Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world

Angeliki Lazaridou and **Elia Bruni** and **Marco Baroni**
Center for Mind/Brain Sciences
University of Trento
{angeliki.lazaridou|elia.bruni|marco.baroni}@unitn.it

Abstract

Following up on recent work on establishing a mapping between vector-based semantic embeddings of words and the visual representations of the corresponding objects from natural images, we first present a simple approach to cross-modal vector-based semantics for the task of *zero-shot learning*, in which an image of a previously unseen object is mapped to a linguistic representation denoting its word. We then introduce *fast mapping*, a challenging and more cognitively plausible variant of the zero-shot task, in which the learner is exposed to new objects and the corresponding words in very limited linguistic contexts. By combining prior linguistic and visual knowledge acquired about words and their objects, as well as exploiting the limited new evidence available, the learner must learn to associate new objects with words. Our results on this task pave the way to realistic simulations of how children or robots could use existing knowledge to bootstrap grounded semantic knowledge about new concepts.

1 Introduction

Computational models of meaning that rely on corpus-extracted context vectors, such as LSA (Landauer and Dumais, 1997), HAL (Lund and Burgess, 1996), Topic Models (Griffiths et al., 2007) and more recent neural-network approaches (Collobert and Weston, 2008; Mikolov et al., 2013b) have successfully tackled a number of lexical semantics tasks, where context vector similarity highly correlates with various indices of semantic relatedness (Turney and Pantel, 2010). Given that these models are learned from naturally occurring data using simple associative techniques, various authors have advanced the claim that they might be also capturing some crucial aspects of how humans acquire and use language (Landauer and Dumais, 1997; Lenci, 2008).

However, the models induce the meaning of words entirely from their co-occurrence with other words, without links to the external world. This constitutes a serious blow to claims of cognitive plausibility in at least two respects. One is the *grounding problem* (Harnad, 1990; Searle, 1984). Irrespective of their relatively high performance on various semantic tasks, it is debatable whether models that have no access to visual and perceptual information can capture the holistic, grounded knowledge that humans have about concepts. However, a possibly even more serious pitfall of vector models is *lack of reference*: natural language is, fundamentally, a means to communicate, and thus our words must be able to *refer* to objects, properties and events in the outside world (Abbott, 2010). Current vector models are purely language-internal, solipsistic models of meaning. Consider the very simple scenario in which visual information is being provided to an agent about the current state of the world, and the agent's task is to determine the truth of a statement similar to *There is a dog in the room*. Although the agent is equipped with a powerful context vector model, this will not suffice to successfully complete the task. The model might suggest that the concepts of *dog* and *cat* are semantically related, but it has no means to determine the visual appearance of dogs, and consequently no way to verify the truth of such a simple statement.

Mapping words to the objects they denote is such a core function of language that humans are highly optimized for it, as shown by the so-called *fast mapping* phenomenon, whereby children can learn to associate a word to an object or property by a single exposure to it (Bloom, 2000; Carey, 1978; Carey and Bartlett, 1978; Heibeck and Markman, 1987). But lack of reference is not

only a theoretical weakness: Without the ability to refer to the outside world, context vectors are arguably useless for practical goals such as learning to execute natural language instructions (Branavan et al., 2009; Chen and Mooney, 2011), that could greatly benefit from the rich network of lexical meaning such vectors encode, in order to scale up to real-life challenges.

Very recently, a number of papers have exploited advances in automated feature extraction form images and videos to enrich context vectors with visual information (Bruni et al., 2014; Feng and Lapata, 2010; Leong and Mihalcea, 2011; Regneri et al., 2013; Silberer et al., 2013). This line of research tackles the grounding problem: Word representations are no longer limited to their linguistic contexts but also encode visual information present in images associated with the corresponding objects. In this paper, we rely on the same image analysis techniques but instead focus on the reference problem: We do not aim at enriching word representations with visual information, although this might be a side effect of our approach, but we address the issue of automatically mapping objects, as depicted in images, to the context vectors representing the corresponding words. This is achieved by means of a simple neural network trained to project image-extracted feature vectors to text-based vectors through a hidden layer that can be interpreted as a cross-modal semantic space.

We first test the effectiveness of our cross-modal semantic space on the so-called *zero-shot learning* task (Palatucci et al., 2009), which has recently been explored in the machine learning community (Frome et al., 2013; Socher et al., 2013). In this setting, we assume that our system possesses linguistic and visual information for a set of concepts in the form of text-based representations of words and image-based vectors of the corresponding objects, used for vision-to-language-mapping training. The system is then provided with visual information for a previously unseen object, and the task is to associate it with a word by cross-modal mapping. Our approach is competitive with respect to the recently proposed alternatives, while being overall simpler.

The aforementioned task is very demanding and interesting from an engineering point of view. However, from a cognitive angle, it relies on strong, unrealistic assumptions: The learner is asked to establish a link between a new object and a word for which they possess a full-fledged text-based vector extracted from a billion-word corpus. On the contrary, the first time a learner is exposed to a new object, the linguistic information available is likely also very limited. Thus, in order to consider vision-to-language mapping under more plausible conditions, similar to the ones that children or robots in a new environment are faced with, we next simulate a scenario akin to fast mapping. We show that the induced cross-modal semantic space is powerful enough that sensible guesses about the correct word denoting an object can be made, even when the linguistic context vector representing the word has been created from as little as 1 sentence containing it.

The contributions of this work are three-fold. First, we conduct experiments with simple image- and text-based vector representations and compare alternative methods to perform cross-modal mapping. Then, we complement recent work (Frome et al., 2013) and show that zero-shot learning scales to a large and noisy dataset. Finally, we provide preliminary evidence that cross-modal projections can be used effectively to simulate a fast mapping scenario, thus strengthening the claims of this approach as a full-fledged, fully inductive theory of meaning acquisition.

2 Related Work

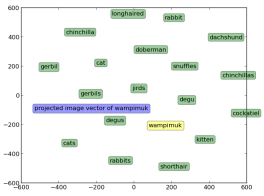
The problem of establishing word reference has been extensively explored in computational simulations of cross-situational learning (see Fazly et al. (2010) for a recent proposal and extended review of previous work). This line of research has traditionally assumed artificial models of the external world, typically a set of linguistic or logical labels for objects, actions and possibly other aspects of a scene (Siskind, 1996). Recently, Yu and Siskind (2013) presented a system that induces word-object mappings from features extracted from short videos paired with sentences. Our work complements theirs in two ways. First, unlike Yu and Siskind (2013) who considered a limited lexicon of 15 items with only 4 nouns, we conduct experiments in a large search space containing a highly ambiguous set of potential target words for every object (see Section 4.1). Most importantly, by projecting visual representations of objects into a shared *semantic space*, we do not limit ourselves to establishing a link between ob-

jects and words. We induce a rich semantic representation of the multimodal concept, that can lead, among other things, to the discovery of important properties of an object even when we lack its linguistic label. Nevertheless, Yu and Siskind's system could in principle be used to initialize the vision-language mapping that we rely upon.

Closer to the spirit of our work are two very recent studies coming from the machine learning community. Socher et al. (2013) and Frome et al. (2013) focus on zero-shot learning in the vision-language domain by exploiting a shared visual-linguistic semantic space. Socher et al. (2013) learn to project unsupervised vector-based image representations onto a word-based semantic space using a neural network architecture. Unlike us, Socher and colleagues train an outlier detector to decide whether a test image should receive a known-word label by means of a standard supervised object classifier, or be assigned an unseen label by vision-to-language mapping. In our zero-shot experiments, we assume no access to an outlier detector, and thus, the search for the correct label is performed in the full concept space. Furthermore, Socher and colleagues present a much more constrained evaluation setup, where only 10 concepts are considered, compared to our experiments with hundreds or thousands of concepts.

Frome et al. (2013) use linear regression to transform vector-based image representations onto vectors representing the same concepts in linguistic semantic space. Unlike Socher et al. (2013) and the current study that adopt simple unsupervised techniques for constructing image representations, Frome et al. (2013) rely on a supervised state-of-the-art method: They feed low-level features to a deep neural network trained on a supervised object recognition task (Krizhevsky et al., 2012). Furthermore, their text-based vectors encode very rich information, such as $\vec{king} - \vec{man} + \vec{woman} = \vec{queen}$ (Mikolov et al., 2013c). A natural question we aim to answer is whether the success of cross-modal mapping is due to the high-quality embeddings or to the general algorithmic design. If the latter is the case, then these results could be extended to traditional distributional vectors bearing other desirable properties, such as high interpretability of dimensions.

(a)



(b)

Figure 1: A potential *wampimuk* (a) together with its projection onto the linguistic space (b).

3 Zero-shot learning and fast mapping

"We found a cute, hairy *wampimuk* sleeping behind the tree." Even though the previous statement is certainly the first time one hears about *wampimuks*, the linguistic context already creates some visual expectations: Wampimuks probably resemble small animals (Figure 1a). This is the scenario of *zero-shot learning*. Moreover, if this is also the first linguistic encounter of that concept, then we refer to the task as *fast mapping*.

Concretely, we assume that concepts, denoted for convenience by word labels, are represented in linguistic terms by vectors in a text-based distributional semantic space (see Section 4.3). Objects corresponding to concepts are represented in visual terms by vectors in an image-based semantic space (Section 4.2). For a subset of concepts (e.g., a set of animals, a set of vehicles), we possess information related to both their linguistic and visual representations. During training, this cross-modal vocabulary is used to induce a projection function (Section 4.4), which – intuitively – represents a mapping between visual and linguistic dimensions. Thus, this function, given a visual vector, returns its corresponding linguistic representation. At test time, the system is presented with a previously unseen object (e.g., *wampimuk*). This object is projected onto the linguistic space and associated with the word label of the nearest neighbor in that space (*degus* in Figure 1b).

The fast mapping setting can be seen as a special case of the zero-shot task. Whereas for the latter our system assumes that all concepts have rich linguistic representations (i.e., representations estimated from a large corpus), in the case of the former, new concepts are assumed to be encounted in a limited linguistic context and therefore lacking rich linguistic representations. This is operationalized by constructing the text-based vector for these

Figure 2: Images of *chair* as extracted from CIFAR-100 (left) and ESP (right).

concepts from a context of just a few occurrences. In this way, we simulate the first encounter of a learner with a concept that is new in both visual and linguistic terms.

4 Experimental Setup

4.1 Visual Datasets

CIFAR-100 The CIFAR-100 dataset (Krizhevsky, 2009) consists of 60,000 32x32 colour images (note the extremely small size) representing 100 distinct concepts, with 600 images per concept. The dataset covers a wide range of concrete domains and is organized into 20 broader categories. Table 1 lists the concepts used in our experiments organized by category.

ESP Our second dataset consists of 100K images from the ESP-Game data set, labeled through a "game with a purpose" (Von Ahn, 2006).[1] The ESP image tags form a vocabulary of 20,515 unique words. Unlike other datasets used for zero-shot learning, it covers adjectives and verbs in addition to nouns. On average, an image has 14 tags and a word appears as a tag for 70 images. Unlike the CIFAR-100 images, which were chosen specifically for image object recognition tasks (i.e., each image is clearly depicting a single object in the foreground), ESP contains a random selection of images from the Web. Consequently, objects do not appear in most images in their prototypical display, but rather as elements of complex scenes (see Figure 2). Thus, ESP constitutes a more realistic, and at the same time more challenging, simulation of how things are encountered in real life, testing the potentials of cross-modal mapping in dealing with the complex scenes that one would encounter in event recognition and caption generation tasks.

[1] `http://www.cs.cmu.edu/~biglou/resources/`

4.2 Visual Semantic Spaces

Image-based vectors are extracted using the unsupervised bag-of-visual-words (BoVW) representational architecture (Sivic and Zisserman, 2003; Csurka et al., 2004), that has been widely and successfully applied to computer vision tasks such as object recognition and image retrieval (Yang et al., 2007). First, low-level visual features (Szeliski, 2010) are extracted from a large collection of images and clustered into a set of "visual words". The low-level features of a specific image are then mapped to the corresponding visual words, and the image is represented by a count vector recording the number of occurrences of each visual word in it. We do not attempt any parameter tuning of the pipeline.

As low-level features, we use Scale Invariant Feature Transform (SIFT) features (Lowe, 2004). SIFT features are tailored to capture object parts and to be invariant to several image transformations such as rotation, illumination and scale change. These features are clustered into vocabularies of 5,000 (ESP) and 4,096 (CIFAR-100) visual words.[2] To preserve spatial information in the BoVW representation, we use the spatial pyramid technique (Lazebnik et al., 2006), which consists in dividing the image into several regions, computing BoVW vectors for each region and concatenating them. In particular, we divide ESP images into 16 regions and the smaller CIFAR-100 images into 4. The vectors resulting from region concatenation have dimensionality $5000 \times 16 = 80,000$ (ESP) and $4,096 \times 4 = 16,384$ (CIFAR-100), respectively. We apply Local Mutual Information (LMI, (Evert, 2005)) as weighting scheme and reduce the full co-occurrence space to 300 dimensions using the Singular Value Decomposition.

For CIFAR-100, we extract distinct visual vectors for single images. For ESP, given the size and amount of noise in this dataset, we build vectors for visual *concepts*, by normalizing and summing the BoVW vectors of all the images that have the relevant concept as a tag. Note that relevant literature (Pereira et al., 2010) has emphasized the importance of learners self-generating multiple views when faced with new objects. Thus, our multiple-image assumption should not be considered as problematic in the current setup.

[2] For selecting the size of the vocabulary size, we relied on standard settings found in the relevant literature (Bruni et al., 2014; Chatfield et al., 2011).

Category	Seen Concepts	Unseen (Test) Concepts
aquatic mammals	beaver, otter, seal, whale	dolphin
fish	ray, trout	shark
flowers	orchid, poppy, sunflower, tulip	rose
food containers	bottle, bowl, can ,plate	cup
fruit vegetable	apple, mushroom, pear	orange
household electrical devices	keyboard, lamp, telephone, television	clock
household furniture	chair, couch, table, wardrobe	bed
insects	bee, beetle, caterpillar, cockroach	butterfly
large carnivores	bear, leopard, lion, wolf	tiger
large man-made outdoor things	bridge, castle, house, road	skyscraper
large natural outdoor scenes	cloud, mountain, plain, sea	forest
large omnivores and herbivores	camel, cattle, chimpanzee, kangaroo	elephant
medium-sized mammals	fox, porcupine, possum, skunk	raccoon
non-insect invertebrates	crab, snail, spider, worm	lobster
people	baby, girl, man, woman	boy
reptiles	crocodile, dinosaur, snake, turtle	lizard
small mammals	hamster, mouse, rabbit, shrew	squirrel
vehicles 1	bicycle, motorcycle, train	bus
vehicles 2	rocket, tank, tractor	streetcar

Table 1: Concepts in our version of the CIFAR-100 data set

We implement the entire visual pipeline with VSEM, an open library for visual semantics (Bruni et al., 2013).[3]

4.3 Linguistic Semantic Spaces

For constructing the text-based vectors, we follow a standard pipeline in distributional semantics (Turney and Pantel, 2010) without tuning its parameters and collect co-occurrence statistics from the concatenation of ukWaC[4] and the Wikipedia, amounting to 2.7 billion tokens in total. Semantic vectors are constructed for a set of 30K target words (lemmas), namely the top 20K most frequent nouns, 5K most frequent adjectives and 5K most frequent verbs, and the same 30K lemmas are also employed as contextual elements. We collect co-occurrences in a symmetric context window of 20 elements around a target word. Finally, similarly to the visual semantic space, raw counts are transformed by applying LMI and then reduced to 300 dimensions with SVD.[5]

4.4 Cross-modal Mapping

The process of learning to map objects to the their word label is implemented by training a projection function $\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}}$ from the visual onto the linguistic semantic space. For the learning, we use a set of N_s *seen* concepts for which we have both image-based visual representations $\mathbf{V}_s \in \mathbb{R}^{N_s \times d_v}$ and text-based linguistic representations $\mathbf{W}_s \in \mathbb{R}^{N_s \times d_w}$. The projection function is subject to an objective that aims at minimizing some cost function between the induced text-based representations $\hat{\mathbf{W}}_s \in \mathbb{R}^{N_s \times d_w}$ and the gold ones $\mathbf{W}_s$. The induced $\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}}$ is then applied to the image-based representations $\mathbf{V}_u \in \mathbb{R}^{N_u \times d_v}$ of N_u unseen objects to transform them into text-based representations $\hat{\mathbf{W}}_u \in \mathbb{R}^{N_u \times d_w}$. We implement 4 alternative learning algorithms for inducing the cross-modal projection function $\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}}$.

Linear Regression (lin) Our first model is a very simple linear mapping between the two modalities estimated by solving a least-squares problem. This method is similar to the one introduced by Mikolov et al. (2013a) for estimating a translation matrix, only solved analytically. In our setup, we can see the two different modalities as if they were different languages. By using least-squares regression, the projection function $\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}}$ can be derived as

$$\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}} = (\mathbf{V}_s^T \mathbf{V}_s)^{-1} \mathbf{V}_s^T \mathbf{W}_s \qquad (1)$$

Canonical Correlation Analysis (CCA) CCA (Hardoon et al., 2004; Hotelling, 1936) and variations thereof have been successfully used in the past for annotation of regions (Socher and Fei-Fei, 2010) and complete images (Hardoon et al., 2006; Hodosh et al., 2013). Given two paired observation matrices, in our case $\mathbf{V}_s$ and $\mathbf{W}_s$, CCA aims at capturing the linear relationship that exists between these variables. This is achieved by finding a pair of matrices, in our

[3] http://clic.cimec.unitn.it/vsem/

[4] http://wacky.sslmit.unibo.it

[5] We also experimented with the image- and text-based vectors of Socher et al. (2013), but achieved better performance with the reported setup.

case $\mathbf{C}_V \in \mathbb{R}^{d_v \times d}$ and $\mathbf{C}_W \in \mathbb{R}^{d_w \times d}$, such that the correlation between the projections of the two multidimensional variables into a common, lower-rank space is maximized. The resulting multimodal space has been shown to provide a good approximation to human concept similarity judgments (Silberer and Lapata, 2012). In our setup, after applying CCA on the two spaces $\mathbf{V}_s$ and $\mathbf{W}_s$, we obtain the two projection mappings onto the common space and thus our projection function can be derived as:

$$\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\to\mathrm{w}}} = \mathbf{C}_V {\mathbf{C}_W}^{-1} \quad (2)$$

Singular Value Decomposition (SVD) SVD is the most widely used dimensionality reduction technique in distributional semantics (Turney and Pantel, 2010), and it has recently been exploited to combine visual and linguistic dimensions in the multimodal distributional semantic model of Bruni et al. (2014). SVD smoothing is also a way to infer values of unseen dimensions in partially incomplete matrices, a technique that has been applied to the task of inferring word tags of unannotated images (Hare et al., 2008). Assuming that the concept-representing rows of $\mathbf{V}_s$ and $\mathbf{W}_s$ are ordered in the same way, we apply the (k-truncated) SVD to the concatenated matrix $[\mathbf{V}_s \mathbf{W}_s]$, such that $[\hat{\mathbf{V}}_s \hat{\mathbf{W}}_\mathbf{s}] = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{Z}_k^T$ is a k-rank approximation of the original matrix.[6] The projection function is then:

$$\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\to\mathrm{w}}} = \mathbf{Z}_k \mathbf{Z}_k^T \quad (3)$$

where the input is appropriately padded with 0s ($[\mathbf{V}_u \mathbf{0}_{Nu \times W}]$) and we discard the visual block of the output matrix $[\hat{\mathbf{V}}_u \hat{\mathbf{W}}_u]$.

Neural Network (NNet) The last model that we introduce is a neural network with one hidden layer. The projection function in this model can be described as:

$$\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\to\mathrm{w}}} = \mathbf{\Theta}_{\mathrm{v}\to\mathrm{w}} \quad (4)$$

where $\mathbf{\Theta}_{v\to w}$ consists of the model weights $\boldsymbol{\theta}^{(1)} \in \mathbb{R}^{d_v \times h}$ and $\boldsymbol{\theta}^{(2)} \in \mathbb{R}^{h \times d_w}$ that map the input image-based vectors $\mathbf{V}_s$ first to the hidden layer and then to the output layer in order to obtain text-based vectors, i.e., $\hat{\mathbf{W}}_s = \sigma^{(2)}(\sigma^{(1)}(\mathbf{V}_s \boldsymbol{\theta}^{(1)})\boldsymbol{\theta}^{(2)})$, where $\sigma^{(1)}$ and $\sigma^{(2)}$ are the non-linear activation functions. We experimented with sigmoid, hyperbolic tangent and linear; hyperbolic tangent yielded the highest performance. The weights are estimated by minimizing the objective function

$$J(\mathbf{\Theta}_{\mathrm{v}\to\mathrm{w}}) = \frac{1}{2}(1 - sim(\mathbf{W}_s, \hat{\mathbf{W}}_s)) \quad (5)$$

where sim is some similarity function. In our experiments we used *cosine* as similarity function, so that $sim(\mathbf{A}, \mathbf{B}) = \frac{AB}{\|A\|\|B\|}$, thus penalizing parameter settings leading to a low cosine between the target linguistic representations $\mathbf{W}_s$ and those produced by the projection function $\hat{\mathbf{W}}_s$. The cosine has been widely used in the distributional semantic literature, and it has been shown to outperform Euclidean distance (Bullinaria and Levy, 2007).[7] Parameters were estimated with standard backpropagation and L-BFGS.

5 Results

Our experiments focus on the tasks of zero-shot learning (Sections 5.1 and 5.2) and fast mapping (Section 5.3). In both tasks, the projected vector of the unseen concept is labeled with the word associated to its cosine-based nearest neighbor vector in the corresponding semantic space.

For the zero-shot task we report the *accuracy* of retrieving the correct label among the top k neighbors from a semantic space populated with the *union of seen and unseen* concepts. For fast mapping, we report the *mean rank* of the correct concept among fast mapping candidates.

5.1 Zero-shot Learning in CIFAR-100

For this experiment, we use the intersection of our linguistic space with the concepts present in CIFAR-100, containing a total of 90 concepts. For each concept category, we treat all concepts but one as seen concepts (Table 1). The 71 seen concepts correspond to 42,600 distinct visual vectors and are used to induce the projection function. Table 2 reports results obtained by averaging the performance on the 11,400 distinct vectors of the 19 unseen concepts.

Our 4 models introduced in Section 4.4 are compared to a theoretically derived baseline **Chance** simulating selecting a label at random. For the neural network **NN**, we use prior knowledge

[6] We denote the right singular vectors matrix by $\mathbf{Z}$ instead of the customary $\mathbf{V}$ to avoid confusion with the visual matrix.

[7] We also experimented with the same objective function as Socher et al. (2013), however, our objective function yielded consistently better results in all experimental settings.

Model \ k	1	2	3	5	10	20
Chance	1.1	2.2	3.3	5.5	11.0	22.0
SVD	1.9	5.0	8.1	14.5	29.0	48.6
CCA	3.0	6.9	10.7	17.9	31.7	51.7
lin	2.4	6.4	10.5	18.7	33.0	55.0
NN	3.9	6.6	10.6	21.9	37.9	58.2

Table 2: Percentage accuracy among top k nearest neighbors on CIFAR-100.

about the number of concept categories to set the number of hidden units to 20 in order to avoid tuning of this parameter. For the **SVD** model, we set the number of dimensions to 300, a common choice in distributional semantics, coherent with the settings we used for the visual and linguistic spaces.

First and foremost, all 4 models outperform **Chance** by a large margin. Surprisingly, the very simple **lin** method outperforms both **CCA** and **SVD**. However, **NN**, an architecture that can capture more complex, non-linear relations in features across modalities, emerges as the best performing model, confirming on a larger scale the recent findings of Socher et al. (2013).

5.1.1 Concept Categorization

In order to gain qualitative insights into the performance of the projection process of **NN**, we attempt to investigate the role and interpretability of the *hidden layer*. We achieve this by looking at which visual concepts result in the *highest* hidden unit activation.[8] This is inspired by analogous qualitative analysis conducted in Topic Models (Griffiths et al., 2007), where "topics" are interpreted in terms of the words with the highest probability under each of them.

Table 3 presents both seen and unseen concepts corresponding to visual vectors that trigger the highest activation for a subset of hidden units. The table further reports, for each hidden unit, the "correct" unseen concept for the category of the top seen concepts, together with its rank in terms of activation of the unit. The analysis demonstrates that, although prior knowledge about categories was not explicitly used to train the network, the latter induced an organization of concepts into superordinate categories in which the hidden layer acts as a cross-modal concept categorization/organization system. When the induced projection function maps an object onto the linguistic space, the derived text vector will inherit a mixture of textual features from the concepts that activated the same hidden unit as the object. This suggests a bias towards seen concepts. Furthermore, in many cases of miscategorization, the concepts are still semantically coherent with the induced category, confirming that the projection function is indeed capturing a latent, cross-modal semantic space. A *squirrel*, although not a "*large* omnivore", is still an animal, while *butterflies* are not *flowers* but often feed on their nectar.

[8]For this post-hoc analysis, we include a sparsity parameter in the objective function of Equation 5 in order to get more interpretable results; hidden units are therefore maximally activated by a only few concepts.

Unseen Concept	Nearest Neighbors
tiger	cat, microchip, kitten, vet, pet
bike	spoke, wheel, brake, tyre, motorcycle
blossom	bud, leaf, jasmine, petal, dandelion
bakery	quiche, bread, pie, bagel, curry

Table 4: Top 5 neighbors in linguistic space after visual vector projection of 4 unseen concepts.

5.2 Zero-shot Learning in ESP

For this experiment, we focus on **NN**, the best performing model in the previous experiment. We use a set of approximately 9,500 concepts, the intersection of the ESP-based visual semantic space with the linguistic space. For tuning the number of hidden units of **NN**, we use the MEN-concrete dataset of Bruni et al. (2014). Finally, we randomly pick 70% of the concepts to induce the projection function $f_{proj_{v \rightarrow w}}$ and report results on the remaining 30%. Note that the search space for the correct label in this experiment is approximately 95 times larger than the one used for the experiment presented in Section 5.1.

Although our experimental setup differs from the one of Frome et al. (2013), thus preventing a direct comparison, the results reported in Table 5 are on a comparable scale to theirs. We note that previous work on zero-shot learning has used standard object recognition benchmarks. To the best of our knowledge, this is the first time this task has been performed on a dataset as noisy as ESP. Overall, the results suggest that cross-modal mapping could be applied in tasks where images exhibit a more complex structure, e.g., caption generation and event recognition.

	Seen Concepts	Unseen Concept	Rank of Correct Unseen Concept	CIFAR-100 Category
Unit 1	**sunflower**, **tulip**, pear	butterfly	2 (**rose**)	flowers
Unit 2	**cattle**, **camel**, **bear**	squirrel	2 (**elephant**)	large omnivores and herbivores
Unit 3	**castle**, **bridge**, **house**	bus	4 (**skyscraper**)	large man-made outdoor things
Unit 4	**man**, **girl**, **baby**	**boy**	1	people
Unit 5	**motorcycle**, **bicycle**, tractor	streetcar	2 (**bus**)	vehicles 1
Unit 6	**sea**, **plain**, **cloud**	**forest**	1	large natural outdoor scenes
Unit 7	**chair**, **couch**, **table**	**bed**	1	household furniture
Unit 8	**plate**, **bowl**, **can**	clock	3 (**cup**)	food containers
Unit 9	**apple**, **pear**, **mushroom**	**orange**	1	fruit and vegetables

Table 3: Categorization induced by the hidden layer of the **NN**; concepts belonging in the same CIFAR-100 categories, reported in the last column, are marked in bold. Example: Unit 1 receives the highest activation during training by the category *flowers* and at test time by *butterfly*, belonging to *insects*. The same unit receives the second highest activation by the "correct" test concept, the *flower rose*.

Model \ k	1	2	5	10	50
Chance	0.01	0.02	0.05	0.10	0.5
NN	0.8	1.9	5.6	9.7	30.9

Table 5: Percentage accuracy among top k nearest neighbors on ESP.

5.3 Fast Mapping in ESP

In this section, we aim at simulating a fast mapping scenario in which the learner has been just exposed to a new concept, and thus has limited linguistic evidence for that concept. We operationalize this by considering the 34 concrete concepts introduced by Frassinelli and Keller (2012), and deriving their text-based representations from just a few sentences randomly picked from the corpus. Concretely, we implement 5 models: **context 1**, **context 5**, **context 10**, **context 20** and **context full**, where the name of the model denotes the number of sentences used to construct the text-based representations. The derived vectors were reduced with the same SVD projection induced from the complete corpus. Cross-modal mapping is done via **NN**.

The zero-shot framework leads us to frame fast mapping as the task of projecting visual representations of new objects onto language space for retrieving their word labels ($\mathrm{v} \rightarrow \mathrm{w}$). This mapping from visual to textual representations is arguably a more plausible task than *vice versa*. If we think about how linguistic reference is acquired, a scenario in which a learner *first* encounters a new object and *then* seeks its reference in the language of the surrounding environment (e.g., adults having a conversation, the text of a book with an illustration of an unknown object) is very natural. Furthermore, since not all new concepts in the linguistic environment refer to new objects (they might denote abstract concepts or out-of-scene objects), it seems more reasonable for the learner to be more alerted to linguistic cues about a recently-spotted new object than *vice versa*. Moreover, once the learner observes a new object, she can easily construct a full visual representation for it (and the acquisition literature has shown that humans are wired for good object segmentation and recognition (Spelke, 1994)) – the more challenging task is to scan the ongoing and very ambiguous linguistic communication for contexts that might be relevant and informative about the new object. However, fast mapping is often described in the psychological literature as the opposite task: The learner is exposed to a new word in context and has to search for the right object referring to it. We implement this second setup ($\mathrm{w} \rightarrow \mathrm{v}$) by training the projection function $\mathrm{f}_{\mathrm{proj}_{\mathrm{w}\rightarrow\mathrm{v}}}$ which maps linguistic vectors to visual ones. The adaptation of **NN** is straightforward; the new objective function is derived as

$$J(\mathbf{\Theta}_{\mathrm{w}\rightarrow\mathrm{v}}) = \frac{1}{2}(1 - sim(\mathbf{V}_s, \hat{\mathbf{V}}_s)) \qquad (6)$$

where $\hat{\mathbf{V}}_s = \sigma^{(2)}(\sigma^{(1)}(\mathbf{W}_s\boldsymbol{\theta}^{(1)})\boldsymbol{\theta}^{(2)})$, $\boldsymbol{\theta}^{(1)} \in \mathbb{R}^{d_w \times h}$ and $\boldsymbol{\theta}^{(2)} \in \mathbb{R}^{h \times d_v}$.

Table 7 presents the results. Not surprisingly, performance increases with the number of sentences that are used to construct the textual representations. Furthermore, all models perform better than **Chance**, including those that are based on just **1** or **5** sentences. This suggests that the system can make reasonable inferences about object-word connections even when linguistic evidence is very scarce.

Regarding the sources of error, a qualitative analysis of predicted word labels and objects as

v→w	w→v
cooker→potato	dishwasher→ corkscrew
clarinet→ drum	potato→ corn
gorilla→ elephant	guitar→ violin
scooter→ car	scarf→ trouser

Table 6: Top-ranked concepts in cases where the gold concepts received numerically high ranks.

Context \ Mapping	v → w	w → v
Chance	17	17
context 1	12.6	14.5
context 5	8.08	13.29
context 10	7.29	13.44
context 20	6.02	12.17
context full	5.52	5.88

Table 7: Mean rank results averaged across 34 concepts when mapping an image-based vector and retrieving its linguistic neighbors (v → w) as well as when mapping a text-based vector and retrieving its visual neighbors (w → v). Lower numbers cue better performance.

presented in Table 6 suggests that both textual and visual representations, although capturing relevant "topical" or "domain" information, are not enough to single out the properties of the target concept. As an example, the textual vector of *dishwasher* contains kitchen-related dimensions such as ⟨**fridge**, **oven**, **gas**, **hob**, ..., **sink**⟩. After projecting onto the visual space, its nearest visual neighbours are the visual ones of the same-domain concepts *corkscrew* and *kettle*. The latter is shown in Figure 3a, with a *gas hob* well in evidence. As a further example, the visual vector for *cooker* is extracted from pictures such as the one in Figure 3b. Not surprisingly, when projecting it onto the linguistic space, the nearest neighbours are other kitchen-related terms, i.e., *potato* and *dishwasher*.

6 Conclusion

At the outset of this work, we considered the problem of linking purely language-based distributional semantic spaces with objects in the visual world by means of cross-modal mapping. We compared recent models for this task both on a benchmark object recognition dataset and on a more realistic and noisier dataset covering a wide range of concepts. The neural network architecture emerged as the best performing approach, and our qualitative analysis revealed that it induced a categorical organization of concepts. Most importantly, our results suggest the viability of cross-modal mapping for grounded word-meaning acquisition in a simulation of fast mapping.

Given the success of **NN**, we plan to experiment in the future with more sophisticated neural network architectures inspired by recent work in machine translation (Gao et al., 2013) and multimodal deep learning (Srivastava and Salakhutdinov, 2012). Furthermore, we intend to adopt *visual attributes* (Farhadi et al., 2009; Silberer et al., 2013) as visual representations, since they should allow a better understanding of how cross-modal mapping works, thanks to their linguistic interpretability. The error analysis in Section 5.3 suggests that automated *localization* techniques (van de Sande et al., 2011), distinguishing an object from its surroundings, might drastically improve mapping accuracy. Similarly, in the textual domain, models that extract collocates of a word that are more likely to denote conceptual properties (Kelly et al., 2012) might lead to more informative and discriminative linguistic vectors. Finally, the lack of large child-directed speech corpora constrained the experimental design of fast mapping simulations; we plan to run more realistic experiments with true nonce words and using source corpora (e.g., the Simple Wikipedia, child stories, portions of CHILDES) that contain sentences more akin to those a child might effectively hear or read in her word-learning years.

(a) A kettle

(b) A cooker

Figure 3: Two images from ESP.

Acknowledgments

We thank Adam Liška for helpful discussions and the 3 anonymous reviewers for useful comments. This work was supported by ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

References

Barbara Abbott. 2010. *Reference*. Oxford University Press, Oxford, UK.

Paul Bloom. 2000. *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA.

S. R. K. Branavan, Harr Chen, Luke S. Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of ACL/IJCNLP*, pages 82–90.

Elia Bruni, Ulisse Bordignon, Adam Liska, Jasper Uijlings, and Irina Sergienya. 2013. Vsem: An open library for visual semantics representation. In *Proceedings of ACL*, Sofia, Bulgaria.

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

John Bullinaria and Joseph Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

Susan Carey and Elsa Bartlett. 1978. Acquiring a single new word. *Papers and Reports on Child Language Development*, 15:17–29.

Susan Carey. 1978. The child as a word learner. In M. Halle, J. Bresnan, and G. Miller, editors, *Linguistics Theory and Psychological Reality*. MIT Press, Cambridge, MA.

Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman. 2011. The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of BMVC*, Dundee, UK.

David Chen and Raymond Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of AAAI*, pages 859–865, San Francisco, CA.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of ICML*, pages 160–167, Helsinki, Finland.

Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. 2004. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, Prague, Czech Republic.

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.

Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. 2010. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34:1017–1063.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Proceedings of HLT-NAACL*, pages 91–99, Los Angeles, CA.

Diego Frassinelli and Frank Keller. 2012. The plausibility of semantic properties generated by a distributional model: Evidence from a visual world experiment. In *Proceedings of CogSci*, pages 1560–1565.

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, Nevada.

Jianfeng Gao, Xiaodong He, Wen-tau Yih, and Li Deng. 2013. Learning semantic representations for the phrase translation model. *arXiv preprint arXiv:1312.0482*.

Tom Griffiths, Mark Steyvers, and Josh Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114:211–244.

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.

David R Hardoon, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. 2006. A correlation approach for automatic image annotation. In *Advanced Data Mining and Applications*, pages 681–692. Springer.

Jonathon Hare, Sina Samangooei, Paul Lewis, and Mark Nixon. 2008. Semantic spaces revisited: Investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In *Proceedings of CIVR*, pages 359–368, Niagara Falls, Canada.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Tracy Heibeck and Ellen Markman. 1987. Word learning in children: an examination of fast mapping. *Child Development*, 58:1021–1024.

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.

Colin Kelly, Barry Devereux, and Anna Korhonen. 2012. Semi-supervised learning for automatic conceptual property extraction. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*, pages 11–20, Montreal, Canada.

Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of NIPS*, pages 1106–1114.

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

Thomas Landauer and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

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.

Alessandro Lenci. 2008. Distributional approaches in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1):1–31.

Chee Wee Leong and Rada Mihalcea. 2011. Going beyond text: A hybrid image-text approach for measuring word relatedness. In *Proceedings of IJCNLP*, pages 1403–1407.

David Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2).

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28:203–208.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119, Lake Tahoe, Nevada.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL*, pages 746–751, Atlanta, Georgia.

Mark Palatucci, Dean Pomerleau, Geoffrey Hinton, and Tom Mitchell. 2009. Zero-shot learning with semantic output codes. In *Proceedings of NIPS*, pages 1410–1418, Vancouver, Canada.

Alfredo F Pereira, Karin H James, Susan S Jones, and Linda B Smith. 2010. Early biases and developmental changes in self-generated object views. *Journal of vision*, 10(11).

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

John Searle. 1984. *Minds, Brains and Science*. Harvard University Press, Cambridge, MA.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of EMNLP*, pages 1423–1433, Jeju, Korea.

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

Jeffrey Siskind. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61:39–91.

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

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, Nevada.

Elizabeth Spelke. 1994. Initial knowledge: Six suggestions. *Cognition*, 50:431–445.

Nitish Srivastava and Ruslan Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Proceedings of NIPS*, pages 2231–2239.

Richard Szeliski. 2010. *Computer Vision : Algorithms and Applications*. Springer, Berlin.

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

Koen van de Sande, Jasper Uijlings, Theo Gevers, and Arnold Smeulders. 2011. Segmentation as selective search for object recognition. In *Proceedings of ICCV*, pages 1879–1886, Barcelona, Spain.

Luis Von Ahn. 2006. Games with a purpose. *Computer*, 29(6):92–94.

Jun Yang, Yu-Gang Jiang, Alexander Hauptmann, and Chong-Wah Ngo. 2007. Evaluating bag-of-visual-words representations in scene classification. In James Ze Wang, Nozha Boujemaa, Alberto Del Bimbo, and Jia Li, editors, *Multimedia Information Retrieval*, pages 197–206. ACM.

Haonan Yu and Jeffrey Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of ACL*, pages 53–63, Sofia, Bulgaria.