Towards cognitive saliency

Zoya Bylinskii

ECCV 2016 Tutorial: New directions in saliency research
Oct. 8, 2016
progress in saliency modeling

Bylinskii et al. "Where should saliency models look next?" [ECCV 2016]
progress in saliency modeling

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
where are saliency models getting their knowledge?
where are saliency models getting their knowledge?
where are saliency models getting their knowledge?
where are saliency models getting their knowledge?
object detectors emerge in CNNs

Zhou et al. “Object Detectors Emerge in Deep Scene CNNs” [ICLR 2015]
where are saliency models getting their knowledge?
where are saliency models getting their knowledge?

- reuse
- 100s-1000s data points
- retrain
- 100,000s of data points
where are saliency models getting their knowledge?

**Claim:** saliency models have become great object detectors!
quantifying remaining mispredictions

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
quantifying remaining mispredictions

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
quantifying remaining mispredictions

Describe the region
- Face
- Part of a face
- Head
- Person
- Part of a person
- Crowd of people (or faces)
- Legible text
- Illegible text
- Non-english text
- Symbol
- Animal face
- Part of an animal
- Object
- Background
- Other

(a)

Image 11 of 30

Are any of the people in the image looking at something inside the highlighted region?

Yes

(b)

Submit (17 images left)

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
1. [Object of gaze:] Are any of the people in the image looking at something inside the highlighted region?

2a. [Object of action:] Are any people in the image interacting with something inside the highlighted region?

2b. [Object of action:] Is there an object inside the highlighted region that people are using in some way?

3. [Unusual element:] Is there something unusual about what is inside the highlighted region?

4a. [Part of main subject:] Is the highlighted region part of the main subject of this photograph?

4b. [Part of main subject:] Is the highlighted region part of the photographer’s main focus?

5a. [Possible location for a person:] If this was a video, could a person appear in the highlighted region of the image in the next instant?

5b. [Possible location for a person:] Would you expect to find a person in the highlighted region of an image (even if there’s no one there now)?

6. [Possible location of action/motion:] If this was a video, could something move into the highlighted region of the image in the next instant?

7a. [Location of action/motion:] If this was a video, would there be an action or a motion happening inside the highlighted image?

7b. [Location of action/motion:] Is there an action happening inside the highlighted region?
### quantifying remaining mispredictions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MIT300</th>
<th>CAT2000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DeepFix</td>
<td>SALICON</td>
</tr>
<tr>
<td><strong>Image category</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part of main subject</td>
<td>31%</td>
<td>36%</td>
</tr>
<tr>
<td>Unusual element</td>
<td>18%</td>
<td>16%</td>
</tr>
<tr>
<td>Location of action/motion</td>
<td>16%</td>
<td>16%</td>
</tr>
<tr>
<td>Text</td>
<td>16%</td>
<td>13%</td>
</tr>
<tr>
<td>Part of a person</td>
<td>15%</td>
<td>14%</td>
</tr>
<tr>
<td>Possible location for a person</td>
<td>15%</td>
<td>7%</td>
</tr>
<tr>
<td>Object of action</td>
<td>14%</td>
<td>15%</td>
</tr>
<tr>
<td>Object of gaze</td>
<td>11%</td>
<td>11%</td>
</tr>
<tr>
<td>Part of a face</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td>Part of an animal</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td>40%</td>
<td>40%</td>
</tr>
</tbody>
</table>

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
<table>
<thead>
<tr>
<th>Dataset Model</th>
<th>MIT300</th>
<th>CAT2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image category</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part of main subject</td>
<td>31%</td>
<td>49%</td>
</tr>
<tr>
<td>Unusual element</td>
<td>36%</td>
<td>68%</td>
</tr>
<tr>
<td>Location of action/motion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part of a person</td>
<td>15%</td>
<td>14%</td>
</tr>
<tr>
<td>Possible location for a person</td>
<td>15%</td>
<td>14%</td>
</tr>
<tr>
<td>Object of action</td>
<td>14%</td>
<td>15%</td>
</tr>
<tr>
<td>Object of gaze</td>
<td>11%</td>
<td>11%</td>
</tr>
<tr>
<td>Part of a face</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td>Part of an animal</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>Other</td>
<td>40%</td>
<td>40%</td>
</tr>
</tbody>
</table>

Sample reading list:

- Itti & Baldi, “Bayesian surprise attracts human attention” [NIPS 2005]
### Sample reading list:

- Faivre & Koch, “Inferring the direction of implied motion depends on visual awareness” [JoV 2014]
- Winawer et al., “A motion aftereffect from still photographs depicting motion” [Psych Science 2008]
Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
### Sample reading list:

- Cerf et al., “Faces and text attract gaze independent of the task: Experimental data and computer model.” [JoV 2009]
- Vuilleumier, “Faces call for attention: evidence from patients with visual extinction” [Neuropsychologia 2000]
### Table: Quantifying Remaining Mispredictions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MIT300</th>
<th>CAT2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>DeepFix SALICON</td>
<td>DeepFix</td>
</tr>
<tr>
<td>Image category</td>
<td>All</td>
<td>Social Action Indoor Outdoor</td>
</tr>
<tr>
<td>Part of main subject</td>
<td>31% 36%</td>
<td>49% 68% 12% 24%</td>
</tr>
<tr>
<td>Unusual element</td>
<td>18% 16%</td>
<td>33% 63% 8% 8%</td>
</tr>
<tr>
<td>Location of action/motion</td>
<td>16% 16%</td>
<td>67% 78% 8% 11%</td>
</tr>
<tr>
<td>Text</td>
<td>16% 13%</td>
<td>6% 5% 8% 29%</td>
</tr>
<tr>
<td>Part of a person</td>
<td>15% 14%</td>
<td>23% 37% 8% 5%</td>
</tr>
<tr>
<td>Possible location for a person</td>
<td>15% 7%</td>
<td>6% 24% 10% 11%</td>
</tr>
<tr>
<td>Object of action</td>
<td>14% 15%</td>
<td>27% 51% 0% 3%</td>
</tr>
<tr>
<td>Object of gaze</td>
<td>11% 11%</td>
<td>50% 44% 0% 0%</td>
</tr>
<tr>
<td>Part of a face</td>
<td>6% 8%</td>
<td>46% 7% 0% 0%</td>
</tr>
<tr>
<td>Part of an animal</td>
<td>5% 5%</td>
<td>3% 10% 0% 0%</td>
</tr>
<tr>
<td>Other</td>
<td>40% 40%</td>
<td>3% 2% 61% 37%</td>
</tr>
</tbody>
</table>

### Additionally:

Bruce et al., “A deeper look at saliency: feature contrast, semantics, and beyond” [CVPR 2016]

Bruce et al., “On computational modeling of visual saliency: examining what's right and what's left” [VR 2015]
<table>
<thead>
<tr>
<th>Dataset</th>
<th>MIT300</th>
<th>CAT2000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DeepFix</td>
<td>SALICON</td>
</tr>
<tr>
<td><strong>Image category</strong></td>
<td><strong>Mit300</strong></td>
<td><strong>CAT2000</strong></td>
</tr>
<tr>
<td>Part of main subject</td>
<td>31%</td>
<td>36%</td>
</tr>
<tr>
<td>Unusual element</td>
<td>18%</td>
<td>16%</td>
</tr>
<tr>
<td>Location of action/motion</td>
<td>16%</td>
<td>16%</td>
</tr>
<tr>
<td>Text</td>
<td>16%</td>
<td>13%</td>
</tr>
<tr>
<td>Part of a person</td>
<td>15%</td>
<td>14%</td>
</tr>
<tr>
<td>Possible location for a person</td>
<td>15%</td>
<td>7%</td>
</tr>
<tr>
<td>Object of action</td>
<td>14%</td>
<td>15%</td>
</tr>
<tr>
<td>Object of gaze</td>
<td>11%</td>
<td>11%</td>
</tr>
<tr>
<td>Part of a face</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td>Part of an animal</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td>40%</td>
<td>40%</td>
</tr>
</tbody>
</table>
### Quantifying Remaining Mispredictions

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MIT300</th>
<th>CAT2000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DeepFix</td>
<td>SALICON</td>
</tr>
<tr>
<td><strong>Image category</strong></td>
<td><strong>All</strong></td>
<td><strong>Social</strong></td>
</tr>
<tr>
<td>Part of main subject</td>
<td>31%</td>
<td>36%</td>
</tr>
<tr>
<td>Unusual element</td>
<td>18%</td>
<td>16%</td>
</tr>
<tr>
<td>Location of action/motion</td>
<td>16%</td>
<td>16%</td>
</tr>
<tr>
<td>Text</td>
<td>16%</td>
<td>13%</td>
</tr>
<tr>
<td>Part of a person</td>
<td>15%</td>
<td>14%</td>
</tr>
<tr>
<td>Possible location for a person</td>
<td>15%</td>
<td>7%</td>
</tr>
<tr>
<td>Object of action</td>
<td>14%</td>
<td>15%</td>
</tr>
<tr>
<td>Object of gaze</td>
<td>11%</td>
<td>11%</td>
</tr>
<tr>
<td>Part of a face</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td>Part of an animal</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td>40%</td>
<td>40%</td>
</tr>
</tbody>
</table>
So where should saliency models look next?
<table>
<thead>
<tr>
<th>Input</th>
<th>Ground truth</th>
<th>DeepFix</th>
<th>SALICON</th>
</tr>
</thead>
</table>

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
Recasens et al. “Where are they looking?” [NIPS 2015]
Parks et al. “Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes” [Vision Research 2015]

Recasens et al. “Where are they looking?” [NIPS 2015]
objects of action

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
<table>
<thead>
<tr>
<th>Input</th>
<th>Ground truth</th>
<th>DeepFix</th>
<th>SALICON</th>
</tr>
</thead>
</table>

saliency as storytelling
faces over-predicted

faces under-predicted

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
relative importance of people

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
relative importance of people

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
saliency in a crowd
relative importance of text regions

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
**claim:** we need better metrics for evaluating how well models predict the *relative importance* of image content.
Bylinskii et al. “What do different evaluation metrics tell us about saliency models?” [arXiv 2016]
need for finer-grained evaluations

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
need for finer-grained evaluations

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
relative importance of image regions

Input: image panel

Ground truth

DeepFix

SALICON

Deep Gaze 2

BMS

low

saliency values

high

Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
finger-grained datasets


Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
finger-grained datasets


Bylinskii et al. “Where should saliency models look next?” [ECCV 2016]
What can we do with predicted importance?

**Claim:** image captioning, image understanding, storytelling
attention for image understanding
a girl kissing another girl on the cheek

a warning sign about rip currents
a girl kissing another girl on the cheek

a warning sign about rip currents
a girl kissing another girl on the cheek

a warning sign about rip currents

a girl

a warning sign with a flag
the relationship of object importance to image description

Spain & Perona. “Measuring and predicting object importance” [IJCV 2010]
Spain & Perona. “Some objects are more equal than others: measuring and predicting importance” [ECCV 2008]
the relationship of gaze to object importance

- Objects described by human: 
  - motorbike, person
- Baseline detected objects: 
  - motorbike
- Gaze-enabled detected objects: 
  - motorbike, person

- Objects described by human: 
  - cat, table
- Baseline detected objects: 
  - table
- Gaze-enabled detected objects: 
  - cat, table

- Objects described by human: 
  - person, horse
- Baseline detected objects: 
  - horse
- Gaze-enabled detected objects: 
  - horse, person

- Objects described by human: 
  - cow
- Baseline detected objects: 
  - cow
- Gaze-enabled detected objects: 
  - cow, person

- Objects described by human: 
  - mbike
- Baseline detected objects: 
  - mbike
- Gaze-enabled detected objects: 
  - mbike, person, plant

Yun et al. “Studying Relationships Between Human Gaze, Description, and Computer Vision” [CVPR 2013]
CNN-based feature extraction

RNN-based feature weighting (attention heatmaps)
and caption generation


soft attention

hard attention

A bird flying over a body of water
Figure 5. Examples of mistakes where we can use attention to gain intuition into what the model saw.

A large white bird standing in a forest. A woman holding a clock in her hand. A man wearing a hat and a hat on a skateboard.

A person is standing on a beach with a surfboard. A woman is sitting at a table with a large pizza. A man is talking on his cell phone while another man watches.
CNN-based feature extraction

 question: what are sitting in the bask on a bicycle?

answer: dogs

Yang et al. “Stacked Attention Networks for Image Question Answering” [ArXiv 2016]
object-specific saliency maps
for object proposals, localization and detection


Hong et al. “Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network” [ICML 2015]
object-specific saliency maps
for object proposals, localization and detection

network trained for classification, detection, ...

saliency emerges


Hong et al. “Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network” [ICML 2015]
consider instead:

network trained for saliency classification, detection emerges
consider instead:

network trained for saliency

visual recognition emerges
New data collection methodologies

Webcam-based

Krafka et al. [CVPR 2016]
Xu et al. [ArXiv 2015]
Papoutsaki et al. [IJCAI 2016]

Click-based

Jiang et al. [CVPR 2015]
Kim et al. [CHI EA 2015]

iSUN

SALICON
architectures

saliency datasets

saliency loss
What is the future of cognitive saliency?