NAACL HLT 2021

Visually Grounded Interaction and Language (ViGIL)

Proceedings of the Fourth Workshop

June 10, 2021

FACEBOOK AI

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-954085-27-5

Introduction

Welcome to the Fourth Workshop on Visually Grounded Interaction and Language (ViGIL).

Language is neither learned nor used in a vacuum, but rather grounded within a rich, embodied experience rife with physical groundings (vision, audition, touch) and social influences (pragmatic reasoning about interlocutors, commonsense reasoning, learning from interaction). For example, studies of language acquisition in children show a strong interdependence between perception, motor control, and language understanding. Yet, AI research has traditionally carved out individual components of this multimodal puzzle—perception (computer vision, audio processing, haptics), interaction with the world or other agents (robotics, reinforcement learning), and natural language processing—rather than adopting an interdisciplinary approach.

This fractured lens makes it difficult to address key language understanding problems that future agents will face in the wild. For example, describing "a bird perched on the lowest branch singing in a high pitch trill" requires grounding to perception. Likewise, providing the instruction to "move the jack to the left so it pushes on the frame of the car" requires not only perceptual grounding, but also physical understanding. For these reasons, language, perception, and interaction should be learned and bootstrapped together. In the last several years, efforts to merge subsets of these areas have gained popularity through tasks like instruction-guided navigation in 3D environments, audio-visual navigation, video descriptions, question-answering, and language-conditioned robotic control, though these primarily study disembodied problems via static datasets. As such, there remains considerable scientific uncertainty around how to bridge the gap from current monolithic systems to holistic agents. What are the tasks? The environments? How to design and train such models? To transfer knowledge between modalities? To perform multimodal reasoning? To deploy language agents in the wild?

As in past incarnations, **the goal of this 4th ViGIL workshop is to support and promote this research direction by bringing together scientists from diverse backgrounds—natural language processing, machine learning, computer vision, robotics, neuroscience, cognitive science, psychology, and philosophy—to share their perspectives on language grounding, embodiment, and interaction**. ViGIL provides a unique opportunity for interdisciplinary discussion. We intend to utilize this variety of perspectives to foster new ideas about how to define, evaluate, learn, and leverage language grounding. This one-day session would enable in-depth conversations on understanding the boundaries of current work and establishing promising avenues for future work, with the overall aim to bridge the scientific fields of human cognition and machine learning.

This year, ViGIL will be co-located with the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). We accepted twenty-seven non-archival papers to be presented at our workshop, with topics including instruction following, image captioning, emergent communication, interactive learning, and semantic parsing, among others. The workshop features eight invited speakers with a diverse set of perspectives on language grounding, with research focuses including cognitive science, robotics, computer vision, psycholinguistics, and core natural language processing.

Invited Speakers

Sandra Waxman (Professor, Department of Psychology, Northwestern University) focuses on infant language acquisition and development of concepts and language, and the relation between the two.

Trevor Darrell (Professor, Electrical Engineering and Computer Sciences, UC Berkeley) focuses on computer vision, language, machine learning, graphics, and perception-based human computer interfaces.

Max Garagnani (Lecturer, Department of Computing, University of London) focuses on the implementation of biologically realistic neural-network in language, memory and visual perception.

Roger Levy (Associate Professor, Department of Brain and Cognitive Science, Massachusetts Institute of Technology) focuses on understanding the cognitive underpinning of natural language processing and acquisition.

Yejin Choi (Brett Helsel Associate Professor, Paul G. Allen School of Computer Science and Engineering, University of Washington; Allen Institute for Artificial Intelligence) works at the intersection of natural language and machine learning, with interests in computer vision and digital humanities.

Stefanie Tellex (Associate Professor, Department of Computer Science, Brown University) focuses on constructing robots that seamlessly use natural language to communicate with humans.

Katerina Fragkiadaki (Assistant Professor, Department of Machine Learning, Carnegie Mellon University) explores building machines that understand the stories that videos portray and, using videos to teach machines about the world.

Justin Johnson (Assistant Professor, Department of Electrical Engineering and Computer Science, University of Michigan; Visiting Researcher at Facebook AI Research) focuses on visual reasoning, vision and language, image generation, and 3D reasoning using deep neural networks.

Organizing Committee
Cătălina Cangea, University of Cambridge
Abhishek Das, Facebook AI Research
Drew Hudson, Stanford University
Jacob Krantz, Oregon State University
Stefan Lee, Oregon State University
Jiayuan Mao, Massachusetts Institute of Technology
Florian Strub, DeepMind
Alane Suhr, Cornell University
Erik Wijmans, Georgia Tech

Scientific Committee
Aaron Courville, University of Montreal
Mateusz Malinowski, DeepMind
Olivier Pietquin, Google Brain
Harm de Vries, University of Montreal and Element AI

Program Committee
Adria Recasens, DeepMind
Anna Potapenko, DeepMind
Arjun Majumdar, Georgia Tech
Catherine Wong, Massachusetts Institute of Technology
Christopher Davis, University of Cambridge
Daniel Fried, UC Berkeley
Gabriel Ilharco, University of Washington
Geoffrey Cideron, InstaDeep
Hammad Ayyubi, Columbia University
Hao Tan, University of North Carolina Chapel Hill
Hao Wu, Fudan University
Haoyue Shi, Toyota Technological Institute at Chicago
Hedi Ben-younes, Sorbonne Université
Jack Hessel, Allen Institute for Artificial Intelligence
Jean-Baptiste Alayrac, DeepMind
Joel Ye, Georgia Tech
Johan Ferret, Google Brain
Karan Desai, University of Michigan
Lisa Anne Hendricks, DeepMind
Luca Celotti, Université de Sherbrooke
Mathieu Rita, École Polytechnique
Mathieu Seurin, University of Lille
Meera Hahn, Georgia Institute of Technology
Nicholas Tomlin, UC Berkeley
Olivier Pietquin, Google Brain
Rodolfo Corona, UC Berkeley
Rowan Zellers, University of Washington
Ryan Benmalek, Cornell University
Sanjay Subramanian, Allen Institute for Artificial Intelligence
Sidd Karamcheti, Stanford University
Valts Blukis, Cornell University

Conference Program (all times in EDT)

Thursday, June 10, 2021

8:50–9:00 *Opening Remarks*
ViGIL Organizers

9:00–9:45 *Invited Talk*
Roger Levy

9:45–10:30 *Invited Talk*
Stefanie Tellex

10:30–11:00 ***Break***

11:00–11:45 *Invited Talk*
Katerina Fragkiadaki

11:45–12:30 *Invited Talk*
Max Garagnani

12:30–13:00 ***Break***

13:00–14:00 ***Panel Discussion***

14:00–14:30 ***Break***

14:30–15:15 *Invited Talk*
Yejin Choi

15:15–16:00 *Invited Talk*
Justin Johnson

16:00–16:20 *Results of the 2nd GQA Challenge*
Drew Hudson

16:20–16:30 ***Spotlight Presentations***

Thursday, June 10, 2021 (continued)

16:30–18:00 ***Poster Session***

18:00–18:45 *Invited Talk*
Trevor Darrell

18:45–19:30 *Invited Talk*
Sandra Waxman

19:30–19:40 ***Closing Remarks***