Evaluating machine comprehension of sketch meaning at different levels of abstraction

Kushin Mukherjee1, Xuanchen Lu2, Holly Huey2, Yael Vinker3, Rio Aguina-Kang2, Ariel Shamir4, Judith E. Fan2,5
1University of Wisconsin-Madison, 2University of California San Diego, 3Tel-Aviv University, 4Reichman University, 5Stanford University

**QUESTION**

How well do vision models exhibit human-like understanding of sketches that vary in semantic ambiguity?

**METHODS**

1. Vision Model Selection

<table>
<thead>
<tr>
<th>supervised</th>
<th>self-supervised</th>
<th>semi-supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception-V3</td>
<td>VGG-19</td>
<td>Noise Student</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>ECOSET</td>
<td>CLIP</td>
</tr>
<tr>
<td>CORNet-S</td>
<td>ViT-B</td>
<td>CLIPasso</td>
</tr>
<tr>
<td>Harmonization</td>
<td>MAE</td>
<td>SimCLR</td>
</tr>
<tr>
<td>MLP-Mixer</td>
<td>Noisy Student</td>
<td>IPCL</td>
</tr>
</tbody>
</table>

2. Sketch Dataset Generation

2048 photos from 128 classes from the THINGS database.

- N = 5,563
- Over 90K sketches including sketches of each concept at 4 levels of detail made by humans and CLIPasso.

3. Measuring Sketch Comprehension

- What does this sketch represent?
- Performance: Lions ✓, Cats ×, Taxis ×, Dogs ×
- Alignment with human behavior

**RESULTS**

Humans produce sparser sketches under stronger time constraints.

<table>
<thead>
<tr>
<th>Drawing Time (s)</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 s</td>
<td>0.12</td>
</tr>
<tr>
<td>8 s</td>
<td>0.08</td>
</tr>
<tr>
<td>16 s</td>
<td>0.04</td>
</tr>
<tr>
<td>32 s</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Sparser sketches more semantically ambiguous for models and humans.

- Sketches that are more detailed elicit...
  - ...higher classification accuracy.
  - ...less variable response labels.
- ...guesses that are semantically close to the true label when incorrect.

Models vary in their degree of alignment to human behavior but a large gap remains between human and model sketch understanding.

- Top-1 classification accuracy
- Response entropy
- Semantic neighbor preference

A CLIP-based sketch generation algorithm emulates human sketches at greater levels of detail.

- Human recognition performance on human-made and CLIPasso-made sketches is comparable.
- Labels elicited by humans for human and CLIPasso sketches are more similar for more detailed sketches.

**TAKEAWAYS**

- We introduce a new dataset of >90K sketches at varied abstraction levels made by humans & CLIPasso, an AI-sketch generation model.
- State-of-the-art vision models are sensitive to variation in the semantic information conveyed by sketches under different production budgets.
- A large alignment gap still remains between the most performant vision models and humans.
- CLIPasso-generated and human-made sketches elicit similar responses at greater levels of detail.

**ACKNOWLEDGEMENTS**

- Data & materials available at: https://github.com/cogtoolslab/visual-abstractions-benchmarking_public2023
- Correspondence to: kmukherje2@wisc.edu

- Human model alignment (R2)
- Human recognition performance on human-made and CLIPasso-made sketches is comparable.
- Labels elicited by humans for human and CLIPasso sketches are more similar for more detailed sketches.