From Visual Attributes to Adjectives through Decompositional Distributional Semantics

Angeliki Lazaridou Georgiana Dinu* Adam Liska Marco Baroni
Center for Mind/Brain Sciences
University of Trento
{angeliki.lazaridou|georgiana.dinu|adam.liska|marco.baroni}@unitn.it

Abstract

As automated image analysis progresses, there is increasing interest in richer linguistic annotation of pictures, with *attributes* of objects (e.g., *furry*, *brown*...) attracting most attention. By building on the recent "zero-shot learning" approach, and paying attention to the linguistic nature of attributes as noun modifiers, and specifically adjectives, we show that it is possible to tag images with attribute-denoting adjectives even when no training data containing the relevant annotation are available. Our approach relies on two key observations. First, objects can be seen as bundles of attributes, typically expressed as adjectival modifiers (a *dog* is something *furry*, *brown*, etc.), and thus a function trained to map visual representations of objects to nominal labels can implicitly learn to map attributes to adjectives. Second, objects and attributes come together in pictures (the same thing is a *dog* and it is *brown*). We can thus achieve better attribute (and object) label retrieval by treating images as "visual phrases", and decomposing their linguistic representation into an attribute-denoting adjective and an object-denoting noun. Our approach performs comparably to a method exploiting manual attribute annotation, it outperforms various competitive alternatives in both attribute and object annotation, and it automatically constructs attribute-centric representations that significantly improve performance in supervised object recognition.

* Current affiliation: Thomas J. Watson Research Center, IBM, gdinu@us.ibm.com

1 Introduction

As the quality of image analysis algorithms improves, there is increasing interest in annotating images with linguistic descriptions ranging from single words describing the depicted objects and their properties (Farhadi et al., 2009; Lampert et al., 2009) to richer expressions such as full-fledged image captions (Kulkarni et al., 2011; Mitchell et al., 2012). This trend has generated wide interest in linguistic annotations beyond concrete nouns, with the role of adjectives in image descriptions receiving, in particular, much attention.

Adjectives are of special interest because of their central role in so-called *attribute-centric* image representations. This framework views objects as bundles of properties, or *attributes*, commonly expressed by adjectives (e.g., *furry*, *brown*), and uses the latter as features to learn higher-level, semantically richer representations of objects (Farhadi et al., 2009).[1] Attribute-based methods achieve better generalization of object classifiers with less training data (Lampert et al., 2009), while at the same time producing semantic representations of visual concepts that more accurately model human se-

[1] In this paper, we assume that, just like nouns are the linguistic counterpart of visual objects, visual attributes are expressed by adjectives. An informal survey of the relevant literature suggests that, when attributes have linguistic labels, they are indeed mostly expressed by adjectives. There are some attributes, such as parts, that are more naturally expressed by prepositional phrases (PPs: *with a tail*). Interestingly, Dinu and Baroni (2014) showed that the decomposition function we will adopt here can derive both adjective-noun and noun-PP phrases, suggesting that our approach could be seamlessly extended to visual attributes expressed by noun-modifying PPs.

Transactions of the Association for Computational Linguistics, vol. 3, pp. 183–196, 2015. Action Editor: Patrick Pantel.
Submission batch: 9/2014; Revision batch 1/2015; Revision batch 3/2015; Published 3/2015.
©2015 Association for Computational Linguistics. Distributed under a CC-BY-NC-SA 4.0 license.
Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

mantic intuition (Silberer et al., 2013). Moreover, automated attribute annotation can facilitate finer-grained image retrieval (e.g., searching for a *rocky beach* rather than a *sandy beach*) and provide the basis for more accurate image search (for example in cases of *visual sense disambiguation* (Divvala et al., 2014), where a user disambiguates their query by searching for images of *wooden cabinet* as furniture and not just *cabinet*, which can also mean council).

Classic attribute-centric image analysis requires, however, extensive manual and often domain-specific annotation of attributes (Vedaldi et al., 2014), or, at best, complex unsupervised image-and-text-mining procedures to learn them (Berg et al., 2010). At the same time, resources with high-quality per-image attribute annotations are limited; to the best of our knowledge, coverage of all publicly available datasets containing non-class specific attributes does not exceed 100 attributes,[2] orders of magnitude smaller than the equivalent object-annotated datasets (Deng et al., 2009). Moreover, many visual attributes currently available (e.g., *2D-boxy, furniture leg*), albeit visually meaningful, do not have straightforward linguistic equivalents, rendering them inappropriate for applications requiring natural linguistic expressions, such as the search scenarios considered above.

A promising way to limit manual attribute annotation effort is to extend recently proposed *zero-shot learning* methods, until now applied to object recognition, to the task of labeling images with attribute-denoting adjectives. The zero-shot approach relies on the possibility to extract, through distributional methods, semantically effective vector-based word representations from text corpora, on a large scale and without supervision (Turney and Pantel, 2010). In zero-shot learning, training images labeled with object names are also represented as vectors (of features extracted with standard image-analysis techniques), which are paired with the vectors representing the corresponding object names in language-based distributional semantic space. Given such

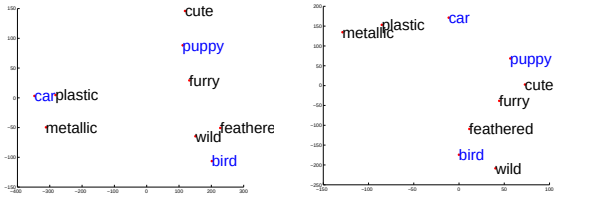


Figure 1: t-SNE (Van der Maaten and Hinton, 2008) visualization of 3 objects together with the 2 nearest attributes in our visual space (left), and of the corresponding nouns and adjectives in linguistic space (right).

paired training data, various algorithms (Socher et al., 2013; Frome et al., 2013; Lazaridou et al., 2014) can be used to induce a *cross-modal projection* of images onto linguistic space. This projection is then applied to map previously unseen objects to the corresponding linguistic labels. The method takes advantage of the similarities in the vector space topologies of the two modalities, allowing information propagation from the limited number of objects seen in training to virtually any object with a vector-based linguistic representation.

To adapt zero-shot learning to attributes, we rely on their nature as (salient) properties of objects, and on how this is reflected linguistically in modifier relations between adjectives and nouns. We build on the observation that visual and linguistic attribute-adjective vector spaces exhibit similar structures: The correlation ρ between the pairwise similarities in visual and linguistic space of all attributes-adjectives from our experiments is 0.14 (significant at $p < 0.05$).[3] While the correlation is smaller than for object-noun data (0.23), we conjecture it is sufficient for zero-shot learning of attributes. We will confirm this by testing a cross-modal projection function from attributes, such as colors and shapes, onto adjectives in linguistic semantic space, trained on pre-existing annotated datasets covering less than 100 attributes (Experiment 1).

We proceed to develop an approach achieving equally good attribute-labeling performance *without* manual attribute annotation. Inspired by linguistic and cognitive theories that characterize objects as attribute bundles (Murphy, 2002), we hypothesize that when we learn to project images of objects to the corresponding noun labels, we implicitly learn to

[2]The attribute datasets we are aware of are the ones of Farhadi et al. (2010), Ferrari and Zisserman (2007) and Russakovsky and Fei-Fei (2010), containing annotations for 64, 7 and 25 attributes, respectively. (This count excludes the SUN Attributes Database (Patterson et al., 2014), whose attributes characterize scenes rather than concrete objects.)

[3]In this paper, we report significance at $\alpha = 0.05$ threshold.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

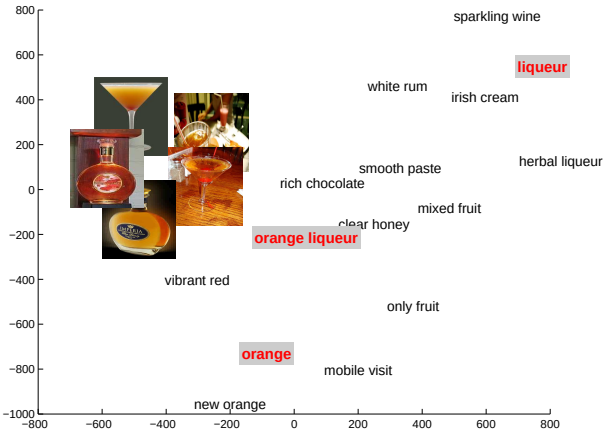


Figure 2: Images tagged with *orange* and *liqueur* are mapped in linguistic space closer to the vector of the phrase *orange liqueur* than to the *orange* or *liqueur* vectors (t-SNE visualization) (the figure also shows the nearest neighbours of phrase, adjective and noun in linguistic space). The mapping is trained using solely noun-annotated images.

associate the visual properties/attributes of the objects to the corresponding adjectives. As an example, Figure 1 (*left*) displays the nearest attributes of *car*, *bird* and *puppy* in the visual space and, interestingly, the relative distance between the noun denoting objects and the adjective denoting attributes is also preserved in the linguistic space (*right*).

We further observe that, as also highlighted by recent work in object recognition, any object in an image is, in a sense, a *visual phrase* (Sadeghi and Farhadi, 2011; Divvala et al., 2014), i.e., the object and its attributes are mutually dependent. For example, we cannot visually isolate the object *drum* from attributes such as *wooden* and *round*. Indeed, within our data, in 80% of the cases the projected image of an object is closer to the semantic representation of a phrase describing it than to either the object or attribute labels. See Figure 2 for an example.

Motivated by this observation, we turn to recent work in distributional semantics defining a vector decomposition framework (Dinu and Baroni, 2014) which, given a vector encoding the meaning of a phrase, aims at decoupling its constituents, producing vectors that can then be matched to a sequence of words best capturing the semantics of the phrase. We adopt this framework to decompose image representations projected onto linguistic space into an adjective-noun phrase. We show that the method yields results comparable to those obtained when using attribute-labeled training data, while only requiring object-annotated data. Interestingly, this *decompositional* approach also doubles the performance of object/noun annotation over the standard zero-shot approach (Experiment 2). Given the positive results of our proposed method, we conclude with an extrinsic evaluation (Experiment 3); we show that attribute-centric representations of images created with the decompositional approach boost performance in an object classification task, supporting claims about its practical utility.

In addition to contributions to image annotation, our work suggests new test beds for distributional semantic representations of nouns and associated adjectives, and provides more in-depth evidence of the potential of the decompositional approach.

2 General experimental setup

2.1 Cross-Modal Mapping

Our approach relies on *cross-modal mapping* from a visual semantic space *V*, populated with vector-based representations of images, onto a linguistic (distributional semantic) space *W* of word vectors. The mapping is performed by first inducing a function $f_{proj} : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$ from data points (v_i, w_i), where $v_i \in \mathbb{R}^{d_1}$ is a vector representation of an image tagged with an object or an attribute (such as *dog* or *metallic*), and $w_i \in \mathbb{R}^{d_2}$ is the linguistic vector representation of the corresponding word. The mapping function can subsequently be applied to any given image $v_i \in V$ to obtain its projection $w'_i \in W$ onto linguistic space:

$$w'_i = f_{proj}(v_i)$$

Specifically, we consider two mapping methods. In the RIDGE regression approach, we learn a linear function $F_{proj} \in \mathbb{R}^{d_2 \times d_1}$ by solving the Tikhonov-Phillips regularization problem, which minimizes the following objective:

$$||W^{Tr} - F_{proj}V^{Tr}||_2^2 - ||\lambda F_{proj}||_2^2,$$

where W^{Tr} and V^{Tr} are obtained by stacking the word vectors w_i and corresponding image vectors v_i, from the training set.[4]

[4]The parameter λ is determined through cross-validation on the training data.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

Second, motivated by the success of Canonical Correlations Analysis (CCA) (Hotelling, 1936) in several vision-and-language tasks, such as image and caption retrieval (Gong et al., 2014; Hardoon et al., 2004; Hodosh et al., 2013), we adapt *normalized Canonical Correlations Analysis* (NCCA) to our setup. Given two paired observation matrices X and Y, in our case W^{Tr} and V^{Tr}, CCA seeks two projection matrices A and B that maximize the correlation between A^TX and B^TY. This can be solved efficiently by applying SVD to

$$\hat{C}_{XX}^{1/2}\hat{C}_{XY}\hat{C}_{YY}^{1/2} = U\Sigma V^T$$

where $\hat{C}$ stands for the covariance matrix. Finally, the projection matrices are defined as $A = \hat{C}_{XX}^{1/2}U$ and $B = \hat{C}_{YY}^{1/2}V$. Gong et al. (2014) propose a *normalized* variant of CCA, in which the projection matrices are further scaled by some power λ of the singular values Σ returned by the SVD solution. In our experiments, we tune the choice of λ on the training data. Trivially, if $\lambda = 0$, NCCA reduces to CCA.

Note that other mapping functions could also be used. We leave a more extensive exploration of possible alternatives to further research, since the details of how the vision-to-text conversion is conducted are not crucial for the current study. As increasingly more effective mapping methods are developed, we can easily plug them into our architecture.

Through the selected cross-modal mapping function, any image can be projected onto linguistic space, where the word (possibly of the appropriate part of speech) corresponding to the nearest vector is returned as a candidate label for the image (following standard practice in distributional semantics, we measure proximity by the *cosine* measure).

2.2 Decomposition

Dinu and Baroni (2014) have recently proposed a general decomposition framework that, given a distributional vector encoding a phrase meaning and the syntactic structure of that phrase, decomposes it into a set of vectors expected to express the semantics of the words that composed the phrase. In our setup, we are interested in a decomposition function $f_{Dec} : \mathbb{R}^{d_2} \rightarrow \mathbb{R}^{2d_2}$ which, given a visual vector projected onto the linguistic space, assumes it represents the meaning of an *adjective-noun phrase*, and decomposes it into two vectors corresponding to the adjective and noun constituents $[w_{adj}; w_{noun}] = f_{Dec}(w_{AN})$. We take f_{Dec} to be a linear function and, following Dinu and Baroni (2014), we use as training data vectors of adjective-noun bigrams directly extracted from the corpus together with the concatenation of the corresponding adjective and noun word vectors. We estimate f_{Dec} by solving a ridge regression problem minimizing the following objective:

$$||[W_{adj}^{Tr}; W_{noun}^{Tr}] - F_{dec}W_{AN}^{Tr}||_2^2 - ||\lambda F_{dec}||_2^2$$

where $W_{adj}^{Tr}, W_{noun}^{Tr}, W_{AN}^{Tr}$ are the matrices obtained by stacking the training data vectors. The λ parameter is tuned through generalized cross-validation (Hastie et al., 2009).

2.3 Representational Spaces

Linguistic Space We construct distributional vectors from text through the method recently proposed by Mikolov et al. (2013), to which we feed a corpus of 2.8 billion words obtained by concatenating English Wikipedia, ukWaC and BNC.[5] Specifically, we used the CBOW algorithm, which induces vectors by predicting a target word given the words surrounding it. We construct vectors of 300 dimensions considering a context window of 5 words to either side of the target, setting the sub-sampling option to 1e-05 and the negative sampling parameter to 5.[6]

Visual Spaces Following standard practice, images are represented as bags of visual words (BoVW) (Sivic and Zisserman, 2003).[7] Local low-level image features are clustered into a set of *visual words* that act as higher-level descriptors. In our case, we use PHOW-color image features, a variant of dense SIFT (Bosch et al., 2007), and a visual vocabulary of 600 words. Spatial information is preserved with a two-level spatial pyramid representation (Lazebnik et al., 2006), achieving a final dimensionality of 12,000. The entire pipeline is implemented using the VLFeat library (Vedaldi and Fulkerson, 2010), and its setup is identical to the

[5] `http://wacky.sslmit.unibo.it`, `http://www.natcorp.ox.ac.uk`

[6] The parameters are tuned on the MEN word similarity dataset (Bruni et al., 2014).

[7] In future research, we might obtain a performance boost simply by using the more advanced visual features recently introduced by Krizhevsky et al. (2012).

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

Category	Attributes
Color	*black, blue, brown, gray, green, orange, pink, red, violet, white, yellow*
Pattern	*spotted, striped*
Shape	*long, round, rectangular, square*
Texture	*furry, smooth, rough, shiny, metallic, vegetation, wooden, wet*

Table 1: List of attributes in the evaluation dataset.

Image	Attributes	Object
	furry white smooth	cat
	green shiny	cocktail

Table 2: Sample annotations from the evaluation dataset.

toolkit's *basic recognition* sample application.[8] We apply Positive Pointwise Mutual Information (Evert, 2005) to the BoVW counts, and reduce the resulting vectors to 300 dimensions using SVD.

2.4 Evaluation Dataset

For evaluation purposes, we use the dataset consisting of images annotated with adjective-noun phrases introduced in Russakovsky and Fei-Fei (2010), which pertains to 384 WordNet/ImageNet synsets with 25 images per synset. The images were manually annotated with 25 attribute-denoting adjectives related to texture, color, pattern and shape, respecting the constraints that a color must cover a significant part of the target object, and all other attributes must pertain to the object as a whole (as opposed to parts). Table 1 lists the 25 attributes and Table 2 illustrates sample annotations.[9]

In order to increase annotation quality, we only consider attributes with full annotator consensus, for a total of 8,449 annotated images, with 2.7 attributes per-image on average. Furthermore, to make the linguistic annotation more natural and avoid sparsity problems, we renamed excessively specific objects with a noun denoting a more general category, following recent work on *entry-level categories* (Ordonez et al., 2013); e.g., *colobus guereza* was relabeled as *monkey*. The final evaluation dataset contains 203 distinct objects.

	Training			Evaluation		
	#im.	#attr.	#obj.	#im.	#attr.	#obj.
Exp. 1	10,749	97	-	*leave-one-attribute-out*		
Exp. 2	23,000	-	750	8,449	25	203

Table 3: Summary of training and evaluation sets.

3 Experiment 1: Zero-shot attribute learning

In Section 1, we showed that there is a significant correlation between pairwise similarities of adjectives in a language-based distributional semantic space and those of visual feature vectors extracted from images labeled with the corresponding attributes. In the first experiment, we test whether this correspondence in attribute-adjective similarity structure across modalities suffices to successfully apply zero-shot labeling. We learn a cross-modal function from an annotated dataset and use it to label images from an evaluation dataset with attributes outside the training set. We will refer to this approach as **DIR$^{\mathbf{A}}$**, for ***Dir**ect **R**etrieval* using ***A**ttribute* annotation. Note that this is the first time that zero-shot techniques are used in the attribute domain. In the present evaluation, we distinguish **DIR$^{\mathbf{A}}$-RIDGE** and **DIR$^{\mathbf{A}}$-NCCA**, according to the cross-modal function used to project from images to linguistic representations (see Section 2.1 above).

3.1 Cross-modal training and evaluation

To gather sufficient data to train a cross-modal mapping function for attributes/adjectives, we combine the publicly available datasets of Farhadi et al. (2009) and Ferrari and Zisserman (2007) with attributes and associated images extracted from MIR-FLICKR (Huiskes and Lew, 2008).[10] The resulting dataset contains 72 distinct attributes and 2,300 images. Each image-attribute pair represents a training data point (v, w_{adj}), where v is the vector representation of the image, and w_{adj} is the linguistic vector of the attribute (corresponding to an adjective). No information about the depicted object is needed.

[8] `http://www.vlfeat.org/applications/apps.html`

[9] Although *vegetation* is a noun, we have kept it in the evaluation set, treating it as an adjective.

[10] We filtered out attributes not expressed by adjectives, such as *wheel* or *leg*.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

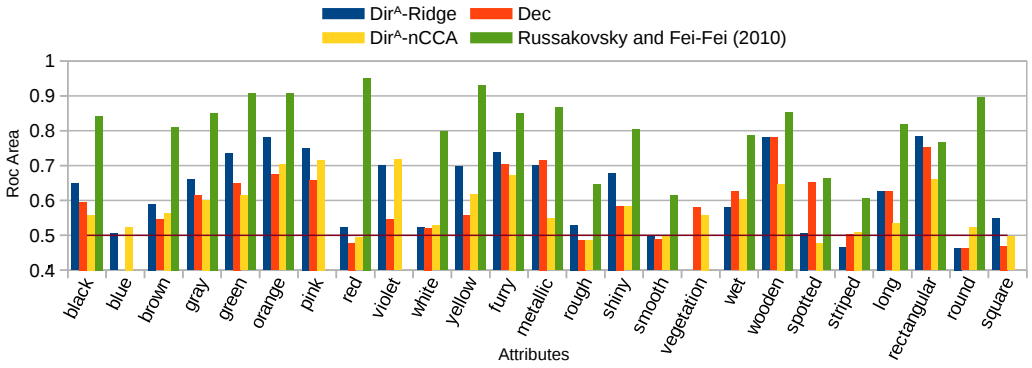


Figure 3: Performance of zero-shot attribute classification (as measured by AUC) compared to the supervised method of Russakovsky and Fei-Fei (2010), where available. The dark-red horizontal line marks chance performance.

To further maximize the amount of training data points, we conduct a leave-one-attribute-out evaluation, in which the cross-modal mapping function is repeatedly learned on all 72 attributes from the training set, as well as all but one attribute from the evaluation set (Section 2.4), and the associated images. This results in $72 + (25 - 1) = 96$ training attributes in total. On average, 45 images per attribute are used. The performance is measured for the single attribute that was excluded from training. A numerical summary of the experiment setup is presented in the first row of Table 3.

3.2 Results and discussion

Russakovsky and Fei-Fei (2010) trained separate SVM classifiers for each attribute in the evaluation dataset in a cross-validation setting. This fully supervised approach can be seen as an ambitious upper bound for zero-shot learning, and we directly compare our performance to theirs using their figure of merit, namely area under the ROC curve (AUC), which is commonly used for binary classification problems.[11] A perfect classifier achieves an AUC of 1, whereas an AUC of 0.5 indicates random guessing. For purposes of AUC computation, DIR$^{\text{A}}$ is considered to label test images with a given adjective if the linguistic-space distance between their mapped representation and the adjective is below a certain threshold. AUC measures the aggregated performance over all thresholds. To get a sense of what AUC compares to in terms of precision and recall, the AUC of DIR$^{\text{A}}$ for *furry* is 0.74, while the precision is 71% and the corresponding recall 14%. For the more difficult *blue* case, AUC is at 0.5, precision and recall are 2% and 55%, respectively.

The AUC results are presented in Figure 3 (ignore red bars for now). We observe first that, of the two mapping functions we considered, RIDGE (blue bars) clearly outperforms NCCA (yellow bars). According to a series of paired permutation tests, RIDGE has a significantly larger AUC in 13/25 cases, NCCA in only 2. This is somewhat surprising given the better performance of NCCA in the experiments of Gong et al. (2014). However, our setup is quite different from theirs: They perform all retrieval tasks by projecting the input visual and language data onto a common multimodal space *different* from both input spaces. NCCA is a well-suited algorithm for this. We aim instead at producing linguistic annotations of images, which is most straightforwardly accomplished by projecting visual representations onto linguistic space. Regression-based learning (in our case, via RIDGE) is a more natural choice for this purpose.

Coming now to a more general analysis of the results, as expected, and analogously to the supervised setting, DIR$^{\text{A}}$-RIDGE performance varies across attributes. Some achieve performance close to the supervised model (e.g., *rectangular* or *wooden*) and, for 18 out of 25, the performance is well above chance (bootstrap test). The exceptions are: *blue, square, round, vegetation, smooth, spotted* and *striped*. Interestingly, for the last 4 attributes in

[11] Table 4 reports hit@k results for DIR$^{\text{A}}$, which will be discussed below in the context of Experiment 2.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

this list, Russakovsky and Fei-Fei (2010) achieved their lowest performance, attributing it to the lower-quality of the corresponding image annotations. Furthermore, Russakovsky and Fei-Fei (2010) excluded 5 attributes due to insufficient training data. Of these, our performance for *blue*, *vegetation* and *square* is not particularly encouraging, but for *violet* and *pink* we achieve more than 0.7 AUC, at the level of the supervised classifiers, suggesting that the proposed method can complement the latter when annotated data are not available.

For a different perspective on the performance of DIR$^{\text{A}}$, we took several objects and queried the model for their most common attribute, based on the average attribute rank across all images of the object in the dataset. Reassuringly, we learn that *sunflowers* are on average *yellow* (mean rank 2.3), *fields* are *green* (4.4), *cabinets* are *wooden* (4) and *vans metallic* (6.6) (*strawberries* are, suspiciously, *blue*, 2.7).

Overall, this experiment shows that, just like object classification, attribute classifiers benefit from knowledge transfer between the visual and linguistic modalities, and zero-shot learning can achieve reasonable performance on attributes and the corresponding adjectives. This conclusion is based on the assumption that *per-image* annotations of attributes are available; in the following section, we show how equal and even better performance can be attained using data sets annotated with objects only, therefore without any hand-coded attribute information.

4 Experiment 2: Learning attributes from objects and visual phrases

Having shown that reasonably accurate annotations of unseen attributes can be obtained with zero-shot learning when a small amount of manual annotation is available, we now proceed to test the intuition, preliminarily supported by the data in Figure 1, that, since objects are bundles of attributes, attributes are implicitly learned together with objects. We thus try to induce attribute-denoting adjective labels by exploiting only widely-available object-noun data. At the same time, building on the observation illustrated in Figure 2 that pictures of objects are pictures of visual phrases, we experiment with a *vector decomposition* model which treats images as composite and derives adjective and noun annotations *jointly*. We compare it with standard zero-shot learning using direct label retrieval as well as against a number of challenging alternatives that exploit gold-standard information about the depicted objects. The second row of Table 3 gives a numerical summary of the setup for this experiment.

4.1 Cross-modal training

We now assume object annotations only, in the form of training data (v, w_{noun}), where v is the vector representation of an image tagged with an object and w_{noun} is the linguistic vector of the corresponding noun. To ensure high imageability and diversity, we use as training object labels those appearing in the CIFAR-100 dataset (Krizhevsky, 2009), combined with those previously used in the work of Farhadi et al. (2009), as well as the most frequent nouns in our corpus that also exist in ImageNet, for a total of 750 objects-nouns. For each object label, we include at most 50 images from the corresponding ImageNet synset, resulting in $\approx 23,000$ training data points. Images containing objects from the evaluation dataset are excluded, so that both adjective and noun retrieval adhere to the zero-shot paradigm.

4.2 Object-agnostic models

DIR$^{\text{O}}$ The ***Dir**ect **R**etrieval using **O**bject annotation* approach projects an image onto the linguistic space and retrieves the nearest adjectives as candidate attribute labels. The only difference with DIR$^{\text{A}}$ (more precisely, DIR$^{\text{A}}$-RIDGE), the zero-shot approach we tested above, is that the mapping function has been trained on object-noun data only.

DEC The ***Dec**omposition* method uses the f_{Dec} function inspired by Dinu and Baroni (2014) (see Section 2.2), to associate the image vector projected onto linguistic space to an adjective and a noun. We train f_{Dec} with about $\approx 50,000$ training instances, selected based on corpus frequency. These data are further balanced by not allowing more than 100 training samples for any adjective or noun in order to prevent very frequent words such as *other* or *new* from dominating the training data. No image data are used, and there is no need for manual annotation, as the adjective-noun tuples are automatically extracted from the corpus.

At test time, given an image to be labeled,

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

we project its visual representation onto the linguistic space and decompose the resulting vector w' into two candidate adjective and noun vectors: $[w'_{adj}; w'_{noun}] = f_{Dec}(w')$. We then search the linguistic space for adjectives and nouns whose vectors are nearest to w'_{adj} and w'_{noun}, respectively.

4.3 Object-informed models

A cross-modal function trained exclusively on object-noun data might be able to capture only prototypical characteristics of an object, as induced from text, independently of whether they are depicted in an image. Although the gold annotation of our dataset should already penalize this image-independent labeling strategy (see Section 2.4), we control for this behaviour by comparing against three models that have access to the gold noun annotations of the image and favor adjectives that are typical modifiers of the nouns.

LM We build a bigram ***L**anguage **M**odel* by using the Berkeley LM toolkit (Pauls and Klein, 2012)[12] on the one-trillion-token Google Web1T corpus[13] and smooth probabilities with the "Stupid" backoff technique (Brants et al., 2007). Given an image with object-**noun** annotation, we score all attributes-adjectives based on the language-model-derived conditional probability $p(adjective|\textbf{noun})$. All images of the same object produce identical rankings. As an example, among the top attributes of *cocktail* we find *heady*, *creamy* and *fruity*.

vLM LM does not exploit visual information about the image to be annotated. A natural way to enhance it is to combine it with DIR$^{\text{O}}$, our cross-modal mapping adjective retrieval method. In the ***v**isually-enriched **L**anguage **M**odel*, we interpolate (using equal weights) the ranks produced by the two models. In the resulting combination, attributes that are both linguistically sensible and likely to be present in the given image should be ranked highest. We expect this approach to be challenging to beat. MacKenzie (2014) recently introduced a similar model in a supervised setting, where it improved over standard attribute classifiers.

[12] https://code.google.com/p/berkeleylm/

[13] https://catalog.ldc.upenn.edu/LDC2006T13

	LM	SP	vLM	DIR$^{\text{O}}$	DEC	DIR$^{\text{A}}$
@1	2	0	5	1	10	7
@5	5	7	16	4	31	23
@10	8	9	29	9	44	37
@20	18	17	50	19	59	51
@50	33	32	72	43	81	68
@100	56	55	82	67	89	77

Table 4: Percentage hit@k attribute retrieval scores.

SP The ***S**electional **P**reference* model robustly captures semantic restrictions imposed by a noun on the adjectives modifying it (Erk et al., 2010). Concretely, for each **noun** denoting a target object, we identify a set of adjectives ADJ_{noun} that co-occur with it in a modifier relation more that 20 times. By averaging the linguistic vectors of these adjectives, we obtain a vector $w_{\textbf{noun}}^{prototypical}$, which should capture the semantics of the *prototypical* adjectives for that **noun**. Adjectives that have higher similarity with this prototype vector are expected to denote typical attributes of the corresponding noun and will be ranked as more probable attributes. Similarly to LM, all images of the same object produce identical rankings. As an example, among the top attributes of *cocktail* we find *fantastic*, *delicious* and *perfect*.

4.4 Results

We evaluate the performance of the models on attribute-denoting adjective retrieval, using a search space containing the top 5,000 most frequent adjectives in our corpus. Tables 4 and 5 present hit@k and recall@k results, respectively ($k \in \{1, 5, 10, 20, 50, 100\}$). *Hit@$k$* measures the percentage of images for which at least one gold attribute exists among the top k retrieved attributes. *Recall@k* measures the proportion of gold attributes retrieved among the top k, relative to the total number of gold attributes for each image.[14]

First of all, we observe that LM and SP – the two models that have access to gold object-noun annotation and are entirely language-based – although well above the random baseline (k/5,000), achieve rather low performance. This confirms that to model our test set accurately, it is not sufficient to predict typical attributes of the depicted objects.

[14] Due to the leave-one-attribute-out approach used to train and test DIR$^{\text{A}}$ (see Section 3), it is not possible to compute recall results for this model.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

	LM	SP	VLM	DIR^{O}	DEC
@1	1	0	2	0	4
@5	2	3	7	2	15
@10	3	5	15	4	23
@20	9	10	30	9	35
@50	20	20	49	22	59
@100	35	34	61	44	70

Table 5: Percentage recall@k attribute retrieval scores.

	DIR^{O}	DEC	DIR^{A}
@1	1	2	0
@5	3	10	0
@10	5	14	1
@20	9	20	2
@50	20	29	6
@100	33	41	12

Table 6: Percentage hit@k noun retrieval scores.

The DIR^{O} method, which exploits visual information, performs numerically similarly to the object-informed models LM and SP, with better hit and recall at high ranks. Although worse than DIR^{A}, the relatively high performance of DIR^{O} is a promising result, suggesting object annotations together with linguistic knowledge extracted in an unsupervised manner from large corpora can replace, to some extent, manual attribute annotations. However, DIR^{O} does not directly model any *semantic compatibility* constraints between the retrieved adjectives and the object present in the image (see examples below). Hence, the object-informed model VLM, which combines visual information wit linguistic co-occurrence statistics, doubles the performance of DIR^{O}, LM and SP.

Our DEC model, which treats images as visual phrases and jointly decouples their semantics, outperforms even VLM by a large margin. It also outperforms DIR^{A}, the standard zero-shot learning approach using attribute-adjective annotated data (see also the attribute-by-attribute AUC comparison between DEC, DIR^{A} and the fully-supervised approach of Russakovsky and Fei-Fei in Figure 3).

Interestingly, accounting for the phrasal nature of visual information leads to substantial performance improvement in object recognition through zero-shot learning (i.e., tagging images with the depicted nouns) as well. Table 6 provides the hit@k results obtained with the DIR^{O} and DEC methods for the **noun** retrieval task in a search space of 10,000 most

Image	**Model**	**Top item**	**Top hit (Rank)**
 A: white, brown N: dog	DEC	A: white N: dog	white (1) dog (1)
	DIR^{O}	A: animal N: goat	white (27) dog (25)
	LM	A: stray	brown (74)
	VLM	A: pet	brown (17)
 A: shiny, round N: syrup	DEC	A: shiny N: flan	shiny (1) syrup (170)
	DIR^{O}	A: crunchy N: ramekin	shiny (15) syrup (113)
	LM	A: chocolate	shiny (84)
	VLM	A: chocolate	shiny (17)

Table 7: Images with gold attribute-adjective and object-noun labels, and highest-ranked items for each model (**Top item**), as well as highest-ranked **correct** item and rank (**Top hit**). Noun results for (V)LM are omitted since these models have access to the gold noun label.

frequent nouns from our corpus. Note that DIR^{O} represents the label retrieval technique that has been standardly used in conjunction with zero-shot learning for objects: The cross-modal function is trained on images annotated with nouns that denote the objects they depict, and it is then used for noun label retrieval of unseen objects through a nearest neighbor search of the mapped image representation (the DIR^{A} column shows that zero-shot noun retrieval using the mapping function trained on adjectives works very poorly). DEC decomposes instead the mapped image representation into two vectors denoting adjective and noun semantics, respectively, and uses the latter to perform the nearest neighbor search for a noun label. Although not directly comparable, the results of DEC reported here are in the same range of state-of-the-art zero-shot learning models for object recognition (Frome et al., 2013).

Annotation examples Table 7 presents some interesting patterns we observed in the results. The first example illustrates the case in which conducting adjective and noun retrieval *independently* results in mixing information, which damages the DIR^{O} approach: Adjectival and nominal properties are not decoupled properly, since the *animal* property of the depicted dog is reflected in both the *animal* adjective and the *goat* noun. At the same time, the *white-*

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

ness of the object (an adjectival property) influences noun selection, since *goats* tend to be white. Instead, DEC unpacks the visual semantics in an accurate and meaningful way, producing correct attribute and noun annotations that form acceptable phrases. LM and VLM are negatively affected by co-occurrence statistics and guess *stray* and *pet* as adjectives, both typical but generic and abstract dog properties.

In the next example, DIR^{O} predicts a reasonable noun label (*ramekin*), focusing on the container rather than the liquid it contains. By ignoring the relation between the adjective and the noun, the resulting adjective annotation (*crunchy*) is semantically incompatible with the noun label, emphasizing the inability of this method to account for semantic relations between attributes-adjectives and object-nouns. DEC, on the other hand, mistakenly annotates the object as *flan* instead of *syrup*. However, having captured the right general category of the object ("smooth gelatinous items that reflect light"), it ranks a semantically appropriate and correct attribute (*shiny*) at the top. Finally, LM and VLM choose *chocolate*, an attribute semantically appropriate for *syrup* but irrelevant for the target image.

Semantic plausibility of phrases The examples above suggest that one fundamental way in which DEC improves over DIR^{O} is by producing semantically coherent adjective-noun combinations. More systematic evidence for this conjecture is provided by a follow-up experiment on the linguistic quality of the generated phrases. We randomly sampled 2 images for each of the 203 objects in our data set. For each image, we let the two models generate 9 descriptive phrases by combining their respective top 3 adjective and noun predictions. From the resulting lists of 3,654 phrases, we picked the 200 most common ones for each model, with only 1/8 of these common phrases being shared by both. The selected phrases were presented (in random order and concealing their origin) to two linguistically-sophisticated annotators, who were asked to rate their degree of semantic plausibility on a 1-3 scale (the annotators were not shown the corresponding images and had to evaluate phrases purely on linguistic/semantic grounds). Since the two judges were largely in agreement ($\rho = 0.63$), we averaged their ratings. The mean averaged plausibility score

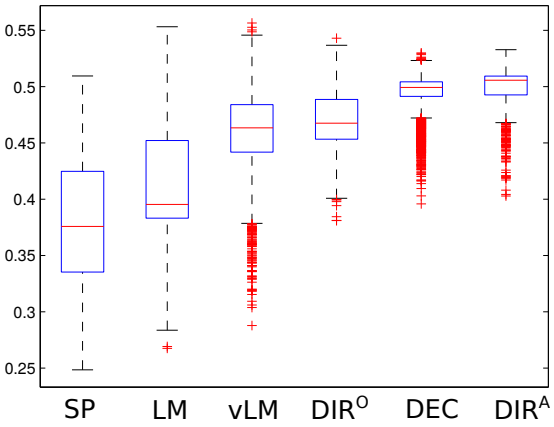


Figure 4: Distributions of (per-image) concreteness scores across different models. Red line marks median values, box edges correspond to 1st and 3rd quartiles, the wiskers extend to the most extreme data points and outliers are plotted individually.

for DIR^{O} phrases was 1.74 (s.d.: 0.76), for DEC it was 2.48 (s.d.: 0.64), with the difference significant according to a Mann-Whitney test. The two annotators agreed in assigning the lowest score ("completely implausible") to more than 1/3 of the DIR^{O} phrases (74/200; e.g., *tinned tostada, animal bird, hollow hyrax*), but they unanimously assigned the lowest score to only 7/200 DEC phrases (e.g., *cylindrical bed-sheet, sweet ramekin, wooden meat*). We thus have solid quantitative support that the superiority of DEC is partially due to how it learns to jointly account for adjective and noun semantics, producing phrases that are linguistically more meaningful.

Adjective concreteness We can gain further insight into the nature of the adjectives chosen by the models by considering the fact that phrases that are meant to describe an object in a picture should mostly contain concrete adjectives, and thus the degree of concreteness of the adjectives produced by a model is an indirect measure of its quality. Following Hill and Korhonen (2014), we define the concreteness of an adjective as the average concreteness score of the nouns it modifies in our text corpus. Noun concreteness scores are taken, in turn, from Turney et al. (2011). For each test image and model, we obtain a concreteness score by averaging the concreteness of the top 5 adjectives that the model selected for the image. Figure 4 reports the distributions of the resulting scores across models. We con-

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

firm that the purely language-based models (LM, SP) are producing generic abstract adjectives that are not appropriate to describe images (e.g., *cryptographic* key, *homemade* bread, *Greek* salad, *beaten* yolk). The image-informed VLM and DIR^{O} models produce considerably more concrete adjectives. Not surprisingly, DIR^{A}, that was directly trained on concrete adjectives, produces the most concrete ones. Importantly, DEC, despite being based on a cross-modal function that was not explicitly exposed to adjectives, produced adjectives that are approaching the concreteness level of those of DIR^{A} (both differences between DEC and DIR^{O}, DEC and DIR^{A} are significant as by paired Mann-Whitney tests).

Image	Object	Predicted Attributes
	aeroplane	thick, wet, dry, cylindrical, motionless, translucent
	dog	cuddly, wild, cute, furry, white, coloured

Table 8: Two VOC images with some top attributes assigned by DEC: these attributes, together with their cosine similarities to the mapped image vectors, serve as attribute-centric representations.

5 Using DEC for attribute-based object classification

As discussed in the introduction, attributes can effectively be used for attribute-based object classification. In this section, we show that classifiers trained on attribute representations created with DEC – which does not require any attribute-annotated training data nor training a battery of attribute classifiers – outperform (and are complementary to) standard BoVW features.

We use a subset of the Pascal VOC 2008 dataset.[15] Specifically, following Farhadi et al. (2009), we use the original VOC training set for training/validation, and the VOC validation set for testing. One-vs-all linear-SVM classifiers are trained for all VOC objects, using 3 alternative image representations.

First, we train directly on BoVW features (**PHOW**, see Section 2.3), as in the classic object recognition pipeline. We compare PHOW to an attribute-centric approach with attribute labels automatically generated by **DEC**. All VOC images are projected onto the linguistic space using the cross-modal mapping function trained with object-noun data only (see Section 4.1), from which we further removed all images depicting a VOC object. Each image projection is then decomposed through DEC into two vectors representing adjective and noun information. The final attribute-centric vector representing an image is created by recording the cosine similarities of the DEC-generated adjective vector with all the adjectives in our linguistic space. Informally, this representation can be thought of as a vector of weights describing the appropriateness of each adjective as an annotation for the image.[16] This is comparable to standard attribute-based classification (Farhadi et al., 2009), in which images are represented as distributions over attributes estimated with a set of *ad hoc* supervised attribute-specific classifiers. Table 8 show examples of top attributes automatically assigned by DEC. While not nearly as accurate as manual annotation, many attributes are relevant to the objects, both as specifically depicted in the image (the aeroplane is *wet*), but also more prototypically (aeroplanes are *cylindrical* in general).

We also perform feature-level fusion (**FUSED**) by *concatenating* the PHOW and DEC features, and reducing the resulting vector to 100 dimensions with SVD (Bruni et al., 2014) (SVD dimensionality determined by cross-validation on the training set).

5.1 Results

There is an improvement over PHOW visual features when using DEC-based attribute vectors, with accuracy raising from 30.49% to 32.76%. The confusion matrices in Figure 5 show that PHOW and DEC do not only differ in quantitative performance, but make different kinds of errors, in part pointing at the different modalities the two models tap into. PHOW, for example, tends to confuse *cats* with *sofas*, probably because the former are often pictured lying on

[15] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/

[16] Given that the resulting representations are very dense, we sparsify them by setting to zeros all adjective dimensions with cosine below the global mean cosine value.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

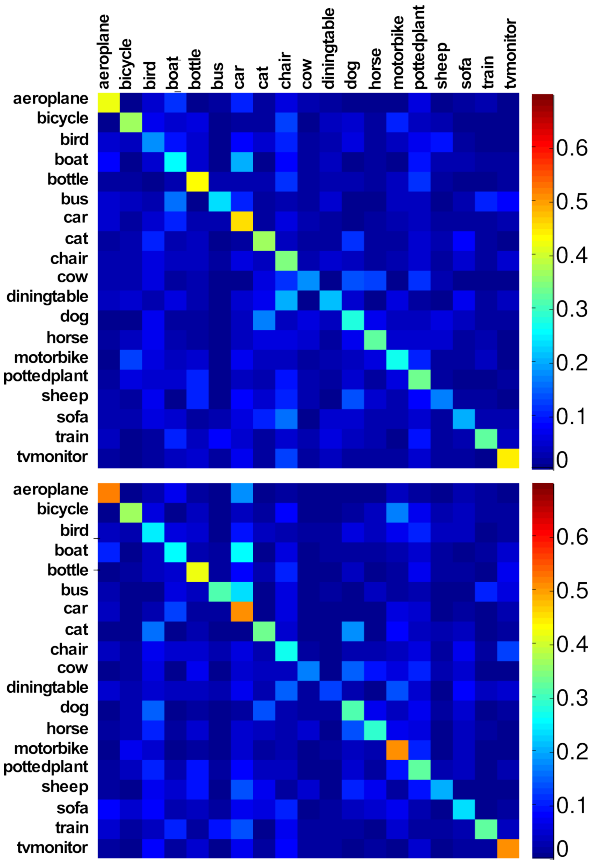


Figure 5: Confusion matrices for PHOW (**top**) and DEC (**bottom**). Warmer-color cells correspond to higher proportions of images with gold row label tagged by an algorithm with the column label (e.g., the first cells show that DEC tags a larger proportion of aeroplanes correctly).

the latter. DEC, on the other hand, tends to confuse *chairs* with *TV monitors*, partially misguided by the taxonomic information encoded in language (both are pieces of furniture). Indeed, the combined FUSED approach outperforms both representations by a large margin (35.81%), confirming that the linguistically-enriched information brought by DEC is to a certain extent complementary to the lower-level visual evidence directly exploited by PHOW. Overall, the performance of our system is quite close to the one obtained by Farhadi et al. (2009) with ensembles of supervised attribute classifiers trained on manually annotated data (the most comparable accuracy from their Table 1 is at 34.3%).[17]

[17]Farhadi and colleagues reduce the bias for the *people* category by reporting mean per-class accuracy; we directly excluded *people* from our version of the data set.

6 Conclusion

We extended zero-shot image labeling beyond objects, showing that it is possible to tag images with attribute-denoting adjectives that were not seen during training. For some attributes, performance was comparable to that of per-attribute supervised classifiers. We further showed that attributes are *implicitly* induced when learning to map visual vectors of objects to their linguistic realizations as nouns, and that improvements in both attribute *and* noun retrieval are attained by treating images as *visual phrases*, whose linguistic representations must be decomposed into a coherent word sequence. The resulting model outperformed a set of strong rivals. While the performance of the zero-shot decompositional approach in the adjective-noun phrase labeling alone might still be low for practical applications, this model can still produce attribute-based representations that significantly improve performance in a supervised object recognition task, when combined with standard visual features.

By mapping attributes and objects to phrases in a linguistic space, we are also likely to produce more *natural* descriptions than those currently used in computer vision (*fluffy kittens* rather than *2-boxy tables*). In future work, we want to delve more into the linguistic and pragmatic naturalness of attributes: Can we predict not just which attributes of a depicted object are true, but which are more *salient* and thus more likely to be mentioned (*red car* over *metal car*)? Can we pick the most appropriate adjective to denote an attribute given the object in the picture (*moist*, rather than *damp lips*)? We should also address attribute dependencies: by ignoring them, we currently get undesired results, such as the aeroplane in Table 8 being tagged as both *wet* and *dry*. More ambitiously, inspired by Karpathy et al. (2014), we plan to associate image *fragments* with phrases of arbitrary syntactic structures (e.g., *PP*s for backgrounds, a *VP*s for main events), paving the way to full-fledged caption generation.

Acknowledgments

We thank the TACL reviewers for their feedback. We were supported by ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

References

Tamara Berg, Alexander Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy Web data. In *Proceedings of ECCV*, pages 663–676, Crete, Greece.

Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Image classification using random forests and ferns. In *Proceedings of ICCV*, pages 1–8, Rio de Janeiro, Brazil.

Thorsten Brants, Ashok Popat, Peng Xu, Franz Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of EMNLP*, pages 858–867, Prague, Czech Republic.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Jia Deng, Wei Dong, Richard Socher, Lia-Ji Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255, Miami Beach, FL.

Georgiana Dinu and Marco Baroni. 2014. How to make words with vectors: Phrase generation in distributional semantics. In *Proceedings of ACL*, pages 624–633, Baltimore, MD.

Santosh Divvala, Ali Farhadi, and Carlos Guestrin. 2014. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of CVPR*, Columbus, OH.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences*. Ph.D dissertation, Stuttgart University.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *Proceedings of CVPR*, pages 1778–1785, Miami Beach, FL.

Ali Farhadi, Mohsen Hejrati, Mohammad A. Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of ECCV*, Crete, Greece.

Vittorio Ferrari and Andrew Zisserman. 2007. Learning visual attributes. In *Proceedings of NIPS*, pages 433–440, Vancouver, Canada.

Andrea Frome, Greg Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A deep visual-semantic embedding model. In *Proceedings of NIPS*, pages 2121–2129, Lake Tahoe, NV.

Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *Proceedings of ECCV*, pages 529–545, Zurich, Switzerland.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning, 2nd edition*. Springer, New York.

Felix Hill and Anna Korhonen. 2014. Concreteness and subjectivity as dimensions of lexical meaning. In *Proceedings of ACL*, pages 725–731, Baltimore, Maryland.

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.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Mark Huiskes and Michael Lew. 2008. The MIR Flickr retrieval evaluation. In *Proceedings of MIR*, pages 39–43, New York, NY.

Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Proceedings of NIPS*, pages 1097–1105, Montreal, Canada.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of NIPS*, pages 1097–1105, Lake Tahoe, Nevada.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. Master's thesis.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander Berg, and Tamara Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of CVPR*, pages 1601–1608, Colorado Springs, CO.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of CVPR*, pages 951–958, Miami Beach, FL.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of ACL*, pages 1403–1414, Baltimore, MD.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of CVPR*, pages 2169–2178, Washington, DC.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024

Calvin MacKenzie. 2014. Integrating visual and linguistic information to describe properties of objects. Undergraduate Honors Thesis, Computer Science Department, University of Texas at Austin.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. `http://arxiv.org/abs/1301.3781/`.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of EACL*, pages 747–756, Avignon, France.

Gregory Murphy. 2002. *The Big Book of Concepts*. MIT Press, Cambridge, MA.

Vicente Ordonez, Jia Deng, Yejin Choi, Alexander Berg, and Tamara Berg. 2013. From large scale image categorization to entry-level categories. In *Proceedings of ICCV*, pages 1–8, Sydney, Australia.

Genevieve Patterson, Chen Xu, Hang Su, and James Hays. 2014. The SUN attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of ACL*, pages 959–968, Jeju Island, Korea.

Olga Russakovsky and Li Fei-Fei. 2010. Attribute learning in large-scale datasets. In *Proceedings of ECCV*, pages 1–14.

Mohammad Sadeghi and Ali Farhadi. 2011. Recognition using visual phrases. In *Proceedings of CVPR*, pages 1745–1752, Colorado Springs, CO.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of semantic representation with visual attributes. In *Proceedings of ACL*, pages 572–582, Sofia, Bulgaria.

Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of ICCV*, pages 1470–1477, Nice, France.

Richard Socher, Milind Ganjoo, Christopher Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Proceedings of NIPS*, pages 935–943, Lake Tahoe, NV.

Peter Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of EMNLP*, pages 680–690, Edinburgh, UK.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605).

Andrea Vedaldi and Brian Fulkerson. 2010. VLFeat – an open and portable library of computer vision algorithms. In *Proceedings of ACM Multimedia*, pages 1469–1472, Firenze, Italy.

Andrea Vedaldi, Siddarth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew Blaschko, David Weiss, Ben Taskar, Karen Simonyan, Naomi Saphra, and Sammy Mohamed. 2014. Understanding objects in detail with fine-grained attributes. In *Proceedings of CVPR*, Columbus, OH.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00132 by guest on 29 March 2024