Evaluating multi-modal deep learning systems with micro-worlds

Alexander Kuhnle, Ann Copestake
University of Cambridge (United Kingdom)
{aok25,aac10}@cam.ac.uk

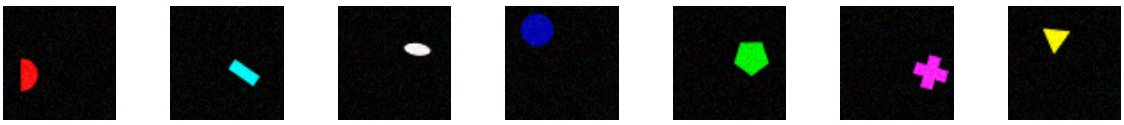
3rd November 2016

Multi-modal deep learning systems recently showed strong performance on the image captioning task (Karpathy and Li, 2015). At the same time, however, some more detailed investigations cast doubt on the quality of the results, or rather, whether the current evaluation practice is sufficient to test for true scene and language understanding (Hodosh and Hockenmaier, 2016; Nguyen et al., 2015).

With the aim of analysing basic linguistic/symbolic capabilities of multi-modal systems, we propose to move away from real-world photos with human-written captions, as well as the task setup of asking the system to generate captions. Regarding the latter, we more closely follow the setup of the recently introduced image question answering task (Antol et al., 2015) and take an image as a natural representation of the world against which statements can be evaluated. In our experiments, a system is trained on pairs of images and statements about them, together with corresponding values indicating the appropriateness of the proposition given the image, while during test time previously unseen object/concept combinations of the same symbolic structure as the train instances are presented. To do well on this task, it is hence crucial for the evaluated system to learn to understand the underlying symbolic principle.

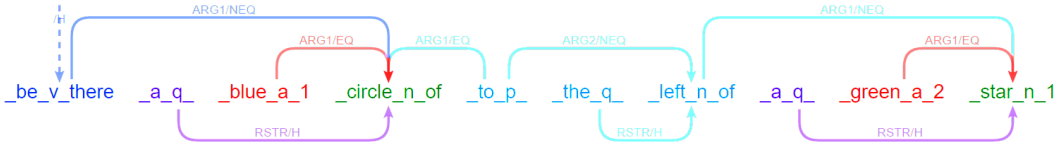
Instead of real-world data, we use automatically generated abstract micro-worlds, similar to other work on formally testing the abilities of deep learning systems (Sorodoc et al., 2016; Joulin and Mikolov, 2015; Bowman et al., 2015; Weston et al., 2015; Vinyals et al., 2015; Sukhbaatar et al., 2015). In doing so, we avoid the problem of visually noisy or otherwise ambiguous instances, and are able to more exhaustively cover the space of possible images and captions. Moreover, it enables us to control the data generation, and so investigate the learning process in a network. For instance, quantifier learning can be analysed by constructing instances specifically targeting interesting quantifier configurations (Pietroski et al., 2009).

Internally, the micro-worlds are explicit representations listing all world objects with their properties, from which both the image and a caption is extracted. The objects are randomly sampled and, for now, consist of coloured shapes. Below some single-shape worlds with example captions generated by our system:



red triangle ⇒ False | *cyan rectangle* ⇒ True | *white ellipse* ⇒ True | *magenta circle* ⇒ False | *green pentagon* ⇒ True | *yellow cross* ⇒ False | *white circle* ⇒ False

For caption generation we use the Dependency Minimal Recursion Semantics (DMRS) formalism (Copestake et al., 2016; Copestake, 2009) to represent the abstract semantic structure of a proposition. Every object and property is annotated with its corresponding DMRS predicate(s), and the compositional framework of DMRS enables us to construct a wide variety of possible sentences from a few general DMRS graph skeletons on this basis. Below an example proposition with coloured compositional components:



"There is a blue circle which is to the left of a green star."

DMRS graphs can be transformed to MRS structures, from which corresponding English sentences can be generated with a bidirectional HPSG-grammar like the English Resource Grammar (Flickinger, 2000; Flickinger et al., 2014) and a parser-generator like ACE (`http://sweaglesw.org/linguistics/ace/`).

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *International Conference on Computer Vision (ICCV)*.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. 2015. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, Beijing. Association for Computational Linguistics.

Ann Copestake, Guy Emerson, Michael W. Goodman, Matic Horvat, Alexander Kuhnle, and Ewa Muszyńska. 2016. Resources for building applications with Dependency Minimal Recursion Semantics. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-16)*, pages 1240–1247, Portorož, Slovenia. European Language Resources Association (ELRA).

Ann Copestake. 2009. Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 1–9, Athens, Greece.

Dan Flickinger, Emily M. Bender, and Stephan Oepen. 2014. Towards an encyclopedia of compositional semantics: Documenting the interface of the English Resource Grammar. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-14)*, pages 875–881, Reykjavik, Iceland. European Language Resources Association (ELRA).

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28.

Micah Hodosh and Julia Hockenmaier. 2016. Focused evaluation for image description with binary forced-choice tasks. In *Proceedings of the 5th Workshop on Vision and Language*, Berlin, Germany.

Armand Joulin and Tomas Mikolov. 2015. Inferring algorithmic patterns with stack-augmented recurrent nets. In *Advances in Neural Information Processing Systems 28*, pages 190–198. Curran Associates, Inc.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 3128–3137, Boston, MA, USA.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Paul Pietroski, Jeffrey Lidz, Tim Hunter, and Justin Halberda. 2009. The meaning of 'most': Semantics, numerosity and psychology. *Mind and Language*, 24(5):554–585.

Ionut Sorodoc, Angeliki Lazaridou, Gemma Boleda, Aurélie Herbelot, Sandro Pezzelle, and Raffaella Bernardi. 2016. "Look, some green circles!": Learning to quantify from images. In *Proceedings of the 5th Workshop on Vision and Language*, Berlin, Germany.

Sainbayar Sukhbaatar, Arthur Szlam, Gabriel Synnaeve, Soumith Chintala, and Rob Fergus. 2015. MazeBase: A sandbox for learning from games. *CoRR*, abs/1511.07401.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 2692–2700, Montreal, Canada. MIT Press.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards AI-complete question answering: A set of prerequisite toy tasks. *CoRR*, abs/1502.05698.