Pattern Recognition is Not Enough: Representing Language, Action and Perception with Modular Neural Networks

Simon Dobnik
Department of Philosophy, Linguistics
and Theory of Science (FLoV)
University of Gothenburg, Sweden
simon.dobnik@gu.se*

John D Kelleher
Information, Communications
and Entertainment Institute (ICE)
Dublin Institute of Technology, Ireland
john.d.kelleher@dit.ie

Abstract

Current deep learning approaches to modelling of spatial language in generating image captions have shortcomings because they are focused on recognition of visual patterns. The multiplicity of factors that influence spatial language which also include aspects of interaction between speakers and between speakers and their environment invites a modular approach where the solution can be built in a piece-wise manner and then integrated. We call this approach where deep learning is assisted with domain knowledge expressed as modules that are trained on data a top-down or mechanistic approach to otherwise a bottom-up phenomenological approach.

In recent years deep learning approaches have made significant breakthroughs. An exciting aspect of deep learning is learning inter/multi-modal representations from data that includes discrete information (e.g. words) and continuous representations (e.g. word embeddings and visual features), such as those used in automatic image captioning systems. A number of shortcomings with current deep learning architectures have been identified with respect to their application to spatial language such as "the chair is to the left and close to the table" or "go down the corridor until the large painting on your right, then turn left". For example, in (Kelleher and Dobnik, 2017) we argue that contemporary image captioning networks have been configured in a way that they capture visual properties of objects ("what" in terms of (Landau and Jackendoff, 1993)) rather than spatial relations between them ("where"). Consequently, within the captions generated by these systems the relation between the preposition and the object is not grounded in geometric representation of space but only in the linguistic sequences through the decoder language model where the co-occurrence of particular words in a sequence is estimated.[1] This is because neural networks are typically used as generalised learning mechanisms that learn with as little supervision through architecture design as possible. We call this data-driven approach a *bottom-up* or *phenomenological approach*. The problem is that the chosen architecture may not be optimal for every aspect of the cognitive representations that we want to learn.

We do not argue that language model is not informative for predicting spatial relations. In addition to (i) scene geometry (Logan and Sadler, 1996; Dobnik and Åstbom, 2017) they also rely on (ii) perspective and perceptual context (Kelleher and Kruijff, 2005; Dobnik et al., 2015), (iii) functional world knowledge about dynamic kinematic routines of objects (Coventry et al., 2005), and (iv) interaction between agents through language and dialogue and with the environment through perception (Schutte et al., 2017; Dobnik and de Graaf, 2017). In (Dobnik et al., 2018) we show that a language model is useful in predicting functional relations between objects. The system can learn something about object interaction without visually observing these objects and such knowledge is used as background knowledge when generating and interpreting spatial descriptions. The information expressed in a language model or visual features of the scene is therefore just one of the modalities that must be taken into account. This provides a challenge for computational modelling

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

[1] The over-reliance of deep learning models on the language model has been criticised recently for example, in relation to visual question answering and an attempts have been made to make the systems give a greater weight to images in predicting the caption, for example by balancing different answers in datasets (Agrawal et al., 2017).

of spatial descriptions because (i) it is difficult to provide and integrate that kind of knowledge and (ii) its contextual underspecification. A computational system taking into account these meaning components in the context would be able to understand and generate better, more human-like, spatial descriptions and engage in more efficient communication in the domain of situated agents and humans. Furthermore, it could exploit the synergies between different knowledge sources to compensate missing knowledge in one source from another (Schutte et al., 2017).

In (Dobnik and Kelleher, 2017) we argue that the multiplicity of factors that influence spatial language invites a modular approach where the solution can be built in a piece-wise manner and then integrated (Feldman, 1989; Regier, 1996; Andreas et al., 2016; Johnson et al., 2017). We call this approach where deep learning is assisted with domain knowledge expressed as modules that are trained on data a *top-down* or *mechanistic approach*. One challenge to spatial language is the lack of an overarching theory explaining how these different factors should be integrated but (Herskovits, 1987) and (Coventry et al., 2005) appear to be promising candidates. Early work on neural networks includes some examples of neural models that could provide a basis for the design of specific modules. For example, (Regier, 1996) captures geometric factors and paths of motion. The system in (Coventry et al., 2005) processes dynamic visual scenes containing three objects: a teapot pouring water into a cup and the network learns to optimise, for each temporal snapshot of a scene, the appropriateness score of a spatial description obtained in subject experiments. The idea behind these experiments is that descriptions such as *over* and *above* are sensitive to a different degree of geometric and functional properties of a scene, the latter arising from the functional interactions between objects. The model is split into three modules: (i) a vision processing module that deals with detection of objects from image sequences that show the interaction of objects, (ii) an Elman recurrent network that learns the dynamics of the attended objects in the scene over time, and (iii) a dual feed-forward vision and language network to which representations from the hidden layer of the Elman network are fed and which learns how to predict the appropriateness score of each description for each temporal configuration of objects. Each module of this network is dedicated to a particular task: (i) to recognition of objects, (ii) to follow motion of attended objects in time and (iii) to integration of the attended object locations with language to predict the appropriateness score, factors that have been identified to be relevant for computational modelling of spatial language and cognition in previous experimental work (Coventry et al., 2005). The example shows the effectiveness of representing networks as modules and their possibility of joint training where individual modules constrain each other.

The model could be extended in several ways. For example, contemporary CNNs and RNNs could be used which have become standard in neural modelling of vision and language due to their state-of-the-art performance. Secondly, the approach is trained on a small dataset of artificially generated images of a single interactive configuration of three objects. An open question is how the model scales on a large corpus of image descriptions (Krishna et al., 2017) where considerable noise is added: the appearance and location of objects is distorted by the angle at which the image is taken. Furthermore, there are no complete temporal sequences of objects and the corpora mostly do not contain human judgement scores on how appropriate a description is given an image. Finally, (Coventry et al., 2005)'s model integrates three modalities used in spatial cognition, but as we have seen there are several others. An important aspect is grounded linguistic interaction and adaptation between agents. For example, (Lazaridou et al., 2016) describe a system where two networks are trained to perform referential games (dialogue games performed over some visual scene) between two agents. In this context, the agents develop their own language interactively. An open research question is whether parameters such as frame of reference intended by the speaker of a description could also be learned this way.

Due to their dependence on several modalities spatial descriptions therefore provide a good testbed for the requirements of modelling language, action and perception with neural networks. While it is hard to capture these modalities with a general learning framework, using our expert domain knowledge and splitting the networks into modules that can be specialised for a purpose reduces the complexity of the learning task and makes it more tractable.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2017. Don't just assume; look and answer: Overcoming priors for visual question answering. *arXiv*, arXiv:1712.00377 [cs.CV]:1–15.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. In *Proceedings of NAACL-HLT 2016*, pages 1545–1554, San Diego, California. Association for Computational Linguistics.

Kenny R. Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V. Richards. 2005. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In Christian Freksa, Markus Knauff, Bernd Krieg-Brückner, Bernhard Nebel, and Thomas Barkowsky, editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, volume 3343 of *Lecture Notes in Computer Science*, pages 98–110. Springer Berlin Heidelberg.

Simon Dobnik and Amelie Åstbom. 2017. (Perceptual) grounding as interaction. In *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*, pages 17–26, Saarbrücken, Germany.

Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Simon Dobnik and Erik de Graaf. 2017. KILLE: a framework for situated agents for learning language through interaction. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 162–171, Gothenburg, Sweden. Northern European Association for Language Technology (NEALT), Association for Computational Linguistics.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32, Gothenburg, Sweden.

Simon Dobnik and John D. Kelleher. 2017. Modular mechanistic networks: On bridging mechanistic and phenomenological models with deep neural networks in natural language processing. In *CLASP Papers in Computational Linguistics: Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017), Gothenburg, 12–13 June 2017*, volume 1, pages 1–11, Gothenburg, Sweden.

Jerome A. Feldman. 1989. Structured neural networks in nature and in computer science. In Rolf Eckmiller and Christoph v.d. Malsburg, editors, *Neural Computers*, pages 17–21. Springer, Berlin, Heidelberg.

Annette Herskovits. 1987. *Language and Spatial Cognition*. Cambridge University Press, New York, NY, USA.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Fei-Fei Li, C. Lawrence Zitnick, and Ross B. Girshick. 2017. Inferring and executing programs for visual reasoning. *arXiv*, arXiv:1705.03633v1 [cs.CV]:1–13.

John D. Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. In *CLASP Papers in Computational Linguistics: Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017)*, volume 1, pages 41–52, Gothenburg, Sweden.

John D. Kelleher and Geert-Jan M. Kruijff. 2005. A context-dependent algorithm for generating locative expressions in physically situated environments. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, pages 1–7, Aberdeen, Scotland. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Barbara Landau and Ray Jackendoff. 1993. "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238, 255–265.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. *arXiv*, arXiv:1612.07182v2 [cs.CL]:1–11.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.

Terry Regier. 1996. *The human semantic potential: spatial language and constrained connectionism*. MIT Press, Cambridge, Massachusetts, London, England.

Niels Schutte, Brian Mac Namee, and John D. Kelleher. 2017. Robot perception errors and human resolution strategies in situated human–robot dialogue. *Advanced Robotics*, 31(5):243–257.