Semantic Scene Statistics Using a Novel Computational Method

Dylan Rose & Peter Bex

Northeastern University, Psychology

Introduction

- Eye movement driven by bottom-up (salience)[1] & top-down (semantic) factors [2]
- The latter needs methods for assessing scene content relations without disrupting scene statistics by image manipulation (e.g. [3])
- Previous efforts (e.g. [2],[4]) don't directly address object-contextual relations,need relatively large sets of high quality, cleaned, observer responses (reports of image-patch "meaningfulness" [2] or object labels/segmentations [4])
- Our approach focuses on object-context relations with increased flexibility & application scope for the method through automation, automatic object labeling/segmentation using pre-trained neural net

Objectives

1. Develop quantitative object-contextual semantic relatedness maps. Test automatic approach to generate needed object position, label data against user-generated ground-truth maps from **LabelMe**[5]
2. Examine spatial distributions of object-contextual semantic content for maps generated from each data source
3. Compare semantic maps to image-salience maps
4. Compare semantic and salience maps for predicting observer eye movements

Methods

Stimuli

- Ten thousand randomly selected images from LabelMe database
- Images contained a minimum of two labeled objects; no other constraints

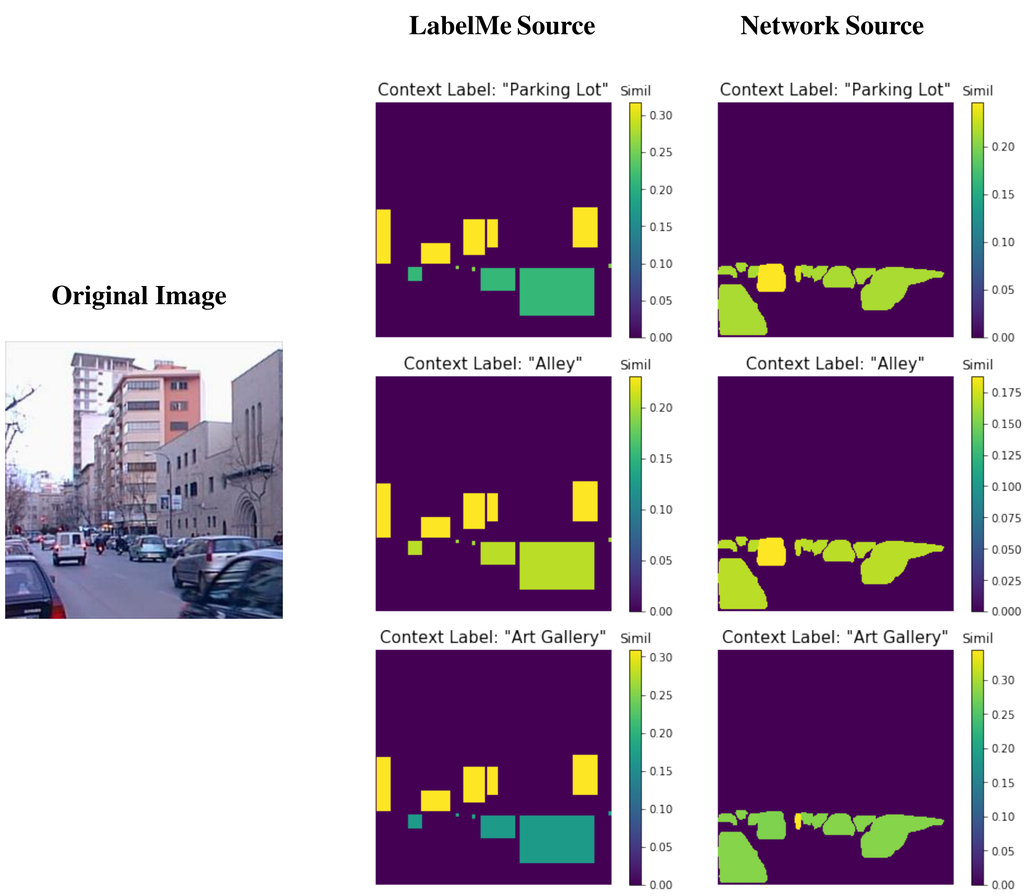
Semantic Similarity Map Calculation

- Generate five scene descriptors for each image with AlexNet implementation trained on MIT Places 365 using Keras
- Embed scene descriptors and object labels into pre-computed English Wikipedia language corpus
- Object-context label semantic similarity quantified as cosine distance using vector-space language model
- Compare semantic-similarity/spatial relationships between maps generated with LabelMe object segmentation/tags with network generated equivalents

Gaze Data

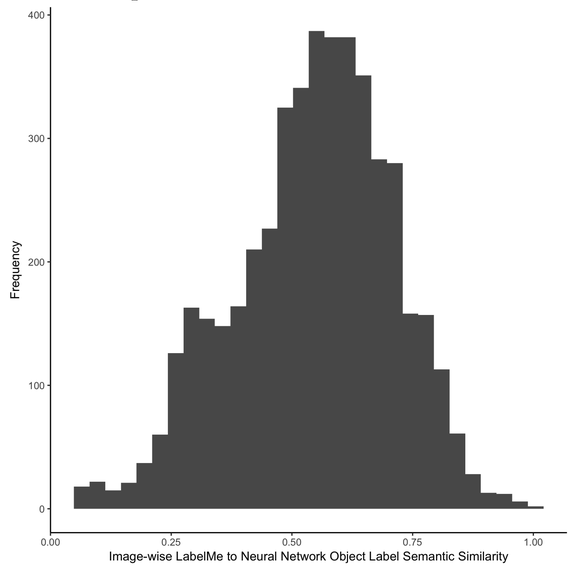
- 15 visually healthy, neuro-typical undergraduates
- Each viewed 210 randomly selected images from stimuli pool
- Gaze data: Eyelink 1000, 1000Hz binocular sampling
- Random task/trial across subjects
- Each trial: one of three tasks (free viewing, scene description, object counting)
- 5s viewing time, unlimited response time (scene description, object counting only)

Example Maps

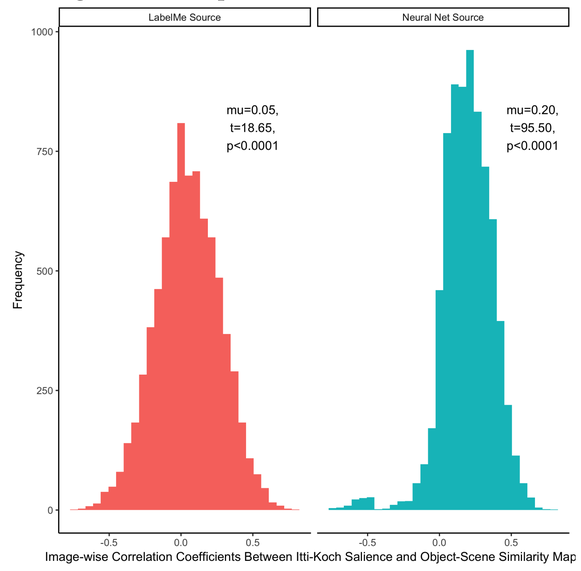


Results

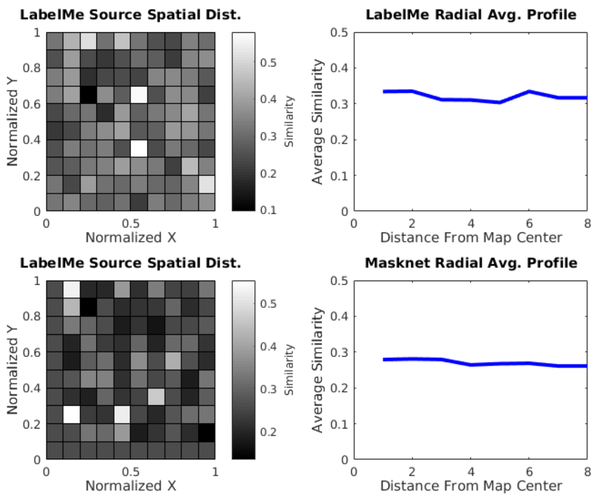
1. Highly Semantically Similar Scene Objects Across Map Data Source



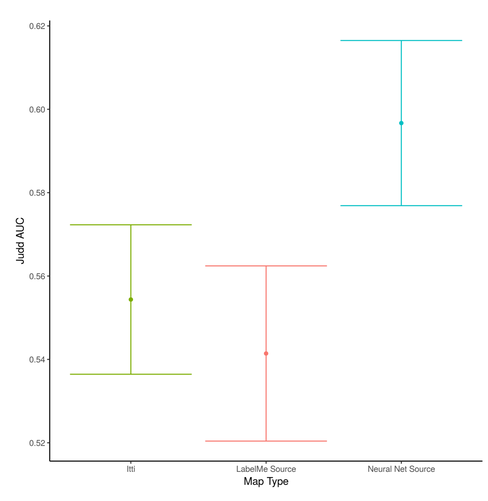
2. Semantic Maps Correlate With Image Salience Maps



3. Relatively Flat, Even Spatial Distributions of Similarity Across Map Data Source



4. Salience and Semantic Image Content Predict Gaze



Conclusions

1. Items labeled and identified by LabelMe users and a neural network are highly semantically similar
2. Semantic similarity maps have small but significant pixel-wise correlations with image saliency maps
   - Significance of t-test results for these correlations: inflated by the large numbers of images tested
   - Correlation between neural net generated semantic and image salience maps possibly because of increased dependence of neural net on image-level features to identify and segment objects.
3. Spatial distributions of semantically similar scene objects is relatively uniform and is consistent for both human and network labeled sources
4. Semantic relationship maps: predict gaze position as well as or better than image salience
   - The weak correlation between these estimates suggests that additional improvements in gaze prediction will be possible with combined salience and semantic maps.

Semantic relationship maps can be generated automatically, connect bottom-up and top-down image content, and can help accurately predict gaze behavior in natural scenes

References

[1] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, 2001.

[2] John M. Henderson and Taylor R. Hayes. Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, September 2017.

[3] Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.

[4] Alex D. Hwang, Hsueh-Cheng Wang, and Marc Pomplun. Semantic guidance of eye movements in real-world scenes. *Vision Research*, 51(10):1192–1205, May 2011.

[5] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. LabelMe a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, May 2008.