Towards Robust and Socially-Adept Autonomous Vehicles Through Vehicle Trajectory Sensing for Safety Assessment

‘YZ’ Yezhou Yang, Assistant Professor, SCAI, ASU
Group Lead, Active Perception Group,
School of Computing and AI, Arizona State University
Projects Tech co-Lead, The Institute of Automated Mobility (IAM)

@Yezhou_Yang

Sep 30th 2021 @ ITS AZ
56%: would not ride in an autonomous vehicle.
16%: feel safe to let an autonomous vehicle drive them without the option of taking control.
16%: feel autonomous cars will eventually eliminate the need for car insurance.

https://www.erieinsurance.com/blog/multi-gen-car-survey
So, what are these gaps?
And, how are we going to fill (or attempt to fill) these gaps (or a few of them) from AI/CV perspectives?

1) The signal to semantic gap $\leftarrow$ Visual Recognition with Knowledge

2) From lab to the society gap $\leftarrow$ Socially adept autonomous driving

3) The equipment gap $\leftarrow$ AV Performance Evaluation with Existing Traffic Cameras;

4) From tech to transportation practitioner gap $\leftarrow$ ARGOS Vision.
1) The signal to semantic gap ← Visual Recognition with Knowledge

Visual Question Answering

Q: how many people are waiting for bus?
A: Two? or Three?
Visual Recognition as Pattern Matching:

“Visual recognition is a cognitive process that involves identification of a visible CATEGORY from previous encounters.”

Visual Recognition as it is:

“Visual recognition is a cognitive process that involves identification of a visible CONCEPT from previous encounters or KNOWLEDGE.”

What is a concept?

“… A theory of concepts should describe the kind of knowledge stored in concepts, the way they are used in agents’ cognitive processes, their format, their acquisition, and their neural localization…”

Categories ≠ Concepts
1) The signal to semantic gap: the representation gap

<table>
<thead>
<tr>
<th>Experiment</th>
<th>BRNN-Karpathy</th>
<th>Our Method</th>
<th>Gold Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>R ± D(8k)</td>
<td>2.08 ± 1.35</td>
<td>2.82 ± 1.56</td>
<td>4.69 ± 0.78</td>
</tr>
<tr>
<td>T ± D(8k)</td>
<td>2.24 ± 1.33</td>
<td>2.62 ± 1.42</td>
<td>4.32 ± 0.99</td>
</tr>
<tr>
<td>R ± D(30k)</td>
<td>1.93 ± 1.32</td>
<td>2.43 ± 1.42</td>
<td>4.78 ± 0.61</td>
</tr>
<tr>
<td>T ± D(30k)</td>
<td>2.17 ± 1.34</td>
<td>2.49 ± 1.42</td>
<td>4.52 ± 0.93</td>
</tr>
<tr>
<td>R±D(COCO)</td>
<td>2.69 ± 1.49</td>
<td>2.14 ± 1.29</td>
<td>4.71 ± 0.67</td>
</tr>
<tr>
<td>T±D(COCO)</td>
<td>2.55 ± 1.41</td>
<td>2.06 ± 1.24</td>
<td>4.37 ± 0.92</td>
</tr>
</tbody>
</table>

Table 1: Sentence generation relevance (R) and thoroughness (T) human evaluation results with gold standard and BRNN-Karpathy on Flickr 8k, 30k and MS-COCO datasets. D: Standard Deviation.
1) The signal to semantic gap: the explicit reasoning for interpretation
1) The signal to semantic gap: the fundamental logic-based reasoning

NEGATION

Intelligence?!

Never mind...
1) The signal to semantic gap: the fundamental logic-based reasoning

VQA-LOL: Visual Question Answering under the Lens of Logic

<table>
<thead>
<tr>
<th>Image</th>
<th>Question</th>
<th>Predicted Answer</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q_1$: Is there beer?</td>
<td>YES</td>
<td>(0.96)</td>
</tr>
<tr>
<td></td>
<td>$Q_2$: Is the man wearing shoes?</td>
<td>NO</td>
<td>(0.90)</td>
</tr>
</tbody>
</table>

VQA-Compose

<table>
<thead>
<tr>
<th>Question</th>
<th>Predicted Answer</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\neg Q_2$: Is the man not wearing shoes?</td>
<td>NO</td>
<td>(0.80)</td>
</tr>
<tr>
<td>$\neg Q_2 \land Q_1$: Is the man not wearing shoes and is there beer?</td>
<td>NO</td>
<td>(0.62)</td>
</tr>
<tr>
<td>$Q_1 \land C$: Is there beer and does this seem like a man bending over to look inside of a fridge?</td>
<td>NO</td>
<td>(1.00)</td>
</tr>
</tbody>
</table>

VQA-Supplement

<table>
<thead>
<tr>
<th>Question</th>
<th>Predicted Answer</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\neg Q_2 \lor B$: Is the man not wearing shoes or is there a clock?</td>
<td>NO</td>
<td>(1.00)</td>
</tr>
<tr>
<td>$Q_1 \land \text{anto}(B)$: Is there beer and is there a wine glass?</td>
<td>YES</td>
<td>(0.84)</td>
</tr>
</tbody>
</table>

ECCV 20'

VQA-LOL
1) The signal to semantic gap: out-of-domain (OOD) generalization

Visual Question Answering
Q: Is it a fast vehicle?
A: Yes

Visual Question Answering
Q: Is it a fast vehicle?
A: No
1) The signal to semantic gap: out-of-domain (OOD) generalization

A: Green

A: I think it is still green?...

Intelligence?!

Never mind...
1) The signal to semantic gap: out-of-domain (OOD) generalization

Mutant: A training paradigm for out-of-distribution generalization in visual question answering

What is the color of the frisbee?
### Analysis: Effect of Mutant Samples

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>VQA-CP v2 test ↑ (%)</th>
<th>Increase in Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>Yes/No</td>
</tr>
<tr>
<td>UpDn</td>
<td>VQA-CP</td>
<td>39.74</td>
<td>42.27</td>
</tr>
<tr>
<td>UpDn</td>
<td>VQA-CP + Mutant</td>
<td>50.16</td>
<td>61.45</td>
</tr>
<tr>
<td>LXMERT</td>
<td>VQA-CP</td>
<td>46.23</td>
<td>42.84</td>
</tr>
<tr>
<td>LXMERT</td>
<td>VQA-CP + Mutant</td>
<td>59.69</td>
<td>73.19</td>
</tr>
<tr>
<td>LXM + Ours</td>
<td>VQA-CP + Img. Mut.</td>
<td>64.85</td>
<td>85.68</td>
</tr>
<tr>
<td>LXM + Ours</td>
<td>VQA-CP + Que. Mut.</td>
<td>67.92</td>
<td>91.64</td>
</tr>
<tr>
<td>LXM + Ours</td>
<td>VQA-CP + Both Mut.</td>
<td>69.52</td>
<td>93.15</td>
</tr>
</tbody>
</table>

- **Comparison of Backbone models (UpDn, LXMERT) trained with VQA-CP data augmented with MUTANT samples.**

- **Comparison of our best model when trained with: image mutations, question mutations, and both types of mutations.**
1) The signal to semantic gap: perceiving beyond appearance

EMNLP 11’
Sen. Gen. from Img, Captioning

ACS 16’
DeepLU
Scene
Description
Graph (SDG)

CVIU 17’
Image Under.
w/ SDG

V2C: Video to Commonsense

https://asu-active-perception-group.github.io/Video2Commonsense/index.html
2) From lab to the society gap ← Socially adept autonomous driving

How do we define a good driver?

[1. They drive predictably](https://jalopnik.com/how-to-recognize-a-good-driver-5947854)

10. They move over after passing.

7. They are not overly polite at intersections.

6. They can park.

5. They use their turn signals.

3. They make confident lane changes.

How to create a driver that is naturally good? How to evaluate whether a driver is naturally good?

Irrationally courteous AV:

AV recognizes that its best action from the driver’s perspective is to wait. Thus it waits...
“Self-driving cars need to be nice, but not overly nice”: Simply behave to satisfy others does not make a good driver.

Solution: **Rational courtesy** (through recognition of Nash Equilibria)

**Brief explanation of our method**

Method: Find hypothesis that matches with the reality. Update belief using Bayesian update.

Challenge: Hypotheses change all the time. (Don’t have enough time to update them in reality).

Solution: Create intuition through experience (Offline fictitious self-play)
2) From lab to the society gap ← Socially adept autonomous driving

Irrationally courteous AV:

AV recognizes that its best action from the driver’s perspective is to wait. Thus it waits...

Rationally courteous AV:

AV recognizes that from the driver’s perspective, its best action among all Nash Equilibria is to leave as soon as possible.
• Scalability and other natural driving scenarios
• Human modeling and prediction
• Safety metrics & corner cases via human experiments

https://interaction-dataset.com/
3) The equipment gap ↔ AV Performance Monitoring with Existing Traffic Cameras;

Boss (CMU, DARPA Grand Challenge, 2002-2007)
CAROM - **CARs On the Map**

- **CAROM** is a framework to track and localize vehicles using monocular traffic monitoring cameras on road infrastructures.

- The vehicle localization results are stored in files or in a database as records.

- Using the results, a traffic scene can be reconstructed and replayed on a map.
Vehicle Tracking

- Original image
- Instance segmentation mask
- Optical flow
- Camera parameters
- 3D bounding box
- Vehicle localization in 3D
We drove a test vehicle with differential GPS in the first site for evaluation.
Evaluation - GPS
# Evaluation - Drone

<table>
<thead>
<tr>
<th>Video</th>
<th>L-Diff (m)</th>
<th>V-Diff (m/s)</th>
<th>#Vehicles (w/ Ref)</th>
<th>Coverage (m)</th>
<th>Ref Device</th>
</tr>
</thead>
<tbody>
<tr>
<td>Track 1A</td>
<td>2.05</td>
<td>1.01</td>
<td>1</td>
<td>25 ~120</td>
<td>GPS</td>
</tr>
<tr>
<td>Track 1B</td>
<td>1.57</td>
<td>0.69</td>
<td>1</td>
<td>25 ~120</td>
<td>GPS</td>
</tr>
<tr>
<td>Track 2</td>
<td>1.68</td>
<td>1.47</td>
<td>69</td>
<td>15 ~110</td>
<td>Drone</td>
</tr>
</tbody>
</table>
Applications

Minimum Safety Distance Violation

Traffic Rule Violation
AvaCAR: avatar of vehicles
Pick-up Truck
CAROM - Vehicle Localization and Traffic Scene Reconstruction from Monocular Cameras on Road Infrastructures

Demo Video Submission

Duo Lu¹, Varun C Jammula¹, Steven Como¹, Jeffrey Wishart², Yan Chen¹, Yezhou Yang¹

¹{duolu, vjammula, scomo, yanchen, yz.yang}@asu.edu
²jwishart@exponent.com
ARGOS project provides a full stack software + hardware intelligent camera solution with on-board CV/AI processing that performs semantic-level understanding of the environment and generate a vast amount of privacy-preserved, real-time, semantic DATA.
Thank you and Acknowledgements

NSF CAREER 18’ VR-K
2 NSF RI SMALLs
NSF NRI
NSF CPS
NSF SaTC
NSF CCRI (planning)
NSF I-Corps
DARPA KAIROS
LESTAT project
And
GAILA ADAM-E
ONR Social Interaction
DARPA KAIROS
LESTAT project
And
GAILA ADAM-E
ONR Social Interaction

and ASU close collaborating groups (C. Baral [KR & NLP], M. Ren/W. Zhang [Optimization & ML & Control], IAM collaborators: Jeff Wishart, Duo Lv, Mohammad Farhadi, Maria Elli, Yan Chen, Larry Head, Greg Leeming, Prabal Dutta, Rahul Varma and many more).
Public/Business/Academia to APG: yz.yang@asu.edu

@Yezhou_Yang

Public/Business/Academia to ARGOS: argos.vision.co@gmail.com
www.argos.vision

Check out our live demo @ ITS AZ!