Abstract

Imagination, creating new images in the mind, is a fundamental capability of humans, studies of which date back to Plato's ideas about memory and perception. Through imagery, we form mental images, picture-like representations in our mind, that encode and extend our perceptual and linguistic experience of the world. Recent work in neuroscience attempts to generate reconstructions of these mental images, as encoded in vector-based representations of fMRI patterns (Nishimoto et al., 2011). In this work, we take the first steps towards implementing the same paradigm in a computational setup, by generating images that reflect the imagery of distributed word representations.

We introduce *language-driven image generation*[1], the task of visualizing the contents of a linguistic message, as encoded in word embeddings, by generating a real image. Language-driven image generation can serve as evaluation tool providing intuitive visualization of what computational representations of word meaning encode. More ambitiously, effective language-driven image generation could complement image search and retrieval, producing images for words that are not associated to images in a certain collection, either for sparsity, or due to their inherent properties (e.g., artists and psychologists might be interested in images of abstract or novel words). In this work, we focus on generating images for distributed representations encoding the meaning of *single* words. However, given recent advances in compositional distributed semantics (Socher et al., 2013a) that produce embeddings for arbitrarily long linguistic units, we also see our contribution as the first step towards generating images depicting the meaning of phrases (e.g., *blue car*) and sentences. After all, language-driven image generation can be seen as the symmetric goal of recent research (e.g., (Karpathy and Li, 2014; Kiros, Salakhutdinov, and Zemel, 2014)) that introduced effective methods to generate linguistic descriptions of the contents of a given image.

To perform language-driven image generation, we combine various recent strands of research. Tools such as word2vec (Mikolov et al., 2013) and Glove (Pennington, Socher, and Manning, 2014) have been shown to produce extremely high-quality vector- based word embeddings. At the same time, in computer vision, images are effectively represented by vectors of abstract visual features, such as those extracted by Convolutional Neural Networks (CNNs) (Krizhevsky, Sutskever, and Hinton, 2012). Consequently, the problem of translating between linguistic and visual representations has been coached in terms of learning a *cross-modal mapping* function between vector spaces (Frome et al., 2013; Socher et al., 2013b). Finally, recent work in computer vision, motivated by the desire to achieve a better understanding of what the layers of CNNs and other deep architectures have

[1]Our work and this material have been published in (Lazaridou, Nguyen, and Baroni, 2015; Lazaridou et al., 2015)

really learned, has proposed *feature inversion* techniques that map a representation in abstract visual feature space (e.g., from the top layer of a CNN) back onto pixel space, to produce a real image (Zeiler and Fergus, 2014; Mahendran and Vedaldi, 2015).

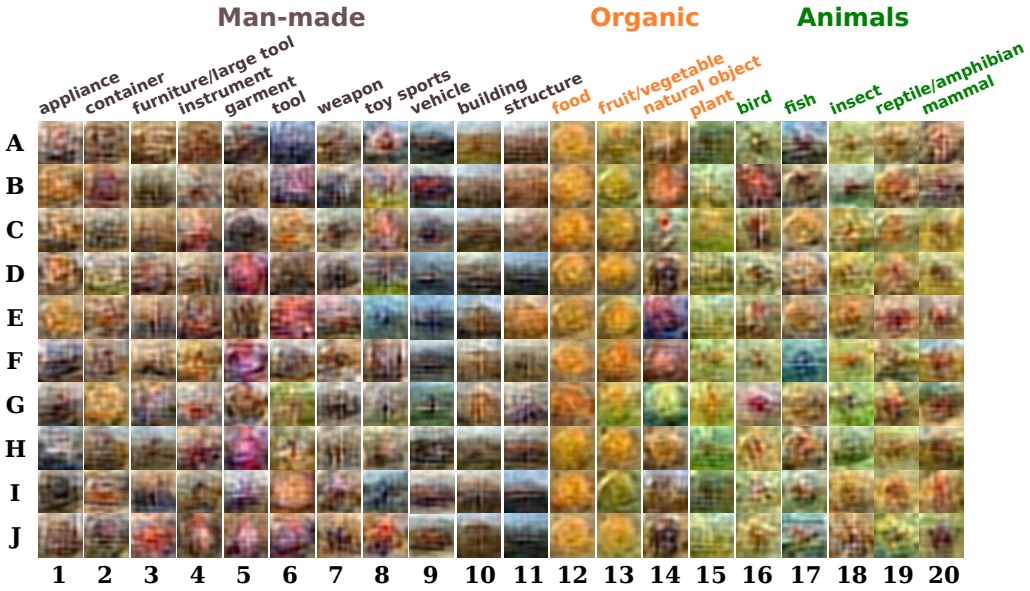

FIGURE 1: Generated images of 10 concepts per category for 20 basic categories, grouped my macro- category. See supplementary materials for the answer key.

Our language-driven image generation system takes a word embedding as input (e.g., the word2vec vector for *grasshopper*), projects it with a cross-modal function onto visual space (e.g., onto a representation in the space defined by a CNN layer), and then applies feature inversion to it (using the method HOGgles method of (Vondrick et al., 2013)) to generate an actual image (cell A18 in Figure 1). We test our system in a rigorous zero-shot setup, in which words and images of tested concepts are neither used to train cross-modal mapping, nor employed to induce the feature inversion function. So, for example, our system mapped *grasshopper* onto visual and then pixel space without having ever been exposed to *grasshopper* pictures.

Figure 1 illustrates our results ("answer key" for the figure provided as supplementary material). While it is difficult to discriminate among similar objects based on these images, the figure shows that our language-driven image generation method already captures the broad gist of different domains (food looks like food, animals are blobs in a natural environment, and so on).

Keywords: text2image, Cross-model Mapping, Distributed Semantics, Convolutional Neural Networks, Visual Feature Inversion.

This thesis is organized as follows:

Chapter 2 is a literature review of extracting semantic representations from text and images and related works. It starts by briefly introducing recent advanced research in word and image embedding. We also describe the task of cross-modal mapping in zero-shot manner. The last section is about recent research in visual feature inversion.

Chapter 3 presents our language-driven image generation system. It describes in detail our pipeline of image generation from word embeddings. The chapter first sketches out the system and then specifies materials which used to do training and evaluation.

Chapter 4 provides results on model selection and experimental evaluation. We carry our pre-experiments to determine the best model and parameters for our system in the first section. Subsequently, evaluation section covers four different experiments which estimate visual properties of the generated images. In each experiment, the task description is first described; hence the experimental results and discussion.

Chapter 5 is a summary of our achievements throughout the previous chapters. Some future research directions are also proposed to continue our recent work of image generation.

Bibliography

Frome, Andrea et al. (2013). "DeViSE: A Deep Visual-Semantic Embedding Model". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*. Pp. 2121–2129.

Karpathy, Andrej and Fei-Fei Li (2014). "Deep Visual-Semantic Alignments for Generating Image Descriptions". In: *CoRR* abs/1412.2306. URL: http://arxiv.org/abs/1412.2306.

Kiros, Ryan, 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*.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Proceeding of Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 1106–1114.

Lazaridou, Angeliki, Dat Tien Nguyen, and Marco Baroni (2015). "Do Distributed Semantic Models Dream of Electric Sheep? Visualizing Word Representations through Image Synthesis". In: *Proceedings of VL'2015, co-located with EMNLP*. Association for Computational Linguistics.

Lazaridou, Angeliki et al. (2015). "Unveiling the Dreams of Word Embeddings: Towards Language-Driven Image Generation". In: *Multimodal Machine Learning Workshop NIPS*.

Mahendran, Aravindh and Andrea Vedaldi (2015). "Understanding Deep Image Representations by Inverting Them". In: *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Mikolov, Tomas et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *Proceedings of Workshop at International Conference on Learning Representations (ICLR)*.

Nishimoto, S. et al. (2011). "Reconstructing visual experiences from brain activity evoked by natural movies". In: *Current Biology* 21.19, pp. 1641–1646.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove:Global vectors for word representation". In: *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, pp. 1532–1543.

Socher, Richard et al. (2013a). "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1631–1642. URL: http://www.aclweb.org/anthology/D13-1170.

Socher, Richard et al. (2013b). "Zero-shot learning through cross-modal transfer". In: *Proceedings of Annual Conference on Neural Information Processing Systems NIPS*. Lake Tahoe, NV, pp. 935–943.

Vondrick, Carl et al. (2013). "HOGgles: Visualizing Object Detection Features". In: *Proceedings of International Conference on Computer Vision (ICCV)*.

Zeiler, Matthew D. and Rob Fergus (2014). “Visualizing and Understanding Convolutional Networks”. In: *Proceeding of European Conference on Computer Vision (ECCV)*, pp. 818–833.