Quantifying Context Effects on Image Memorability

bigdata @CSAIL

Zoya Bylinskii, Phillip Isola, Antonio Torralba, Aude Oliva
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

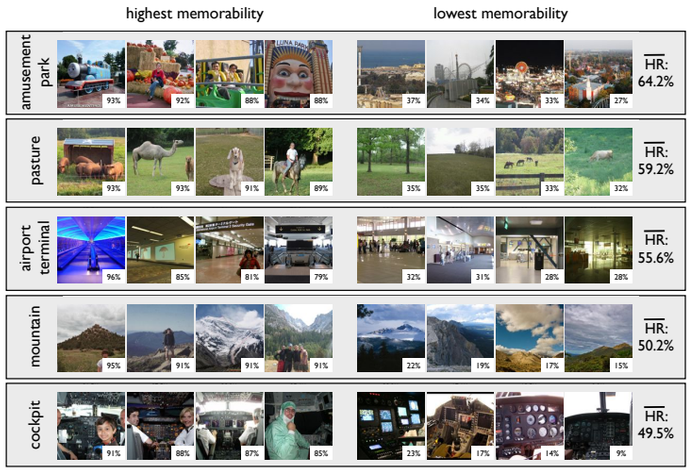
NSF NSERC CRSNG Google xerox

Intrinsic Effects on Memorability

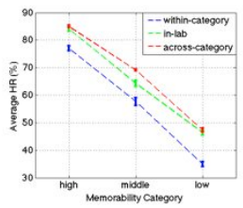
Crowd-sourced (AMT) memory (recognition) games

FIGRIM Dataset (Fine-Grained Image Memorability)

21 scene categories, more than 300 instances/category



Pairwise correlations
b/w experiments r > 0.60



amusement park	64.2%
playground	63.3%
bridge	61.2%
pasture	59.2%
bedroom	58.9%
house	58.0%
dining room	57.8%
conference room	57.1%
bathroom	57.1%
living room	57.0%
castle	56.4%
kitchen	56.3%
airport terminal	55.6%
badlands	52.9%
golf course	52.9%
skyscraper	52.8%
tower	52.8%
lighthouse	52.1%
mountain	50.2%
highway	50.0%
cockpit	50.0%

Memorability rank of images is consistent across participants and experiments
Spearman r = 0.69-0.86
(for each of 21 categories)

Memorability rank of categories is stable
Spearman r = 0.68
(across splits of images)

HR (hit rate): ratio of hits to total of hits and misses

* figures include scores for within-category experiment

Dataset: http://figrim.mit.edu

Paper: Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., Oliva, A. "Intrinsic and Extrinsic Effects on Image Memorability", Vision Research 2015.

Extrinsic Effects of Image Context

context: set of images from which experimental sequence is sampled

Information theoretic model for context

ran 21 separate experiments: one per scene category → within-category experiment

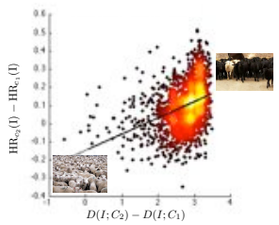
AND

an additional experiment, combining images from all categories → across-category experiment

$$D(I;C) = -\log P_c(f_i)$$

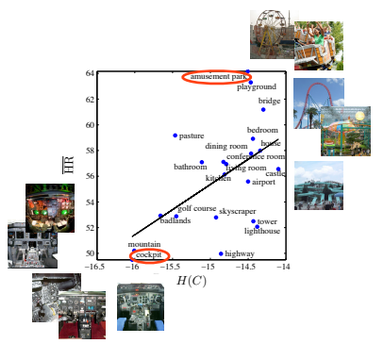
$$P_c(f_i) = \frac{1}{\|C\|}\sum_{j \in C} K(f_i - f_j)$$

Modeling is done with CNN (convolutional neural net) features which encode image semantics.

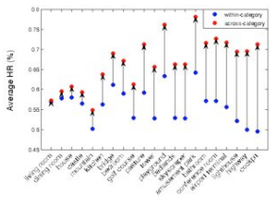


$$H(C) = \mathbb{E}_c[-\log P_c(f_i)]$$

Contextually distinct images are more memorable
Pearson r(D_1,HR_1) = 0.26
Pearson r(D_2,HR_2) = 0.24
Pearson r(ΔD,ΔHR) = 0.35



More varied image contexts are more memorable overall
Pearson r(H,HR) = 0.53
Pearson r(ΔH,ΔHR) = 0.74



Memorable within categories

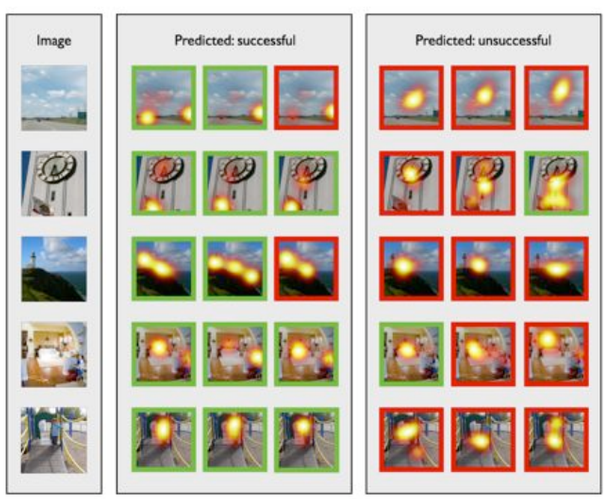
Memorable across categories

more likely to look like images from other categories

more likely to be memorable across different contexts

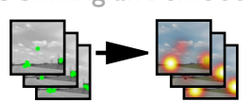
Extrinsic Effects of Individual Observer

We train a classifier to predict whether a set of eye movements will lead to a successful encoding

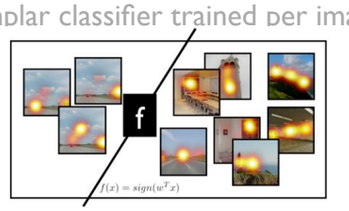


* we only use fixation position for prediction

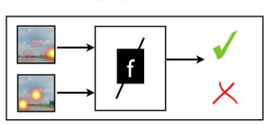
coarse binning and smoothing



exemplar classifier trained per image



test-time classification



images sorted by expected classifier confidence

Eye movements are predictive of whether an image will be remembered later