Do Distributed Semantic Models Dream of Electric Sheep? Visualizing Word Representations through Image Synthesis

Angeliki Lazaridou and **Dat Tien Nguyen** and **Marco Baroni**
Center for Mind/Brain Sciences
University of Trento
{angeliki.lazaridou|tiendat.nguyen|marco.baroni}@unitn.it

Abstract

We introduce the task of visualizing distributed semantic representations by generating images from word vectors. Given the corpus-based vector encoding the word *broccoli*, we convert it to a visual representation by means of a cross-modal mapping function, and then use the mapped representation to generate an image of broccoli as "dreamed" by the distributed model. We propose a baseline dream synthesis method based on averaging pictures whose visual representations are topologically close to the mapped vector. Two experiments show that we generate dreams that generally belong to the the right semantic category, and are sometimes accurate enough for subjects to distinguish the intended concept from a related one.

1 Introduction

When researchers "visualize" distributed/distributional semantic models, they typically present 2D scatterplots illustrating the distances between a set of word representations (Van der Maaten and Hinton, 2008). We propose a much more direct approach to visualization. Given a vector representing a word in a corpus-derived distributed space, we generate a picture depicting how the denotatum of the word looks like, according to the model. Given, say, the word2vec vector of *broccoli*, we want to know how broccoli looks like to word2vec (see Figure 1 for the answer).

Besides the inherent coolness of the task, it has many potential applications. Current qualitative analysis of distributed semantic models is limited to assessing the *relation* between words, e.g., by looking at, or plotting, nearest neighbour sets, but it lacks methods to inspect the properties of a specific word directly. Our image synthesis approach will allow researchers to "see", in a very literal sense, how a model represents a single word. Moreover, in the spirit of the "*A picture is worth a thousand words*" adage, the generated images will allow researchers to quickly eyeball the results, getting the gist of what a model is capturing much faster than from textual neighbour lists. For example, a more "topical" model might produce pictures depicting the wider scenes in which objects occur (a ball being dribbled by soccer players), whereas a model capturing strictly conceptual aspects might produce narrow views of the denoted objects (a close-up of the ball). Image synthesis could also be used to explore the effect of different input corpora on representations: e.g., given a historical corpus, generate images for the *car* word representations induced from early 20th-century vs. 21st-century texts. As a last example, Aletras and Stevenson (2013) proposed to examine the topics of Topic Models by associating them with images retrieved from the Web. Given that topics are represented by vectors, we could directly *generate* images representing these topics.

In cognitive science, there is a lively debate on whether abstract words have embodied representations, (Barsalou and Wiemer-Hastings, 2005; Lakoff and Johnson, 1999), an issue that has recently attracted the attention of the distributed semantics community (Hill and Korhonen, 2014; Kiela et al., 2014; Lazaridou et al., 2015). An intriguing application of image synthesis would be to produce and assess imagery for abstract concepts. Recent work in neuroscience attempts to generate images of "what people think", as encoded in vector-based representations of fMRI patterns (Naselaris et al., 2009; Nishimoto et al., 2011). With our method, we could then directly compare images produced from corpus-based representations to what humans visualize when thinking of the same words.

Proceedings of the 2015 Workshop on Vision and Language (VL'15), pages 81–86,
Lisbon, Portugal, 18 September 2015.

In the long term, we would like to move beyond words, towards generating images depicting the meaning of phrases (e.g., an angry cat vs. a cute cat vs. a white cat) and sentences. This would nicely complement current work on generating verbal descriptions of images (Karpathy and Fei-Fei, 2015; Kiros et al., 2014) with the inverse task of generating images from verbal descriptions.

Generating images from vectorial word representations is of course extremely challenging. However, various relevant strands of research have reached a level of maturity that makes it a realistic goal to pursue. First, tools such as word2vec (Mikolov et al., 2013a) and Glove (Pennington et al., 2014) produce high-quality word representations, making us confident that we are not trying to generate visual signals from semantic noise. Second, there is very promising recent work on learning to map between word representations and an (abstract) image space, for applications such as image retrieval and annotation (Frome et al., 2013; Karpathy and Fei-Fei, 2015; Kiros et al., 2014; Lazaridou et al., 2014; Socher et al., 2014). Finally, the computer vision community is starting to explore the task of image generation (Gregor et al., 2015), typically in an attempt to understand the inner workings of visual feature extraction algorithms (Zeiler and Fergus, 2014).

The main aim of this paper is to present proof-of-concept evidence that the task is feasible. To this end, we rely on state-of-the-art word representation and cross-modality mapping methods, but we adopt an image synthesis strategy that could be seen as an interesting baseline to compare other approaches against. Briefly, our pipeline works as follows. Our input is given by pre-computed word representations (word2vec) and a set of labeled images together with their pre-compiled representations in a high-level visual feature space (specifically, we use activations on one of the top layers (fc7) of a convolutional neural network as high-level image representations). Given an input word vector, we use a linear *cross-modal function* to map it into visual space, and we retrieve the n nearest image representations. Finally, we overlay the actual images corresponding to these nearest neighbours in order to derive a visualization of the mapped word, a method we refer to as *averaging*. For example, the first image in Figure 1 below is our visualization of broccoli, obtained by projecting the *broccoli* word vector onto visual space, retrieving the 20 nearest images and averaging them.

Importantly, we apply this synthesis method to words that are not used to train the cross-modal mapping function, and that do not match the label of any picture in the image data set. So, for example, our system had to map *broccoli* onto visual space without having ever been exposed to labeled broccoli images (*zero-shot* setting), and it generated the *broccoli* image by averaging pictures that do not depict broccoli.

2 General setup

We refer to the words we generate images for as *dreamed* words, and to the corresponding images as *dreams*. We refer to the set of words that are associated to real pictures as *seen* words. The real picture set contains approximately 500K images extracted from ImageNet (Deng et al., 2009) representing 5.1K distinct seen words. The dreamed word set includes 510 concrete, base-level concepts from the semantic norms of McRae et al. (2005) (we excluded 31 McRae concepts because they were marked as ambiguous there, or for technical reasons).

Linguistic and Visual Representations For all seen and dreamed concepts, we build 300-dimensional word vectors with the word2vec toolkit,[1] choosing the CBOW method.[2] CBOW, which learns to predict a target word from the ones surrounding it, produces state-of-the-art results in many linguistic tasks (Baroni et al., 2014). Word vectors are induced from a corpus of 2.8 billion words.[3] The 500K images are represented by 4096-dimensional visual vectors, extracted with the pre-trained convolutional neural network model of Krizhevsky et al. (2012) through the Caffe toolkit (Jia et al., 2014).

Cross-modal mapping We use 5.1K training pairs $(\mathbf{w}_c, \mathbf{v}_c) = \{\mathbf{w}_c \in \mathbb{R}^{300}, \mathbf{v}_c \in \mathbb{R}^{4096}\}$, where $\mathbf{w}_c$ is the word vector and $\mathbf{v}_c$ the visual vector for (seen) concept c, the latter obtained by averaging all visual representations labeled with the concept (no dreamed concept is included in the training

[1] https://code.google.com/p/word2vec/

[2] Other hyperparameters, adopted without tuning, include a context window size of 5 words to either side of the target, setting the sub-sampling option to 1e-05 and estimating the probability of target words by negative sampling, drawing 10 samples from the noise distribution (Mikolov et al., 2013b).

[3] Corpus sources: http://wacky.sslmit.unibo.it, http://www.natcorp.ox.ac.uk

set, given the zero-shot setup). Following previous work on cross-modal mapping (Frome et al., 2013; Lazaridou et al., 2014), we assume a linear mapping function. To estimate its parameters $\mathbf{M} \in \mathbb{R}^{300\times4096}$, given word vectors $\mathbf{W}$ paired with visual vectors $\mathbf{V}$, we use L1-penalized least squares (Lasso) regression:[4]

$$\hat{\mathbf{M}} = \underset{\mathbf{M}\in\mathbb{R}^{300\times4096}}{\operatorname{argmin}} \|\mathbf{WM} - \mathbf{V}\|_F + \lambda\|\mathbf{M}\|_1$$

Image synthesis Suppose you have never seen cougars, but you know they are big cats. You might reasonably visualize a cougar as resembling a combination of lions, cheetahs and other felines. One simple way to simulate this process is through *image averaging*. Specifically, given the word representation $\mathbf{w}_c$ of a dreamed concept c, we apply cross-modal mapping $\mathbf{M}$ to obtain an estimate of its visual vector $\hat{\mathbf{v}}_c$. Following that, we search for the top $k = 20$ nearest images in 4096-dimensional visual space. Finally, the dream of concept c is obtained by averaging the colors in each pixel position (x, y) across the 20 images. These images do *not* contain the dreamed concept, and they will typically depict *several* distinct concepts (e.g., with a fairly accurate mapping $\mathbf{M}$, we might get the dream of cougar by averaging images of 5 cheetahs and 15 lions).[5]

3 Experiment 1: Naming the dream

Task definition and data collection In this experiment we presented a dream, and asked subjects if they thought it was more likely to denote the correct dreamed word or a confounder randomly picked from the seen word set (we did not use the "dream" terminology to explain the task to subjects). Since the confounder is a randomly picked term, the task is relatively easy. At the same time, since the confounders are picked from a set of concrete concepts, just like the dreamed words, it sometimes happens that the two concepts are quite related, as illustrated in Figure 1. Moreover, all confounders were used to train the mapping function, and their pictures are present in the averaging pool. These factors could introduce a bias in favour of them. We tested all 510 McRae words, collecting 20 ratings for each. We randomized word order both across and within trials. We used the CrowdFlower[6] platform to collect the judgments, limiting participation to subjects from English-speaking countries who self-declared English as their native language.

Figure 1: **Experiment 1:** Example dreams with correct dreamed word and confounder. Subjects showed a significant preference for the colored word (green if right, red if wrong).

Results Subjects show a consistent preference for the correct (dreamed) word (median proportion of votes in favor of it: 90%). Preference for the correct word is significantly different from chance in 419/510 cases (two-sided exact binomial tests, corrected for multiple comparisons with the false discovery rate method, $\alpha = .05$). Subjects expressed a significant preference for the confounder in only 5 cases (*budgie/parakeet*, *cake/pie*, *camel/ox*, *shotgun/revolver*, *squid/octopus*).

For the first two dreams in Figure 1, subjects showed a significant preference for the dreamed word, despite the fact that the confounder is a related term. Still, when the two words are closely related, it is more likely that subjects will be at random. The figure also shows two interesting examples in which dreamed word and confounder are related, and subjects significantly preferred the latter. The *tongs/utensil* case is very challenging, because any tongs picture would also be an utensil picture (and the dreamed object does not look like tongs to start with). For *zebra/baboon*, we conjecture that subjects could make up an animal in the dream, but one lacking the salient black-and-white pattern of zebras.

4 Experiment 2: Picking the right dream

Task definition and data collection In this experiment, we matched each dreamed word with its own dream and a confounder dream generated from the *most similar* dreamed term (see Figure 2 for examples). Word similarity was

[4] λ is 10-fold cross-validated on the training data.

[5] The idea of generating a more abstract depiction of something by averaging a number of real pictures is popular in contemporary art (Salavon, 2004) and it has recently been adopted in computer vision, as a way to visualize large sets of images of the same concept, e.g., averaging across different cat breeds (Zhu et al., 2014).

[6] http://www.crowdflower.com

measured in a space defined by subject-generated properties describing the concepts of interest (this method is known to produce high-quality similarity estimates, better than those obtained with text-based distributional models, see, e.g., Baroni et al. (2010)). Subjects were asked which of the two images is more likely to contain the thing denoted by the word. This is a very challenging task, as in most cases target and confounder are closely related concepts, and thus their dreams must have considerable granularity to allow subjects to make the correct choice. Again, we used CrowdFlower to collect 20 votes per item, with the same modalities of Experiment 1.

Results We expected the simple averaging method to fail completely at the level of accuracy required by this task. The results, however, suggest at least a trend in the right direction. This time, the median proportion of votes for the correct dream is at 60%. In 165/510 cases, there is a significant preference for the correct dream (same statistical testing setup as above), and in 57 cases for the confounder. A manual annotation of higher-level categories of dreamed word and confounder (e.g., *garment*, *mammal*, etc.) revealed that the proportion of votes for the correct dream was much higher in the 100 cases in which the two items belonged to different categories (80% vs. 55% for same-category pairs). The top row of Figure 2 illustrates cases where the pairs belong to the same category, and yet subjects still showed a strong preference for the correct dream. In the *tractor/truck* case, both dreams represent vehicles, but the correct one is evoking the rural environment a tractor. For *swan/dove*, we can make out birds in both dreams, but the *swan* dream is clearly of a larger, aquatic bird. Still, the more common case is the one where, if the two concepts are closely related, subjects assign random preferences, as they did for the examples in the second row.

5 Discussion

Averaging lets common visual properties in the source images emerge, as discussed in the next paragraphs in relation to the examples of Figure 3.

Shape The typical position and orientation of objects in images is an important factor determining dream quality. For example, weapons often appear in opposite orientations, which gives the averaged *bayonet* dream an improbable X-like shape. Other concepts, like *umbrella*, whose dream averages circular objects, are not so strongly affected by the orientation problem.

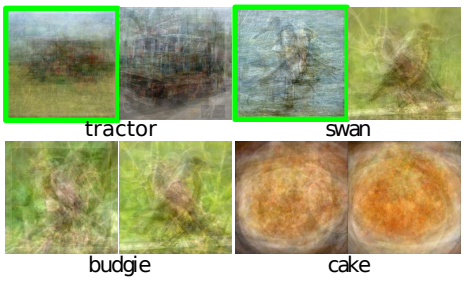

Figure 2: **Experiment 2:** Example dream pairs: the one on the left was generated from the word below the pair, the other from a confounder (clockwise from top left: *truck, dove, pie, parakeet*). Subjects showed significant preference for the green-framed correct dreams, and were at chance level in the other cases.

Context Even when bad object alignment leads to blurry dreams with unrecognizable concepts, averaging might highlight a shared context, sufficient to reveal the general category the dreamed concept belongs to. While both dreams in the 2nd column of Figure 3 are blurry, we can guess that the first one is related to water or to the sea, while the second is related to forest nature (dreams of a *mackerel* and *bison*, respectively).[7]

Color Visual averaging can differentiate concepts by capturing characteristics that are not typically verbalized. In black and white, the *skirt* and *trousers* dreams look almost identical (and they wrongly depict an upper-body garment). What differentiates the two images is color, red for *skirt* black for *trousers*. Indeed, a Google image search reveals that skirts tend to be colorful and trousers dark. The McRae norms list `is_colorful` as a property of *skirts*, but not trousers. We thus conjecture that image synthesis could provide fine-grained perceptual information complementing linguistic properties encoded in classic nearest neighbour lists.

6 Conclusion

We presented a proof-of-concept study taking the first steps toward generation of novel images from text-based word vectors. Obviously, the next step is to use genuine image generation methods in-

[7] Interestingly, Torralba (2003) used same-object image averaging to illustrate contextual priming during object detection.

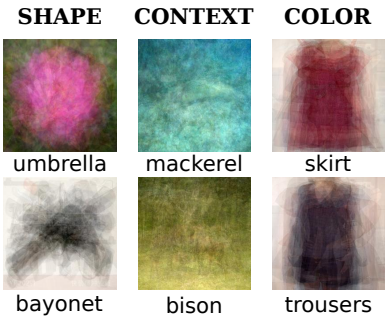

Figure 3: Examples illustrating properties of dream synthesis by image averaging.

stead of averaging (Gregor et al., 2015; Mahendran and Vedaldi, 2015; Vondrick et al., 2014; Zeiler and Fergus, 2014).

We would also like to consider alternative evaluation methods: for example, as suggested by a reviewer, asking subjects to label the generated dreams, and then measuring distance between the volunteered labels and the ground truth.

In a relatively short-term application perspective, given the intriguing results on context and other visual properties we reported, a natural first step would be to see how such properties change when different embeddings are used as input.

References

Nikolaos Aletras and Mark Stevenson. 2013. Representing topics using images. In *In Proceedings of NAACL-HLT*, pages 158–167.

Marco Baroni, Eduard Barbu, Brian Murphy, and Massimo Poesio. 2010. Strudel: A distributional semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*.

Lawrence Barsalou and Katja Wiemer-Hastings. 2005. Situating abstract concepts. In D. Pecher and R. Zwaan, editors, *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought*, pages 129–163. Cambridge University Press, Cambridge, UK.

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

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

Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. 2015. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proceedings of EMNLP*, pages 255–265, Doha, Qatar.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of CVPR*, Boston, MA. In press.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL*, pages 835–841, Baltimore, MD.

Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. In *Proceedings of the NIPS Deep Learning and Representation Learning Workshop*.

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

George Lakoff and Mark Johnson. 1999. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books, New York.

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

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of NAACL*, pages 153–163, Denver, CO.

Aravindh Mahendran and Andrea Vedaldi. 2015. Understanding deep image representations by inverting them. In *Proceedings of CVPR*.

Ken McRae, George Cree, Mark Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

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

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

Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. 2009. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915.

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543, Doha, Qatar.

Jason Salavon. 2004. http://cabinetmagazine.org/issues/15/salavon.php.

Richard Socher, Quoc Le, Christopher Manning, and Andrew Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Antonio Torralba. 2003. Contextual priming for object detection. *International journal of computer vision*, 53:169–191.

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

Carl Vondrick, Hamed Pirsiavash, Aude Oliva, and Antonio Torralba. 2014. Acquiring visual classifiers from human imagination. *arXiv preprint arXiv:1410.4627*.

Matthew Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of ECCV (Part 1)*, pages 818–833, Zurich, Switzerland.

Jun-Yan Zhu, Yong Jae Lee, and Alexei A Efros. 2014. Averageexplorer: Interactive exploration and alignment of visual data collections. *ACM Transactions on Graphics (TOG)*, 33:160.