Year 2 of Interactive Grounded Language Understanding

chist-era Projects Seminar 2018

IGLU

Jérôme Abdelnour[1], Simon Brodeur[1], Mathilde Brousmiche[1,3], Luca Celotti[1], Javier Civera[6], Aaron Courville[5], Jean-Benoit Delbrouck[3], Harm de Vries[5], Stéphane Dupont[3], Florian Golemo[7], Olivier Mastropietro[5], Pablo Azagra[6], Ana C. Murillo[6], Pierre-Yves Oudeyer[7], Olivier Pietquin[2], Bilal Piot[2], Gueorgui Pironkov[3], Jean Rouat[1], Giampiero Salvi[4], Omar Seddati[3], Florian Strub[2] and Sean Wood[1]

[1] Université de Sherbrooke, Canada
[2] Université de Lille 1, France
[3] Université de Mons, Belgium
[4] Royal Institute of Technology, Sweden
[5] Université de Montréal, Canada
[6] Universidad de Zaragoza, Spain
[7] Inria Bordeaux Sud-Ouest, France

Authors by alphabetical order

Summary

What was done during the second year of the IGLU project:

- Beginning of the integration of developed algorithms on multiple robotic platforms (e.g. IRL-1, Baxter).
- Study of the effects of grounding in multimodal neural machine translation.
- Speech-related processing in terms of source separation and visual detection of the active speaker.
- Generative modeling for language learning using probabilistic frameworks.
- Goal-oriented visual and dialogue tasks for language learning and grounding.
- Evaluation framework for grounded language understanding in cognitive agents.

Evaluation of Language Learning Agents

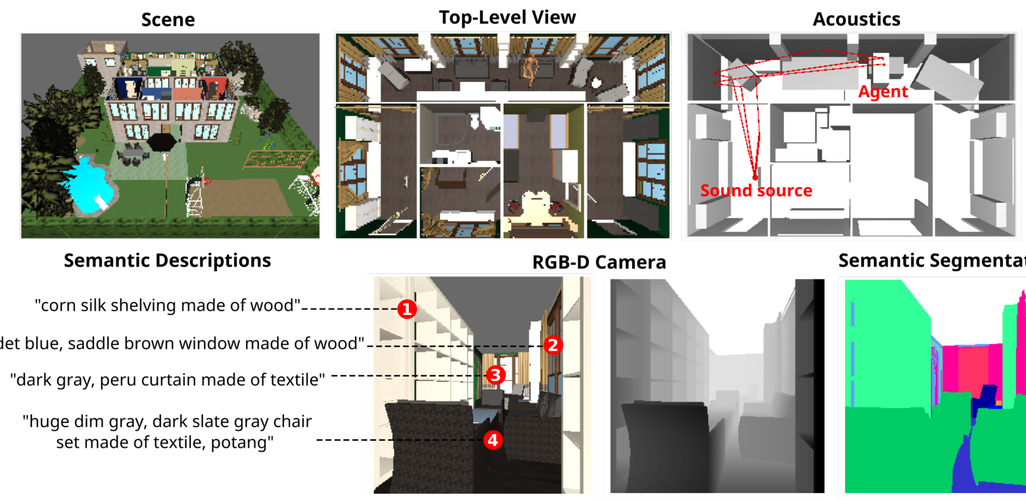

Figure: HoME: a Household Multimodal Environment

Research aims:

- Objective evaluation of the grounding abilities of artificial agents.
- Use goal-oriented dialogue games to learn language and ground it in multimodal perception.
- Realistic and complex environments, yet controllable and reproducible research.

Language Learning with Goal-oriented Visual Tasks

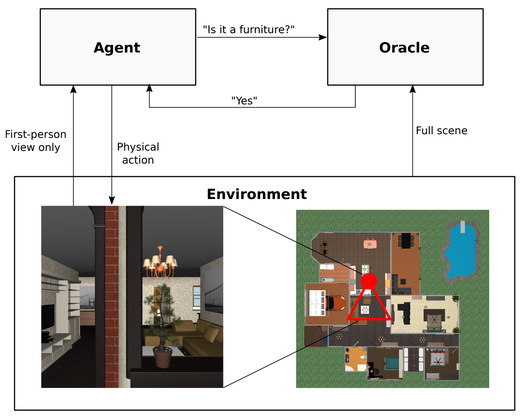

Figure: Situated dialogue on HoME

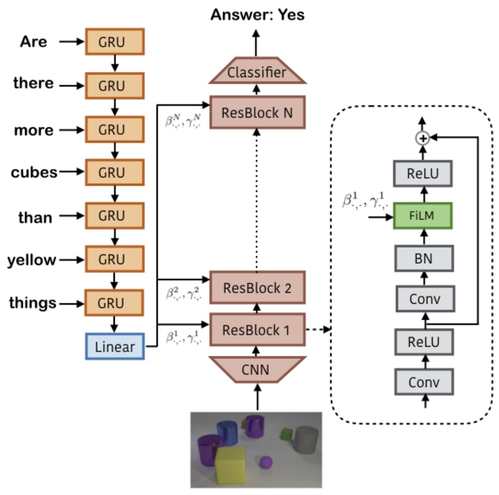

Figure: Visual question answering on CLEVR

Research aims:

- Situated dialogue and visual question answering that require effective language and vision integration.
- General reasoning over visual scenes with attention mechanisms (CBN and FiLM).

Generative Modeling for Language Learning

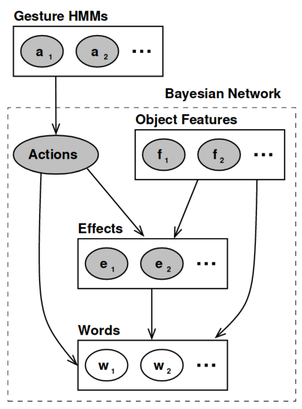

Figure: Learning of gestures, language and affordances

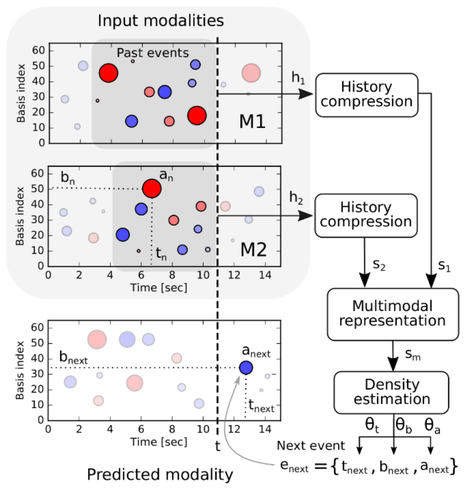

Figure: Event-based generative modeling

Research aims:

- Probabilistic frameworks for cognitive agents, with reasoning by building internal model of the environment.
- Joint learning of robot affordances and word descriptions with statistical recognition of human gestures.
- Multimodal event-based representations and probabilistic generative modeling of robot sensory data.

Integration on Robotic Platforms

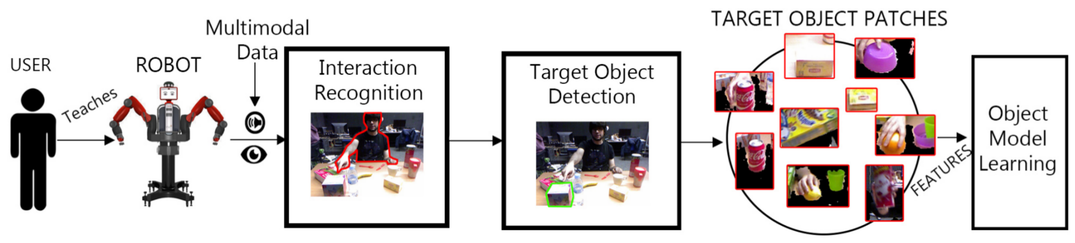

Figure: Object model learning from human-robot interactions

Research aims:

- Develop an API for robot incremental learning that supports multiple robotic platforms.
- Build a partially annotated dataset for object modeling from natural human-robot interactions.
- Acquire object models from interaction data (point, show and speak) using an end-to-end pipeline.
- Perform knowledge transfer from simulation to real robots using generative models.

Multimodal Neural Machine Translation

Figure: Translation with attentional mechanism for sentence “Einkleiner schwarzer Hund springt über Hindernisse”.

Source: A person in a red jacket with black pants holding rainbow ribbons .
Reference: Eine Person in einer roten Jacke und schwarzen Hosen hält Regenbogenbänder .
M2: Eine Person in roter Jacke mit schwarzen Hosen hält eine Znde in der Hand .
M3: Eine Person in einer roten Jacke und schwarzen Hosen hält einen Zebnde .

Figure: Multimodal translation task

Research aims:

- Solve ambiguities in machine translation by providing the visual context.
- Develop better attentional mechanisms (CBP and CBN) to find objects in the image.
- Improve visual and word representations (DenseCap and Glove).

Speech Enhancement and Processing

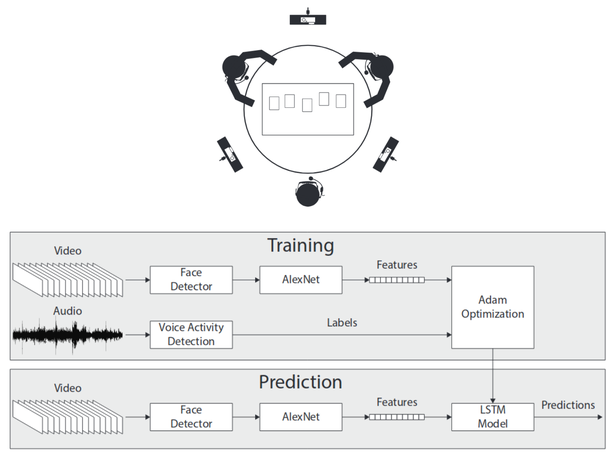

Figure: Active speaker detection

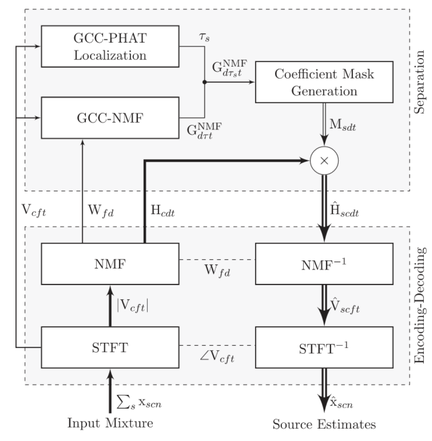

Figure: GCC-NMF source separation

Research aims:

- Automatic visual detection of the active speaker in multiparty interactions.
- Real-time source separation system with GCC-NMF, running on embedded systems.

References

P. Azagra, J. Civera, and A. C. Murillo. Finding regions of interest from multimodal human-robot interactions. In *Proc. GLU 2017 International Workshop on Grounding Language Understanding*, pages 73–77, 2017.

P. Azagra, F. Golemo, Y. Mollard, M. Lopes, J. Civera, and A. C. Murillo. A multimodal dataset for object model learning from natural human-robot interaction. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2017)*, 2017.

S. Brodeur, L. Celotti, and J. Rouat. Proposal of a generative model of event-based representations for grounded language understanding. In G. Salvi, editor, *Proceedings of the First International Workshop on Grounding Language Understanding*, pages 81–86. KTH, 2017.

S. Brodeur, E. Perez, A. Anand, F. Golemo, L. Celotti, F. Strub, J. Rouat, H. Larochelle, and A. C. Courville. Home: a household multimodal environment. *CoRR*, abs/1711.11017, 2017.

J.-B. Delbrouck and S. Dupont. Multimodal compact bilinear pooling for multimodal neural machine translation. *arXiv preprint arXiv:1703.08084*, 2017.

J.-B. Delbrouck, S. Dupont, and O. Seddati. Visually grounded word embeddings and richer visual features for improving multimodal neural machine translation. *arXiv preprint arXiv:1707.01009*, 2017.

A. Kumar Dhaka and G. Salvi. Sparse autoencoder based semi-supervised learning for phone classification with limited annotations. In *Proc. GLU 2017 International Workshop on Grounding Language Understanding*, pages 22–26, 2017.

E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. *arXiv preprint arXiv:1709.07871*, 2017.

H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6597–6607, 2017.

F. Strub, H. de Vries, J. Mary, B. Piot, A. C. Courville, and O. Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

S. Brodeur, S. Carrier and J. Rouat. Create: Multimodal dataset for unsupervised learning and generative modeling of sensory data from a mobile robot. *IEEE Dataport*, http://dx.doi.org/10.21227/H2M94J, 2018.

G. Saponaro, L. Jamone, A. Bernardino, and G. Salvi. Interactive robot learning of gestures, language and affordances. In *Proc. GLU 2017 International Workshop on Grounding Language Understanding*, pages 83–87, 2017.

K. Stefanov, J. Beskow, and G. Salvi. Vision-based active speaker detection in multiparty interaction. In *Proc. GLU 2017 International Workshop on Grounding Language Understanding*, pages 47–51, 2017.

S. Wood, J. Rouat, S. Dupont, and G. Pironkov. Blind speech separation and enhancement with GCC-NMF. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(4):745–755, 2017.

CONTACT: jean.rouat@usherbrooke.ca