Modelling Visual Properties and Visual Context in Multimodal Semantics

Davis, C.; Bulat, L.; Vero, A.; Shutova, E.

Publication date
2018

Document Version
Final published version

Published in
Visually Grounded Interaction and Language (ViGIL)

Modelling Visual Properties and Visual Context in Multimodal Semantics

Christopher Davis, Luana Bulat, Anita Vero
Computer Laboratory
University of Cambridge
United Kingdom
ccd38, luana.bulat, alv34@cl.cam.ac.uk

Ekaterina Shutova
ILLC
University of Amsterdam
The Netherlands
e.shutova@uva.nl

Abstract

Multimodal semantic models that extend linguistic representations with additional perceptual input have proved successful in a range of natural language processing (NLP) tasks. However, existing research has extracted visual features from complete images, and has not examined how different kinds of visual information impact performance. We construct multimodal models that differentiate between internal visual properties of the objects and their external visual context. We evaluate the models on the task of decoding brain activity associated with the meanings of nouns, demonstrating their advantage over those based on complete images.

1 Introduction

Multimodal models combining linguistic and visual information have enjoyed a growing interest in the field of semantics. Recent research has shown that such models outperform purely linguistic models on a range of NLP tasks, including modelling semantic similarity [Silberer and Lapata, 2014], lexical entailment [Kiela et al., 2015], and metaphor identification [Shutova et al., 2016]. Despite this success, little is known about the nature of semantic information learned from images and why it is useful. For instance, some concepts may be better characterised by their own (internal) visual properties and others by the (external) visual context, in which they appear. However, existing multimodal semantic approaches use entire images to learn visual word representations, without differentiating between these two kinds of visual information. In contrast, we investigate whether differentiating between internal visual properties and external visual context is beneficial compared to learning visual representations from complete images. We construct three multimodal models combining linguistic and visual information: using (1) *internal* visual features extracted from an object's bounding box, (2) *external* visual features outside the bounding box, i.e. the visual context, and (3) visual features extracted from *complete* images. We use skip-gram [Mikolov et al., 2013] as our linguistic model and extract visual representations from a convolutional neural network (CNN) pre-trained on the ImageNet classification task [Fei-Fei, 2010].

We evaluate the models in their ability to decode patterns of brain activity associated with the meanings of nouns, obtained via brain imaging. This choice of task allows us to assess the importance of each type of visual information in human semantic processing. Specifically, we perform two experiments: (1) using the Visual Genome [Krishna et al., 2016] dataset of images where objects are manually annotated with bounding boxes, and (2) using images retrieved from Google Image Search and automatically segmenting them using a Faster R-CNN (FRCNN) model [Ren et al., 2015]. We find that all of our multimodal models are able to decode brain activity patterns and that the models relying on internal visual properties are superior to all others.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

2 Related Work

Multimodal Semantics Multimodal models are inspired by cognitive science research, suggesting that human semantic knowledge relies on perceptual and sensori-motor experience [Louwerse, 2011]. Contemporary approaches use deep CNNs trained on image classification tasks to extract visual representations of words. Kiela and Bottou [2014] extract visual word representations from feature extraction layers in CNNs and concatenate them with linguistic representations obtained from a skip-gram model. Their results presented empirical improvements over the previous bag-of-visual-words method [Bruni et al., 2012]. Other approaches use restricted Boltzmann machines [Srivastava and Salakhutdinov, 2012], recursive neural networks [Socher et al., 2014] and autoencoders [Silberer and Lapata, 2014].

Decoding Brain Activity Research in neuroscience supports the view that concepts are represented as patterns of neural activation and, similarly to distributed semantic representations, are naturally encoded in neural semantic vector space [Haxby et al., 2001, Huth et al., 2012, Anderson et al., 2013]. Mitchell et al. [2008] were the first to employ distributional semantic models to predict neural activation in the human brain using data obtained via functional Magentic Resonance Imaging (fMRI). Murphy et al. [2012], Devereux et al. [2010], Pereira et al. [2013] have since successfully tested a wider range of distributional models in this task. Recent research shows that multimodal models grounded in the visual modality strongly correlate with neural activation patterns associated with word meaning [Anderson et al., 2013, Bulat et al., 2017]. We aim to further our understanding of the role of vision in semantic processing by evaluating our models on this task.

3 Data

Visual Data In the first experiment, we used the Visual Genome [Krishna et al., 2016] dataset of images manually-annotated for objects and their bounding boxes. In the second experiment, we trained Faster-RCNN networks on manually annotated images from ImageNet [Deng et al., 2009, Fei-Fei, 2010], and then processed images retrieved from Google Images to construct a dataset of automatically-annotated images. Both Visual Genome and ImageNet were selected as they contain bounding box annotations around objects.

Brain Imaging Data We used a dataset of brain activity patterns associated with the meanings of nouns created by Mitchell et al. [2008] (MITCHELL). The dataset includes 60 concrete nouns from 12 semantic categories, such as *vehicles* or *vegetables*. fMRI images were recorded when participants were presented with line drawings of the objects and the corresponding nouns. We use 50 nouns from the dataset in our experiments, since 10 of the nouns were not covered by the Visual Genome and ImageNet datasets.

Following Mitchell et al. [2008], we select the 500 voxels with the most stable activation profile across concepts. We perform leave-two-out cross validation and select voxels independently for each of the cross validation folds during training. The stability score for a voxel is measured across six presentations of a word and is approximated as the average pairwise Pearson correlation among activation profiles over the training words in a cross-validation fold. The 500 voxels with the highest stability score are chosen and combined into a vector, used to evaluate how well the multimodal models can decode brain activity patterns.

4 Methods

We construct three visual models using three types of visual information: the internal features of the object, the external context surrounding it, and the whole image. These representations are then combined with linguistic representations to create the multimodal models.

4.1 Learning linguistic representations

We use the skip-gram model with negative sampling [Mikolov et al., 2013] to learn 100-dimensional word embeddings from a lemmatized 2015 copy of Wikipedia [Rimell et al., 2016].

4.2 Learning visual representations

Object detection and segmentation We use the FRCNN unified object detection model [Ren et al., 2015] to automatically detect objects and their bounding boxes in images associated with our

Model	Manual	Auto
LINGUISTIC	0.79	0.79
VIS-INTERNAL	**0.77**	**0.80**
VIS-EXTERNAL	0.71	0.74
VIS-WHOLE	0.71	0.80
MM-INTERNAL	0.80	0.81
MM-EXTERNAL	0.80	0.81
MM-WHOLE	0.79	0.82
VIS-COMBINED	0.76	0.79
MM-COMBINED	**0.81**	0.82

Table 1: Mean decoding accuracies for models trained on manually annotated images from Visual Genome, and those trained on automatically annotated Google images. *Vis*=visual, *MM*=multimodal, and COMBINED=explicitly differentiates internal and external features. Bolded values indicate statistical significance between the same type of model (visual-only vs multimodal) and models trained on the same annotation type (manual vs automatic annotations).

nouns. FRCNN combines a region proposal network (RPN) with Fast R-CNN, an object detection network, and minimizes computational cost during training and testing by sharing convolutional layers between the networks. To maximize accuracy, we train an FRCNN network for each semantic class in the MITCHELL dataset, starting from a VGG16 network [Simonyan and Zisserman, 2014] pre-trained on the PASCAL VOC 2007 data set.

Extracting visual features We retrieve 60 images per word using Google Image Search. We then create three sets of images for every word: the INTERNAL image (containing the object denoted by the word), an EXTERNAL image (containing its visual context), and the original WHOLE image. To generate the internal images, we crop and extract each object from within the annotated bounding boxes. To generate external images, we fill in the annotated bounding box area with black pixels, leaving only the visual context (black pixels are used as a simple way to represent no information). All images are re-scaled to 256x256 and the original aspect ratios are maintained, padding any remaining area with black pixels.

We use a Caffe [Jia et al., 2014] implementation of a pre-trained AlexNet model [Krizhevsky et al., 2012] to extract a visual representation for each of the images. We first take an image as input to the network, perform a forward pass, and extract the pre-softmax layer in the network (FC7) as a representation of the image. We use the MMfeat toolkit [Kiela, 2016] to load the AlexNet model and extract visual representations for the INTERNAL, EXTERNAL, WHOLE images corresponding to the nouns in our data set.

4.3 Multimodal Models

We construct multimodal models by concatenating L2-normalised linguistic and visual representations. This strategy, known as middle fusion, has been shown successful in previous multimodal semantics research [Kiela and Bottou, 2014]. We combine the linguistic model with each of our visual models, resulting in the three kinds of multimodal models: INTERNAL, EXTERNAL and WHOLE. Furthermore, we construct two *combined* models: a COMBINED visual-only model concatenating the internal and external models, and a COMBINED multimodal model concatenating the internal, external, and linguistic models.

4.4 Decoding Brain Activity

We evaluate our models in their ability to decode brain activity associated with unseen words, i.e. to predict the correct label associated with their fMRI patterns. We follow the same procedure as Anderson et al. [2016], computing a *semantic model similarity matrix* consisting of semantic model similarity codes for each of the 50 nouns from the MITCHELL dataset. Similarly, we construct a *brain activity similarity matrix* consisting of brain activity similarity codes of the 50 nouns. We perform leave-two-out cross validation, selecting the semantic model similarity codes $(\overrightarrow{s}_i, \overrightarrow{s}_j)$ and brain activity similarity codes $(\overrightarrow{a}_i, \overrightarrow{a}_j)$ for two nouns. We remove the *i*-th and *j*-th elements from each of the similarity codes as these entries correspond to the nouns being tested. Decoding is successful if the sum of Pearson correlations for the correct pairings is greater than the sum of Pearson correlations for the incorrect pairings, resulting in decoding accuracy of 1 for this pair and 0 otherwise. The expected chance-level decoding accuracy is 50% if a model were to match word labels with similarity vectors at random.

5 Experiments

We first experiment with a set of manually-annotated images from Visual Genome and then with images where objects and their bounding boxes have been automatically detected using FRCNN networks.

5.1 Manually annotated images

Experimental Setup We use 50 nouns from the MITCHELL dataset and assess each model's ability to decode brain activity vectors using leave-two-out cross validation, resulting in 1225 (50 choose 2) cross-validation folds per model.

Results The results, presented in Table 4.1, show all semantic models trained on manually annotated images decode brain activity patterns significantly above chance levels[1]. The INTERNAL visual-only model achieves a mean accuracy of 0.77, significantly[2] outperforming (V={36, 43}, all p<0.015) the EXTERNAL and WHOLE visual-only models, using the paired Wilcoxon signed rank test. The INTERNAL and EXTERNAL multimodal models both achieve a mean accuracy of 0.80, outperforming the WHOLE multimodal model with a mean of 0.79. Finally, the COMBINED multimodal model outperforms the INTERNAL and EXTERNAL multimodal models, and significantly outperforms (V=35, p<0.02) the WHOLE multimodal model with a mean accuracy of 0.81. These results demonstrate that it is beneficial to differentiate between internal and external visual information, but that both are useful for semantic processing, with the internal visual features having the most prominent influence for visual-only models.

5.2 Automatically annotated images

Experimental Setup For each of our 50 nouns from the MITCHELL dataset, we retrieve 60 images using Google Image Search. The images are annotated using FRCNNs and then processed to create INTERNAL, EXTERNAL and WHOLE models. We follow the same evaluation procedure as in the previous experiment, performing 1225 (50 choose 2) cross-validation folds.

Results Table 4.1 also presents decoding accuracies for the semantic models trained on the automatically annotated images, averaged over nine participants. We again find that all models decode brain activity vectors significantly above chance level. The results also show multimodal models constructed with automatic object detection perform on par with representations learned from manually annotated images. Overall, we observe a similar trend; the INTERNAL visual-only model significantly outperforms (V=43, p<0.015) the EXTERNAL visual-only model (mean accuracies of 0.80 and 0.74).

Besides corroborating the findings of the previous experiment on the importance of the internal visual features, these results show that high quality visual representations capturing the objects' internal properties and their visual context can be learned through automatic object detection techniques, decreasing the reliance on human annotated datasets (albeit some annotated data is required to train the object detection system) and allowing for a greater scalability of the models.

6 Conclusion

Our results show that multimodal semantic models correlate with human neural semantic representations associated with concrete concepts, and the visual-only model using internal visual features outperforms the other visual-only models in most cases. Similar performance across models using manual and automatically annotated images demonstrates progress in object detection systems, presenting opportunities to expand to other evaluation tasks where evaluation datasets may not be covered by manually annotated image datasets. Future research might investigate alternative fusion strategies, or whether NLP tasks benefit from visual feature differentiation.

[1] Using permutation testing with 1000 repeats, we found all models perform significantly above chance level. We follow the same shuffling procedure detailed in Anderson et al. [2017] to obtain a null distribution of chance-level decoding accuracies. The p-value of decoding accuracy is the proportion of chance-level accuracies greater than or equal to the observed cross-validated decoding accuracy.

[2] When comparing two models, we used paired Wilcoxon signed rank tests (two-tailed) to tell us whether their mean accuracy scores significantly differ from each other.

References

Andrew J Anderson, Elia Bruni, Ulisse Bordignon, Massimo Poesio, and Marco Baroni. Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. In *EMNLP*, pages 1960–1970, 2013.

Andrew J Anderson, Benjamin D Zinszer, and Rajeev DS Raizada. Representational similarity encoding for fmri: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, 128:44–53, 2016.

Andrew J Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30, 2017.

Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1219–1228. ACM, 2012.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

Barry Devereux, Colin Kelly, and Anna Korhonen. Using fmri activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 70–78. Association for Computational Linguistics, 2010.

Li Fei-Fei. Imagenet: crowdsourcing, benchmarking & other cool things. In *CMU VASC Seminar*, 2010.

James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 675–678, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3063-3. doi: 10.1145/2647868.2654889. URL `http://doi.acm.org/10.1145/2647868.2654889`.

Douwe Kiela. Mmfeat: A toolkit for extracting multi-modal features. In *Proceedings of ACL*, 2016.

Douwe Kiela and Leon Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*, pages 36–45. Citeseer, 2014.

Douwe Kiela, Laura Rimell, Ivan Vulic, and Stephen Clark. Exploiting image generality for lexical entailment detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, pages 119–124. ACL, 2015.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. URL `https://arxiv.org/abs/1602.07332`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Max M Louwerse. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2):273–302, 2011.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

Brian Murphy, Partha Talukdar, and Tom Mitchell. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 114–123. Association for Computational Linguistics, 2012.

Francisco Pereira, Matthew Botvinick, and Greg Detre. Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial intelligence*, 194: 240–252, 2013.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

Laura Rimell, Jean Maillard, Tamara Polajnar, and Stephen Clark. Relpron: A relative clause evaluation data set for compositional distributional semantics. *Computational Linguistics*, 42(4): 661–701, 2016.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. Black holes and white rabbits: Metaphor identification with visual features. In *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, 2016.

Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *ACL (1)*, pages 721–732, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.