The How2 Challenge

New Tasks for Vision and Language

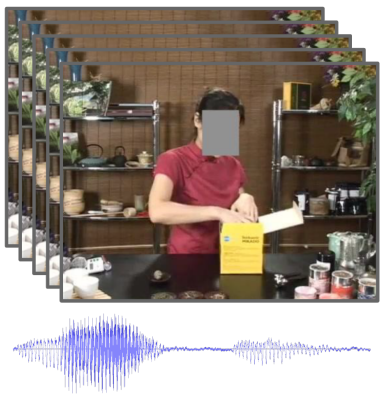
ICML 2019 Workshop, Long Beach, California

Research at the intersection of vision and language has attracted an increasing amount of attention over the last ten years. Current topics include the study of multi-modal representations, translation between modalities, bootstrapping of labels from one modality into another, visually-grounded question answering, embodied question-answering, segmentation and storytelling, and grounding the meaning of language in visual data. Still, these tasks may not be sufficient to fully exploit the potential of vision and language data.

To support research in this area, we recently released the How2 data-set, containing 2000 hours of how-to instructional videos, with audio, subtitles, Brazilian Portuguese translations, and textual summaries, making it an ideal resource to bring together researchers working on different aspects of multimodal learning. We hope that a common dataset will facilitate comparisons of tools and algorithms, and foster collaboration.

We are organizing a workshop at ICML 2019, to bring together researchers and foster the exchange of ideas in this area. We invite papers that participate in the How2 Challenge and also those that use other public or private multimodal datasets.

For any questions, contact us at how2challenge@gmail.com.



When you are selecting a teapot that is glass, it is always best to go with the best brand because glass already is very fragile.

Quando você está escolhendo um bule que é de vidro, é sempre mais eficiente escolher a melhor marca porque vidro já é muito frágil.

Glass teapots are useful for brightly colored and flowered teas. Learn to use a glass teapot with tips from a tea lounge owner in this free tea brewing video.

How2 contains a large variety of instructional videos with utterance-level English subtitles (in bold), aligned Portuguese translations (in italics), and video-level English summaries (in the box). Multimodality helps resolve ambiguities and improves understanding.

Get the How2 dataset and start exploring!

Paper | **Dataset** | **Learn more**

Important Dates

Please note these dates are tentative and may be changed. Refer to this website for more information.

Challenge starts: **March 15 2019**

Deadline for paper submission: **May 15 2019** (11:59pm Pacific Standard Time)

Author Notification: **May 22 2019**

Workshop: **June 14 or 15 2019**

Call For Papers

We seek submissions in the following two categories:

- Papers that describe work on the How2 data, either on the shared challenge tasks, e.g. multi-modal speech recognition (Palaskar et al. ICASSP 2018, Caglayan et al. ICASSP 2019), machine translation (Shared task on Multimodal MT), or video summarization (Libovicky et al. ViGIL, NeurIPS 2018), or creating “un-shared”, novel tasks that create language, speech and/or vision.

- Examples of novel tasks could be spoken language translation, cross-modal multimodal learning, unsupervised representation learning (Holzenberger et al. ICASSP 2019), reasoning in vision and language, visual synthesis from language, vision and language interaction for humans, learning from in-the-wild videos (How2 data or others), lip reading, audio-visual scene understanding, sound localization, multimodal fusion, visual question answering, and many more.

- Papers that describe other related and relevant work to further vision and language ideas by proposing new tasks, or analyzing the utility of existing tasks and data sets in interesting ways

We encourage both the publication of novel work that is relevant to the topics of discussion, and late-breaking results on the How2 tasks in a single format. We aim to stimulate discussion around new tasks that go beyond image captioning and visual question answering, and which could form the basis for future research in this area.

Challenge Leaderboard

The How2 Challenge has three tasks: Speech Recognition, Machine Translation, and Summarization. For more information on baseline code, evaluation and submission, visit https://github.com/srvk/how2-challenge/wiki

You could email the results files to us at how2challenge@gmail.com or submit your evaluated files through the Google Forms below. See more at https://github.com/srvk/how2-challenge/wiki

---

Planned Activities (Tentative)

- A series of presentations on the How2 Challenge Tasks, based on invited papers
- Posters from How2 challenge participants to encourage in-depth discussion
- Invited speakers to present other viewpoints and complementary work
- A moderated round table discussion to develop future directions

Schedule

Coming soon!

Invited Speakers

Katerina Fragkiadaki is an Assistant Professor in the Machine Learning Department at Carnegie Mellon. She is interested in building machines that understand the stories that videos portray, and, inversely, in using videos to teach machines about the world.

Lisa Anne Hendricks is a graduate student researcher at UC Berkeley studying computer vision with Professor Trevor Darrell. Her PhD work has focussed on building systems which can express information about visual content using natural language and retrieve visual information given natural language.

Qin Jin is an associate professor in School of Information at Renmin University of China where she leads the multimedia content analysis research group. She received her Ph.D. degree in Language and Information Technologies from Carnegie Mellon University and had research experiences in both academia and industry.

Angeliki Lazaridou is a research scientist at DeepMind. Her primary research interests are in the area of natural language processing (NLP), and specifically, in multimodal semantics.

Devi Parikh is an Assistant Professor in the School of Interactive Computing at Georgia Tech, and a Research Scientist at Facebook AI Research (FAIR). Her research interests include computer vision and AI in general and visual recognition problems in particular.

Kate Saenko is an Associate Professor at the Department of Computer Science at Boston University, and the director of the Computer Vision and Learning Group and member of the IVC Group. Her research interests are in the broad area of Artificial Intelligence with a focus on Adaptive Machine Learning, Learning for Vision and Language Understanding, and Deep Learning.

Bernt Schiele is a Max Planck Director at MPI for Informatics and Professor at Saarland University since 2010, both in Saarbrucken, Germany. Before that he had academic positions at ETH Zurich, TU Darmstadt, MIT, INP Grenoble and CMU. His current research interests cover a wide range of topics in computer vision, including how to leverage language for visual scene understanding and how to leverage vision for language understanding.

Organizers


Florian Metze


Lucia Specia


Desmond Elliott


Loïc Barrault


Ramon Sanabria


Shruti Palaskar

For any questions, contact us at how2challenge@gmail.com.

References

1. Palaskar et al. "End-to-End Multimodal Speech Recognition", ICASSP 2018
2. Caglayan et al. "Multimodal Grounding for Sequence-to-Sequence Speech Recognition", ICASSP 2019
3. Elliott et al. "Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description", WMT 2017
4. Shared Task on Multimodal Machine Translation, Workshop on Machine Translation [Webpage]
5. Libovicky et al. "Multimodal Abstractive Summarization for Open-Domain Videos", ViGIL Workshop, NeurIPS 2018
6. Holzenberger et al. "Learning from Multiview Correlations in Open-Domain Videos", ICASSP 2019
7. Sanabria et al. "How2: A Large-scale Dataset for Multimodal Language Understanding", ViGIL Workshop, NeurIPS 2018