Visually Grounded Meaning Representations

Carina Silberer, *Member, IEEE,* Vittorio Ferrari, *Member, IEEE,* Mirella Lapata, *Member, IEEE*

Abstract—In this paper we address the problem of grounding distributional representations of lexical meaning. We introduce a new model which uses stacked autoencoders to learn higher-level representations from textual and visual input. The visual modality is encoded via vectors of *attributes* obtained automatically from images. We create a new large-scale taxonomy of 600 visual attributes representing more than 500 concepts and 700K images. We use this dataset to train attribute classifiers and integrate their predictions with text-based distributional models of word meaning. We evaluate our model on its ability to simulate word similarity judgments and concept categorization. On both tasks, our model yields a better fit to behavioral data compared to baselines and related models which either rely on a single modality or do not make use of attribute-based input.

Index Terms—Cognitive simulation, Computer vision, Distributed representations, Concept learning, Connectionism and neural nets, Natural Language Processing

1 Introduction

RECENT years have seen a surge of interest in single word vector spaces ([1], [2], [3]) and their successful use in many natural language applications. Examples include information retrieval [4], search query expansions [5], document classification [6], and question answering [7]. Vector spaces have been also popular in cognitive science figuring prominently in simulations of human behavior involving semantic priming, deep dyslexia, text comprehension, synonym selection, and similarity judgments (see [8] and the references therein). In general, these models specify mechanisms for constructing semantic representations from text corpora based on the *distributional hypothesis* [9]: words that appear in similar linguistic contexts are likely to have related meanings. Despite their success in modeling semantic phenomena, vector spaces have been criticized as "disembodied" in that they represent word meaning as the result of statistical text analysis only, falling short of binding words to their real-world referents [10]. Many behavioral studies suggest that word meaning is *grounded* in the external environment and relates to sensorimotor experience ([11], [12], [13]).

To account for this, new types of perceptually grounded distributional models have emerged. These models conceptualize the problem of meaning representation as one of learning from multiple views corresponding to different modalities (i.e., textual and perceptual input). They still represent words as vectors resulting from the combination of representations with different statistical properties that do not necessarily have a natural correspondence (e.g., text and images). A few models use feature norms (e.g., [14]) as a proxy for sensorimotor experience ([15], [16]). These are obtained by asking native speakers to write down attributes they consider important in describing the meaning of a word. The attributes represent perceived physical and functional properties associated with the referents of words. For example, *apples* are typically green or red, round, shiny, smooth, crunchy, tasty, and so on; *dogs* have four legs and bark, whereas *chairs* are used for sitting. Numerous theories and models in cognitive science are based on representations involving semantic attributes ([17], [18]) which are thought to represent salient psychological aspects of word meaning including multisensory information from the nonlinguistic environment. Other approaches focus on the visual modality as a major source of perceptual data and extract information automatically from images ([19], [20], [21]), or combine norming data and image feature extraction techniques [22].

In this article, we introduce a model, illustrated in Figure 1, which draws elements from connectionist, attribute-based, and distributional models to learn grounded meaning representations. Our model uses (stacked) autoencoders [23], a variant of multilayer neural networks, to learn high-level meaning representations by mapping words and images into a common hidden space. The literature describes several successful approaches to multimodal learning using different variants of deep networks ([24], [25]) and data sources including text, images, audio, and video. Unlike most previous work, our model computes representations for *individual* words and is unique in its use of *attributes* as a means of representing the visual modality.

To this end, we created a large-scale dataset consisting of nearly 700K images and a taxonomy of 636 visual attributes. We used this dataset to train attribute classifiers and extract attribute predictions for new images. We obtained textual attributes from Strudel [26], a distributional model akin to other vector-based models except that collocates of a concept are established by relations to other concepts interpreted as properties. It is important to note that our model is not attribute-specific, any type of visual and textual representation (encoded as a vector) can serve as input. Nonetheless, we argue that attributes are well-suited to describing visual phenomena (e.g., objects, scenes, actions), beyond providing a natural and cognitively motivated way of expressing salient properties of word meaning. They

- *The authors are with the School of Informatics, University of Edinburgh, Edinburgh, United Kingdom. C. Silberer and M. Lapata are with the Institute for Language, Cognition and Computation; V. Ferrari is with the Institute of Perception, Action and Behaviour*

Manuscript received March 4, 2016; revised July 19, 2016; accepted November 18, 2016

allow to generalize to new instances when there are no training examples available; it is possible to say something about new objects even though we cannot identify them (e.g., it has a beak and a long tail). In other words, attributes transcend category[1] and task boundaries whilst offering a generic description of visual data [27] (e.g., animals have stripes and so do clothing items; balls are round, and so are oranges and coins). They are also expedient from a modeling perspective allowing for easier integration of different modalities. In our case, visual and textual knowledge are rendered in the same medium, namely, language. The attributes we use are similar to those provided by participants in norming studies, but importantly *learned* from training data and thus applicable to new instances without additional human involvement.

We experimentally evaluated the meaning representations our model produces using attributes as well as more standard encodings based on neural network architectures (e.g., CNN features for representing images and skip-gram embeddings for representing words). We present results on two tasks, namely word similarity and concept categorization. In the first task, model estimates of word similarity (e.g., *gem–jewel* are similar but *glass–peanut* are not) are compared against elicited similarity ratings. We performed a large-scale evaluation on a new dataset consisting of human judgments for 7,576 word pairs. The dataset contains ratings for *visual* and *textual* similarity, thus allowing to study the two modalities (and their contribution to meaning representation) together and in isolation. In the second task, we assessed whether the learned representations are appropriate for categorization, i.e., grouping a set of objects into meaningful semantic categories (e.g., *peach* and *apple* are members of FRUIT, whereas *chair* and *table* are FURNITURE).

Our contributions in this work are threefold: we introduce a novel modeling framework for grounded meaning representations based on semantic attributes; we demonstrate that the proposed model learns representations which are meaningful and more accurate compared to models based on individual modalities, different modality integration mechanisms, and non-attribute based visual encodings; we create two novel resources which we hope will be of use to the computer vision and NLP communities: a large-scale taxonomy of visual attributes based on ImageNet, and a large dataset of (visual and semantic) word similarity judgments for the evaluation of grounded semantic spaces.

In the remainder of this article, we first present an overview of related work. We then describe our model, explain how higher-level meaning representations are learned and discuss the type of visual and textual input it expects. Experimental results and discussion conclude the paper.

2 Related Work

The presented model has connections to several lines of research in neural networks, natural language processing, computer vision, and more generally multimodal learning. We review related work in these areas below.

Neural Network Models of Meaning Connectionist models of semantic representations have a long tradition in cognitive science (see [28] for an overview) where they are typically employed to study semantic memory and its impairments. Many early models were trained using feature norms (e.g., [18]) or hand-crafted attributes which either explicitly correspond to the units of the semantic representation (e.g., [29]) or are presented as input (or output) in order to learn abstract (distributed) representations in the network's hidden layer (e.g., [30], [31]). In Rogers et al. [32], the hidden layer unifies the representations of verbal and visual units corresponding to linguistic and perceptual properties of objects. A few models make use of neural network architectures based on autoencoders ([18], [31]). In contrast to these models, we learn deeper representations by means of more than one hidden layer. More importantly, to the best of our knowledge, we present the first model to use as input attribute activations automatically extracted from text and image data.

There has been a recent surge of interest in the development of neural network architectures that learn word representations corresponding to vectors of activation of network units, a.k.a. *word embeddings*. A notable difference between these models and earlier models outlined above is that learning is performed on unlabeled text corpora using various methods inspired from neural-network language modeling. The *continuous skip-gram model* [33] is one of the best-known word-embedding models. It not only produces useful word representations, it is also efficient to train, and scales to very large corpora (billions of words).

Grounded Semantic Spaces Grounded semantic spaces are essentially distributional models augmented with perceptual information. Existing models mainly differ with respect to the type of perceptual information used and the way it is integrated with linguistic information.

Some models ([15], [16], [34]) use feature norms as an approximation of the perceptual environment. Other models focus on the visual modality and exploit image databases, such as ImageNet [35] or ESP [36]. A few approaches ([19], [37]) use *visual words* which they derive by clustering SIFT descriptors [38] extracted from images, or combine both feature norms and visual words [22]. Drawing inspiration from the successful application of attribute classifiers in object recognition, Silberer et al. [21] show that automatically predicted visual attributes from images can act as substitutes for feature norms without any critical information loss. In other work ([39], [40]) representations for the visual modality are obtained directly from image pixels using the feature extraction layers of a deep convolutional neural network (CNN) trained on a large labeled object recognition data set. Finally, some models use human generated image tags as a proxy for visual information ([20], [34]).

As far as the integration mechanism is concerned, the simplest method is to concatenate the vectors corresponding to a word's perceptual and linguistic representation ([39], [41]). Other approaches infer bimodal representations over latent variables responsible for the co-occurrence of words over featural dimensions. Bruni et al. [37] concatenate two independently constructed textual and visual spaces and subsequently project them onto a lower-dimensional space using Singular Value Decomposition (SVD). Several models ([15], [19], [22]) present extensions of Latent Dirich-

[1] We use the term concept to refer to basic-level concepts (e.g., *chair*), and the term category for superordinate concepts (e.g., FURNITURE).

let Allocation (LDA, [42]) where topic distributions are learned from words *and* other perceptual units treating them both as observed variables. Hill and Korhonen [34] extend the skip-gram network model [33] in a similar fashion, perceptual input is encoded verbally and treated as a word's linguistic context, whereas Lazaridou et al. [40] modify skip-gram's learning objective so that representations are trained to predict linguistic and visual features. In most cases the visual and textual modalities are decoupled and obtained independently, i.e., from text corpora and feature norms or image databases (but see [19] for an exception).

Our model uses stacked autoencoders to learn higher-level vector representations from textual and visual input. Rather than simply adding perceptual information to textual data it integrates both modalities *jointly* in a single representation which is desirable, at least from a cognitive perspective. It is unlikely that we have separate representations for different aspects of word meaning [32]. Following earlier work, we also train our model on independently collated linguistic and visual data. However, in our case, the two modalities are unified in their representation by natural language attributes.

Multimodal Deep Learning Our work employs deep learning to project linguistic and visual information onto a unified representation that fuses the two modalities together. The use of stacked autoencoders to extract a shared lexical meaning representation is new to our knowledge, although related to a large body of work on deep learning.

Previous work has focused on projecting words and images onto a common space using a variety of methods including deep and restricted Boltzman machines [43], autoencoders [44], and recursive neural networks [45]. Similar methods were employed to combine other modalities such as speech and video or images ([24], [43]). Although our model is conceptually similar to these studies (especially those applying stacked autoencoders), it differs in at least two aspects. Firstly, many former models learn bimodal representations with the aim to reason about one modality given the other ([24], [46]); in contrast, we aim to learn an optimal representation unifying complimentary and redundant information from different modalities. Secondly, most approaches deal with a specific end task (e.g., image classification, but see [43] for an exception). Different modalities are unified in a task-independent representation by means of an unsupervised criterion and subsequently fine-tuned with a supervised criterion [47] or used as features for training a conventional classifier ([24], [46]); in contrast, we aim to learn generic representations and fine-tune our autoencoder with a semi-supervised criterion. We use a combined objective comprising the reconstruction of the attribute-based representation and object classification of the input. Furthermore, our model is defined at a finer level of granularity than most previous work—it computes representations for *individual* words—and leverages information from decoupled data sources, i.e., image collections and text corpora. Existing work on multimodal representation learning exploits images and their associated tags ([43], [46]) or captions with the aim of performing image description generation or retrieval ([48], [49], [50], [51]). These models yield *task-specific* bimodal representations by using a training criterion and an architecture suited to the task at hand. Finally, there are instances of cross-modal learning methods ([52], [53]) which are similar to our work in that they operate on the word-level using disjoint data. But unlike us, they learn a *mapping* between two modalities to tackle an image-based task such as zero-shot classification.

Extracting Attributes from Data A key prerequisite in describing images by their attributes is the availability of training data for learning attribute classifiers. Initial work [54] on automatic attribute extraction from images focused on simple color and texture attributes (e.g., blue, stripes) and showed that these can be learned in a weakly supervised setting from images returned by a search engine when using the attribute as a query. Farhadi et al. [27] developed a dataset representing 20 objects from the PASCAL Visual Object Classes Challenge 2008 [55] and annotated approximately 12,000 instances with attributes from an inventory of 64 attribute labels. The dataset created by Lampert et al. [56] contains over 30,000 images representing 50 animal concepts. It describes each concept using the 85 attributes from the norming study of [57]. Other work focuses on face descriptions [58], scenes [59], specific animal categories [60], or human actions [61]. Although our database shares many features with previous work ([27], [56]), it differs in focus and scope. Since our goal is to develop distributional models that are applicable to many words, it contains a considerably larger number of concepts (i.e., more than 500), images (i.e., 700K), and attributes (i.e., 636) based on a detailed taxonomy which we argue is cognitively plausible and beneficial for image and NLP tasks.

Several methods have been developed for extracting norm-like attributes from text using pattern-based approaches and co-occurrence association measures [62], dependency parsing and WordNet [63] as well as manual extraction rules [64]. We obtain textual attributes from Strudel [26], an unsupervised pattern-based system which extracts weighted concept-property pairs (e.g., *swan*–bird:n) from a text corpus. We opted for Strudel due to its knowledge-lean approach—it merely expects part-of-speech (PoS) tagged input—and the fact that it has a bias towards non-perceptual properties such as actions, functions or situations [26].

3 Autoencoders for Grounded Semantic Representations

Our model builds upon autoencoders to learn higher-level meaning representations for single words. We first briefly review autoencoders placing emphasis on aspects relevant to our model which we then describe in Section 3.2.

3.1 Background

An autoencoder is an unsupervised feed-forward neural network which is trained to reconstruct a given input from its hidden distributed representation ([65], [66]). A basic autoencoder consists of an encoder f_θ which maps an input vector $\mathbf{x}^{(\mathbf{i})}$ to a hidden representation $\mathbf{y}^{(\mathbf{i})} = f_\theta(\mathbf{x}^{(\mathbf{i})}) = s(\mathbf{W}\mathbf{x}^{(\mathbf{i})} + \mathbf{b})$, with s being a non-linear activation function, such as a sigmoid function. A decoder $g_{\theta'}$ then aims to reconstruct input $\mathbf{x}^{(\mathbf{i})}$ from $\mathbf{y}^{(i)}$, i.e., $\hat{\mathbf{x}}^{(i)} = g_{\theta'}(\mathbf{y}^{(i)}) = s(\mathbf{W}'\mathbf{y}^{(i)} + \mathbf{b}')$. The training objective is to determine the parameters $\hat{\theta} = \{\mathbf{W}, \mathbf{b}\}$ and $\hat{\theta}' = \{\mathbf{W}', \mathbf{b}'\}$ which minimize the average reconstruction error over a set of input vectors $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\}$:

$$\hat{\theta}, \hat{\theta}' = \underset{\theta, \theta'}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(\mathbf{x}^{(i)}, g_{\theta'}(f_{\theta}(\mathbf{x}^{(i)}))), \tag{1}$$

where L is a loss function, such as cross-entropy. Parameters θ and θ' can be optimized by gradient descent methods.

Autoencoders are a means to learn representations of some input by retaining useful features in the encoding phase which help reconstruct (an approximation of) the input, whilst discarding useless or noisy ones. A bottleneck hidden layer that has a smaller number of units compared to the input is commonly used to guide parameter learning. Other strategies include constraining the hidden layer to yield sparse representations, or *denoising* [67]. The training criterion with denoising autoencoders is the reconstruction of clean input $\mathbf{x^{(i)}}$ given a corrupted version $\tilde{\mathbf{x}}^{(i)}$. The reconstruction error for input $\mathbf{x^{(i)}}$ with loss L then is:

$$L(\mathbf{x^{(i)}}, g_{\theta'}(f_{\theta}(\tilde{\mathbf{x}}^{(\mathbf{i})}))) \tag{2}$$

One possible corruption process is *masking noise*, where the corrupted version $\tilde{\mathbf{x}}^{(\mathbf{i})}$ results from randomly setting a fraction v of $\mathbf{x^{(i)}}$ to 0.

The underlying idea of denoising autoencoders is that if a latent representation is capable of reconstructing the original input from its corruption, it has presumably learned to capture its structure and can be thus deemed a good representation. From a cognitive perspective, denoising can be construed as learning to activate knowledge about a concept when being exposed to partial information. An example is the ability of humans to recognize objects which are partially occluded or depicted in corrupted images [68].

Several (denoising) autoencoders can be used as building blocks to form a deep neural network [67]. They are often pre-trained layer by layer, with the current layer being fed the hidden representation of the previous autoencoder as input. Using this unsupervised pre-training procedure, initial parameters are found which approximate a good solution. Subsequently, the original input layer and hidden representations of all the autoencoders are stacked and all network parameters are fine-tuned with backpropagation.

To further optimize the parameters of the network, a supervised criterion can be imposed on top of the last hidden layer such as the minimization of a prediction error on a supervised task [66]. Another approach is to unfold the stacked autoencoders and fine-tune them with respect to the minimization of the global reconstruction error [69]. Alternatively, a semi-supervised criterion [70] can be used through combination of the unsupervised training criterion (global reconstruction) with a supervised criterion (prediction of some target given the hidden representation).

3.2 Model Details

To learn meaning representations from textual and visual input, our model employs stacked (denoising) autoencoders (AEs). Both input modalities are vector-based representations of words, or, more precisely, of the objects they refer to (e.g., *canary*, *trolley*). The vector dimensions correspond to textual and visual attributes, examples of which are shown in Table 1 (Details on how these input representations are derived are given in Section 4). We first train AEs with two hidden layers (encodings) for each modality separately. Then, we join these two AEs by feeding their respective second encoding simultaneously to another AE. Its hidden layer $\check{\mathbf{y}}$ yields representations that capture the meaning of words across both modalities. In the final training phase, we stack all layers and unfold them in order to fine-tune a bimodal stacked AE (SAE). Figure 1 illustrates the architecture of the model. As can be seen, we additionally add a softmax-layer on top of the bimodal encoding layer (shown above the bimodal layer in Figure 1), which outputs predictions with respect to the object label of the input (e.g., *dog*). This serves as a supervised training criterion in addition to the unsupervised reconstruction objective during fine-tuning with the aim of guiding the learning towards (bimodal) representations that not only capture the structure of the input patterns within and across the two modalities, but also discriminate between different objects.

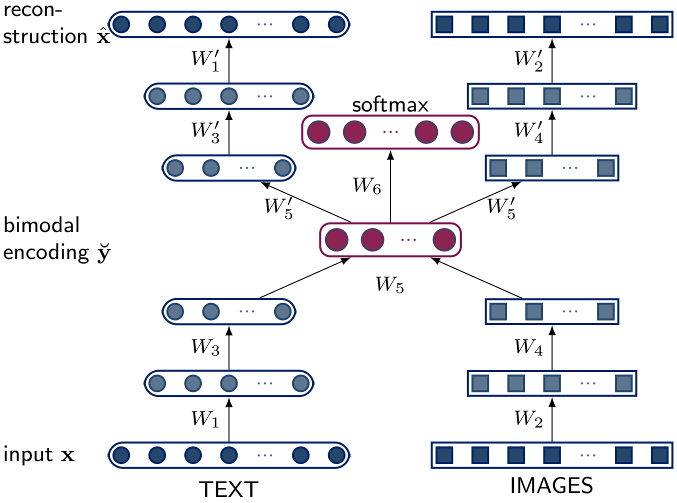


Fig. 1. Stacked autoencoder trained with semi-supervised objective. Input to the model are attribute-based vector representations of single words obtained from text and images (see Table 1).

After training, a word is represented by its encoding in the bimodal layer, corresponding to a vector $\check{\mathbf{y}}$ of distributed unit activations. Note, that an individual unit of $\check{\mathbf{y}}$ does not represent a nameable attribute; rather, it is part of a pattern formed by the interplay between the visual and linguistic characteristics of the word it represents. Two words can then be compared on the basis of their encoding vectors; the more their activation patterns coincide, the more similar the words are assumed to be.

Unimodal Autoencoders For both modalities, we use the hyperbolic tangent as activation function for encoder f_θ and decoder $g_{\theta'}$ and an entropic loss function for L. The weights of each AE are tied, i.e., $\mathbf{W}' = \mathbf{W}^T$. We employ denoising AEs for pre-training the textual modality. Regarding the visual autoencoder, we treat $\mathbf{x}^{(i)}$ itself as corrupted input and construct a new ('denoised') target vector for each input vector $\mathbf{x}^{(i)}$: Each object o (or concept) is represented by multiple images. Each image in turn is rendered in a visual attribute vector $\mathbf{x}^{(i)}$. The target vector is the aggregation of $\mathbf{x}^{(i)}$ and the centroid $\mathbf{x}^{(j)}$ of all attribute vectors col-

Visual		eats_seeds	has_beak	has_claws	has_handlebar	has_wheels	has_wings	is_yellow	made_of_wood
	canary	0.05	0.24	0.15	0.00	–0.10	0.19	0.34	0.00
	trolley	0.00	0.00	0.00	0.30	0.32	0.00	0.00	0.25

Textual		bird:n	breed:v	cage:n	chirp:v	fly:v	track:n	ride:v	run:v	rail:n	wheel:n
	canary	0.16	0.19	0.39	0.13	0.13	0.00	0.00	0.00	0.00	–0.05
	trolley	–0.40	0.00	0.00	0.00	0.00	0.14	0.16	0.33	0.17	0.20

TABLE 1
Examples of attribute-based representations provided as input to our autoencoders.

lectively representing object o. This denoising procedure compensates for prediction errors made by the attribute classifiers on individual images. Moreover, not all attributes which are true for a concept are necessarily observable from a relevant image. Attribute predictions for individual images therefore introduce corruption with respect to the overall *concept* they represent.

Bimodal Autoencoder The bimodal AE is fed with the concatenated final hidden codings of the visual and textual modalities as input and maps these inputs to a joint hidden layer $\check{\mathbf{y}}$ with B units. We normalize both unimodal input codings to unit length. Again, we use tied weights for the bimodal AE. We actively encourage the AE to detect dependencies between the two modalities while learning the mapping to the bimodal hidden layer, and therefore apply masking noise to one modality with a factor v so that the corrupted modality has to optimally rely on the other modality in order to reconstruct its missing input features.

Stacked Bimodal Autoencoder We finally build an SAE with all pre-trained layers and fine-tune them with respect to a semi-supervised criterion. That is, we unfold the stacked autoencoder and furthermore add a softmax output layer on top of the bimodal layer $\check{\mathbf{y}}$ that outputs predictions $\hat{\mathbf{t}}$ with respect to the inputs' object labels (e.g., boat):

$$\hat{\mathbf{t}}^{(i)} = \frac{\exp(\mathbf{W}^{(6)}\check{\mathbf{y}}^{(i)} + \mathbf{b}^{(6)})}{\sum_{k=1}^{O}\exp(\mathbf{W}^{(6)}_{k.}\check{\mathbf{y}}^{(i)} + \mathbf{b}^{(6)}_{k})}, \qquad (3)$$

with weights $\mathbf{W}^{(6)} \in \mathbb{R}^{O\times B}$, $\mathbf{b}^{(6)} \in \mathbb{R}^{O\times 1}$, where O is the number of unique object labels. The overall objective to be minimized is therefore the weighted sum of the reconstruction error L_r and the classification error L_c:

$$L = \frac{1}{n}\sum_{i=1}^{n}\Big(\delta_r L_r(\mathbf{x}^{(i)}, \hat{\mathbf{x}}^{(i)}) + \delta_c L_c(\mathbf{t}^{(i)}, \hat{\mathbf{t}}^{(i)})\Big) + \lambda R \qquad (4)$$

where δ_r and δ_c are weighting parameters that give different importance to the partial objectives, L_c and L_r are entropic loss functions, and R is a regularization term with $R = \sum_{j=1}^{5} 2||\mathbf{W}^{(j)}||^2 + ||\mathbf{W}^{(6)}||^2$. Finally, $\hat{\mathbf{t}}^{(i)}$ is the object label vector predicted by the softmax layer for input vector $\mathbf{x}^{(i)}$, and $\mathbf{t}^{(i)}$ is the correct object label, represented as an O-dimensional one-hot vector[2].

4 Visual and Textual Representations

In this section we describe in more detail how the visual and textual modalities are represented in our model. To obtain visual attribute vectors, we created VISA, a large-scale dataset, consisting of 700K images and attribute descriptions for approximately 500 concepts. In the following we describe VISA and explain how it was used to train SVM-based classifiers that predict visual attributes for images [27]. We also describe how textual attributes were extracted from a corpus.

4.1 Visual Attributes

The VISA Dataset[3] We created the dataset for the nouns contained in the McRae [14] feature norms[4] which cover a wide range of concrete concepts including animate and inanimate things (e.g., animals, clothing, vehicles, utensils, fruits, and vegetables). The norms were elicited by asking participants to list properties (e.g., barks, an_animal, has_legs) describing the nouns they were presented with. We harvested images representing McRae's concepts from ImageNet[5] [35], an ontology of images based on the nominal hierarchy of WordNet [71]. ImageNet has more than 14 million images spanning 21K WordNet synsets. We chose this database due to its high coverage and the high quality of its images (i.e., cleanly labeled and high resolution). McRae et al.'s norms contain 541 concepts out of which 516 appear in ImageNet[6] and are represented by nearly 700K images overall. The average number of images per concept is 1,310 with the most popular being closet (2,149 images) and the least popular prune (5 images).

Our aim was to develop a set of visual attributes that are both discriminating and cognitively plausible, i.e., humans would generally use them to describe a concrete concept. As a starting point, we thus used the visual attributes from McRae's norming study. Attributes capturing other primary sensory information (e.g., smell, sound), functional or motor properties, or encyclopedic information were not taken into account. For example, is_purple is a valid visual attribute for an *eggplant*, whereas a_vegetable is not, since it cannot be visualized. Collating all the visual attributes in the norms resulted in a total of 673 which we further modified and extended during the annotation process explained below.

The annotation was conducted on a *per-concept* rather than a *per-image* basis (as for example in [27]). For each concept (e.g., bear or eggplant), we inspected the images in the development set and chose all visual attributes that applied. If an attribute was generally true for the concept, but the images did not provide enough evidence, the attribute was

[2] In a one-hot vector, the element corresponding to the object label is one and the others are zero.

[3] Available at homepages.inf.ed.ac.uk/csilbere/resources.html.

[4] Available at sites.google.com/site/kenmcraelab/norms-data.

[5] ImageNet is available at http://www.image-net.org.

[6] Some words had to be modified in order to match the correct synset, e.g., *tank_(container)* was found as *storage_tank*.

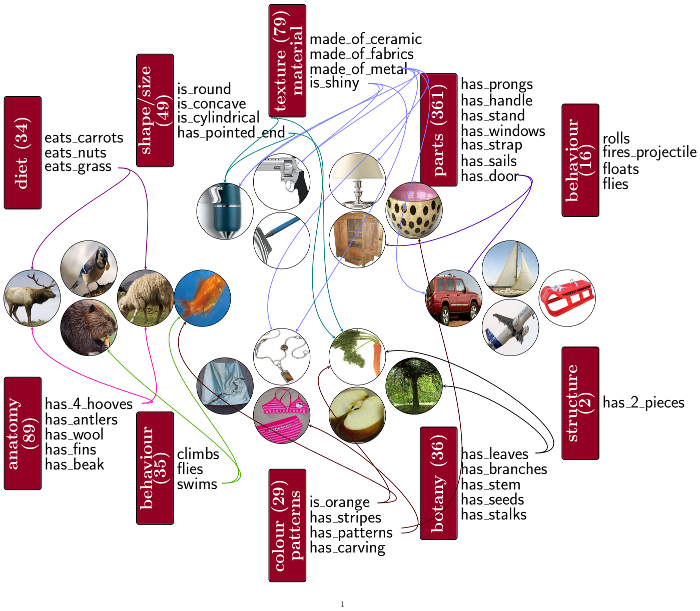


Fig. 2. Attribute categories and examples of attribute instances and images. Parentheses denote the number of attributes per category.

nevertheless chosen and labeled with `<no_evidence>`. For example, a *plum* has_a_pit, but most images in ImageNet show plums where only the outer part of the fruit is visible. We added attributes which were supported by the image data but missing from the initial set as given by the norms. For example, has_lights and has_bumper are attributes of *cars* but are not included in the norms.

Our methodology is slightly different from [56] in that we did not simply transfer the attributes from the norms to the concepts in question but refined and extended them according to the visual data. There are several reasons for this. Firstly, it makes sense to select attributes corroborated by the images. Secondly, by looking at the actual images, we could eliminate errors in McRae's norms. For example, eight participants erroneously thought that a *catfish* has_scales. Thirdly, during the annotation process, we normalized synonymous attributes (e.g., has_pit and has_stone) and attributes that exhibited negligible variations in meaning (e.g., has_stem and has_stalk). Finally, our aim was to collect an exhaustive list of visual attributes for each concept which is consistent across all members of a category. This is unfortunately not the case in McRae's norms. Participants were asked to list up to 14 different properties that describe a concept. As a result, the attributes of a concept denote the set of properties humans consider most salient. For example, both, *lemons* and *oranges* have_pulp. But the norms provide this attribute only for the second concept.

Annotation proceeded on a category-by-category basis, e.g., first all food-related concepts were annotated, then animals, vehicles, and so on. Two annotators (both co-authors of this paper) developed the set of attributes for each category. One annotator first labeled concepts with their attributes, and the other annotator reviewed the annotations, making changes if needed. Annotations were revised and compared per category in order to ensure consistency across all concepts of that category. Attributes were grouped in ten general classes shown in Figure 2 (e.g., anatomy, diet).

Overall, we discarded or modified 262 visual attributes in McRae's norms and added 294 attributes. On average, each concept was annotated with 15 attributes; approximately 11.5 of these were not part of the semantic representation created by McRae's participants for that concept even though they figured in the representations of other concepts. Furthermore, on average two McRae attributes per concept were discarded.

Automatic Extraction of Visual Attributes Following previous work ([27], [56]) we learned one classifier per attribute provided that the attribute had been assigned to at least two concepts in our dataset. This resulted in 414 classifiers in total.[7] We used an L2-regularized L2-loss linear SVM [72] to learn the attribute predictions, and adopted the training procedure of [27].[8]

For each concept in our dataset, we partitioned the corresponding images into a training, development, and test set. For most concepts the development set contained a maximum of 100 images and the test set a maximum of 200 images. Concepts with less than 800 images in total were split into $^1/_8$ test and development set each, and $^3/_4$ training set. The splits were done randomly, regardless of the test set, to which we assigned images for which bounding box annotations (if any) were provided. To learn a classifier for a particular attribute, we used all images in the training data, totaling to approximately 550K images.

Images of concepts annotated with the attribute were used as positive examples, and the rest as negative examples. We optimized cost parameter C on the training data, randomly partitioning it into a split of 70% for training, and 30% for validation. The final SVM for the attribute was trained on the entire training data, i.e., on all positive and negative examples. The SVM learners used the four different feature types proposed in [27], namely color, texture, visual words, and edges. Texture descriptors were computed for each pixel and quantized to the nearest 256 k-means centers. Visual words were constructed with a HOG spatial pyramid. HOG descriptors were quantized into 1000 k-means centers. Edges were detected using a standard Canny detector and their orientations were quantized into eight bins. Color descriptors were sampled for each pixel and quantized to the nearest 128 k-means centers. Shapes and locations were represented by generating histograms for each feature type for each cell in a grid of three vertical and horizontal blocks. Our classifiers used 9,688 features in total. Figure 3 shows classifier predictions for seen (i.e., encountered during training) and unseen concepts, respectively.

We quantitatively evaluated the attribute classifiers by measuring the Average Precision (AP, [73]) on the test set. Since gold annotations in VISA are concept-based, evaluation was performed on concept-level predictions (computed as the centroid of all attribute predictions for images belonging to the same concept—see next paragraph for details on how we compute the concept-level predictions); specifically, we plot precision against recall based on a threshold.[9] Recall is the proportion of correct attribute predictions whose

[7] We furthermore only trained classifiers for attributes corroborated by the images and excluded those labeled with `<no_evidence>`.

[8] Code is available at http://vision.cs.uiuc.edu/attributes/.

[9] Threshold values ranged from 0 to 0.9 with 0.1 stepsize.

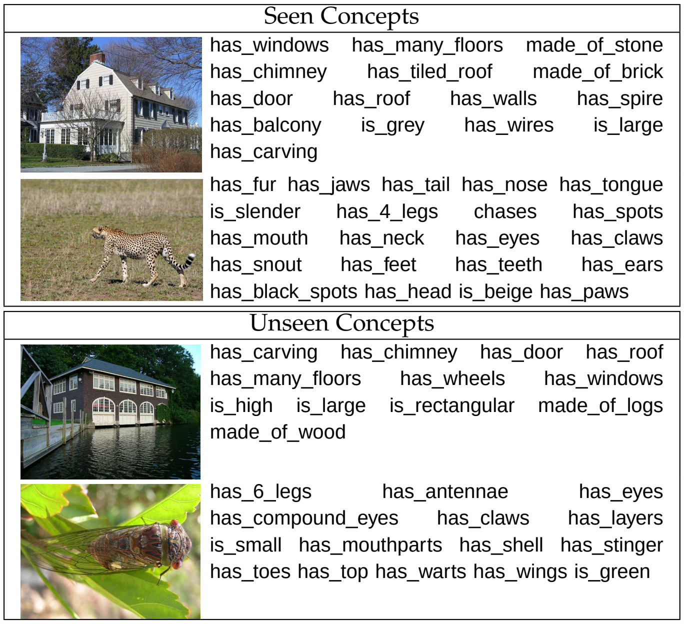

Seen Concepts
has_windows has_many_floors made_of_stone
has_chimney has_tiled_roof made_of_brick
has_door has_roof has_walls has_spire
has_balcony is_grey has_wires is_large
has_carving
has_fur has_jaws has_tail has_nose has_tongue
is_slender has_4_legs chases has_spots
has_mouth has_neck has_eyes has_claws
has_snout has_feet has_teeth has_ears
has_black_spots has_head is_beige has_paws
Unseen Concepts
has_carving has_chimney has_door has_roof
has_many_floors has_wheels has_windows
is_high is_large is_rectangular made_of_logs
made_of_wood
has_6_legs has_antennae has_eyes
has_compound_eyes has_claws has_layers
is_small has_mouthparts has_shell has_stinger
has_toes has_top has_warts has_wings is_green

Fig. 3. Attribute predictions for concepts seen during training (top; *house*, *cheetah*) and unseen concepts (bottom; *boathouse*, *cicada*).

prediction score exceed the threshold to the true attribute assignments given by the dataset, and precision is the fraction of correct attribute predictions to all predictions exceeding the threshold. The interpolated average precision is then the mean of the maximum precision at eleven recall levels $[0, 0.1, ..., 1]$. The precision/recall curve is shown in Figure 4; the attribute classifiers achieved a mean AP of 0.52.

Computing Visual Representations of Concepts The classifiers predict attributes on an image-by-image basis, and a concept w is represented by multiple images. To derive a single representation for w we need to aggregate all related attributes. We use a vector-based representation where each attribute corresponds to a dimension of an underlying semantic space. Just as in text-based semantic spaces, we can then quantify similarity between two concepts by measuring the geometric distance of their vectors. Since we encode visual attributes, however, the underlying semantic space is perceptual, and so is the similarity we measure. For each image $i_w \in I_w$ of concept w, we output an F-dimensional vector containing prediction scores $\mathrm{score}_\mathrm{a}(i_w)$ for attributes $a = 1, ..., F$. We transform these attribute vectors into a single vector $\mathbf{p}_w \in \mathbb{R}^{1\times F}$, by computing the centroid of all vectors for concept w. That is, we average the scores for the various attributes:

$$\mathbf{p}_w = \left(\frac{1}{|I_w|} \sum_{i_w \in I_w} \mathrm{score}_\mathrm{a}(i_w)\right)_{a=1,...,F} \tag{5}$$

The construction of the visual representation for concept *chick* is exemplified in Figure 5.

Table 2 (left) shows the 5 nearest neighbors for seven example concepts from our dataset using the visual attribute vectors. Neighbors for a concept were found by measuring the cosine similarity between the attribute vectors $\mathbf{p}$ of that concept and all other concepts in our dataset and choosing the five concepts with the highest similarity. The examples show that our representation is able to capture visual and semantic similarity discovering neighbors of the same semantic category as the target. For comparison, we show the five nearest neighbors when the example concepts are represented by their textual attribute vectors and by their bimodal vectors as learnt with our SAE model, respectively (Table 2, middle and right, respectively).

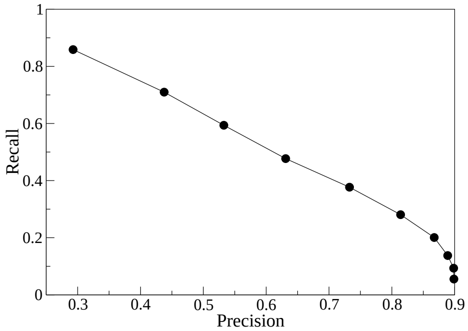


Fig. 4. Attribute classifier performance for thresholds δ (on the test set).

4.2 Textual Attributes

Recall that the SAE can be augmented with any type of input data. Analogously to representing the visual modality through attributes extracted from images, we can represent the textual modality through attributes from text data. To obtain textual attributes we use Strudel[10] [26], a system for automatically extracting weighted concept–attribute pairs (e.g., *chick*–bird:n (60.1), *chick*–brood:v (67.5), *chick*–precocial:j (45.8)) from a lemmatized and PoS-tagged corpus. Strudel takes as input a set of target concepts and a set of patterns, and extracts a list of attributes for each concept. The attributes are not known a priori, but are directly extracted from the corpus. Strudel induces meaning representations that describe a concept via its properties instead of a bag of co-occurring words. Each concept-attribute pair is weighted with a log-likelihood ratio expressing the pair's strength of association. Baroni et al. [26] show that the learned representations can be used as a basis for various tasks such as typicality rating, categorization, or clustering features into types. To obtain a textual semantic space from Strudel's output, we represent each target word as a vector in a high-dimensional space, where each component corresponds to some textual attribute (entries are set to word-attribute log-likelihood ratio scores). Example representations are shown in Table 1.

4.3 Word Embeddings

In addition to attribute-based textual representations, we present experiments with textual embeddings obtained from the continuous skip-gram model [33]. Because of this and the fact that two competitor models we compare against are essentially extensions of skip-gram, we briefly outline how these embeddings are learned. Given a text corpus, skip-gram uses a neural network to learn embeddings for words (and phrases) by optimizing the training objective of predicting the context words of a target word. The network

[10] The software is available from http://clic.cimec.unitn.it/strudel.

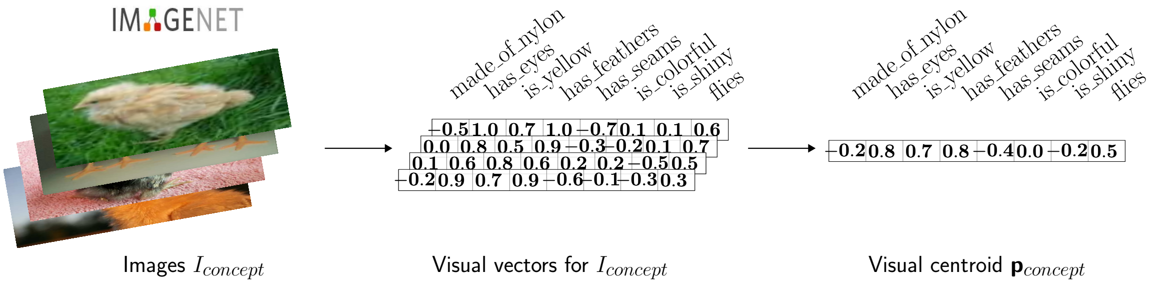


Fig. 5. Visual representation for concept *chick*. Attribute classifiers predict attributes for example images depicting *chicks*. These prediction scores are then converted into vectors (first arrow). To compute a single visual attribute vector for a concept, all vectors are aggregated into $\mathbf{p}_{chick}$.

Concept	Visual Neighbors	Textual Neighbors	SAE Neighbors
ambulance	*van, truck, taxi, bus, limousine*	*helicopter, trolley, van, taxi, train*	*taxi, van, truck, bus, train*
bison	*ox, bull, pony, elephant, bear*	*buffalo, bear, elephant, caribou, deer*	*elk, buffalo, deer, caribou, bear*
brush	*paintbrush, pencil, ladle, hammer, screwdriver*	*comb, paintbrush, vest, scissors, doll*	*comb, paintbrush, pencil, scissors, razor*
dress	*robe, blouse, camisole, nightgown, vest*	*gown, shirt, skirt, blouse, jacket*	*gown, blouse, robe, skirt, nightgown*
hut	*shed, shack, barn, cabin, house*	*shack, cottage, bungalow, cabin, tent*	*shack, cabin, house, cottage, bungalow*
microwave	*oven, shelves, stove, cabinet, freezer*	*stove, oven, freezer, radio, pot*	*radio, stove, oven, freezer, stereo*
scarf	*gloves, shawl, socks, sweater, veil*	*shawl, sweater, cloak, veil, gown*	*shawl, sweater, pyjamas, skirt, socks*

TABLE 2
Nearest neighbors represented by visual, textual and bimodal (SAE) vectors. Neighbors are shown in order of decreasing cosine similarity.

architecture consists of an input layer encoding target w_t, a continuous projection (embedding) layer, and an output layer encoding contexts $w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c}$, within a window surrounding the target w_t. During training, the model uses negative sampling where the objective is to distinguish a correct context word of w_t from k randomly sampled negative examples using logistic regression.

5 Experiment 1: Similarity

Vector-based models aimed at representing the meaning of individual words are commonly evaluated against human similarity judgments. The ability to judge similarity underlies many cognitive tasks such as as semantic priming [74] and practical applications such as document retrieval [4]. In the following we give details on the dataset used for evaluating word similarity, explain how the SAE model was trained, and describe the approaches used for comparison with our own work.

5.1 Method

To evaluate how well our model predicts word similarity ratings, we created a new dataset based on the concepts represented in VISA. Although several related datasets exist, such as the widely used WordSim353 [75] or the more recent Rel-122 norms [76], they contain many abstract words, (e.g., *love–sex* or *arrest–detention*) which are not covered by our visual attributes database. This is for a good reason, as most abstract words do not have discernible attributes, or at least attributes that participants would agree upon. The new dataset we created consists exclusively of concrete nouns which we hope will be useful for the development and evaluation of grounded semantic space models.[11] Initially, we created all possible pairings over McRae's nouns and computed their semantic relatedness using Patwardhan et al.'s [77] WordNet-based measure. We opted for this specific measure as it achieves high correlation with human ratings and has a high coverage on our nouns. Next, for each word we randomly selected 30 pairs under the assumption that they are representative of the full variation of semantic similarity. This resulted in 7,576 word pairs. We split the pairs into overall 255 similarity rating tasks; each task consisted of 32 pairs covering examples of weak to very strong semantic relatedness, and furthermore contained at most one instance of each target word. Two control pairs from Miller and Charles (M&C, [78]) were included in each task to potentially help identify and eliminate data from participants who assigned random scores.

We obtained similarity ratings using Amazon Mechanical Turk (AMT). Participants were first presented with instructions that explained the task and gave examples. They were asked to rate a pair on two dimensions, visual and semantic similarity using a Likert scale of 1 (highly dissimilar) to 5 (highly similar). Each task was completed by five volunteers, all self-reported native English speakers. Participants were allowed to complete as many tasks as they wished. A total of 46 subjects (27 women, 18 men, 1 unspecified, mean age: 38.5 years, age range: 18–67) took part in the study and completed between one and 147 tasks each. Participants were paid $0.5 per task. Examples of the stimuli and elicited mean ratings are shown in Table 3.

The similarity data was post-processed so as to identify and remove outliers. An outlier was any individual whose mean pairwise correlation fell outside two standard deviations from the mean correlation. 11.5% of the annotations were detected as outliers and removed. After outlier removal, we examined how well the participants agreed in their judgments. We measured inter-subject agreement as the average pairwise correlation coefficient (Spearman's ρ) between the ratings of all annotators for

[11] Available at homepages.inf.ed.ac.uk/csilbere/resources.html.

Word Pairs	Semantic	Visual
couch–sofa	5.0	5.0
frog–toad	5.0	5.0
cup–mug	5.0	4.3
gloves–mittens	5.0	4.2
missile–rocket	4.8	5.0
tortoise–turtle	4.8	5.0
bat_(baseball)–baton	2.8	4.0
pencil–wand	1.8	4.0
bracelet–chain	2.8	4.0
pencil–wand	1.8	4.0
car–scooter	4.0	1.7
gun–missile	4.0	1.0
screwdriver–wrench	3.6	1.4
airplane–truck	3.4	1.2

TABLE 3
Mean semantic and visual similarity ratings using a scale of 1 (highly dissimilar) to 5 (highly similar); averaged across AMT participants.

each task. For semantic similarity, the mean correlation was $\rho = 0.76$ (Min=0.34, Max=0.97, StD=0.11) and for visual similarity $\rho = 0.63$ (Min=0.19, Max=0.90, StD=0.14). These results indicate that the participants found the task relatively straightforward and produced similarity ratings with a reasonable level of consistency. The correlation between the average ratings of the AMT annotators and the M&C dataset was $\rho = 0.91$.

5.2 Comparison Models

The SAE model learned meaning representations for the McRae nouns covered by the VISA dataset (see Section 4). As shown in Figure 1, the autoencoder (AE) takes as input two (real-valued) vectors representing the visual and textual modalities. We maintained the partition of the VISA image data into training, validation, and test set and acquired visual vectors for each of the sets by means of our attribute classifiers (see Section 4). We used the visual vectors of the training and development set for training the AEs, and the vectors of the test set for evaluation. Visual vectors were scaled to the $[-1, 1]$ range. Textual attributes were extracted by running Strudel [26] on WaCkypedia, a 2009 dump of the English Wikipedia of about 800M tokens.[12] We only retained the ten attributes with highest log-likelihood ratio scores for each target word which resulted in 2,362 dimensions for the textual vectors. Analogously to the visual representations, association scores were scaled to the $[-1, 1]$ range. We also trained an SAE with skip-gram embeddings (and visual attributes). We obtained 500-dimensional skip-gram embeddings from the same WaCkypedia corpus using hierarchical softmax, negative sampling with $k = 5$, and a window size of $c = 5$.

Parameters for the SAE model were optimized on a subset of the free word association norms collected by Nelson et al. [79]. These were established by presenting participants with a cue word (e.g., *canary*) and asking them to name an associate word in response (e.g., *bird, sing, yellow*). For each cue, the norms provide a set of associates and the frequencies with which they were named. The dataset contains a very large number of cue-associate pairs (63,619 in total) some of which luckily are covered in McRae's norms and by extension in VISA.[13] During training we used correlation analysis (Spearman's ρ) to monitor the degree of linear relationship between model cue-associate (cosine) similarities and human probabilities. The best autoencoder on the word association task obtained a correlation coefficient of 0.33.

The resulting SAE model has the following architecture: the textual autoencoder (see Figure 1, left-hand side) consists of 700 hidden units which are then mapped to the second hidden layer with 500 units (the corruption parameter was set to $v = 0.1$); the visual AE (see Figure 1, right-hand side) has 170 and 100 hidden units, in the first and second layer, respectively. The 500 textual and 100 visual hidden units were fed to a bimodal AE containing 500 hidden units, and masking noise was applied to the textual modality with $v = 0.2$. The weighting parameters for the joint training objective of the stacked autoencoder were set to $\delta_r = 0.8$ and $\delta_c = 1$ (see Equation (4)). The textual AE trained on skip-gram embeddings, in turn, has 490 and 450 hidden units, respectively, and $v = 0.1$. The corresponding bimodal SAE was trained with $v = 0.4$ and $\delta_r = \delta_c = 1$. The other hyperparameters are the same as described above.

Throughout our experiments we compare the meaning representations obtained from the output of the bimodal hidden layer of the SAE against unimodal autoencoders based solely on textual and visual input (left- and right-hand sides in Figure 1 respectively). We also evaluated our model against three latent inference models that differ in their modality integration mechanisms. The first one is based on kernelized canonical correlation analysis (kCCA, [80]) with a linear kernel and was the best performing model in Silberer et al. [21]. The main assumption underlying CCA is two (or more) heterogeneous representations contain some joint information that is reflected in their correlation. Given two random variables $\mathbf{x}$ and $\mathbf{y}$ (or two sets of vectors), CCA can be seen as determining two sets of basis vectors in such a way, that the correlation between the projections of the variables onto these bases is mutually maximized [81]. The second model is a deep learning-based variant of kCCA (DCCA, [82]) which computes representations by passing the two views through a deep network which is fine-tuned to maximize the total correlation of the output layers. The third model emulates Bruni et al.'s [37] integration mechanism. Specifically, we concatenated the textual and visual vectors and projected them onto a lower dimensional latent space using singular value decomposition, a mathematical technique for reducing the dimensionality of semantic spaces [83]. All these models were run on the same data and were given input identical to our model, namely attribute-based visual representations and textual information represented as attributes or skip-gram embeddings.

We further report results with Bruni et al.'s [37] bimodal distributional model using their publicly available system [84]. Their textual modality is represented by a 30K-dimensional co-occurrence matrix[14] extracted from the ukWaC corpus (2 billion tokens)[15] and WaCkypedia. Note that our attribute-based input relies solely on the latter. The entries of the matrix correspond to the weighted co-

[12] From http://wacky.sslmit.unibo.it/doku.php?id=corpora.

[13] 435 word pairs constitute the overlap between Nelson et al.'s norms [79] and McRae et al.'s [14] nouns.

[14] We thank Elia Bruni for providing us with their data.

[15] From http://wacky.sslmit.unibo.it/doku.php?id=corpora.

#	Pair	#	Pair
1	*pliers–tongs*	11	*cello–violin*
2	*cathedral–church*	12	*cottage–house*
3	*cathedral–chapel*	13	*horse–pony*
4	*pistol–revolver*	14	*gun–rifle*
5	*chapel–church*	15	*cedar–oak*
6	*airplane–helicopter*	16	*bull–ox*
7	*dagger–sword*	17	*dress–gown*
8	*pistol–rifle*	18	*bolts–screws*
9	*cloak–robe*	19	*salmon–trout*
10	*nylons–trousers*	20	*oven–stove*

TABLE 4
Word pairs with highest semantic and visual similarity according to SAE model. Pairs are ranked from highest to lowest similarity.

occurrence frequency of a target word (rows) and a context word (columns). Two words are considered co-occurring if one of them occurs in the window of two content words on each side of the other word. Moreover, they extract visual information from the ESP game dataset [36] which comprises 100K images randomly downloaded from the Internet and tagged by humans (the average number of images per tag is 70). The visual modality is represented by bag-of-visual-words histograms built on the basis of clustered SIFT descriptors [38].

Finally, we also compare to two multimodal extensions of Mikolov et al.'s [33] continuous skip-gram model. Kiela and Bottou [39] concatenate skip-gram textual embeddings with vectors based on CNN image features (the vectors representing both modalities are normalized prior to concatenation). The CNN was trained on about 1.6M ImageNet images associated with 1,512 categories ([85], [86]). They obtain 6,144-dimensional feature vectors for arbitrary images using the seventh layer of the pre-trained CNN. A visual representation for an individual word is then computed as the aggregated feature vectors of its associated images. We compare against four variants of their model using either 500-dimensional textual attribute vectors obtained from WaCkypedia or skip-gram embeddings obtained from the same corpus (also with 500 dimensions), and image vectors extracted from ESP and aggregated through averaging[16] or our own visual attributes.

Lazaridou et al.'s [40] model takes into account visual information during training by adding a visual objective[17] to the text-based skip-gram objective. The visual objective is to maximize the similarity between the fixed visual vector of a target word (when available) and its textual embedding to be learned using a max-margin framework. We re-implemented their model and trained it on WaCkypedia representing the visual modality by our visual attribute vectors (as we did not have access to their visual features).

5.3 Results

We evaluated the models described above on the word similarity dataset introduced in Section 5.1. We measure how well model predictions (cosine similarities) correlate with (mean) human similarity ratings using Spearman's ρ. Table 5 summarizes our results. The table is divided into three parts. The upper part (see McRae row) report results of a distributional model induced from McRae's original norms as an indicator of how well automatically extracted attributes can approach the performance of clean human generated attributes. Each noun is represented as a vector with dimensions corresponding to attributes elicited from participants of the norming study. Vector components are set to the (normalized) frequency with which participants generated the corresponding attribute. The middle part reports the performance of models which integrate the two modalities into a joint representation. We report results with our SAE model, SVD, and the CCA models using (automatically obtained) textual and visual attributes (tAttrib, vAttrib) or skip-gram embeddings and visual attributes (skip-gram, vAttrib). We also compare SAE to Lazaridou et al.'s [40] multimodal skip-gram model and Bruni et al. [37]. The third section of the table presents concatenation models using our textual and visual attributes (tAttrib, vAttrib), skip-gram embeddings, CNN features, and combinations thereof. We show results using both textual and visual modalities (T+V) and each of them individually (T or V), wherever possible.[18]

We observe that amongst models trained on attribute-based input the bimodal SAE (tAttrib, vAttrib, T+V) performs best on both similarity tasks. Table 4 shows examples of word pairs with highest semantic and visual similarity according to this model. We also observe that simply concatenating textual and visual attributes (tAttrib+vAttrib, T+V) performs competitively with SVD and better than the CCA models. This indicates that the attribute-based representation is a powerful predictor on its own. Moreover, the visual attributes outperform Kiela and Bottou's [39] CNN-based features (skip-gram+CNN, V) on both similarity tasks, and the concatenation of the latter with skip-gram vectors is as effective as skip-gram alone (skip-gram+CNN, T and T+V). On the other hand, the textual attributes fall short compared to the skip-gram embeddings (tAttrib vs. skip-gram, T). The bimodal SAE trained on the latter (skip-gram, vAttrib; T+V) is the overall best model, outperforming SVD and CCA (skip-gram, vAttrib, T+V), Lazaridou et al. [40], Bruni et al. [37], and all concatenation models. It yields a correlation coefficient of ρ =.77 on semantic similarity and $\rho = 0.66$ on visual similarity. Human agreement is 0.77 on the former task and 0.63 on the latter. We present information on the statistical significance of the reported results in Appendix A, which is available in the online supplemental material.

Moreover, we would expect the textual modality to be more dominant when modeling semantic similarity and conversely the perceptual modality to be stronger with respect to visual similarity. This is borne out in our unimodal SAEs for both types of textual input. The textual SAEs correlate better with semantic similarity judgments ($\rho = 0.67$ and $\rho = 0.74$) than their visual equivalent ($\rho = 0.61$). And the visual SAEs correlate better with visual similarity judgments ($\rho = 0.60$) compared to the textual SAEs ($\rho = 0.55$, $\rho = 0.59$). Interestingly, the bimodal SAEs are better than the unimodal variants on both types of similarity judgments. This suggests that both modalities contribute complemen-

[16] Downloaded from the author's website http://www.cl.cam.ac.uk/~dk427/imgembed.html. Computing the maximum instead of the average vector performed worse in our experiments.

[17] Among the two visual objectives they propose, we chose the one that performed best on our similarity dataset.

[18] Classification of attributes into categories is provided by McRae et al. [14] in their dataset.

Models	Semantic Similarity			Visual Similarity		
	T	V	T+V	T	V	T+V
McRae	0.71	0.49	0.68	0.58	0.52	0.61
SAE (tAttrib, vAttrib)	0.67	0.61	0.72	0.55	0.60	0.65
SVD (tAttrib, vAttrib)	—	—	0.70	—	—	0.59
kCCA (tAttrib, vAttrib)	—	—	0.58	—	—	0.56
DCCA (tAttr, vAttr)	—	—	0.65	—	—	0.59
SAE (skip-gram, vAttrib)	0.74	0.61	0.77	0.59	0.60	0.66
SVD (skip-gram, vAttrib)	—	—	0.75	—	—	0.63
kCCA (skip-gram, vAttrib)	—	—	0.59	—	—	0.57
DCCA (skip-gram, vAttr)	—	—	0.68	—	—	0.56
Lazaridou et al.	0.70	0.62	0.70	0.55	0.57	0.61
Bruni et al.	—	—	0.50	—	—	0.44
tAttrib+vAttrib	0.63	0.62	0.71	0.49	0.57	0.60
tAttrib+CNN	0.63	0.34	0.67	0.49	0.34	0.53
skip-gram+CNN	0.71	0.34	0.71	0.56	0.34	0.55
skip-gram+vAttrib	0.71	0.62	0.75	0.56	0.57	0.62

TABLE 5
Correlation of model predictions against similarity ratings for [14]'s noun pairs (using Spearman's ρ).

Models	Categorization		
	T	V	T+V
McRae	0.52	0.31	0.42
SAE (tAttrib, vAttrib)	0.36	0.35	0.43
SVD (tAttrib, vAttrib)	—	—	0.39
kCCA (tAttrib, vAttrib)	—	—	0.37
DCCA (tAttrib, vAttrib)	—	—	0.36
SAE (skip-gram, vAttrib)	0.44	0.35	0.48
SVD (skip-gram, vAttrib)	—	—	0.43
kCCA (skip-gram, vAttrib)	—	—	0.35
DCCA (skip-gram, vAttrib)	—	—	0.35
Lazaridou et al.	0.37	0.37	0.39
Bruni et al.	—	—	0.34
tAttrib+vAttrib	0.35	0.37	0.33
tAttrib+CNN	0.35	0.30	0.37
skip-gram+CNN	0.37	0.30	0.42
skip-gram+vAttrib	0.37	0.37	0.45

TABLE 6
F-score results on concept categorization.

tary information and that the bimodal SAE model is able to extract a shared representation which improves generalization performance across tasks via joint learning.

6 Experiment 2: Categorization

The task of categorization (i.e., grouping objects into meaningful categories) is a classic problem in the field of cognitive science, central to perception, learning, and the use of language (see [87] for an overview). Existing models typically focus on a single modality, either perception or language (but see [31], [37] for exceptions). For example, perceptual information is represented in form of hand-coded (binary) values on a few dimensions such as color or shape (e.g., [88]), via artificial stimuli (e.g., [89]), geometric shapes (e.g., [90]) or real-world images (e.g., [91]). And linguistic representations are often derived from large text corpora (e.g., [92], [93]). In our second experiment, we induce semantic categories following a clustering-based approach which uses the bimodal word representations learned by our model.

6.1 Method

To obtain a clustering of nouns into categories, we used Chinese Whispers (CW, [94]), a randomized agglomerative graph-clustering algorithm. In the categorization setting, CW produces a hard clustering over a weighted graph whose nodes correspond to words and edges to cosine similarity scores between vectors representing their meaning. At the beginning, each word forms an own, basic-level category. All words are then iteratively processed for a few repetitions in which each word is assigned to the category (i.e., cluster) of the most similar neighbor words, as determined by the maximum sum of (edge) weights between the word and the neighbor nodes pertaining to the same category. CW is a non-parametric model, it induces the number of clusters from the data as well as which nouns belong to these clusters. We initialized CW with different graphs resulting from different vector-based representations of the McRae nouns. We evaluated model output against a gold standard set of categories created by Fountain and Lapata [95]. The dataset contains a classification (produced by human participants) of the McRae nouns into (possibly multiple) semantic categories (40 in total).[19] We transformed the dataset into hard categorizations by assigning each noun to its most typical category as extrapolated from human typicality ratings (see [95] for details).

We used the same SAE model described in Experiment 1. While some performance gains could be expected if parameter optimization took place separately for each task, we wanted to avoid overfitting, and show that our parameters are robust across tasks and datasets. The SAE model was evaluated against the same comparison models described in Section 5.2 (Experiment 1). We evaluated the clusters produced by CW using the F-score measure introduced in the SemEval 2007 task [96]; it is the harmonic mean of precision and recall defined as the number of correct members of a cluster divided by the number of items in the cluster and the number of items in the gold standard class, respectively.

6.2 Results

Our results on the categorization task are given in Table 6. Again, we observe that amongst models based on visual and textual attributes, SAE is the better model (tAttrib, vAttrib; T+V) outperforming the CCA models and SVD by a large margin as well as the related models of Lazaridou et al. [40] and Bruni et al. [37]. Table 7 shows examples of clusters produced by CW when using vector representations provided by the bimodal SAE model (the cluster labels are added by the authors for illustration purposes). Overall, the SAE with skip-gram embeddings (and visual attributes) performs best, delivering clustering performance similar to McRae's gold standard norms. We also observe that simple concatenation of visual and textual attributes does not yield improved performance over the individual modalities (tAttrib+vAttrib) and that the concatenation of textual attributes and CNN-based features (tAttrib+CNN) do not improve over visual attributes alone. As observed in Experiment 1, concatenation models gain a substantial boost when using skip-gram embeddings to represent the

[19] Available at http://homepages.inf.ed.ac.uk/s0897549/data/.

Category	Words
STICK-LIKE UTENSILS	*baton, ladle, peg, spatula, spoon*
RELIGIOUS BUILDINGS	*cathedral, chapel, church*
WIND INSTRUMENTS	*clarinet, flute, saxophone, trombone, trumpet, tuba*
AXES	*axe, hatchet, machete, tomahawk*
ENTRY POINTS	*door, elevator, gate*
UNGULATES	*bison, buffalo, bull, calf, camel, cow, donkey, elephant, goat, horse, lamb, ox, pig, pony, sheep*
BIRDS	*crow, dove, eagle, falcon, hawk, ostrich, owl, penguin, pigeon, raven, stork, vulture, woodpecker*

TABLE 7
Examples of clusters produced by CW using the semantic representations obtained from the bimodal SAE model.

textual modality. We present information on the statistical significance of the reported results in Appendix B, which is available in the online supplemental material.

7 Conclusions

In this paper, we presented a model that uses stacked autoencoders to learn grounded meaning representations by simultaneously combining textual and visual modalities. The two modalities are encoded as vectors of *natural language attributes* and are obtained automatically from text and image data. To the best of our knowledge, our model is novel in its use of attribute-based input in a deep neural network. Experimental results in two tasks, namely simulation of word similarity and word categorization show that our model outperforms competitive baselines and related models trained on the same attribute-based input. Our evaluation also reveals that the bimodal models are superior to their unimodal counterparts and that higher-level unimodal representations are better than the raw input. Since the attribute-based representation is general and text-based, it can be conveniently integrated with any type of distributional model or embeddings such as those obtained from skip-gram.

Compared to related bimodal models, with the exception of DCCA [82], the SAE has a deeper architecture, and thus obtains meaning representations from multiple layers. The first layers operate on individual modalities, whereas the final hidden layer combines them to create a bimodal representation. This architecture allows us to test different hypotheses with respect to word meaning. Specifically, we can disentangle the contribution of visual or textual information (e.g., by contrasting words based on their unimodal against their bimodal representation). Models using SVD [37], LDA [22], or kCCA [21] project the input data into a joint space *directly*. There is no hierarchy of representations with potentially increasing complexity, nor an intermediate unimodal representation naturally connecting the input to the bimodal representation. The semi-supervised architecture of our SAE model affords flexibility allowing it to adapt to specific tasks. For example, by setting the corruption parameter v for the textual modality to one and δ_r to zero, a standard object classification model for images can be trained.

Similarly to models employing SVD, kCCA or the bimodal skip-gram [40], our model and DCCA perform dimensionality-reduction in the course of representation learning, but integrate the different modalities *non-linearly* which we argue allows to model complex relationships between visual and textual data. Importantly, our SAE can be augmented with any type of input and can derive bimodal representations for out-of-vocabulary words when being trained on meaningful input data (e.g., attributes). In addition, our model can perform inductive inference [97] when faced with a concept for which only one modality is available. For example, when presented only with visual information for the new word *currant* the SAE predicts textual attributes *fruit:n*, *sugar:n*, *pickled:j*, *flavor:n*, *salad:n*, *pick:v*, *sour:j*, *juice:n*, *ripe:j*, *cultivate:v*; whereas for *jellyfish* it predicts *silver:j*, *color:v*, *fisherman:n*, *catch:v*, *swim:v*, *fish:v*, *carpet:n*, *white:j*, *ocean:n*, *fishing:n*. This inference ability follows directly out of the model, without additional assumptions or modifications. Related models either do not have a simple way of projecting one modality into a joint space [15], altogether lack a mechanism of inferring missing modalities [16], or can at most provide a generic representation for new words not seen during training. The latter is the case for the multimodal skip-gram models ([39], [40]) which learn word embeddings from randomly initialized textual input.

Directions for future work are many and varied. We would like to extend our database to actions and show that an attribute-centric representation is useful to to other tasks, such as image and text retrieval, zero-shot learning, and word learning.

References

[1] P. D. Turney and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics," *J. Artificial Intelligence Research*, vol. 37, no. 1, pp. 141–188, 2010.

[2] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (almost) from Scratch," *J. Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[3] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations," in *Proc. 2013 Conf. North American Chapter Assoc. Computational Linguistics: Human Language Technologies*, 2013, pp. 746–751.

[4] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY: Cambridge University Press, 2008.

[5] R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating Query Substitutions," in *Proc. 15th Int'l Conf. World Wide Web*, 2006, pp. 387–396.

[6] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, pp. 1–47, 2002.

[7] W.-t. Yih, M.-W. Chang, C. Meek, and A. Pastusiak, "Question Answering Using Enhanced Lexical Semantic Models," in *Proc. 51st Ann. Meeting Assoc. Computational Linguistics*, 2013, pp. 1744–1753.

[8] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in Semantic Representation," *Psychological Review*, vol. 114, no. 2, pp. 211–244, 2007.

[9] Z. Harris, "Distributional Structure," in *Papers in Structural and Transformational Linguistics*, 1970, pp. 775–794.

[10] A. M. Glenberg and M. P. Kaschak, "Grounding Language in Action," *Psychonomic Bulletin & Review*, vol. 9, no. 3, pp. 558–565, 2002.

[11] T. Regier, *The Human Semantic Potential*. Cambridge, Massachusetts: MIT Press, 1996.

[12] B. Landau, L. Smith, and S. Jones, "Object Perception and Object Naming in Early Development," *Trends in Cognitive Sciences*, vol. 27, pp. 19–24, 1998.

[13] L. W. Barsalou, "Grounded Cognition," *Ann. Review of Psychology*, vol. 59, pp. 617–845, 2008.

[14] K. McRae, G. S. Cree, M. S. Seidenberg, and C. McNorgan, "Semantic Feature Production Norms for a Large Set of Living and Nonliving Things," *Behavior Research Methods*, vol. 37, no. 4, pp. 547–559, 2005.

[15] M. Andrews, G. Vigliocco, and D. Vinson, "Integrating Experiential and Distributional Data to Learn Semantic Representations," *Psychological Review*, vol. 116, no. 3, pp. 463–498, 2009.

[16] C. Silberer and M. Lapata, "Grounded Models of Semantic Representation," in *Proc. 2012 Conf. Empirical Methods in Natural Language Processing*, 2012, pp. 1423–1433.

[17] D. P. Vinson and G. Vigliocco, "Semantic Feature Production Norms for a Large Set of Objects and Events," *Behavior Research Methods*, vol. 40, no. 1, pp. 183–190, 2008.

[18] G. S. Cree, K. McRae, and C. McNorgan, "An Attractor Model of Lexical Conceptual Processing: Simulating Semantic Priming," *Cognitive Science*, vol. 23, no. 3, pp. 371–414, 1999.

[19] Y. Feng and M. Lapata, "Visual Information in Semantic Representation," in *Human Language Technologies: The 2010 Ann. Conf. North American Chapter Assoc. Computational Linguistics*, 2010, pp. 91–99.

[20] E. Bruni, G. Boleda, M. Baroni, and N. Tran, "Distributional Semantics in Technicolor," in *Proc. 50th Ann. Meeting Assoc. Computational Linguistics*, 2012, pp. 136–145.

[21] C. Silberer, V. Ferrari, and M. Lapata, "Models of Semantic Representation with Visual Attributes," in *Proc. 51st Ann. Meeting Assoc. Computational Linguistics*, 2013, pp. 572–582.

[22] S. Roller and S. Schulte im Walde, "A Multimodal LDA Model integrating Textual, Cognitive and Visual Modalities," in *Proc. 2013 Conf. Empirical Methods in Natural Language Processing*, 2013, pp. 1146–1157.

[23] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-Wise Training of Deep Networks," in *Advances in Neural Information Processing Systems 19*, 2006, pp. 153–160.

[24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal Deep Learning," in *Proc. 28th Int'l Conf. Machine Learning*, 2011, pp. 689–696.

[25] N. Srivastava and R. Salakhutdinov, "Multimodal Learning with Deep Boltzmann Machines," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 2231–2239.

[26] M. Baroni, B. Murphy, E. Barbu, and M. Poesio, "Strudel: A Corpus-Based Semantic Model Based on Properties and Types," *Cognitive Science*, vol. 34, no. 2, pp. 222–254, 2010.

[27] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing Objects by their Attributes," in *Proc. 2009 IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1778–1785.

[28] M. N. Jones, J. A. Willits, and S. Dennis, "Models of Semantic Memory," in *Oxford Handbook of Computational and Mathematical Psychology*, J. Busemeyer, J. Townsend, Z. Wang, and A. Eidels, Eds. Oxford University Press, 2015, pp. 232–254.

[29] M. J. Farah and J. L. McClelland, "A Computational Model of Semantic Memory Impairment: Modality Specificity and Emergent Category Specificity," *J. Experimental Psychology: General*, vol. 120, no. 4, pp. 339–357, 1991.

[30] G. E. Hinton and T. Shallice, "Lesioning an Attractor Network: Investigations of Acquired Dyslexia," *Psychological Review*, vol. 98, pp. 74–95, 1991.

[31] G. Westermann and D. Mareschal, "From Perceptual to Language-mediated Categorization," *Philosophical Trans. Royal Society B: Biological Sciences*, vol. 369, no. 1634, p. 20120391, 2014.

[32] T. T. Rogers, M. A. Lambon Ralph, P. Garrard, S. Bozeat, J. L. Mcclelland, J. R. Hodges, and K. Patterson, "Structure and Deterioration of Semantic Memory: A Neuropsychological and Computational Investigation." *Psychological Review*, vol. 111, no. 1, pp. 205–235, 2004.

[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 3111–3119.

[34] F. Hill and A. Korhonen, "Learning Abstract Concept Embeddings from Multi-Modal Data: Since You Probably Cant See What I Mean," in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing*, 2014, pp. 255–265.

[35] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. 2009 IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[36] L. von Ahn and L. Dabbish, "Labeling Images with a Computer Game," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2004, pp. 319–326.

[37] E. Bruni, N. Tran, and M. Baroni, "Multimodal distributional semantics," *J. Artificial Intelligence Research*, vol. 49, pp. 1–47, 2014.

[38] D. Lowe, "Distinctive Image Features from Scale-invariant Keypoints," *Int'l J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[39] D. Kiela and L. Bottou, "Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics," in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing*, 2014, pp. 36–45.

[40] A. Lazaridou, N. T. Pham, and M. Baroni, "Combining Language and Vision with a Multimodal Skip-gram Model," in *Human Language Technologies: The 2015 Ann. Conf. North American Chapter Assoc. Computational Linguistics*, May–June 2015, pp. 153–163.

[41] E. Bruni, G. Tran, and M. Baroni, "Distributional Semantics from Text and Images," in *Proc. GEMS 2011 Workshop GEometrical Models of Natural Language Semantics*, 2011, pp. 22–32.

[42] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[43] N. Srivastava and R. Salakhutdinov, "Multimodal Learning with Deep Boltzmann Machines," *J. Machine Learning Research*, vol. 15, pp. 2949–2980, 2014.

[44] P. Wu, S. C. H. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, "Online Multimodal Deep Similarity Learning with Application to Image Retrieval," in *Proc. 21st ACM Int'l. Conf. Multimedia*, 2013, pp. 153–162.

[45] R. Socher, Q. V. Le, C. D. Manning, and A. Y. Ng., "Grounded Compositional Semantics for Finding and Describing Images with Sentences," *Trans. Assoc. Computational Linguistics*, 2013.

[46] K. Sohn, W. Shang, and H. Lee, "Improved Multimodal Deep Learning with Variation of Information," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2141–2149.

[47] J. Huang and B. Kingsbury, "Audio-visual Deep Learning for Noise Robust Speech Recognition," in *Proc. 38th Int'l Conf. Acoustics, Speech, and Signal Processing*, 2013, pp. 7596–7599.

[48] R. Kiros, R. Salakhutdinov, and R. Zemel, "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models," in *Deep Learning and Representation Learning Workshop: NIPS 2014*, Montréal, Canada, 2014.

[49] R. Socher, A. Karpathy, Q. V. Le, C. Manning, and A. Ng, "Grounded Compositional Semantics for Finding and Describing Images with Sentences," *Trans. Assoc. Computational Linguistics*, pp. 113–124, 2014.

[50] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)," in *Proc. Int'l Conf. Learning Representations*, 2015.

[51] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[52] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "DeViSE: A Deep Visual-Semantic Embedding Model," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 2121–2129.

[53] R. Socher, M. Ganjoo, C. D. Manning, and A. Y. Ng, "Zero-Shot Learning Through Cross-Modal Transfer," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 935–943.

[54] V. Ferrari and A. Zisserman, "Learning Visual Attributes," in *Advances in Neural Information Processing Systems 20*, 2007, pp. 433–440.

[55] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results," 2008.

[56] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer," in *Proc. 2009 IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 951–958.

[57] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith, "Default Probability," *Cognitive Science*, vol. 2, no. 15, pp. 251–269, 1991.

[58] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Describable Visual Attributes for Face Verification and Image Search," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1962–1977, October 2011.

[59] G. Patterson and J. Hays, "SUN Attribute Database: Discovering, Annotating and Recognizing Scene Attributes," in *Proc. 25th IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2751–2758.

[60] K. Duan, D. Parikh, D. Crandall, and K. Grauman, "Discovering Localized Attributes for Fine-grained Recognition," in *Proc. 2012*

IEEE Conf. Computer Vision and Pattern Recognition, 2012, pp. 3474–3481.

[61] J. Liu, B. Kuipers, and S. Savarese, "Recognizing Human Actions by Attributes," in *Proc. 2011 IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 3337–3344.

[62] E. Barbu, "Combining Methods to Learn Feature-norm-like Concept Descriptions," in *Proc. ESSLLI Workshop Distributional Lexical Semantics*, 2008, pp. 9–16.

[63] B. Devereux, N. Pilkington, T. Poibeau, and A. Korhonen, "Towards Unrestricted, Large-Scale Acquisition of Feature-Based Conceptual Representations from Corpus Data," *Research on Language and Computation*, vol. 7, no. 2-4, pp. 137–170, 2009.

[64] C. Kelly, B. Devereux, and A. Korhonen, "Acquiring Human-like Feature-based Conceptual Representations from Corpora," in *Proc. NAACL HLT 2010 1st Workshop Computational Neurolinguistics*, 2010, pp. 61–69.

[65] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Parallel Distributed Processing: Explorations in the Microstructure of Cognition," D. E. Rumelhart and J. L. McClelland, Eds. MIT Press, 1986, vol. 1: Foundations, ch. Learning Internal Representations by Error Propagation, pp. 318–362.

[66] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[67] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *J. Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[68] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders," in *Proc. 25th Int'l Conf. Machine Learning*, 2008, pp. 1096–1103.

[69] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[70] M. Ranzato and M. Szummer, "Semi-supervised Learning of Compact Document Representations with Deep Networks," in *Proc. 25th Int'l Conf. Machine Learning*, 2008, pp. 792–799.

[71] C. Fellbaum, Ed., *WordNet: an Electronic Lexical Database*. MIT Press, 1998.

[72] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "LIBLINEAR: A Library for Large Linear Classification," *J. Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[73] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill, 1986.

[74] S. L. Thompson-Schill, K. J. Kurtz, and J. D. E. Gabrieli, "Effects of Semantic and Associative Relatedness on Automatic Priming," *J. Memory and Language*, vol. 38, no. 4, 1998.

[75] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing Search in Context: The Concept Revisited," *ACM Trans. Information Systems*, vol. 20, no. 1, pp. 116–131, 2002.

[76] S. R. Szumlanski, F. Gomez, and V. K. Sims, "A New Set of Norms for Semantic Relatedness Measures," in *Proc. 51st Ann. Meeting Assoc. Computational Linguistics*, 2013, pp. 890–895.

[77] S. Patwardhan and T. Pedersen, "Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts," in *Proc. EACL 2006 Workshop Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, 2006, pp. 1–8.

[78] G. A. Miller and W. G. Charles, "Contextual Correlates of Semantic Similarity," *Language and Cognitive Processes*, vol. 6, no. 1, 1991.

[79] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, "The University of South Florida Word Association, Rhyme, and Word Fragment Norms," 1998. [Online]. Available: http://w3.usf.edu/FreeAssociation/

[80] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[81] M. Borga, "Canonical Correlation - a Tutorial," January 2001.

[82] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep Canonical Correlation Analysis," in *Proc. 30th Int'l Conf. Machine Learning*, 2013, pp. 1247–1255.

[83] T. Landauer and S. T. Dumais, "A Solution to Plato's Problem: the Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge," *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.

[84] E. Bruni, U. Bordignon, A. Liska, J. Uijlings, and I. Sergienya, "Vsem: An open library for visual semantics representation," in *Proc. 51st Ann. Meeting Assoc. Computational Linguistics: System Demonstrations*, August 2013, pp. 187–192.

[85] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks," in *Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1717–1724.

[86] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[87] R. L. Goldstone, A. Kersten, and P. F. Cavalho, "Concepts and Categorization," in *Comprehensive Handbook of Psychology*, A. F. Healy and R. W. Proctor, Eds. New Jersey: Wiley, 2012, vol. 4: Experimental psychology, pp. 607–630.

[88] J. R. Anderson, "The Adaptive Nature of Human Categorization," *Psychological Review*, vol. 98, no. 3, pp. 409–429, 1991.

[89] A. N. Sanborn, T. L. Griffiths, and D. J. Navarro, "A More Rational Model of Categorization," in *Proc. 28th Ann. Conf. Cognitive Science Society*, 2006.

[90] S. C. McKinley and R. M. Nosofsky, "Investigations of exemplar and decision bound models in large, ill-defined category structures," *J. Experimental Psychology, Human Perception and Performance*, vol. 21, no. 1, pp. 128–48, 1995.

[91] A. S. Hsu, J. B. Martin, A. N. Sanborn, and T. L. Griffiths, "Identifying representations of categories of discrete items using Markov chain Monte Carlo with People," in *Proc. 34th Ann. Conf. Cognitive Science Society*, 2012, pp. 485–490.

[92] T. Fountain and M. Lapata, "Incremental Models of Natural Language Category Acquisition," in *Proc. 32nd Ann. Conf. Cognitive Science Society*, C. Carlson, Hölscher, and T. Shipley, Eds., 2011.

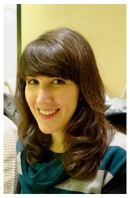
[93] L. Frermann and M. Lapata, "Incremental Bayesian Learning of Semantic Categories," in *Proc. 14th Conf. European Chapter Assoc. Computational Linguistics*, 2014, pp. 249–258.

[94] C. Biemann, "Chinese Whispers – an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems," in *Proc. TextGraphs: the 1st Workshop Graph Based Methods for Natural Language Processing*, 2006, pp. 73–80.

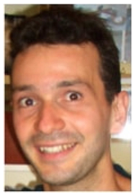
[95] T. Fountain and M. Lapata, "Meaning Representation in Natural Language Categorization," in *Proc. 31st Ann. Conf. Cognitive Science Society*, 2010, pp. 1916–1921.

[96] E. Agirre and A. Soroa, "SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems," in *Proc. 4th Int'l Workshop Semantic Evaluations*, 2007, pp. 7–12.

[97] B. T. Johns and M. N. Jones, "Perceptual Inference through Global Lexical Similarity," *Topics in Cognitive Sciences*, vol. 4, no. 1, pp. 103–120, 2012.



Carina Silberer is a research associate at the School of Informatics, University of Edinburgh. Her research interests are on grounded language learning and neural network models of lexical meaning.



Vittorio Ferrari is a Reader at the School of Informatics, University of Edinburgh, where he leads the CALVIN research group on visual learning. His research interests are in weakly supervised learning of object classes, semantic segmentation, and large-scale auto-annotation.

Mirella Lapata is a professor in the School of Informatics, University of Edinburgh. Her research interests include machine learning techniques for natural language understanding, generation, and grounded language acquisition.