relation between known classes $y_1, \ldots, y_K$, unknown classes $z_1, \ldots, z_L$, and descriptive attributes $a_1, \ldots, a_M$ is given by

a matrix of binary associations values a_m^y resp. a_m^z (see Fig. 1) which encodes whether an attribute is active or inactive for a given class. While this association matrix is provided by human supervision in [2], it is derived from semantic relatedness measured between class and attribute concepts in [4]. At training time, attribute classifiers are trained using the known classes $y_1, \ldots, y_K$. At test time, the activation of an individual attribute a_m in an image x is measured by its posterior probability $p(a_m|x)$, estimated from its classifier output. Multiple attribute activations are then combined to yield the posterior probability of the (unknown) object class z being present in the image

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

3.2 Direct Similarity-Based Classification

The direct similarity model is structurally similar to the attribute-based model. It can be interpreted as a DAP with $M = K$ attributes, where attributes correspond to the known classes $y_1, \ldots, y_K$. The posterior probability of the (unknown) object class z being present in image x is then $p(z|x) \propto \prod_{k=1}^{K} \left(\frac{p(y_k|x)}{p(y_k)}\right)^{y_k^z}$, where y_k^z represents the semantic relatedness between known class y_k and unkown class z.

4 Experimental Setup

In the following sections we apply the attribute- and direct similarity-based object class models to the zero-shot classification task defined by the publicly available Animals with Attributes (AwA) data set [2]. It consists of 50 mammal classes, each containing at least 92 images, together with a human-provided inventory of 85 attributes and corresponding object class-attribute associations [19, 20]. We follow the experimental protocol of [4] based on [2]. We use the provided split into 40 training and 10 test classes (24,295 training, 6,180 test images) and the provided pre-computed feature descriptors, namely, RGB color histograms, SIFT, rgSIFT, PHOG, SURF, and local self-similarity histograms. We concatenate all features to a single vector and train histogram intersection kernel SVMs for classification, down-sampling all training images to the minimum number of 92 images available per class. We use libSVM with the built-in probability estimates (based on [21]) and a fixed cost parameter C=10.

5 Individual Semantic Relatedness (SR) Measures

We commence by determining the strength of object class-attribute associations (in the case of the attribute-based model) or object class-object class similarity (for the direct similarity-based model) by individual semantic relatedness measures.[1]

[1] All software for computing object class-attribute associations from linguistic knowledge bases is publicly available on our web page.

5.1 Semantic Relatedness Measures as Introduced in [4]

We recapitulate briefly the linguistic knowledge bases and semantic relatedness (SR) measures used in [4], since these constitute the starting point of our extensions. We put more emphasis on the description of those measures which we newly introduce, namely, Yahoo Snippets and Yahoo Near.

WordNet (Path) [22] is the largest machine readable expert-created language ontology. Similarity of concepts is usually defined on its hierarchical graph structure, as, e.g., in the Lin measure [23].

Wikipedia (Vector) is the largest community built online encyclopedia. The Explicit Semantic Analysis (ESA) measure is considered state-of-the-art [24], representing each term as a vector of frequencies over all articles. Similarity of two terms is computed by the cosine between the two respective vectors.

Yahoo Web (HC). The web itself is apparently the largest collection of textual content. For semantic relatedness computation, actual content is usually summarized in the form of search engine (Yahoo) hit counts (HC). The Dice coefficient then measures similarity of two terms by the relative number of co-occurrences, inferred from hit counts $sim_{DICE}(t_1, t_2) = \frac{HC(t_1,t_2)}{HC(t_1)+HC(t_2)}$.

Yahoo Img / Flickr Img (HC). In order to compensate for noise of full web page content, we restrict general web search to image search (Yahoo Img), or to a proper subset of the web devoted to collaborative photo sharing (Flickr Img).

5.2 Novel Semantic Relatedness Measures

Yahoo Near (HC). Restricting search engine queries to holonym patterns [25] significantly improves the performance of Yahoo Web (HC) [4] but is limited to part attributes. Similar in spirit, we suggest to impose proximity constraints on the occurrences of queried terms. The intuition is that requiring two terms to occur in proximity of one another in a document increases the likelihood of the co-occurrence being non-incidental and possibly even referring to the same physical entity. While Exalead [18] offers a built-in Near operator providing this functionality, we implemented these constraints for the Yahoo search engine, using its wildcard operator ("*"). The above defined Dice coefficient can then be applied by letting $HC(t_1, t_2) \equiv HC(t_1\ NEAR_k\ t_2)$, where $t_1\ NEAR_k\ t_2$ limits the number of words occurring between t_1 and t_2 to at most k. We found $2 \leq k \leq 4$ to work best and thus consistently report results for $k = 4$ in all experiments.

Yahoo Snippets. A robust variation of hit count-based measures has recently been proposed by [17], relying on short summary texts (snippets) accompanying the actual links returned by search engine (Yahoo) queries. In order to determine the relatedness of terms t_1 and t_2, the search engine is queried for t_1, measuring the frequency of occurrences of t_2 in the returned snippets, which we denote $f(t_2@t_1)$ and vice versa $f(t_1@t_2)$, explaining its common name "Web Search with Double Checking". The snippet-based approach has two intuitive advantages. First, a term has to qualify for its appearance in a snippet according to some

notion of importance, implemented by the search engine. Second, the ranking of search results can be taken into account when crawling snippets, which we do by restricting them to the 1,000 highest ranked pages. The resulting semantic relatedness measure is computed in analogy to the Dice coefficient:

$$sim_{Snippets}(t_1, t_2) = \frac{f(t_1@t_2) + f(t_2@t_1)}{f(t_1@t_1) + f(t_2@t_2)} \quad (2)$$

We note that [17] found CODC (Co-Occurrence Double Check) to outperform $sim_{Snippets}$. However, we found that CODC is not appropriate for the specific case of determining object class-attributes associations. It assumes a symmetric relation between two terms (by requiring $f(t_1@t_2)$ and $f(t_2@t_1)$ to simultaneously be greater than zero), which clearly does not hold for object class-attribute associations.

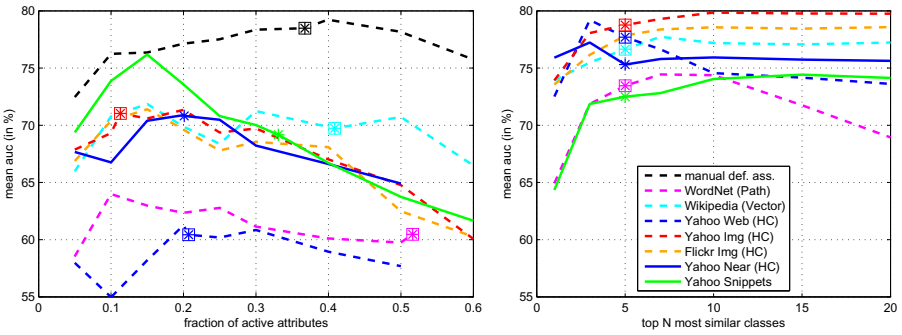
5.3 Discretizing Semantic Relatedness

Both attribute-based and direct similarity-based models for knowledge transfer require the discretization of semantic relatedness values. For the attribute-based model, semantic relatedness values have to be binarized to form an object class-attribute association matrix. This is typically done by applying a threshold t [2, 4]. For the direct similarity-based model, discretization is achieved through ranking: determining whether a test image contains an instance of test class z involves combining the classifier outputs corresponding to the N most similar training classes. In both cases, the choice of t or N can have a direct impact on performance (see below).

While [2, 4] use the mean over all continuous-valued object class-attribute association matrix entries as the threshold t, we suggest to sample different points from the space of meaningful thresholds, according to the fraction of matrix entries becoming 1 after binarization. Likewise, we suggest to vary N for the direct similarity-based model instead of fixing it to $N = 5$ as done in [4].

5.4 Experimental Results for Individual SR Measures

We start with the discussion of zero-shot classification results on the AwA data set [2] using individual semantic relatedness measures, for both attribute-based and direct similarity-based models. Fig. 2 plots the average classification performance over all 10 test object classes, measured as the mean area under the ROC curve (AUC), for attribute-based (Fig. 2 (a)) and direct similarity-based models (Fig. 2 (b)). Each curve corresponds to a distinct experiment using an individual semantic relatedness measure, varying either the applied binarization threshold t (Fig. 2 (a)) or the number of considered most similar classes N (Fig. 2 (b)). Additionally, we mark with an asterisk (*) the curve points for choices of t and N according to [4] and with a box (□) the curve points actually reported by [4]. We give results for the measures of [4] (dashed curves) and the two novel measures that we propose in this paper (solid curves). We also give the performance of the human-provided attribute association matrix (black dashed curve).



(a) Attributed-based (b) Direct similarity-based

Fig. 2. Zero-shot classification results for individual SR measures

We begin with the general observation that varying the threshold t has a non-negligible impact on performance (Fig. 2 (a)). E.g., for Yahoo Snippets (green solid curve), the performance difference is 14.5% between minimum (at a fraction of 0.6 active attributes) and maximum (at 0.15). The second general observation is that we can improve the results reported in [4] for all measures by varying t. The third general observations is that performance peaks are mostly located between 0.1 to 0.2 of active attributes, while performance drops beyond 0.2. This is contrary to human-provided associations and can be explained by the observation that the top-ranked associations are more reliable than lower ranks for semantic relatedness. As concerns the relative performance of the different measures, we note that the newly introduced Yahoo snippets (HC) (solid green curve) performs overall best (76.2%), outperforming all other measures by a large margin (in particular the ones proposed in [4]). The newly introduced Yahoo Near (HC) measure (solid blue curve) improves significantly (9.6% measured between the maxima of both curves) over its natural base line, Yahoo Web (HC) (dashed blue curve). We conclude that we can improve the results of the attribute-based model significantly already at the level of individual semantic relatedness measures.

For the direct similarity-based model (Fig. 2 (b)), we observe similar general tendencies as for the attribute-based model. Choosing N different from its default value $N = 5$ always improves performance. Performance increases but saturates for higher values of N. As concerns the performance of the newly proposed measures, they tend to perform worse (Yahoo Snippets, solid green curve) or equal (Yahoo Near (HC), solid blue curve) to the ones used in [4]. The reasons for the limited improvements of the new measures for the direct similarity-based model are two-fold. First, the room for improvement is limited as, apart from WordNet, the measures used in [4] provided already very reliable ranking for the most similar classes. Second, in contrast to attribute-based classification Yahoo Snippets and Yahoo Near are now required to estimate relatedness of object classes instead of objects and their attributes. The proximity requirement

both measures place between the compared terms, however, is especially present for objects and their attributes, e.g. in phrases such as "*white sheep*" (color attributes), "*elephant*'s *tusks*" (part attributes), or "*swim* with *dolphins* in the *ocean*" (activity and context attributes).

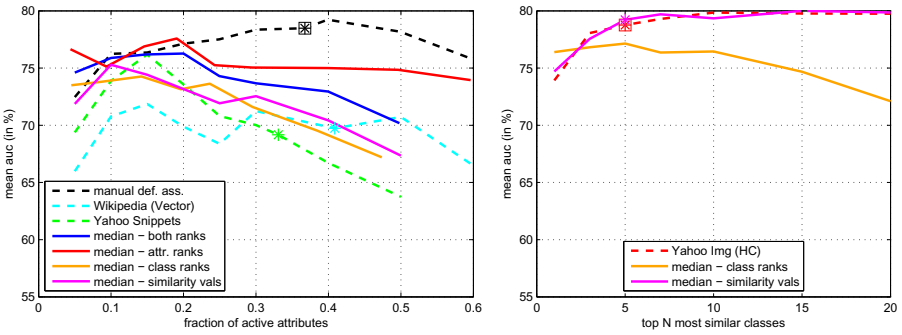
6 Combined Semantic Relatedness Measures

While we showed improved performance for two newly proposed individual semantic relatedness measures over the ones used in [4] in Sec. 5, we observe there is still room for improvement. In particular, we hope to benefit from the complementary nature of different knowledge bases and semantic relatedness measures, by combining individual measures to yield composite measures. As an example, consider the false positive associations between the attribute *big* and various object classes. While Wikipedia (Vector) and Yahoo Snippets list *chihuahua* among the 10 most strongly related classes, Yahoo Web (HC) lists *mouse*, Yahoo Img (HC) *mole*, Flickr Img (HC) *rat*, and Yahoo Near (HC) *beaver*. This diverse set of true positive associations is a clear hint towards complementary. In this section, we propose a strategy for exploiting these complementarities by combining measures, namely using median ranks.

Since semantic relatedness values computed by means of different measures are not per se comparable, an obvious pre-processing step for combination is to replace those values by a corresponding integer rank. For a given continuous-valued object class-attribute association matrix, this can be done either row-wise (producing an attribute ranking for each class) or column-wise (producing a class ranking for each attribute). Additionally, we can join both by first computing both attribute and class ranks, scaling them to the range $[0, 1]$, and multiplying the resulting values, yielding three different meaningful alternatives for rank computation. Having computed corresponding ranks for a number of individual semantic relatedness measures, a robust combination is the median over these ranks (i.e., the median over all corresponding entries in the object class-attribute association rank matrices of all measures).

Experimental Results for Median Rank Combined SR Measures. Fig. 3 gives results for the different variants of combining measures described above (solid curves), replicating the best curves of Sec. 5 as a reference (dashed curves). Again, each curve denotes a single experiment, varying the threshold t used for binarization of the object class-attribute association matrix. For the combinations, we consistently combine the five measures Wikipedia (Vector), Yahoo Img (HC), Flickr Img (HC), Yahoo Near (HC), and Yahoo Snippets.

As can be seen in Fig. 3 (a), median attribute ranks (solid red curve) perform best, outperforming the best individual measures Wikipedia (Vector) (dashed cyan curve) and Yahoo Snippets (dashed green curve) consistently for all thresholds. The maximum performance is reached at a threshold of 0.19 with 77.6% mean AUC, which is close to human-provided associations (dashed black curve, attaining a maximum of 79.2%). At the same time, and in contrast to all other measures, median attribute ranks achieve stable performance beyond 0.3 active



(a) Attributed-based (b) Direct similarity-based

Fig. 3. Zero-shot classification results for combined SR measures

attributes. The median of both ranks (solid blue curve) is second best. The third best combination is an unranked version (solid magenta curve), where we directly compute the median over the original semantic relatedness values. It shows clearly inferior performance to the median attribute ranks and median of both ranks, and is even inferior to the individual measure Yahoo Snippets (dashed green curve). Median class ranks performs worst (solid orange curve). We attribute this drop in performance to the fact that using class ranks as object class-attribute associations results in all classes having the same number of active attributes. This is in stark contrast to the typically imbalanced number of active attributes which we observed in our experiments.

As an example of successful recovery from errors in individual measures by median attribute rank combination, consider the attribute *long leg*: while all individual measures wrongfully assign high ranks to classes such as *mole* (Yahoo Img (HC), rank 3), *seal* (Yahoo Near (HC), rank 3), *rat* (Yahoo Snippets, rank 5), *hippopotamus* (Flicker Img (HC), rank 3), and *bat* (Wikipedia (Vector), rank 2), the first erroneous rank for median attribute ranks is *bat* at rank 9.

For the direct similarity-based model, median class ranks are inferior to the unranked version, whose performance is very close to the best individual measure Yahoo Img (HC).

7 Expanded Attribute Inventory

Combining different linguistic knowledge bases and semantic relatedness measures by ranking enables us to achieve higher performance than using individual measures alone and can almost match human performance. This section takes a very different route compared to previous sections, by expanding the inventory of descriptive attributes provided as part of the AwA data set [2]. While this inventory apparently provides a valid encoding of common and discriminating aspects between the various animal classes, intuition suggests two potential ways

of increasing the overall robustness of the attribute-based model. The first way is obviously to improve robustness of individual attribute classifiers. The second way is to expand the inventory of attributes, similar in spirit to building strong ensemble classifiers from a plethora of weak ones, in order to solidify the basis on which class level decisions are taken. In the following, we pursue both directions, by explicitly expanding the given inventory of attributes by new ones, which we generate on the basis of the existing ones. In this way, we hope to benefit from increased robustness while preserving the valuable knowledge encoded in the original attribute inventory.

We start from the observation that each attribute in the attribute-based model induces a 2-partitioning of object classes and vice versa: one partition of classes where the attribute is active and another partition of classes where it is inactive. Based on this observation, we suggest to form new partitions (i.e., generate new attributes) by clustering object classes in some feature space. Each cluster then induces a partitioning: the cluster itself constitutes one partition, its complement the other partition. As features, we choose the semantic relatedness values computed between object classes and the original attribute inventory, thus preserving the inherent information encoded in the original attributes. By clustering, we effectively replace individual measurements of semantic relatedness by multiple measurements, which we hope will improve the robustness of the resulting attribute classifiers. Likewise, we vary the parameters of the clustering such that it produces varying numbers of induced attributes, thereby expanding the original attribute inventory also quantitatively.

Prior to clustering, we split the original attributes into a set of distinct categories, namely colors (8: red, green, ...), texture (3: patches, spotted, stripe), skintype (3: furry, hairless, tough skin), stature (4: big, bulbous,...), parts (17: flipper, horn), locomotion (7: fly, hop,...), strength (3: strong, weak,...), moving behavior (5: active, agile,...), nutrition (5: meat, plankton,...), hunting style (6: grazer, scavenger,...), context (17: arctic, coastal,...), behavior (7: fierce, timid,...). k-means is then performed on a per-category basis to form aggregate attributes from semantically similar ones (e.g. black&white for the giant panda bear class).

In order to measure the qualitative differences to the original inventory of 85 attributes, we first generate an expanded inventory of size 85. We then further expand this inventory by merging it with additional clusterings of varying k, resulting in an expanded inventory of 164 attributes. Please note that our clustering result is a hard assignment, corresponding to a single binarization threshold (0.14 for 85 and 0.22 for 164 attributes).

Experimental Results for Expanded Attribute Inventories. In Table 1, we give the results for two different clustering variants generating 85 (second rightmost column) and 164 attributes, respectively (rightmost column). The leftmost column lists the average performance of individual measures over varying thresholds between 0.1 and 0.3 as a reference (for the complete results refer to Fig. 2a and 3a). Examining Table 1 we make two important observations. First, 164 clusters consistently outperform 85 clusters. Second, 164 clusters perform

Table 1. Zero-shot classification results for expanded attribute inventories in comparison to average performance over thresholds [0.1 0.3] from Sec. 5 and 6; discussion in Sec. 7

	mean AUC in %		
	original attributes	clustered attributes	
	average	85	164
Individual SR measures			
Wikipedia (Vector)	70.4	69.1 (-1.3)	72.4 (+2.0)
Yahoo Img (HC)	70.1	**75.2** (+5.1)	**77.2** (+7.1)
Flickr Img (HC)	69.5	73.7 (+4.2)	74.0 (+4.5)
Yahoo Near (HC)	69.4	70.6 (+1.2)	74.1 (+4.7)
Yahoo Snippets	72.9	72.0 (-0.9)	73.8 (+0.9)
Combined SR measures			
median attribute ranks	76.0	74.8 (-1.2)	76.6 (+0.6)
median both ranks	75.3	73.5 (-1.8)	76.8 (+1.5)

always better than the corresponding original attributes. For hit count-based measures (Yahoo Img (HC), Flicker Img (HC), Yahoo Near (HC)) and Wikipedia, this improvement is particularly pronounced. Notably, Yahoo Img (HC) improves to 77.2% mean AUC, which is very close to the performance of human-provided associations (79.2%). In summary, our results confirm the intuition given in the beginning of this section.

8 Classifier Level Fusion

In Sec. 6 we showed the success of combining different measures on the level of semantic relatedness values. As a final step, this section explores fusing the different measures on classifier level. We achieve this by combining the class probabilities (i.e. the $p(z|x)$ values of Equation 1) returned by different models. We use the product of the class probabilities for combination. We fuse the top 5 measures already combined in Sec. 6 for the attribute-based and direct similarity-based model (Sec. 5), as well as for expanded attribute inventories (Sec. 7).

Experimental Results. Table 2 shows the results of fusion (rightmost columns) in comparison to the best results achieved without fusion for the respective settings (middle columns). As can be seen in lines 2 and 3 of Table 2, a significant improvement is achieved when fusing the classifier probabilities of the expanded attribute inventories (85 and 164 clustered). The combined model achieves a mean AUC of 79.0% and 79.5%, respectively, which is on the level of human-provided associations (79.2%). For the direct similarity-based model, fusion does not improve performance (line 4).

Fusing the predictions of models based on the original AwA attribute set (75.9% mean AUC) cannot exceed the best performing single measure Yahoo Snippets with 76.2% mean AUC (Table 2, line 1). However, we note that the

Table 2. Classifier level fusion. Details in Sec. 8.

#	Setting	*respective best without fusion* reference	thresh	mean auc (%)	*fused* thresh	mean auc (%)
1	Attribute-based: AwA attributes	Sec. 5, Fig. 2a	0.15	76.2	0.15	75.9 (-0.3)
2	Attribute-based: 85 clustered	Sec. 7, Table 1	0.14	75.2	0.22	79.0 (+3.8)
3	Attribute-based: 164 clustered	Sec. 7, Table 1	0.22	77.2	0.22	**79.5** (+2.3)
4	Direct similarity	Sec. 5, Fig. 2b	10	79.9	10	75.9 (-4.0)

fused measure provides consistently higher performance than the individual measures for the non-peak locations on the respective curves (not shown in the table). The fused measure is apparently not as sensitive to the selection of the binarization threshold, which is a valuable characteristic on its own. We consider this a highly promising result, as we managed to reach a performance level on par with using human-provided associations. As this is achieved for an attribute-based model, we expect better generalization than for direct similarity-based models, which we will explore in the next section.

9 Extending Test Set with Images from Known Classes

In all previous experiments, following the experimental protocol of [2], the set of object *classes* used for training and test were disjoint. This setting assumes that no images belonging to the known (training) classes are present at testing time. As discussed in [4], this setting is less challenging, as it does not require the zero-shot classifier to reject images from classes it already knows (i.e. the training classes). Using images from the training classes (that were not used for training) as additional negative examples for testing is an especially difficult (adversary) setting, as it requires the classifier to generalize over the known classes. We argue that this more difficult setting is also more realistic and allows us to draw conclusions that are more appropriate to a real-life object recognition setting. Thus, following [4], we report results using all images from the test classes not used for training as additional negatives in the test set.

Experimental Results. Table 3 lists the best results from [4] the best measures and combinations of the previous sections. The second last column gives results when including training class images as negatives in comparison to the performance reported in the previous sections (third last column).

The most important observations based on the results in Table 3 are: First, while human-provided associations show stable results (line 1), performance of direct similarity significantly drops when including training class images (line 6). We could slightly increase overall performance by varying thresholds (line 7), but direct similarity does not level with human-provided associations for the more

Table 3. Effect of images from known classes in the test set. Selection of respective best from sections 6-8. Discussion in Sec. 9.

#	Setting / measure	Sec.	threshold	mean auc in % imgs: test	+ train cls	diff.
Object - Attribute Associations						
1	manually defined associations	5	0.40	79.2	79.4	+0.2
2	Yahoo Img (HC) [4]	5	0.11	71.0	73.2	+2.2
3	median: attribute ranks	6	0.19	77.6	**79.2**	+2.4
4	164 clustered: Yahoo Img (HC)	7	0.22	77.2	76.9	-0.3
5	classifier fusion: 164 clustered	8	0.22	**79.5**	78.9	-0.6
Direct Similarity						
6	Yahoo Img (HC) [4]	5	5	78.8	76.0	-2.8
7	Yahoo Img (HC)	5	10	**79.9**	**76.4**	-2.5
8	classifier fusion	8	10	75.9	72.3	-3.6

difficult adversary setting, even when fusing on classifier level (line 8). Second, in contrast to direct similarity, we found attribute-based measures, e.g. Yahoo Img (line 2), to slightly improve in most cases, i.e. generalize well. Third, the best combined models, median attribute ranks (line 3) and classifier fusion with the 164 clustered attribute inventory (line 5) are not only very competitive in terms of performance, but also perform well in this adversary setting (79.2%, 78.9%), on par with the model using human-provided assocations (79.4%). This property makes these measures favorable to those based on direct similarities that are less suited to recognize (and reject) training classes at testing stage.

10 Conclusions

In this paper, we propose several tools to increase the robustness of semantic relatedness for use in attribute-based zero-shot object class recognition, leading to performance on par with human supervision. First, on the level of individual measures we find Yahoo Snippets to provide significantly higher performance than the measures used in [4]. Second, combining individual measures on the level of semantic relatedness values achieves performance close to human-provided associations using attribute ranks. Third, expanding the attribute inventory using clustering also reaches performance close to human supervision for the Yahoo Image (HC) measure. Finally, fusing measures on classifier level achieves performance on par with human supervision for expanded attribute inventories. This is particularly valuable, since the attribute-based model generalizes well even for the difficult setting when images from known classes are added to the test set.

Acknowledgments. The authors would like to thank Torsten Zesch for providing the ESA measure index for Wikipedia [24], Giuseppe Pirrò for the Java WordNet Similarity Library [26], and Yahoo for the BOSS API.[1]

References

1. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS (2007)
2. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
3. Wang, G., Forsyth, D.: Joint learning of visual attributes, object classes and visual saliency. In: ICCV (2009)
4. Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., Schiele, B.: What helps where – and why? semantic relatedness for knowledge transfer. In: CVPR (2010)
5. Fink, M.: Object classification from a single example utilizing class relevance pseudo-metrics. In: NIPS (2004)
6. Marszalek, M., Schmid, C.: Semantic hierarchies for visual object recognition. In: CVPR (2007)
7. Zweig, A., Weinshall, D.: Exploiting object hierarchy: Combining models from different category levels. In: ICCV (2007)
8. Bart, E., Ullman, S.: Cross-generalization: Learning novel classes from a single example by feature replacement. In: CVPR (2005)
9. Thrun, S.: Is learning the n-th thing any easier than learning the first. In: NIPS 1996 (1996)
10. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. PAMI 28 (2006)
11. Stark, M., Goesele, M., Schiele, B.: A shape-based object class model for knowledge transfer. In: ICCV (2009)
12. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV (2009)
13. Farhadi, A., Endres, I., Hoiem, D.: Attribute-centeric recognition for cross-category generalization. In: CVPR (2010)
14. Palatucci, M., Pomerleau, D., Hinton, G., Mitchell, T.: Zero-shot learning with semantic output codes. In: NIPS (2009)
15. Wang, J., Markert, K., Everingham, M.: Learning models for object recognition from natural language descriptions (2009)
16. Bart, E., Ullman, S.: Single-example learning of novel classes using representation by similarity. In: BMVC (2005)
17. Chen, H.H., Lin, M.S., Wei, Y.C.: Novel association measures using web search with double checking. In: ACL-44 (2006)
18. Delezoide, B., Pitel, G., Borgne, H.L., Greffenstette, G., Moëllic, P.A., Millet, C.: Object/background scene classification in photographs using linguistic statistics from the web. In: OntoImage (2008)
19. Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T., Ueda, N.: Learning systems of concepts with an infinite relational model. In: AAAI (2006)
20. Osherson, D.N., Stern, J., Wilkie, O., Stob, M., Smith, E.E.: Default probability. Cognitive Science 15, 251–269 (1991)
21. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. JMLR 2004 (2004)
22. Fellbaum, C.: WordNet: An Electronical Lexical Database. The MIT Press (1998)
23. Lin, D.: An information-theoretic definition of similarity. In: ICML (1998)
24. Zesch, T., Gurevych, I.: Wisdom of crowds versus wisdom of linguists - measuring the semantic relatedness of words. JNLE 16 (2010)
25. Berland, M., Charniak, E.: Finding parts in very large corpora. In: ACL (1999)
26. Pirrò, G., Seco, N.: Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In: ODBASE (2008)