NeuroImage xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text

Andrew James Anderson [a,b,*], Elia Bruni [b], Alessandro Lopopolo [b], Massimo Poesio [b,c], Marco Baroni [b]

[a] Brain and Cognitive Sciences, University of Rochester, NY 14627, USA
[b] Center for Mind/Brain Sciences, 38068, Rovereto, Italy
[c] University of Essex, CO4 3SQ, UK

ARTICLE INFO

Article history:
Received 26 November 2014
Accepted 30 June 2015
Available online xxxx

Keywords:
Concept representation
Embodiment
Mental imagery
Perceptual simulation
Language
Multimodal semantic models
Representational similarity

ABSTRACT

Embodiment theory predicts that mental imagery of object words recruits neural circuits involved in object perception. The degree of visual imagery present in routine thought and how it is encoded in the brain is largely unknown. We test whether fMRI activity patterns elicited by participants reading objects' names include embodied visual-object representations, and whether we can decode the representations using novel computational image-based semantic models. We first apply the image models in conjunction with text-based semantic models to test predictions of visual-specificity of semantic representations in different brain regions. Representational similarity analysis confirms that fMRI structure within ventral-temporal and lateral-occipital regions correlates most strongly with the image models and conversely text models correlate better with posterior-parietal/lateral-temporal/inferior-frontal regions. We use an unsupervised decoding algorithm that exploits commonalities in representational similarity structure found within both image model and brain data sets to classify embodied visual representations with high accuracy (8/10) and then extend it to exploit model combinations to robustly decode different brain regions in parallel. By capturing latent visual-semantic structure our models provide a route into analyzing neural representations derived from past perceptual experience rather than stimulus-driven brain activity. Our results also verify the benefit of combining multimodal data to model human-like semantic representations.

Introduction

Embodiment theory predicts that visual aspects of object-related semantic knowledge are simulated in the visual system even when reading, but there is little quantitative information on the depth and spatial distribution of visual content in imagined representations. This article uses novel image-based computational models of generic concepts to distinguish visual from non-visual aspects of fMRI activity patterns cued by reading object nouns, without the participant having been explicitly asked to mentally visualize specific object images (as in previous studies classifying mental imagery e.g. Lee et al., 2012; Reddy et al., 2010) and quantitatively demonstrates detailed visual imagery by decoding multiple categories from its content using the models. We target three questions:

(1) Are visually embodied object representations elicited when people read and contemplate object words?
(2) How are visually embodied representations distributed throughout the cortex and how does this distribution relate to language-based semantic representations?
(3) How detailed is the visual-object mental imagery - can our models correctly assign labels to multiple object categories in unlabeled brain representations?

Why might we expect reading to induce visual object representations in the brain?

Building successively on behavioral evidence supporting both Paivio (1971) dual-coding theory and Glaser's (1992) lexical hypothesis theory, behavioral/fMRI evidence for Barsalou et al.'s (2008) language and situated simulation theory, recent EEG results support the view that word comprehension involves initial activation of a shallow language-based conceptual representation (for rapid semantic evaluation) that is later complemented by a deeper simulation of the visual properties of the concept (Louwerse and Hutchinson, 2012). More generally speaking an increasing body of evidence supports the notion that

* Corresponding author at: Brain and Cognitive Sciences, University of Rochester, NY 14627, USA.
E-mail address: andrewanderson@bcs.rochester.edu (A.J. Anderson).

http://dx.doi.org/10.1016/j.neuroimage.2015.06.093
1053-8119/

conceptual representations are embodied in sensory and motor systems, as opposed to being purely language-based (recent reviews are Binder and Desai, 2011; Pulvermüller, 2013). Conceptual representations of object words (*animals*, *tools*) are associated to brain regions linked to visual object perception (Martin, 2007), reading the names of objects that have acoustic properties (*telephone*) activates auditory processing regions (Kiefer et al., 2008) and reading action words elicits activity in representations of the body (e.g., *kick* activates foot/leg-related brain regions, Hauk et al., 2004). Beyond this relation between object words and perceptual/motor brain regions linked to experiencing the objects, little is known about the nature of embodied representations. Are they internally synthesized in some detail as a valuable component of cognition (Barsalou, 2009; Pulvermüller et al., 2010; Trumpp et al., 2013) or is perceptual and/or motor overlap a non-essential epiphenomena of thought (Bedny and Caramazza, 2011)? Regarding the depth of simulation Zwaan et al. (2002) observe behavioral differences consistent with perceptual representations of the overall shape of objects being automatically activated in sentence comprehension and Kellenbach et al. (2000) observe associated EEG signal differences using a perceptual semantic priming task. That detailed visually embodied aspects of concept representations elicited in language tasks can be extracted from fMRI data has not previously been demonstrated.

Why might we expect to identify detailed visually embodied representations in fMRI concepts?

Building on pioneering fMRI studies (Ishai et al., 2000; O'Craven et al., 1999), several recent multivariate analyses targeting relationships between visual perception and visual mental imagery (Lee et al., 2012; Reddy et al., 2010; Stokes et al., 2009a; Stokes et al., 2011)/visual attention (Peelen et al., 2009; Stokes et al., 2009b) provide evidence that internally induced visual object representations can be distinguished. More specifically, Stokes et al. (2009a), Reddy et al. (2010), Stokes et al. (2011), and Lee et al. (2012) demonstrate that fMRI activity evoked by explicit mental visualization of shape/object images can be discriminated by classifiers trained on fMRI data recorded when the respective images were perceived. Image-perception driven fMRI object representations correlate with computer-vision models of the specific experimental images (Leeds et al., 2013), which adds to the argument that mental-imagery representations discriminated in similar brain regions in the previous studies are indeed visually grounded. Peelen et al. (2009) suggest the existence and approximate anatomical location of object-category-specific abstract visual templates invoked in natural scene categorization that are "invariant to geometric and photometric changes and spatially unspecific". As such visual templates must be regularly, rapidly and efficiently evoked in routine activities, it is reasonable that they would also be recruited in cognitive processing that is not driven by sensory input.

How do we detect visually embodied object representations in fMRI activity patterns?

We introduce a new method to probe embodied representations with computational semantic models. By piggybacking on extensive research in computer vision, we build *image-based* models of generic object representations derived from natural image statistics, and use them to search for brain regions encoding similar information. Specifically, our approach is grounded on the argument that representational similarity structure (see Kriegeskorte et al., 2008a) between image-based models and visually embodied brain representations should correlate (better than other aspects of brain representation and better than with non-visual models).

How do we build an image-based model?

We construct image-based models through a computational procedure (schematically illustrated in Fig. 1) that links object names to combinations of abstract visual features automatically extracted from large collections of non-curated images (Bruni et al., 2014; Sivic and Zisserman, 2003). The images are non-curated in sense that they have not been picked nor edited to be particularly representative of the depicted objects: they are pictures that were independently taken for other purposes, retrieved from the internet using the object name as search term, and only filtered out if they don't contain the object. However, the object can be occluded, it might occur in multiple instances, it might be only present in the background, etc. In this sense, it is a very "natural" data set. Fig. 2 provides examples of the visual features that our algorithm spots and uses. Since the source images are natural pictures capturing widely differing instances of the same object (Fig. 3), we incorporate real-world levels of visual variability.

What is the difference between our image-based models and those previously applied?

Although image-based models have previously been used in multi-voxel pattern analysis studies, unlike our generic object representations, they have all been stimulus-specific, and used to track visual processing of a pictorial/text stimuli (e.g., Connolly et al., 2012; Devereux et al., 2013; Hiramatsu et al., 2011; Kriegeskorte et al., 2008a; Leeds et al., 2013). These have been models of relatively low-level visual processing (in particular V1 and V4, from the HMAX algorithm of Serre et al., 2007), with the exception of Leeds et al. (2013) and Khaligh-Razavi and Kriegeskorte (2014) who used various contemporary computer vision object recognition algorithms to build stimulus-specific image-based models, and our pilot analysis of Mitchell et al.'s (2008) image/text cued fMRI data with prototype models that did not differentiate imagery from other aspects of semantic representation (Anderson et al., 2013).

How can we distinguish visually embodied object representations from other aspects of embodied and disembodied representation?

Our image-based models provide a means to spot visually grounded representations. However, as language can describe similarities/differences between visual patterns (*bananas* and *lemons* are *yellow* and *cucumbers* are *green*), and indeed, due to the tendency of words to co-vary with the percepts they denote, sensorial knowledge is deeply embedded in language (Connell and Lynott, 2014; Louwerse, 2008), the models risk detecting visual aspects of language-based brain patterns. To deal with this we introduce a strategy where image-based models are used in combination with text-based semantic models to interpret fMRI activity patterns. These text-based models, induced from patterns of co-occurrence of words in large text corpora, are extensively used in cognitive and computer science as proxies to the linguistic aspects of human semantic memory (Landauer and Dumais, 1997; Lund and Burgess, 1996; Turney and Pantel, 2010). Thus, they should capture those aspects of conceptual similarity that people could extract by statistical generalization over their linguistic experience. Mitchell et al. (2008) demonstrated that such text-based semantic models encode linguistically-based semantic information sufficient to predict widely distributed activity patterns elicited when people think about object nouns (see section on Computational text-based semantic models for details about our text-based model and e.g. Devereux et al., 2010; Murphy et al., 2012; for a comparison of different text-based models applied to neural decoding). If there is a significant difference in correlation strength between a brain region and the image- or text-based model, it suggests that semantic representation in that region is visually/linguistically dominated (even if both models significantly correlate with it, because of the systematic word-percept covariance patterns

Please cite this article as: Anderson, A.J., et al., Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induc..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.06.093

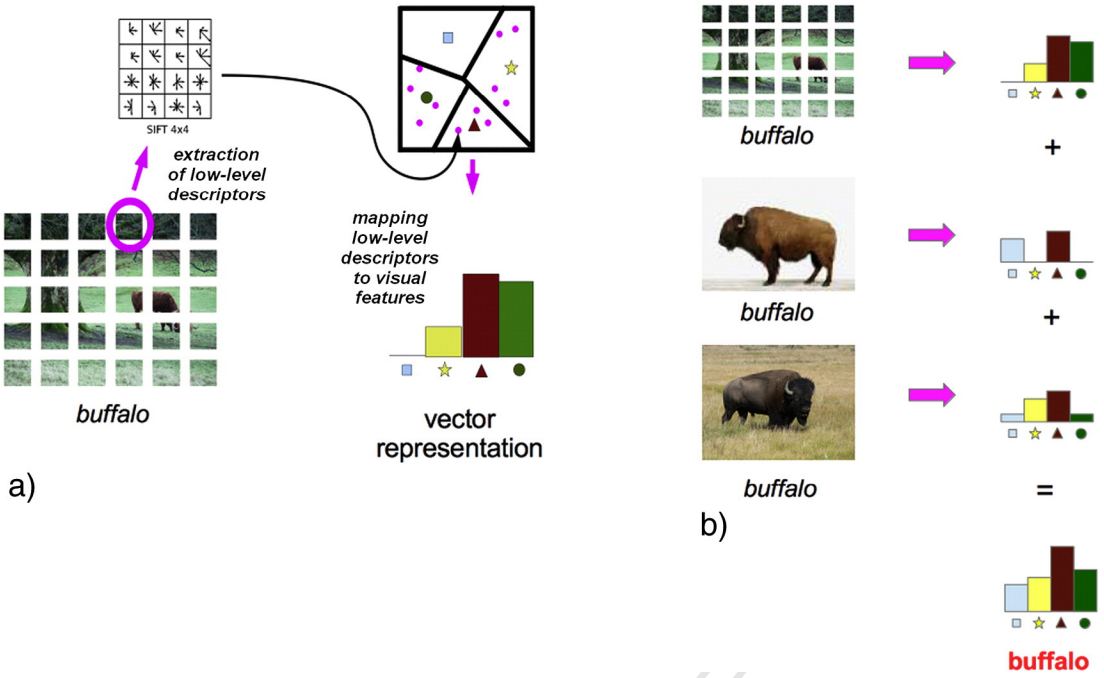

Fig. 1. (a) Visual representation of a single image. Our pipeline extracts low-level SIFT descriptors from equally spaced regions of the image. These low-level descriptors are then discretized by mapping them to a set of higher-level visual features (that have been determined in advance by clustering low-level descriptors from a larger image collection). The image is represented by a vector that records (a function of) how often each visual feature occurs in it (akin to the "bags-of-words" representation of documents in information retrieval, see, e.g., Manning et al., 2008). (b) Visual representation of a concept. Given a set of images depicting the same concept (e.g., a *buffalo*), the concept representation is obtained by summing the vectors representing all the input images.

we mentioned above). In particular, if a region correlates significantly more strongly with the visual model, we can be reasonably confident that it is sensitive to visual imagery information beyond what might be encoded in language. With only two models, the technique still does not safeguard against detecting embodied representations in other modalities. Since we have not developed motor/audio/smell/etc.-based computational models yet, we deal with this risk by taking into account prior knowledge of brain region modal-specificity through specific hypothesis tests, e.g., image-based models should correlate best with brain representations in known visual processing areas, and text-based models with linguistic/amodal brain regions. We make no claim about correlation in brain areas strongly linked to non-visual modalities.

Having detected visually embodied objects how can we quantify representational detail?

If we can discriminate object categories in visually embodied neural representations, we have proof of principle that imagery in a verbal task, whether it is conscious or unconscious, contains at least category-level visual detail (and classification accuracy provides a simple metric of this detail). This in turn would suggest that category-based decision making relying on internally simulated visual properties (e.g. "which image-category would look best…") is at least a possibility, which would extend beyond the view that grounded representations are functionally irrelevant epiphenomena of linguistic activity (though our study does not attempt to distinguish between epiphenomenal visual processing and functionally relevant processing). We use an unsupervised representational-similarity-based decoding algorithm (Raizada and Connolly, 2012) to classify visual-brain patterns using the image-based model. In addition, in line with recent proposals about how both perceptual and linguistic evidence is characterizing conceptual knowledge (Barsalou et al., 2008), and given that the previous analyses suggest that image- and text-based models provide complementary sources of information, we tailor the algorithm for multi-model brain-wide decoding. In a mixture-of-experts strategy, we use the image-based model to decode visually-specific brain regions in parallel with the text-based model decoding more linguistic regions.

In summary, we present:

(1) A hypothesis-driven representational similarity analysis (RSA) targeting specific brain regions and distinguishing visual/linguistic representations by differential correlation with image/text-based models.
(2) An exploratory searchlight RSA charting the spread and overlap of localized correlation with image/text-based models throughout the brain.
(3) Image-model driven unsupervised classification of visually embodied representations, and image/text-model decoding of multiple brain regions in parallel.

Materials and methods

fMRI experiment

We re-analyzed fMRI data for a popular set of stimuli (Table 1) consisting of 60 concrete nouns belonging to 12 taxonomic categories (because of limitations in the ImageNet database we use to construct our image-based computational model, we had coverage for 51 words and 11 classes) originally recorded by Just et al. (2010), but also forming the basis of, e.g., Mitchell et al. (2008); Leeds et al. (2013). Eleven consenting adults (consent approved by the University of Pittsburgh and Carnegie Mellon Institutional Review Boards), eight female, all right handed and all from the Carnegie Mellon community were shown each word six times over, and were asked to actively think about the properties of the object to which the word referred as further described below. Just et al.'s (2010) original experiment demonstrated latent, ecologically interpretable and spatially distributed semantic dimensions in the brain activity of participants thinking about the nouns. Details of data collection and preprocessing are as follows.

Please cite this article as: Anderson, A.J., et al., Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induc..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.06.093

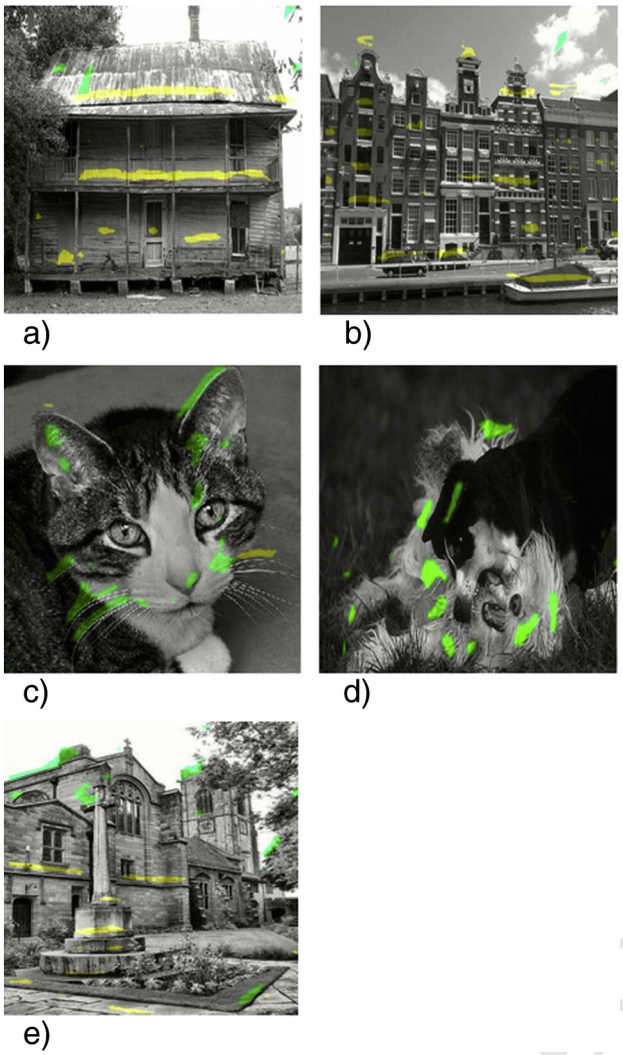

Fig. 2. Backprojection of visual features mainly associated to horizontal (yellow) and curvy/oblique (green) lines.

Fig. 3. Example pictures whose BoVF representations are nearest to (left) viz. farthest from (right) the average vector representing the corresponding concept.

Experimental paradigm and task

The sequence of words was presented in a random order in six runs. Individual words were displayed for 3 s, followed by a 7 s rest, during which time a fixation cross was shown. The fixation cross was also displayed 12 extra times for periods of 31 s to give a baseline.

On word presentation, participants actively thought about the properties of the object to which the word referred. To prompt consistent responses across all presentations of the same word participants had previously listed a set of word properties that they personally and freely associated with the noun. Different to Fairhall and Caramazza (2013) and Devereux et al. (2013), participants were not required to implicitly or explicitly reference the superordinate object category and different to Reddy et al. (2010) and Lee et al. (2012) participants were not asked to rehearse vivid mental imagery of the stimuli prior to the experiment or to specifically visualize the object.

fMRI acquisition and preprocessing

Just et al. (2010) recorded functional images on a Siemens Allegra 3.0 T scanner, with a gradient echo EPI pulse sequence (TR = 1000 ms, TE = 30 ms and 60° degree flip angle). Seventeen 5 mm thick oblique-axial slices were imaged with a 1 mm gap between slices and the acquisition matrix was 64*64 with 3.125 mm*3.125 mm*5 mm voxels. They subsequently corrected data for slice timing, participant motion and linear trend, before normalizing to MNI space and resampling to a 3*3*6 mm^3 grid. The voxel-wise percentage signal change between stimulus and fixation conditions was calculated, and the mean of the 4 images acquired 4 s after stimulus onset was taken to represent each word. Voxel values within each word were normalized by subtracting the mean and dividing by the standard deviation. To create a single brain representation per word per participant, we took the mean per voxel of the six presentations of each word. To create a single category representation per participant, we took the mean per voxel of all presentations of all words in that category.

Table 1
Classes and words from Just et al. (2010) covered by ImageNet.

Animals	Bear, Cat, Cow, Dog Horse
Buildings	Apartment, Barn, Church, House
Building parts	Arch, Chimney, Closet, Door, Window
Clothing	Coat, Dress, Pants, Shirt, Skirt
Furniture	Bed, Chair, Desk, Dresser, Table
Insects	Ant, Bee, Beetle, Butterfly, Fly
Kitchen utensils	Bottle, Cup, Glass, Knife, Spoon
Man-made objects	Bell, Key, Refrigerator, Telephone, Watch
Tools	Chisel, Hammer, Screwdriver
Vegetables	Celery, Corn, Lettuce, Tomato
Vehicles	Airplane, Bicycle, Car, Train, Truck

Please cite this article as: Anderson, A.J., et al., Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induc..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.06.093

Computational image-based semantic models

We describe here the procedure we adopted to construct our image-based representations of concepts. First we extract low-level visual descriptors from a large image collection, inducing a higher-level vocabulary of discrete features capturing simple visual properties that, following standard usage in the computer vision literature are called with the slightly misleading term "visual words". To avoid confusion here between "visual words" and the visually presented written word experimental stimuli, we will use the term "visual features" in place of "visual words". We represent the contents of a single image with a vector recording how many times each visual feature occurs in it. This is the so-called *Bag-of-Visual-Words* representation of the image (Csurka et al., 2004; Sivic and Zisserman, 2003), which we here shall refer to as *Bag-of-Visual-Features* (BoVF). Following up on our recent work (Bruni et al., 2014), we approximate the visual meaning of a *concept* (as opposed to a single depiction of it) by averaging the BoVF vectors of all images that are captioned with the relevant concept name in a large image corpus. The pipeline to extract concept representations from sets of images is summarized in Fig. 1. The model construction parameters described in this section were picked without tuning, based on their effectiveness in our earlier experiments.

Source data

We retrieved pictures from ImageNet (Deng et al., 2009), one of the largest freely available image databases (14 M pictures), with each picture manually captioned. We chose ImageNet for its high-quality images and labels, and because it provides bounding boxes localizing the object within a picture. Despite its size, ImageNet does not contain pictures with bounding boxes for 9 concepts used by Just et al. (2010), namely *arm, carrot, eye, foot, hand, igloo, leg, pliers, saw* (in Just et al., 2010), so we run our experiments on 51 concepts with a total of 17,765 images (350 images per concept on average).

Low-level feature vector extraction

Scale Invariant Feature Transform (SIFT) is one of the most effective low-level descriptor extraction methods for computer vision tasks such as object recognition and image retrieval (Grauman and Leibe, 2011; Lowe, 2004). Its success is due to its robustness to image scale, noise, distortion and partial invariance to illumination changes. SIFT produces descriptors that are actually 128-dimensional vectors encoding the magnitude gradients (directional changes in intensity) within a given image patch. In our implementation, we extract SIFT vectors from the whole image at regular pixel intervals. In particular, we densely sample the image on a regular grid, iterating through the image in steps of 5 pixels, at 4 different scales (10, 15, 20, 25 pixel radii), zeroing descriptors that correspond to very low contrast regions (which results in a small portion of the selected pixels not being associated to SIFT vectors). SIFT features are extracted for each component of the HSV (Hue, Saturation and Value) color space, resulting in 3x128-dimensional vectors, with 128 features per channel. Thus, in general, from an image with N pixels, we extract in the order of N/5 SIFT vectors.

Visual feature dictionary construction and bag-of-visual-features representation of specific images

We construct a visual dictionary by clustering all SIFT vectors extracted from our dataset into 25 K clusters (visual features), using a Gaussian mixture model (GMM), which can be seen as a *probabilistic* visual vocabulary that approximates the SIFT vector distribution via soft-clustering (Chatfield et al., 2011). To take uncertainty into account, each SIFT vector in the collection is associated to a probability distribution over clusters (visual features), instead of being assigned to the single nearest centroid. We use a large visual feature dictionary to account for the wealth of visual information encountered in the 18 K images from our collection. We will later reduce the full feature space to a small set of more general features, sufficient to represent the 51 concepts of interest, as described below (*Visual concept representations* paragraph).

For the representation of a specific image, each of its SIFT vectors is matched to the dictionary, and counted as an instance of the visual feature with the nearest centroid. For this last operation we use Fisher encoding, which captures average first and second order (mean and covariance) differences between GMM centers and SIFT vectors. The image as a whole is then represented in "visual feature space" by a higher-level vector that records how many times each visual feature occurs in it.

The original BoVF method completely discards information about the relative location of visual features in the image. In our implementation, we preserve partial spatial information by dividing the image into 8 regions, repeating the BoVF pipeline for each region and concatenating the resulting vectors (Grauman and Darrell, 2005; Lazebnik et al., 2006). The final vectors thus contain 200 K dimensions (25 K visual features $\times$ 8 regions). While the choice of preserving just approximate spatial information might be surprising, it insures that the BoVF approach is robust to the infinity of possible variations in part locations. The head of an animal will generally be above its legs, but a model attempting to encode precise head location would fail to generalize across pictures in which the head is on one or other side, from far or from near, etc. (Grauman and Leibe, 2011). Leeds et al. (2013) have observed that BoVF representations of images correlate with neural representations in the ventral visual pathway, suggesting that spatial information might not play such a crucial role as intuitively expected in human object recognition either.

To give intuition about the information that visual features are capturing, we utilize the backprojection technique of Yanulevskaya et al. (2012), allowing us to highlight all the pixels in an image whose SIFT vectors were assigned to a specific visual feature. Figs. 2a-e show all the pixels associated to two frequent visual features (with arbitrary yellow and green color coding) in a few pictures. "Yellow" visual features cluster on horizontal lines, "green" ones along somewhat oblique curvy lines. Consequently, the first visual feature is discriminative of buildings (Figs. 2a,b) and the second of animals (Figs. 2c,d), but there are exceptions, such as the building in Fig. 2e, which features both shapes, and consequently many instances of both visual features. As these examples illustrate, it is not possible to directly map visual features to any obvious high-level semantic denotation (e.g., object parts). However, image patches linked to the same visual feature capture basic but interpretable visual properties, such as simple shapes.

Visual concept representations

Following our previous work (Bruni et al., 2014), we derive a BoVF vector representation of a concept (e.g., a *dog*) by summing the (normalized) BoVF vectors of all images in our collection that are labeled with the target concept (e.g., all dog pictures). We then apply two transformations to the resulting aggregated count vector. First, raw counts are transformed into non-negative Pointwise Mutual Information (PMI) scores (Church and Hanks, 1990), assigning larger weights to visual dimensions that are more discriminative across concepts. Finally, we apply the Singular Value Decomposition technique (Manning et al., 2008) to the concept vectors, reducing them to 50 dimensions that generalize across the original BoVF features, with almost no loss of variance in the representation of the 51 concepts of interest, but a much more compact encoding of the relevant information.

The resulting "average" concept vectors abstract away from more idiosyncratic features of specific images, and capture their commonalities, thus approximating a data-induced prototype representation of the visual aspects of a concept. Fig. 3 illustrates how our aggregated vectors naturally encode prototypical representations. We present there, side

Please cite this article as: Anderson, A.J., et al., Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induc..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.06.093

by side, a specific picture whose vector is very near the average for the corresponding concept and a picture (still labeled with the relevant concept) that is quite far from the average vector (note that both kinds of pictures were used to construct the aggregated concept vectors). Clearly, pictures near the concept vector depict instances that are more prototypical than those that are far from it.

Image-based vectors were constructed with VSEM, an open library for visual semantics (Bruni et al., 2013; http://clic.cimec.unitn.it/vsem/).

Computational text-based semantic models

A long tradition of studies in computational linguistics and cognitive science has shown that it is possible to extract empirically effective representations of word meaning by using other words (or other linguistic units) that tend to naturally co-occur with a target term as semantic features a vector-based representation of its meaning (the resulting representations are often called distributional semantic models; see, e.g., Clark, 2013; Landauer and Dumais, 1997; Lenci, 2008; Lund and Burgess, 1996; Turney and Pantel, 2010). Again, the parameters of our linguistic vectors were picked without tuning, based on their effectiveness in our earlier work (Bruni et al., 2014).

We record co-occurrences with collocates within a window of a fixed size of 2 to left and right of each target word. Co-occurrence statistics are gathered from the freely available ukWaC and Wackypedia corpora combined, containing about 3 billion words in total (http://wacky.sslmit.unibo.it/). As collocates, we select a subset of 30 K words, composed by the top 20 K most frequent nouns, 5 K most frequent adjectives and 5 K most frequent verbs. Similar to the image-based vectors, we reweight and reduce the dimensionality of text vectors. However, unlike the approximately normally distributed visual feature counts, word frequencies in word corpora are heavily skewed with a prevalence of very rare types (the so-called "Zipfian" distribution). Since PMI is known to severely overestimate the importance of rare types, we correct non-negative PMI scores multiplying them by the corresponding raw co-occurrence counts (Evert, 2005; Manning and Schütze, 1999). As with the visual vectors, we then apply Singular Value Decomposition down to 50 dimensions.

Representational similarity analysis: general procedure and terminology

RSA compares the way the same referents are organized in different representational spaces (such as activation patterns in different areas of the brain and feature-based computational representations). We might expect a snake to occupy a similar position to a belt in visual space, but be far separated in linguistic semantic space, and RSA follows this intuition by systematically comparing paired item similarities. In the resulting common pairwise similarity space, we anticipate that visual similarity structuring (image-based models and visual brain areas) would be differentiable from linguistic structuring (text-based models and linguistic brain areas). In other words, that the matrix of paired similarities derived from a brain region of interest (ROI) that is visually specialized should correlate more with the matching matrix derived from image-based models (*Im*) than text-based models (*Tx*), and vice-versa.

The abstraction to similarity space allows multiple participants' brain data to be easily combined in an average similarity matrix, thus capturing group-level commonalities (and side stepping some problems surrounding imprecisions in spatial normalization of cross-participant data to the same anatomical space - in our case all participants were normalized to MNI space). This is beneficial if there are group-level commonalities in representational similarity (as opposed to individual representational schemes), in which case averaging across participants will bring out these regularities and cancel out noise in individual participants' data. It is also relevant that our computational semantic models are at group level, built from photographs taken by many people and text written by many authors. Although individual-level computational models are a theoretical possibility (e.g., built from an individual's documents and photography), their creation was infeasible for the current study.

We conduct hypotheses tests at both group-level (on mean similarity matrices) and individual-level (where hypotheses are first tested on individual's similarity matrices and then in a second-level analysis a one sample t-test is used to test whether the resultant set of individual-level test statistics differs from zero). Whilst group-level analyses are more likely to detect a pattern in the data (because the effects of noise are reduced), as a reviewer pointed out, we are still treating participants as fixed, rather than random effects, meaning that the inferences we draw might not generalize to other individuals from the population. We leave a more complex mixed-effects analysis to further work. We summarize results of both group and individual hypothesis tests in the main article and list all results in detail in Supplementary Materials. The more succinctly reportable group-level results covers all experimental and exploratory ROIs in both hemispheres (as defined in the next section). Individual-level results, which lead to the same conclusions, are provided for the key experimental regions (because results are lengthier to report, tabulation in detail is in Supplementary Materials 6).

Specifically, for each computational model or ROI, the representational similarity structure (denoted by $_{ss}$) was estimated by taking Pearson's correlation coefficient between all 1275 unique word pair combinations, given by the lower/upper off diagonal triangle of the 51*51 symmetric paired similarity matrix (e.g., Im_{ss} is the list of 1275 image-based paired similarities). Group-level similarity structures for each ROI were obtained by taking item-wise mean similarities.

The similarity structure of models and ROIs was compared using Spearman's correlation ($n = 1275$). Statistical significance was evaluated using a permutation test as described in Kriegeskorte et al. (2008a): The word labels of one of the two pairwise similarities matrices under comparison were randomly shuffled, rows and columns reordered accordingly, and the resulting similarity structure correlated with the other. This process was repeated 10,000 times and the p-value taken as the proportion of times the permuted correlation coefficient was greater than or equal to the observed correlation coefficient.

We repeated analyses at class-level (with classes defined by the original data set) following the reasoning that averaging representations within classes should let more general patterns emerge in the inherently noisy fMRI data. As the class *man-made objects* is in essence a superordinate category, that could reasonably subsume other test classes such as *tool* and *kitchen utensils*, it was left out of the class-level analyses. In the class-level RSA, there were 10 classes and 45 unique class pairs (RSA results including the left-out class are documented in Supplementary Materials Figure S2 and Table S2 and do not change interpretation). To create a group class-level similarity structure for each ROI, the item-wise mean similarity structure was taken across participants (45 mean similarities for 10 classes).

Hypotheses of visual/linguistic semantic representational dominance in different brain regions

Hypotheses predicting whether image- vs. text-based similarity structure would correlate more with specific ROIs were predominantly inspired by recent multivariate fMRI analyses of object representation, in particular those using RSA.

Visual dominance

Object/shape mental imagery can be discriminated in the ventral temporal cortex and lateral occipital areas and decoding of imagery is possible using classifiers trained on perceptual data (Lee et al., 2012; Reddy et al., 2010; Stokes et al., 2009a, 2009b; Stokes et al., 2011). Q6 This, on top of a number of multivariate analyses demonstrating object classification in the ventral-temporal stream when participants were

Please cite this article as: Anderson, A.J., et al., Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induc..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.06.093

cued by images (Connolly et al., 2012; Fairhall and Caramazza, 2013; Haxby et al., 2001), and Leeds et al.s' (2013) RSA correlating fMRI activation in the ventral-temporal-cortex cued by photographs corresponding to the words we analyze and SIFT-based models of the image stimuli lead to the hypothesis that image-based models would show significantly greater correlation with ventral temporal and lateral occipital areas.

Linguistic dominance

Our starting point for predicting linguistically dominant ROIs was Fairhall and Caramazza (2013) and Devereux et al. (2013), who used RSA to compare fMRI representations of five/six object categories, elicited by both pictures and words to semantic models, as both stimulus modalities should cue linguistic representations. Fairhall and Caramazza (2013) found that the left posterior-mid/inferior-temporal gyrus and posterior-cingulate/precuneus both support cross-modal classification and reflect the category structure of models derived from WordNet. Echoing this, Devereux et al (2013) found semantic category structure elicited by both text and pictures in the left middle temporal gyrus and left posterior cingulate/precuneus, and in addition in the right posterior cingulate/precuneus, inferior parietal lobe, left inferior frontal gyrus (in particular pars triangularis), left/right precentral gyrus and right superior frontal regions. Only representations in the intra-parietal-sulcus maintained a semantic category structure that did not change in response to text or image presentation, however pictures could elicit different or more specific linguistic-semantic representations than words, see also our Discussion, and Glaser (1992) for extensive coverage of behavioral experiments comparing picture/word stimuli in cognitive tasks.

As reading text may evoke visual simulations, viewing images may evoke linguistic representations, and image and text-based semantics may correlate (e.g., in context/situation), decoding brain activity elicited by different stimulus modalities across modalities is no guarantee of a strictly cross/amodal brain representation. However as the above regions were distinct from the regions we hypothesized to be visually dominant we provisionally considered them as candidates for linguistic knowledge representation.

We next referenced these regions to a recent review of semantic memory (Binder and Desai, 2011; Binder et al., 2009). The left mid-temporal gyrus and left posterior inferior parietal lobe have an established role in entity and event knowledge representation, whilst the left inferior frontal gyrus is implicated in knowledge evaluation/selection/retrieval and as such we might expect to detect specific aspects of semantic representation filtering/filtered from a general knowledge base stored elsewhere. We assigned all three regions hypotheses of text dominance. The posterior cingulate/precuneus is comparatively less well understood, with a possible role in episodic memory retrieval, irrespective of the memory's imagery content (Krause et al., 1999) and with stronger ties to encoding scene familiarity (real vs fictitious) as opposed to scene reconstruction (Hassabis et al., 2007). As both Fairhall and Caramazza (2013) and Devereux et al. (2013) detected semantic category structure here (possibly related to context/situation), we also assigned the posterior cingulate/precuneus a hypothesis of text dominance. The role of the superior medial frontal cortex is somewhat obscure (possibly translating affective states to a coordinated plan for knowledge retrieval) and therefore we analyzed it without a prediction.

Equivalent image/text correlation

Strong evidence that object representations in the inferior temporal gyrus interface visual and linguistic semantic knowledge is provided by Kriegeskorte et al. (2008b) and Carlson et al. (2014). The first analysis successfully correlated human fMRI data elicited by 92 color photographs of faces and bodies of animals and humans, and of natural and artificial objects to primate brain data. The second analysis demonstrated that the similarity structure of the same data correlates with text-based semantic models similar to those we use. We therefore predicted that both image- and text-based models would correlate with the left-inferior-temporal gyrus.

Exploratory analysis

Some ROIs were analyzed because of their posited roles in processing semantic knowledge (Binder and Desai, 2011; Binder et al., 2009), but without a dominance hypothesis, either because of their association with other modalities and/or their connection to our models is marginal. These were the precentral and supramarginal gyrii (motor/action), superior temporal and Heschl's gyrii (audition) and the dorsomedial frontal cortex.

Voxel selection in experimental ROIs

ROIs were partitioned using Tzourio-Mazoyer et al.s' (2002) automatic anatomical labeling (AAL) scheme. Individual AAL regions were combined as below (see also Fig. 4) into a total of eleven left hemispheric (L) and right hemispheric (R) sets. Hypotheses were defined to apply primarily to the left hemisphere, in line with the common observation of (right handers') left hemispheric dominance in semantic tasks (e.g., Binder et al., 2009) and mental imagery (e.g., Ishai et al., 2000), and Just et al.s' (2010) observation of left dominant activation on our data set. The hypotheses were reflected to the right hemisphere expecting weaker correlations.

As there was no strong a priori reason to predefine the number of voxels to analyze we repeated analysis within each ROI on the {50,100,200 and 400} most stable voxels: Pearson's correlation of each voxel's activity between matched words in all scanning session pairs (15 unique session pairs giving 15 correlation coefficients of 51 words) was computed and the mean coefficient used as stability measure (as per Mitchell et al., 2008). Experimental hypotheses were the same irrespective of number of voxels. If there were not 400 voxels in the ROI, the entire ROI was analyzed (number of voxels per ROI per participant are in Table S9).

AAL templates were as follows (defined for left hemisphere here, but combinations repeated for right). *Image dominance:* (Fig. 4-Blue): LMOG = {Occipital_Mid_L}, LVTC = {Fusiform_L & ParaHippocampal_L}. *Text dominance* (Fig. 4-Yellow): LMTG = Temporal_Mid_L; LPIP = {Angular_L, Parietal_Inf_L}; LIFG {Frontal_Inf_Tri_L, Frontal_Inf_Oper_L}; LPCP = {Precuneus_L, Cingulum_Post_L}; *Equivalent image/text correlation* (Fig. 4-Green): LITG = Temporal_Inf_L; *Exploratory analysis:* (Fig. 4-Red): LPG = Precentral_L; LSMG = {Supramarginal_L}; LSTG = {Heschl_L, Temporal_Sup_L}; LDMFC = {Frontal_Sup_Medial_L, Frontal_Sup_L}.

Results

Hypothesis-driven analysis: Image-based models detect embodied representations in visual areas

As set out in the introduction, to positively identify a visually embodied representation, the image-based model should show both a significantly stronger correlation than the text-based model in visual regions and weaker correlation in hypothesized non-visual regions (even if both models correlate with the same region, which is not unexpected given that the stimuli were not picked to contrast visual and linguistic similarity, and that visual/linguistic referents may occur in similar contexts). As Im_{ss} and Tx_{ss} were significantly correlated (word-level, ρ = .39, $p < .001$, $n = 1275$, class-level ρ = .42, $p < .001$, $n = 45$), we employed Steiger's (1980) test of difference between dependent correlations using the T_2 statistic.

The hypotheses were tested at both word and class levels. Pooling results for each lateralized anatomical ROI there were 8 tests of image/

Please cite this article as: Anderson, A.J., et al., Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induc..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.06.093

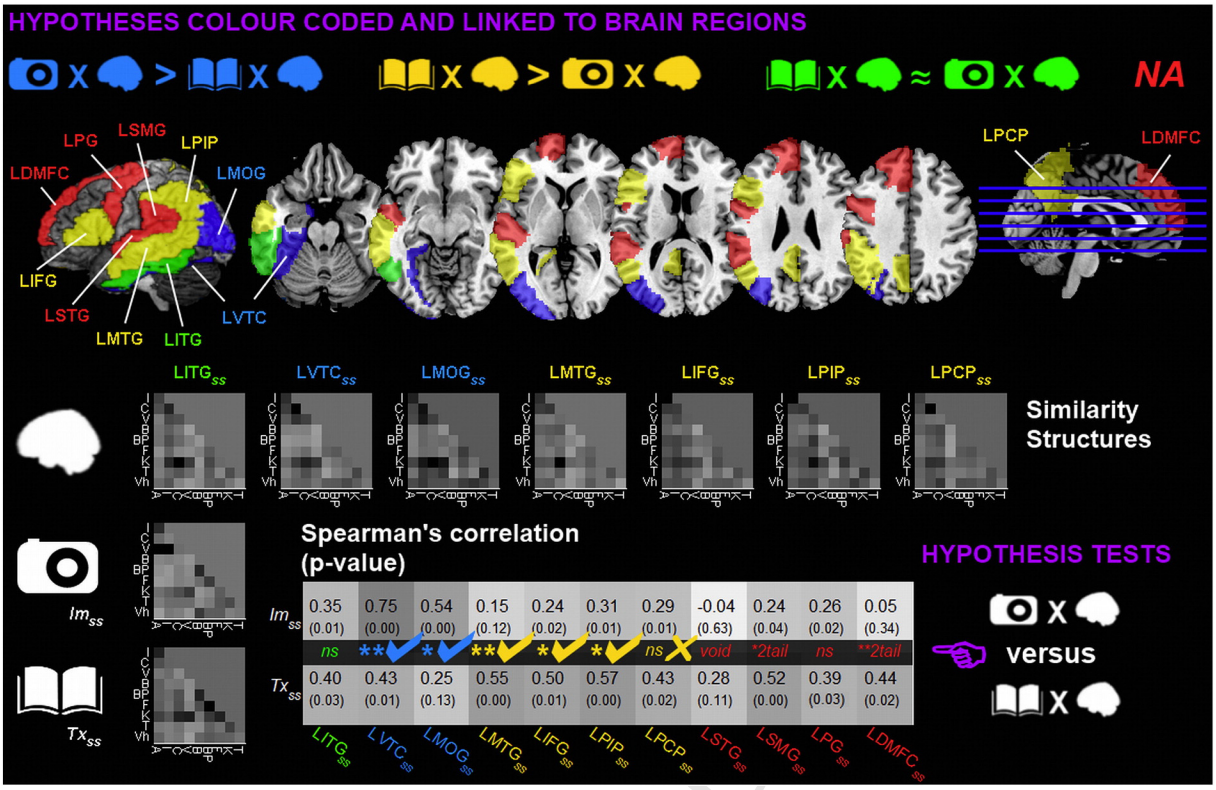

Fig. 4. Representative RSA results. (Top) Hypotheses predicting the representational preference of different brain regions to image/text-based models, color-coded and linked to brain regions. Primary experimental hypotheses (associated with image/text dominance) are blue/yellow. Green denotes hypothesized equivalence and red an exploratory analysis. (Bottom) Results for the class-level 200 voxel test (for all results see Supplementary Materials and Figures S1/S2 and Tables S1/S2). Similarity structures (triangular pixelated plots) are displayed for all models and ROIs. The table reports Spearman's correlations between each similarity structure pair and associated p-values (permutation test). Statistical significance of hypotheses tests are in middle row, ** $p < .01$, * $p < .05$. Blue/yellow ticks indicate confirmation of the associated hypothesis (e.g., blue-tick correlation in upper row is greater than in lower). Void indicates that neither image/text-based model correlated significantly with the ROI.

text modality preference (word/class-level *{50,100,200,400} voxels). If neither computational model significantly correlated with the ROI, the difference test was declared void (9/96 tests). The number of hypotheses supported in non-void tests at group level was as follows (all group-level tests and a word length control analysis is documented in Supplementary Materials 1 Figures S1/S2 and Tables S1-S3): LMOG (8/8), RMOG (5/8), LVTC (8/8), RVTC (3/8), LMTG (7/7), RMTG (4/4), LIFG (6/8), RIFG (7/7), LPIP (5/8), RPIP (6/6), LPCP (1/7) & RPCP (3/8). Correction for false discovery rate ($q = .05$), conducted separately at word-level and class-level, leads to the following amendments RVTC (0/8), LIFG (5/8), LPIP (4/8), LPCP (1/7). Zero tests were contradicted. Significant correlations between LITG and both models were present in all but two tests and there were no significant differences in correlation strength between models (as anticipated). RITG however was found only to correlate with the text-based model at class-level in two tests, where the correlation was significantly greater than with the image-based model.

Of the exploratory ROIs tested, only LSMG showed a significant word-level correlation (with both Im_{ss} and Tx_{ss}). At class-level, we found significant correlations of all exploratory ROIs with Tx_{ss}. Significant differences in the strength of correlations between Im_{ss} and Tx_{ss} were observed, especially in the right hemisphere where correlations with Im_{ss} were around zero.

Representative RSA results comparing Im_{ss} and Tx_{ss} correlations with ROIs at class-level (200 voxels) are in Fig. 4. To support that these results are representative we verified that the overall result pattern were consistent across hemispheres and when tested with different numbers of stable voxels. Specifically, lists of correlation coefficients of Im_{ss} and Tx_{ss} with each $LROI_{ss}$ were stacked ($n = 22$), the same stacking repeated for the right hemisphere, and the two resulting hemispheric lists were compared using Spearman's rank correlation at each test scale ({50,100,200,400}voxels). The correlations are uniformly high (word-level $\rho = \{.71,.84,.82,.77\}$, all $p \leq .001$; class-level, $\rho = \{.65,.75,.81,.77\}$ all $p \leq .002$). As anticipated, correlations were significantly greater in the left hemisphere, where the difference between the above data sets was tested using Wilcoxon signed rank tests: At word-level $W = \{15,8,21,38\}$ all $p \leq .004$. At class-level $W = \{51,52,39,42\}$ all $p \leq .015$. Finally and not surprisingly word and class-level results at each scale were strongly related, as tested by correlating all word-level correlations pooled across hemispheres with all pooled class-level results, giving $\rho = \{.82,.79,.82,.82\}$, all $p \leq .001$, $n = 44$.

The number of hypotheses supported in individual-level tests that focused on the key left-hemispheric experimental ROIs were LMOG (6/8), LVTC (5/8), LMTG (3/8), LIFG (8/8), LPIP (5,8), LPCP (1/8). There were no tests declared void and no contradictions. Tests that were not passed for LMOG, LVTC and LPIP were all at class-level. This is likely attributable to lower analytic power with fewer classes and less data (as the class *man-made objects* was excluded from the class-level analysis). All individual-level results are comprehensively documented in Supplementary Materials 6, Tables S10-S15.

As a further check, we ran a second set of analyses explicitly testing the sensitivity of ROIs to both models, to check for cases where one model has stronger correlation than the other but the weaker model still significantly explains variance in that ROI. We regressed ROI*ss* on one model's similarity structure and tested whether the residuals correlated with that of the other model. This was then repeated exchanging the models. Prior to regression, correlation coefficients in all similarity structures were hyperbolic-arc-tangent transformed according to Fisher's r to z transformation. These secondary analyses were run on the key left hemispheric ROIs that had been assigned modal dominance hypotheses, which were translated as follows. For an ROI hypothesized to be visually dominant the residuals following regression on the text-

Please cite this article as: Anderson, A.J., et al., Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induc..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.06.093

based model were predicted to positively correlate with the image-based model. When regression is on the image-based model, no positive correlation between the resultant residuals and the text-based model was predicted. Image and text-based models were switched accordingly for ROIs hypothesized to be linguistically dominant.

These analyses yielded similar results, with the comparative number of predictions supported being LMOG (16/16), LVTC (15/16), LMTG (14/14), LIFG (16/16), LPIP (13/16), LPCP (16/16), (there are double the number of tests comparative to the previous analysis, because rather than directly comparing image and text-correlations, we test first for visual sensitivity once text-based semantic effects are regressed out, and second for linguistic sensitivity once image-based effects are regressed out). In LVTC, at class-level, there was a single instance of significant correlation with the text-based model following removal of the image-based-model's regression line, and for LPIP one instance of this, and two of the counter-case of significant correlation with the image-based model after removal of the text trend at word-level. Detailed results are in Supplementary Materials 1 (Tables S3 and S4). Individual-level tests returned a similar outcome with the number of supported predictions being: LMOG (15/16), LVTC (14/16), LMTG (16/16), LIFG (16/16), LPIP (14/16), LPCP (13/16). Full documentation is in Supplementary Materials 6 (Tables S16-S19).

In summary hypothesis tests of the relative correlation strengths between Im_{ss} and Tx_{ss} and different ROIs favorably matched expectation. ROIs predicted to show visual dominance (L/RMOG, L/RVTC) did so in the vast majority of tests, and L/RMTG, L/RPIP and L/RIFG showed text-dominance also in line with prediction. We consider this as strong evidence that visually embodied object representations are activated as people read and contemplate object words, without visual object stimulation. L/RPCP were difficult to interpret (detailed results are in Supplementary Materials 1 Figures S1/S2 and Tables S1/S2) and we provisionally consider these ROIs to be weakly defined as regards our hypotheses.

Distribution of local image/text-based correlations throughout the brain

Searchlight RSA (Connolly et al., 2012; Devereux et al., 2013; Fairhall and Caramazza, 2013; Kriegeskorte et al., 2006; Leeds et al., 2013) were run to visualize the spread and relationship between localized image/text correlations throughout the brain. Searchlight was run at word and class-level. For space reasons, and as results are similar, class-level results are in Supplementary Materials 2. Firstly the set of all grey matter voxels in the brain that were common to all participants was identified by intersecting the MNI normalized grey matter voxel masks across participants. A contiguous voxel 'sphere' with a radius of 9 mm (mean ± std voxels: 56.5 ± 12.7) was iteratively centered on each voxel of the common MNI mask of all grey-matter voxels. At each sphere location, neural similarity structures ($SROI_{ss}$) were created individually for each participant in exactly the same way as the previous ROI analysis, but based only on voxels in the searchlight sphere. In order to capitalize on group level regularities, each $SROI_{ss}$ was averaged across all participants at each location. The mean $SROI_{ss}$ at each location was then correlated with Im_{ss} and Tx_{ss} using Spearman's ρ.

A permutation test was used to compute the p-value of the correlation coefficient at each sphere location. Following Leeds et al. (2013), word labels were permuted 500 times. The mean and variance of the correlation between the resultant shuffled similarity structures and the actual brain $SROI_{ss}$ were used to transform correlation coefficients to z-scores. A one-tailed p-value was then obtained using $p = 1 - erf(z)$ where erf is the cumulative density function of a standard normal distribution. Similarity maps of Im_{ss} and Tx_{ss} were thresholded by false discovery rate ($q < .1$) (Genovese et al., 2002). Significant ρ's ($p < .01$ from FDR) were plotted in MNI space.

Fig. 5 displays word-level searchlight results, correlation clusters are in Tables 2 and 3 for the image and text-based models respectively (class-level results are in Figure S3, Tables S6 and S7). Cyan/Orange indicates localized regions significantly correlated with Im_{ss}/Tx_{ss} respectively. Brown indicates overlap in significant correlations. Significant correlations with Im_{ss} only were largely located in the bilateral occipital cortex, the precuneus and in ventral regions of the temporal cortex (with slight left hemispheric bias) in line with the general visual dominance previously confirmed in L/RMOG and L/RVTC. Significant correlations with Im_{ss} were also detected within LIFG (particularly *pars opercularis*) and LSMG. The previous hypothesis testing analysis confirmed LIFG as linguistically-dominant despite significant correlation with Im_{ss}. LIFG might be selectively filtering aspects of broader semantic representations and intermittently processing visual and non-visual information: in this case, the image-correlated representations observed are unlikely to be directly related to visual processing.

Significant correlations with Im_{ss} were more widespread than Tx_{ss}. One interpretation is that visual simulations have more widespread activity traces than linguistic representations (which would be consistent with Glaser, 1992, who observed that pictures produce stronger priming effects than words). However the result is also on face value at odds with the hypothesis-driven ROI analyses (where more ROIs showed linguistic dominance). The difference is attributable to the two analyses operating at different spatial scales (the searchlight looking at all voxels within a small sphere comparative to the ROI analyses which selects stable voxels from a larger volume). As we don't know what spatial scale(s) visual simulations and linguistic cognition operate upon, and how visual/linguistic modalities interact, we remain agnostic on the interpretation of this result.

Significant correlations with Tx_{ss} were observed in similar/neighboring regions to Im_{ss} and infrequently detected in the right hemisphere, consistent with the standard finding of left hemispheric dominance in language tasks. Isolated correlations were observed in LPIP (including the angular gyrus), L/RVTC, L/RPCP, with smaller patches in LIFG and LITG/LMTG. The spread was consistent with the hypothesis testing analysis (and the comparative sparseness is a byproduct of the focal nature of searchlight analysis as identified above).

Areas of overlap between Im_{ss} and Tx_{ss} correlation, suggestive of a transition between visual aspects of object representations and higher-level linguistic semantics, were found in LITG (consistent with our expectation of high-level visual object representations in this region), LPIP, L/RPCP, portions of L/RVTC and within LIFG. Close inspection of the searchlight volume reveals that the multimodal overlap in LVTC is on the boundaries of the perirhinal cortex where Bruffaerts et al.s' (2013) observed that semantic similarity reflects fine-grained within-object-category semantic similarity between words, and more generally this network of areas is strikingly compatible with theories considering semantic memory to incorporate a selection of information convergence zones (e.g., angular/supramarginal gyri and sections of inferior/middle temporal and fusiform gyri) and high-level modulatory areas such as left inferior frontal regions (Binder and Desai, 2011; Binder et al., 2009; Mahon and Caramazza, 2009; Martin, 2007).

Decoding visual/linguistic brain representations using image and text-based models

If we can decode unlabeled embodied brain representations based only on information from the image-based models, we have evidence that the underlying embodied simulation synthesizes distinguishable visual patterns. If this is the case, we should also be able to utilize the complementary information in image/text-based models to decode different ROIs in parallel.

We exploit correspondences between model- and brain-activation-defined similarity spaces in order to guess the stimulus concept classes, without the need for manually labeled training data. Specifically, we adapt the algorithm proposed by Raizada and Connolly (2012) for cross-participant decoding to model-based decoding as follows. We are given the matrix of between-class pairwise similarities produced by a model. For the brain data, we have access to mean activation

Please cite this article as: Anderson, A.J., et al., Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induc..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.06.093

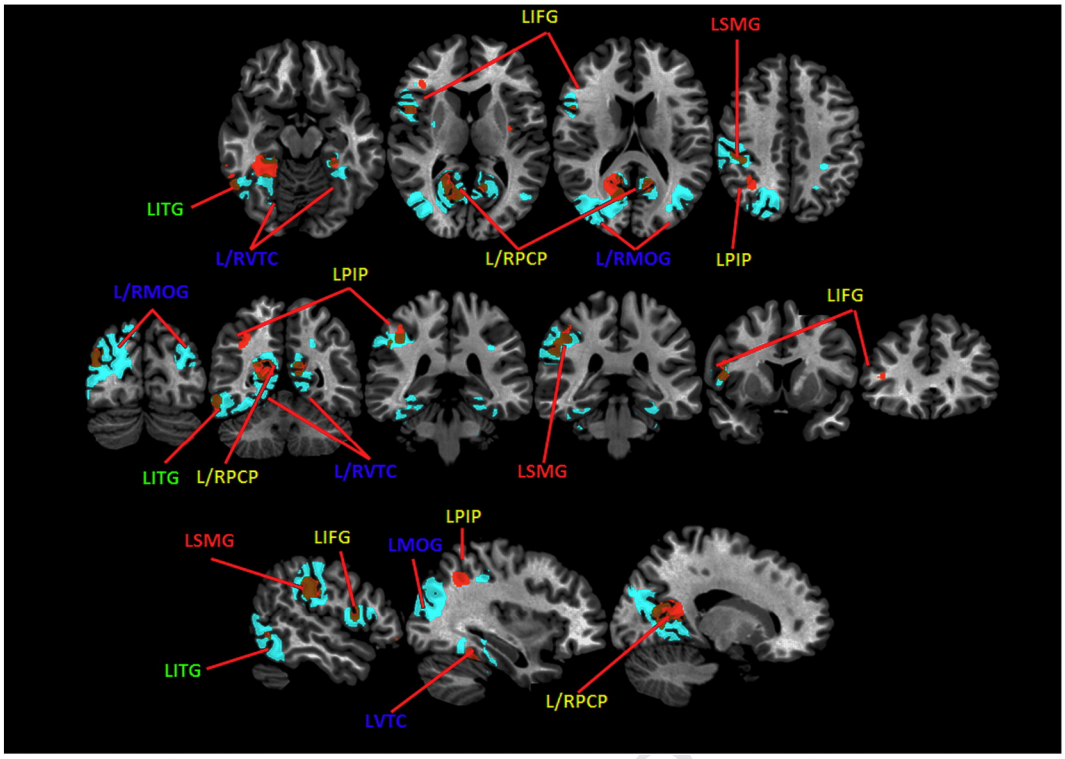

Fig. 5. Word-level searchlight RSA. Cyan/orange indicates significant image/text-based correlation. Brown indicates image/text overlap. Top row slices (MNI coordinates): z = −14, 9, 17, 43; mid row: y = −79, −56, −37, −33, 7, 30; bottom row: x = −50, −33, −23. See Figure S3 for class-level results.

vectors for each class, but not to the corresponding class labels. If the model is accurately capturing the same similarity structure as represented in the brain, we can look for the class label assignments that would make the brain similarity matrix as similar to the model-based matrix as possible. As a toy example, suppose that in model space we measure sim(mammal,insect) = .70, sim(mammal,tool) = .30, sim(insect,tool) = .10. We are given 3 unlabeled vectors of pooled brain activation data with similarities sim(x,y) = .83, sim(x,z) = .29, sim(y,z) = .11. Then, it is most reasonable to assume that x corresponds to mammal trials, y to insects and z to tools, since this assignment maximizes second-order similarity between model and brain similarities. More precisely, we exhaustively evaluate all the possible label permutations of the rows and columns of the brain similarity matrix, and pick the one leading to the largest Pearson correlation of the pairwise similarity scores with those in the model-based matrix (in ongoing work we are targeting word-level classification, which, with 51! possible label permutations necessitates approximate methods to search for the best solution).

We use Im_{ss} to decode class labels in LVTC (which showed greatest correlation to Im_{ss} in the ROI analysis). We repeated decoding on {50,100,200,400} voxels. Chance accuracy was 1/10, and equating conventional statistical thresholds to the number of classes correctly decoded (the same threshold applies using either Binomial or permutation testing as per Raizada and Connolly, 2012) gave [3/10 $p < .05$], [4/10 $p < .01$], [5/10 $p < .001$], [6/10 $p < .0001$]. The number of classes accurately decoded at the four different voxel scales using Im_{ss} was {1,2,0,8} and using Tx_{ss} was {0,0,0,2}. For the image-based model only, at 400 voxels, we could successfully decode eight of the ten classes. The instability in decoding at different scales for the image-based model is possibly due to an inadequate representation of all test classes in the 3 lower voxel scales. If visual simulations are naturally more stable for some words rather than others, then voxel selection would have biased towards covering the subset of most stable words. As the decoding algorithm relies on the inter-relationship between all classes this could seriously disrupt decoding accuracy. Given the local text-based correlation observed in the searchlight analysis of LVTC (Fig. 5), it is almost certain that non-visual aspects of semantic information are incorporated in the decoded representation. However, as successful decoding here was only possible based on entirely visual information, and accomplished without training, it is reasonable to assume that the decoded semantic representations are anchored in visual perception. This confirms that internally induced embodied representations contain at least a class-level degree of visual detail for many classes.

Nevertheless, the previous effect was brittle, being absent from 3 of 4 voxel scales. In line with our expectation that linguistic and visual model similarities would differentially correlate with brain similarities in different regions, we extend the previous procedure to operate by comparing different models with different brain ROIs in parallel. We now assume we have multiple brain region activation vectors for each class (but still don't know their class label). We then select the label assignment that maximizes the (weighted) sum of second-order similarity correlations with the relevant models across regions: That is, we search for a label assignment resulting in both high correlations with the text-based model for regions hypothesized to be mostly devoted to linguistic processing and high correlations with the imaged-based model in visual-processing regions. This "multimodal" decoding strategy thus capitalizes on targeting more diverse brain representations that are spread across greater brain areas, combining different sources of evidence for more robust decoding.

We determined which computational model to apply to each ROI based directly on our previous 6 modal-specificity hypotheses that linked image/text-based models to different ROIs. This was the first multimodal model configuration we tried. Each potential assignment of class labels to the brain data was scored using the mean weighted correlation in Eq. (1), where r(x,y) denotes Pearson's correlation

Please cite this article as: Anderson, A.J., et al., Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induc..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.06.093

Table 2
Image-based model word-level searchlight: Breakdown of AAL regions in significantly correlated searchlight clusters (sampled at 1 mm^3) and MNI coordinates of peak correlations per ROI (AAL regions identified as per Just et al., 2010).

Cluster	Vol (mm^3)	ROI	x	y	z
Bilateral Occipital/Temporal/Parietal	11246	Occipital_Mid_L	−37	−80	23
		Calcarine_L	−16	−62	11
		Lingual_L	−20	−53	−5
		Occipital_Sup_L	−20	−66	23
		Occipital_Inf_L	−54	−67	−11
		Cuneus_L	−15	−61	22
		Precuneus_L	−12	−54	16
		Parietal_Sup_L	−22	−68	44
		Parietal_Inf_L	−33	−79	42
		Angular_L	−46	−76	30
		Temporal_Inf_L	−46	−52	−16
		Fusiform_L	−31	−44	−18
		Temporal_Mid_L	−43	−72	16
		ParaHippocampal_L	−26	−37	−10
		Cerebelum_6_L	−31	−45	−29
		Cerebelum_4_5_L	−22	−41	−25
		Cerebelum_Crus1_L	−45	−44	−31
		Occipital_Mid_R	37	−71	28
		Calcarine_R	17	−58	11
		Lingual_R	14	−56	4
		Cuneus_R	15	−60	23
		Precuneus_R	12	−54	18
		Angular_R	40	−66	22
		Temporal_Mid_R	46	−72	21
		Cerebelum_4_5_R	10	−55	−4
Left Inferior Parietal	2098	SupraMarginal_L	−53	−30	31
		Postcentral_L	−52	−23	29
		Parietal_Inf_L	−50	−28	39
		Temporal_Sup_L	−58	−30	20
Left Inferior Frontal	569	Frontal_Inf_Oper_L	−53	8	11
		Rolandic_Oper_L	−47	3	11
		Frontal_Inf_Tri_L	−58	22	14
		Precentral_L	−48	6	14
Right Ventral Temporal	1684	Fusiform_R	32	−38	−19
		Cerebelum_4_5_R	25	−38	−26
		Cerebelum_6_R	33	−40	−29
		ParaHippocampal_R	28	−36	−13
		Lingual_R	29	−44	−8
		Temporal_Inf_R	44	−42	−20
Right Postcentral	83	Postcentral_R	26	−38	46
		SupraMarginal_R	28	−40	44

Table 3
Text-based model word-level searchlight: Breakdown of AAL regions in significantly correlated searchlight clusters (sampled at 1 mm^3) and MNI coordinates of peak correlations per ROI (AAL regions identified as per Just et al., 2010).

Cluster	Vol (mm^3)	ROI	x	y	z
Bilateral Occipital/Parietal/Temporal	838	Precuneus_L	−17	−53	14
		Calcarine_L	−18	−59	13
		Cuneus_L	−17	−55	22
		Lingual_L	−8	−67	6
		Precuneus_R	16	−56	16
		Calcarine_R	12	−58	18
		Cuneus_R	12	−58	20
		Lingual_R	11	−61	7
Left Occipital/Parietal	252	Occipital_Mid_L	−43	−80	27
		Angular_L	−44	−78	30
Left Inferior Parietal	388	SupraMarginal_L	−50	−27	30
		Parietal_Inf_L	−48	−31	36
		Postcentral_L	−53	−24	29
Left Posterior Parietal	113	Parietal_Inf_L	−30	−55	40
		Angular_L	−35	−55	36
		Parietal_Sup_L	−32	−60	44
Left Ventral Temporal	1200	Fusiform_L	−32	−39	−19
		Cerebelum_4_5_L	−29	−42	−24
		Cerebelum_6_L	−31	−44	−24
		ParaHippocampal_L	−30	−39	−11
		Temporal_Inf_L	−36	−37	−15
Left Lateral Temporal	236	Temporal_Inf_L	−59	−56	−16
		Temporal_Mid_L	−50	−60	−4
Right Ventral Temporal	321	Fusiform_R	31	−33	−18
		ParaHippocampal_R	29	−29	−18
		Cerebelum_4_5_R	28	−35	−24
Left Inferior Frontal	78	Frontal_Inf_Oper_L	−50	10	16
		Precentral_L	−50	6	14
		Frontal_Inf_Tri_L	−48	12	24
		Frontal_Inf_Tri_L	−48	12	24

coefficients of x and y and ss′ denotes the brain-based similarity structure generated by the label assignment to be scored:

$$\begin{aligned}\text{SCORE} = {} & r(\text{LMOG}_{ss'}, Im_{ss})/2 + r(\text{LVTC}_{ss'}, Im_{ss})/2 + r(\text{LMTG}_{ss'}, Tx_{ss})/4 \\ & + r(\text{LPIP}_{ss'}, Tx_{ss})/4 + r(\text{LPCP}_{ss'}, Tx_{ss})/4 + r(\text{LIFG}_{ss'}, Tx_{ss})/4. \end{aligned} \quad (1)$$

The weights of ½ and ¼ were chosen without tuning to give *Im* and *Tx* equal importance.

Decoding results at the four respective voxel scales were {7,7,7,8} (all $p < .0001$). It follows that if we had been given the brain data set blind, and attempted to match the unknown class labels to our known computational model class labels, we would have accurately recovered at least 7/10 classes. In the cases where accuracy was 7/10, errors were as follows: *building-part* was matched to *furniture*, *furniture* to *vehicle*, and *vehicle* to *building-part*. In the 8/10 case *vehicle* and *furniture* were confused. To verify that these results were specific to our hypothesized model/brain modality pairings, a set of control tests were run, firstly reversing Im_{ss} and Tx_{ss} (so that Im_{ss} was matched with ROIs hypothesized to have stronger associations with Tx_{ss} and vice versa) and secondly using either the image-based or text-based model alone to decode all ROIs. Results are in Table 4. None of the control tests were significant. A selection of simpler tests decoding brain region pairs (e.g. LVTC and LMTG; LVTC) that mirror this trend are in Supplementary Materials 3 Table S8.

Summarizing, we achieved high-accuracy decoding of brain representations in an object-selective visual area using our image-based model. This decoding was not possible using the text-based model. We claim that this is evidence that rich visually-embodied object representations are induced by thought in lack of an overt visual stimulus. Next, we showed that we could make results more robust by exploiting multimodal models with a novel adaptation of the decoding algorithm. A decoding test directly encapsulating our original modal-specificity hypotheses had accuracy between 7/10 and 8/10 (consistent with our predictions of image and text-based model dominance in different brain regions).

Discussion

We tested whether fMRI activity patterns elicited by participants reading object names (without visual cues) incorporate embodied visual representations of the objects, and whether we could decode the object class from the representation. To test for visually embodied representations we used novel image-based models derived from object features in conjunction with text-based models to distinguish visual/non-visual aspects of the brain's semantic representation. Different to all previous studies, our image-based models were not modeling a specific visual stimulus, but more generic imagery related to a concept.

Significantly greater correlation with image-based models was observed in visual object-selective brain regions, and with text-based models in brain areas posited to be modality independent. We further demonstrated that the image-, but not the text-based models, could decode object classes from visual brain representations with high accuracy, and, when we applied our models in combination to decode different brain regions in parallel, we got consistently high decoding accuracies. Key points are: (1) This provides evidence that rich embodied visual representations of objects are induced as words are read and contemplated. This mental imagery implicit in conceptual fMRI data contains

Table 4
Model-based decoding results and controls decoding with incongruent model/brain pairings and unimodal models. Scores are out of 10. The first 6 columns fit Eq. (1), e.g., for the incongruent model on the second results row: SCORE = r(LMOGss′,Txss)/2 + r(LVTCss′,Txss)/2 + r(LMTGss′,Imss)/4 + r(LPIPss′,Imss)/4 + r(LPCPss′,Imss)/4 + r(LIFGss′,Imss)/4. Accuracies ≥3 (in bold) are significant ($p < .05$, Binomial test). Further tests with model/ROI pairs are in Table S8.

$LMOG_{ss'}$	$LVTC_{ss'}$	$LMTG_{ss'}$	$LIFG_{ss'}$	$LPIP_{ss'}$	$LPCP_{ss'}$	50vox	100vox	200vox	400vox
Im_{ss}	Im_{ss}	Tx_{ss}	Tx_{ss}	Tx_{ss}	Tx_{ss}	**7**	**7**	**7**	**8**
Tx_{ss}	Tx_{ss}	Im_{ss}	Im_{ss}	Im_{ss}	Im_{ss}	0	0	0	0
Im_{ss}	Im_{ss}	Im_{ss}	Im_{ss}	Im_{ss}	Im_{ss}	0	0	1	1
Tx_{ss}	Tx_{ss}	Tx_{ss}	Tx_{ss}	Tx_{ss}	Tx_{ss}	0	0	0	0

more detailed visual categorical information than has previously been established (Pulvermüller, 2013) even in rehearsed visual imagery where fewer object categories have been successfully discriminated (Lee et al., 2012; Reddy et al., 2010) and is in line with the perceptual representations/mental simulations posited by Paivio (1971), Glaser (1992), Barsalou et al. (2008), Barsalou, 2009; (2) Visual semantic representations of objects are likely to be embodied in lower-level features than are amenable to linguistic representation in line with the depictive view of imagery (Kosslyn and Thompson, 2003) and also Lee et al. (2012); (3) Computational visual models can decode internally induced embodied representations. (4) We bring support to hypotheses derived from recent literature of how brain regions differentially contribute to encode semantic information.

What is the similarity between image-based models and brain representations?

Similarity in sensory input – we sourced the image-based models from a diverse selection of snapshots of objects in natural scenes that might approximate participants' ecological experience.

Similarity in representational features – that words provide a viable basis set for linguistic semantic representations, as per the text-based models, is not difficult to argue. Conversely, our dictionary of visual featuress, which were abstract combinations of local image descriptors was estimated empirically from an image collection. As can be seen from the backprojection of visual features onto natural scenes in Fig. 2, although visual features cluster on apparently similar local visual patterns, and there is evidence of systematic associations between them and classes of depicted objects, it is not straightforward to interpret what each visual feature represents. Obviously, it would be nice to have a neatly interpretable set of visual features, that could be easily defined in parameters such as shape, color and texture, similar to artificial stimuli used to probe visual object representations (e.g., Drucker and Aguirre, 2009; Op de Beeck et al., 2008). However, in hand with the suggestion that visually embodied representations may be grounded in lower level features than are amenable to linguistic description, it may not be easy to verbally define low-level embodied brain features. The computational tools used to build the image-based models draw upon an extensive history of computer vision research attempting to devise algorithms that are robust enough to deal with real world object recognition applications. Similar task demands (e.g., template matching in attention and mental imagery in planning) were presumably fundamental to shape biological concept representations. We thus conjecture that robustness constraints may have streamlined similar computational properties in both biological and artificial models.

Similarity in distributional representation (of features) – Demonstrating that neural object representations are spatially distributed was a founding step for multi-voxel pattern analyses (Haxby et al., 2001) and the principle of feature co-occurrence, used to derive computational semantic models, is akin to Hebbian learning. Assuming that neural associations are formed between combinations of nodes (neural populations receptive to shape fragments/shapes/words) that are frequently co-activated, then co-occurrence constitutes a valid approximation.

The validity of these assertions is amenable to future experimental testing through systematically modifying the diversity of information in source data, visual feature extraction strategies and nature of distributed representation, and comparing the fit to brain data.

Implications for the empirical study of embodied concept representations

An ability to interpret latent visual structure in brain representations adds credibility to an experimental approach complementary to the current "tradition emphasizing stimulus driven brain activity" (Binder and Desai, 2011). Contrary to perceptual experiments that aim to identify brain activity that covaries with controlled change in physical stimulus properties, studies of semantic memory aim to explain brain patterns derived from past perceptual experience. Computational semantic models, as we have shown, can provide a route into the study of internally induced and modality-preferential brain representations that are difficult to interpret otherwise (for instance embodied representations arising from dreaming). Additionally in much the same way as 'the book was better than the film', we conjecture that verbal stimuli may provide the best default means to trigger rich semantic representations even in their visual aspects. It follows that presenting specific modal information (e.g., specific dog pictures) in stimuli designed to probe general semantic representations may paradoxically limit valuable semantic content extracted from memory, by focusing the neural representation on the specific stimulus.

Implications for computational models of conceptual knowledge

(1) There is by now a long line of studies showing that computational models encoding statistical generalizations extracted from large bodies of text can simulate various aspects of human semantic memory (see, e.g., discussion and references in Lenci, 2008). Given the concomitant success of embodied approaches to meaning, a natural question arises about the division of labor, in grounding conceptual knowledge, between linguistic statistics and situated knowledge (e.g., Barsalou et al., 2008; Louwerse, 2008). As part of this debate, various authors have developed computational models that use subject-generated concept property descriptions as proxies to embodied experiential data (e.g., Andrews et al., 2009; Johns and Jones, 2012). While these simulations provide good insights into how the two knowledge sources can be integrated, subject-generated word lists are a rather artificial (and ultimately linguistic!) surrogate of perceptual knowledge. Our multimodal decoding experiments also confirm that integrating perceptual (specifically, visual) and linguistic information provides more human-like semantic representations. However, we approximate perceptual information with a model that is genuinely not verbal, but induced from natural images, combined with a state-of-the-art linguistic model. We thus pave the way to more realistic simulations of how linguistic and perceptual evidence are integrated into human conceptual knowledge.

(2) Brain data provides a useful test-bed for evaluating multimodal semantic models. The traditional approach to appraise semantic models compares human similarity judgments to models. As models incorporate additional modal information it becomes difficult to know how to configure norming questions appropriately. Even with only vision there are many ways to measure

Please cite this article as: Anderson, A.J., et al., Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induc..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.06.093

similarity (e.g. color/shape/texture). Beyond this, introspective judgments may overlook lower level features that our results suggest contribute to embodied representations. Comparing models to brain data circumvents many of these problems and there is obvious mutual benefit to fostering closer ties between biological and computational studies of concept representation.

Acknowledgments

Marcel Just for generously providing fMRI data for reanalysis, two anonymous reviewers for their useful suggestions and Rajeev Raizada for commentary on the curse of knowledge.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2015.06.093.

References

Q18 Anderson, A.J., Bruni, E., Bordignon, U., Poesio, M., Baroni, M., 2013. Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. Proceedings of EMNLP 2013 1960–1970 (Seattle, WA).
Andrews, M., Vigliocco, G., Vinson, D., 2009. Integrating experiential and distributional data to learn semantic representations. Psychol. Rev. 116 (3), 463–498.
Barsalou, L.W., 2009. Simulation, situated conceptualization, and prediction. Philos. Trans. R. Soc. Lond. B Biol. Sci. 364, 1281–1289.
Q19 Barsalou, L.W., Santos, A., Simmons, W.K., Wilson, C.D., 2008. Language and simulation in conceptual processing. In: De Vega, M., Glenberg, A.M., Graesser, A.C. (Eds.), Symbols, embodiment, and meaning. Oxford University Press, Oxford, pp. 245–283.
Bedny, M., Caramazza, A., 2011. Perception, action, and word meanings in the human brain: the case from action verbs. Ann. N. Y. Acad. Sci. 1224, 81–95.
Binder, J.R., Desai, R.H., 2011. The neurobiology of semantic memory. Trends Cogn. Sci. 15, 527–536.
Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb. Cortex 19, 2767–2796.
Bruffaerts, R., Dupont, P., Peeters, R., De Deyne, S., Storms, G., Vandenberghe, R., 2013. Similarity of fMRI activity patterns in left perirhinal cortex reflects similarity between words. J. Neurosci. 33 (47), 18597–18607.
Bruni, E., Bordignon, U., Liska, A., Uijlings, J., Sergienya, I., 2013. Vsem: An open library for visual semantics representation. Proceedings of ACL, pp. 187–192 (Sofia, Bulgaria).
Bruni, E., Tran, N., Baroni, M., 2014. Multimodal distributional semantics. J. Artif. Intell. Res. 49, 1–47.
Q20 Carlson, T.A., Simmons, R.A., Kriegeskorte, N., Slevc, L.R., 2014. The emergence of semantic meaning in the ventral temporal pathway. J. Cogn. Neurosci. 26 (1), 120–131.
Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A., 2011. The devil is in the details: an evaluation of recent feature encoding methods. Proceedings of BMVC.
Church, K., Hanks, P., 1990. Word association norms, mutual information, and lexicography. Comput. Linguist. 16 (1), 22–29 (DC).
Clark, S., 2013. Vector space models of lexical meaning. In: Lappin, S., Fox, C. (Eds.), Handbook of contemporary semantics, 2nd ed. Blackwell, Malden, MA.
Connell, L., Lynott, D., 2014. Principles of representation: why you can't represent the same concept twice. Top. Cogn. Sci. 6 (3), 390–406.
Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.C., et al., 2012. The representation of biological classes in the human brain. J. Neurosci. 32, 2608–2618.
Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints. Workshop on statistical learning in computer vision. ECCV, pp. 1–22.
Q21 Deng, J., Dong, W., Socher, R., Li, L., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. Proceedings of CVPR, pp. 248–255 (Miami Beach, FL).
Devereux, B., Kelly, C., Korhonen, A., 2010. Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In: Murphy, B., Chang, K.K., Korhonen, A. (Eds.), Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics. Association for Computational Linguistics, Los Angeles, USA, pp. 70–78.
Devereux, B.J., Clarke, A., Marouchos, A., Tyler, L.K., 2013. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. J. Neurosci. 33 (48), 18906–18916.
Drucker, D.M., Aguirre, G.K., 2009. Different spatial scales of shape similarity representation in lateral and ventral LOC. Cereb. Cortex 19 (10), 2269–2280.
Evert, S., 2005. The Statistics of Word Cooccurrences. Ph.D dissertation, Stuttgart University.
Fairhall, S., Caramazza, A., 2013. Brain regions that represent amodal conceptual knowledge. J. Neurosci. 33, 10552–10558.
Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15 (4), 870–878.
Glaser, W.R., 1992. Picture naming. Cognition 42, 61–105.
Grauman, K., Darrell, T., 2005. The pyramid match kernel: Discriminative classification Q22 with sets of image features. Proceedings of ICCV, pp. 1458–1465 (Beijing, China).
Grauman, K., Leibe, B., 2011. Visual object recognition. Morgan & Claypool, San Francisco.
Hassabis, D., Kumaran, D., Maguire, E.A., 2007. Using imagination to understand the neural basis of episodic memory. J. Neurosci. 27 (52), 14365–14374.
Hauk, O., Johnsrude, I., Pulvermüller, F., 2004. Somatotopic representation of action words in human motor and premotor cortex. Neuron 41 (2), 301–307.
Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.
Hiramatsu, C., Goda, N., Komatsu, H., 2011. Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. NeuroImage 57 (2), 482–494.
Ishai, A., Ungerleider, L.G., Haxby, J.V., 2000. Distributed neural systems for the generation of visual images. Neuron 28 (3), 979–990.
Johns, B., Jones, M., 2012. Perceptual inference through global lexical similarity. Top. Cogn. Sci. 4 (1), 103–120.
Just, M.A., Cherkassky, V.L., Aryal, S., Mitchell, T.M., 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. PLoS ONE 5 (1), e8622.
Kellenbach, M.L., Wijers, A.A., Mulder, G., 2000. Visual semantic features are activated during the processing of concrete words: Event-related potential evidence for perceptual semantic priming. Cogn. Brain Res. 10 (1), 67–75.
Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised models may explain IT cortical representation. PLoS Comput. Biol. 10 (11), e1003915. http://dx.doi.org/10.1371/journal.pcbi.1003915.
Kiefer, M., Sim, E.-J., Herrnberger, B., Grother, J., Hoenig, K., 2008. The sound of concepts: Four markers for a link between auditory and conceptual brain systems. J. Neurosci. 28 (47), 12224–12230.
Kosslyn, S.M., Thompson, W.L., 2003. When is early visual cortex activated during visual mental imagery? Psychol. Bull. 129, 723–746.
Krause, B.J., Schmidt, D., Mottaghy, F.M., Taylor, J., Halsband, U., Herzog, H., Tellmann, L., Müller-Gärtner, H.-W., 1999. Episodic retrieval activates the precuneus irrespective of the imagery content of word pair associates. Brain 122 (2), 255–263. http://dx.doi.org/10.1093/brain/122.2.255.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868.
Kriegeskorte, N., Mur, M., Bandettini, P., 2008a. Representational similarity analysis – Connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2 (4), 1–28.
Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A., 2008b. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60, 1126–1141.
Landauer, T., Dumais, S., 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. 104 (2), 211–240.
Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid Q23 matching for recognizing natural scene categories. Proceedings of CVPR, pp. 2169–2178 (Washington).
Lee, S.H., Kravitz, D.J., Baker, C.I., 2012. Disentangling visual imagery and perception of real-world objects. NeuroImage 59, 4064–4073.
Leeds, D.D., Seibert, D.A., Pyles, J.A., Tarr, M.J., 2013. Comparing visual representations Q24 across human fMRI and computational vision. J. Vis. 13 (13), 1–27 (25).
Lenci, A., 2008. Distributional semantics in linguistic and cognitive research. Ital. J. Linguist. 20 (1), 1–31.
Louwerse, M., 2008. Embodied representations are encoded in language. Psychon. Bull. Rev. 15, 838–844.
Louwerse, M.M., Hutchinson, S., 2012. Neurological evidence linguistic processes precede perceptual simulation in conceptual processing. Front. Psychol. 3, 385. http://dx.doi.org/10.3389/fpsyg.2012.00385.
Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110.
Lund, K., Burgess, C., 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Methods Instrum. Comput. 28, 203–208.
Mahon, B.Z., Caramazza, A., 2009. Concepts and categories: a cognitive neuropsychological perspective. Annu. Rev. Psychol. 60, 27–51.
Manning, C., Schütze, H., 1999. Foundations of statistical natural language processing. MIT press.
Manning, C., Raghavan, P., Schütze, H., 2008. Introduction to information retrieval. Cambridge University Press, Cambridge, UK.
Martin, A., 2007. The representation of object concepts in the brain. Annu. Rev. Psychol. 58, 25–45.
Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A., Just, M.A., 2008. Predicting human brain activity associated with the meaning of nouns. Science 320, 1191–1195.
Murphy, B., Talukdar, P., Mitchell, T., 2012. Selecting corpus-semantic models for neurolinguistic decoding. Proceedings of First Joint Conference on Lexical and Computational Semantics (*SEM), p. 114.
O'Craven, K., Downing, P., Kanwisher, N., 1999. fMRI evidence for objects as the units of attentional selection. Nature 401, 584–587.
Op de Beeck, H.P., Torfs, K., Wagemans, J., 2008. Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. J. Neurosci. 28, 10111–10123.

Please cite this article as: Anderson, A.J., et al., Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induc..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.06.093

Paivio, A., 1971. Imagery and verbal processes. Holt, Rinehart, and Winston, New York.
Peelen, M.V., Fei-fei, L., Kastner, S., 2009. Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature 460, 94–97.
Pulvermüller, F., 2013. How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. Trends Cogn. Sci. 17 (9), 458–470.
Pulvermüller, F., Cooper-Pye, E., Dine, C., Hauk, O., Nestor, P.J., Patterson, K., 2010. The word processing deficit in semantic dementia: all categories are equal, but some categories are more equal than others. J. Cogn. Neurosci. 22 (9), 2027–2041.
Raizada, R.D.S., Connolly, A.C., 2012. What makes different people's representations alike: neural similarity-space solves the problem of across-subject fMRI decoding. J. Cogn. Neurosci. 24 (4), 868–877.
Reddy, L., Tsuchiya, N., Serre, T., 2010. Reading the mind's eye: decoding category information during mental imagery. NeuroImage 50, 818–825.
Serre, T., Oliva, A., Poggio, T., 2007. A feedforward architecture accounts for rapid categorization. Proc. Natl. Acad. Sci. U. S. A. 104, 6424–6429.
Q25 Sivic, J., Zisserman, A., 2003. Video Google: A text retrieval approach to object matching in videos. Proceedings of ICCV, pp. 1470–1477 (Nice, France).
Steiger, J.H., 1980. Tests for comparing elements of a correlation matrix. Psychol. Bull. 87 (2), 245–251.
Stokes, M., Thompson, R., Cusack, R., Duncan, J., 2009a. Top–down activation of shape-specific population codes in visual cortex during mental imagery. J. Neurosci. 29, 1565–1572.
Stokes, M., Thompson, R., Nobre, A.C., Duncan, J., 2009b. Shape-specific preparatory activity mediates attention to targets in human visual cortex. Proc. Natl. Acad. Sci. U. S. A. 106, 19569–19574.
Stokes, M., Saraiva, A., Rohenkohl, G., Nobre, A.C., 2011. Imagery for shapes activates position-invariant representations in human visual cortex. NeuroImage 56, 1540–1545.
Trumpp, N.M., Kliese, D., Hoenig, K., Haarmeier, T., Kiefer, M., 2013. Losing the sound of concepts: Damage to auditory association cortex impairs the processing of sound-related concepts. Cortex 49 (2), 474–486.
Turney, P., Pantel, P., 2010. From frequency to meaning: Vector space models of semantics. J. Artif. Intell. Res. 37, 141–188.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15, 273–289.
Yanulevskaya, V., Uijlings, J., Bruni, E., Sartori, A., Zamboni, E., Bacci, F., Melcher, D., Sebe, N., 2012. In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings. Proceedings of the 20th ACM international conference on Multimedia. ACM, pp. 349–358.
Zwaan, R.A., Stanfield, R.A., Yaxley, R.H., 2002. Language comprehenders mentally represent the shapes of objects. Psychol. Sci. 13 (2), 168–171.

Please cite this article as: Anderson, A.J., et al., Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induc..., NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.06.093