Research proposal: Evaluating multi-modal deep learning systems with micro-worlds

Alexander Kuhnle
University of Cambridge (United Kingdom)
aok25@cam.ac.uk

26th November 2016

Abstract

My research focuses on the evaluation of multi-modal deep neural networks with respect to various symbolic/linguistic understanding and generalisation abilities. Instead of real-world data, I work on a system which automatically generates artificial toy data by randomly sampling an internal world model representation. Although simple, their structural complexity is still sufficient to generate a broad variety of interesting instances involving many aspects of language. In contrast to the classic image captioning task, a deep network here is asked to decide about the agreement of a given statement and the presented image. By controlling the content of training and test instances, I make sure that achieving good performance clearly indicates the learning of genuine concept understanding. Moreover, such a setup makes a more detailed investigation of the process and content of learning possible.

1 Motivation

Deep learning models have had a major impact on research in natural language processing and raised the performance bar substantially in many of the standard evaluations. Moreover, new multi-modal tasks are tackled which older systems would not have been able to handle (Karpathy and Li, 2015; Antol et al., 2015). But because it is very difficult to comprehend, let alone guide, the process of learning in deep neural networks, there is an air of uncertainty about exactly *what* and *how* these networks learn.

Recent work has identified various unexpected and weird properties in light of the fact that deep networks show strong, often close-to-human performance. On the one hand, Szegedy et al. (2014) show how the decision of an object recognition network can be changed entirely by only shifting a few pixels in an image, invisible to the human eye. Follow-up work by Nguyen et al. (2015) shows how such a network can be perfectly certain about its classification when faced with an image which humans would discount as noise. On the other hand, Sproat and Jaitly (2016) report similarly unexpected behaviour when applying deep learning to text normalisation, with the network dropping digits in numbers or changing the unit of values seemingly at random.

The opaqueness of the inner working is seen as a reason for the strong performance of deep learning and is hence deliberately accepted (Goodfellow et al., 2016, chapter 16). However, the results referred to above illustrate that whatever representations a deep neural network learns, they appear to be fundamentally different from what the network's *"close-to-human" performance* suggests. This emphasises the importance to identify and guarantee aspects about a network's reasoning, because only then one is able to justify its decisions.

Moreover, Zhang et al. (forthcoming) report the surprising result that a deep network does not show increased difficulty when approximating mere noise. This raises other questions about the quality of deep learning and highlights that deep networks are indeed more powerful *universal* approximators than earlier machine learning models. Consequently, unexpected hidden structural regularities and biases specific to a particular dataset can lead to unexpected and undesirable evaluation results not representative of the evaluation goal. The task of creating *good* datasets is hence both of paramount importance and great difficulty.

This is one reason for the recently revived interest in artificial data for deep learning evaluation, in the spirit of traditional formal semantics and early artificial intelligence research. Work inspired by the paradigm of *"block worlds"* includes the MazeBase game environment of Sukhbaatar et al. (2015), the roadmap of Mikolov et al. (2015) and follow-up work, the bAbI dataset of Weston et al. (2015) targeting inference reasoning, the experiments of Bowman et al. (2015) on logic abilities and evaluations of algorithmic capabilities (Joulin and Mikolov, 2015; Vinyals et al., 2015).

The advantage of artificial data over real data is seen in reducing noise and ambiguity unessential to the investigated formal problem, focusing clearer on a certain learning task, and enabling to better control the content and biases of a dataset. Based on these virtues, traditional formal semantics is able to essentially give an interpretable account of reasoning over natural language. The Achilles' heel of these theories, however, is the question of scaling up their symbolic world model to the gigantic extent of the real world.

In contrast, scalability is exactly where the strength of deep learning lies and a reason for its recent successes. My framework attempts to "marry" these two paradigms, by using a symbolic model combined with formal semantics to allow for controlled evaluation and investigation of the learning process of multi-modal deep neural networks. As such, it constitutes a first step towards understanding and consequently justifying the reasoning behind deep learning semantics.

2 My research proposal

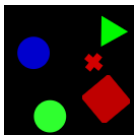
In my research, I am investigating micro-worlds as basis to evaluate the linguistic and symbolic capabilities of multi-modal deep learning systems. However, in contrast to formal semantics, the underlying abstract world model is not directly presented to the evaluated system, but instead a visual representation is generated from it. In addition, a natural language caption expressing a proposition about the micro-world is given, and it is the system's task to decide whether the caption is in agreement with the image/world (see Hodosh and Hockenmaier (2016) for a similar setup involving real world images). I emphasise a clean and natural interface where natural language serves as linguistic representation, images as visual scene representation, and no intermediate representation is involved in the communication with the evaluated deep neural network. This avoids some of the problems of scalability, since such an interface is *universal*, i.e. arbitrarily complex worlds and statements could be "modelled" in this format.

I started working on micro-worlds consisting of coloured two-dimensional shapes located on a plane. Although very simple, their structural complexity is still sufficient to generate a broad variety of interesting instances involving many aspects of language, such as spatial relations, comparative statements, quantification, first-order logic, syntactic composition and underspecification. The system to be evaluated is trained on a set consisting of images, captions and corresponding agreement scores (true: 1.0, false: 0.0).

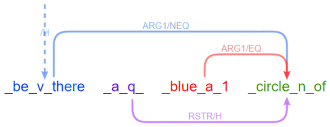
In the following, I describe some important aspects of my proposed setup:

- Artificial data can be generated automatically and in arbitrary quantities. This is an important change for training deep networks as compared to, for instance, the analysis of Nguyen et al. (2015) or Zhang et al. (forthcoming) – instead of iterating over a finite set of data points multiple times, here the same instance is very unlikely to be encountered twice, hence no danger of under-/over-fitting in the classic sense.
- The content of both individual instances and the entire training/test set split can be controlled. By withholding certain concept combinations during training, the generalisation ability of a model is evaluated, i.e. its ability to re-combine previously introduced concepts to handle unseen instances. For instance, colour-shape combinations are withheld to test for learning separate concepts of colour and shape, or a certain number of objects is never shown to test for generalisation independent of the number of objects.
- Toy data consists of clear and structurally rich instances which are non-trivial to handle, at least for deep learning approaches. The implicit object recognition sub-task is at a minimum necessary to evaluate interesting linguistic properties. In contrast to real-world scenes, it allows for the evaluation of certain capabilities in isolation, since the data is not striving to reflect the rich interconnectedness of the real world. Moreover, note that humans often get analysed with similarly abstract scenes in psycholinguistics. For instance, the quantifier experiments in Pietroski et al. (2009) use shape images.
- The language describing abstract worlds avoids any biases real-world language is likely to contain (dogs always chase cats, etc). Since all objects and relations are abstract and equal, there is no natural or common preference of one expression over another, hence focusing on genuine understanding.

I see the aforementioned aspects as clearly advantageous for my goal of what I think of as *unit-testing* multi-modal deep neural networks: Creating artificial test datasets which clearly evaluate to what degree a symbolic capability is learnable in abstract isolation. Given the wealth of structurally rich scenarios that can be generated in such a setup, good performance strongly suggests that a network developed an inference mechanism at least as powerful and expressive as the corresponding underlying rule, for instance, of formal semantics. I will go into more detail later what my generation system can also be used for.

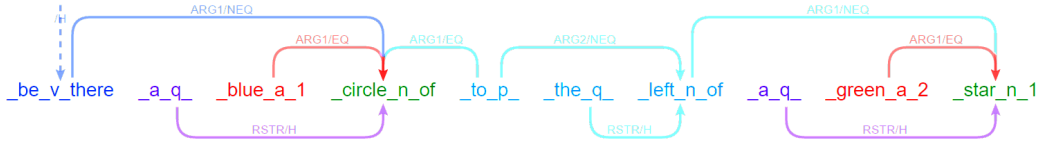


{shape = “triangle”, colour = “green”, x = 0.9, y = 0.1, etc.},
{shape = “circle”, colour = “blue”, x = 0.2, y = 0.3, etc.},
{shape = “cross”, colour = “red”, x = 0.7, y = 0.5, etc.},
{shape = “square”, colour = “red”, x = 0.8, y = 0.8, etc.},
{shape = “circle”, colour = “green”, x = 0.3, y = 0.9, etc.}



“There is a blue circle.”

Figure 1: Visual representation, internal representation, existential statement with corresponding DMRS graph.



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

Figure 2: An example of a more complex statement with its compositional components coloured.

3 Methodology

3.1 Data generation process

The internal data structure representing a shape world is an explicit list keeping record of all world objects together with their attributes like shape, colour, position, etc (figure 1). These structures are automatically generated by randomly sampling relevant generation parameters. Apart from the aforementioned object attributes, these parameters also include world attributes like the number of entities, the spatial arrangement of objects, and information about available caption components (see below). A dataset, finally, is a generator object for worlds and captions which defines the set of relevant generation parameters and their valid value ranges. Image representations in the form of two-dimensional matrices of RGB pixel values are straightforwardly extracted from such internal representations and can be visualised using the Python Imaging Library.

For caption generation I use Dependency Minimal Recursion Semantics (DMRS) (Copestake et al., 2016) as an intermediate representation for the abstract semantic structure of a proposition. In my micro-worlds, a DMRS graph is more a natural annotation of the world rather than an abstracting representation and hence integrates straightforwardly with the internal world model. Every object and property is annotated with its corresponding DMRS predicate(s), and the compositional framework of DMRS semantics enables me to combinatorially construct a wide variety of sentences from a few general DMRS graph skeletons (figure 1 and, using the former, figure 2). DMRS graphs can be transformed to MRS structures, from which corresponding English sentences can be generated with a bi-directional DELPH-IN high-precision grammar like the English Resource Grammar (Flickinger et al., 2014) and a parser-generator like ACE (`http://sweaglesw.org/linguistics/ace/`).

In addition to that, such a caption object directly yields an interpretation as logical formula in the style of formal semantics. Given the symbolic internal model, a caption can thus check whether it applies to a world (figure 3). This is a key property for the generation of false but plausible instances. The generation process creates negative image-caption pairs by internally generating a second world model. This world is then used to extract a new caption, for which is subsequently ensured that it does not apply to the first, actual world.

3.2 Multi-modal deep neural network architecture

I started experimenting with a “generic” multi-modal deep learning architecture implemented in TensorFlow, with a convolutional component processing the visual input and a long short-term memory component for

“All circles are to the left of a red cross.”

$$\forall s_1 \in \text{World}\colon \text{circle}(s_1.\text{shape}) \Rightarrow \Big(\exists s_2 \in \text{World}\colon \text{cross}(s_2.\text{shape}) \wedge \text{red}(s_2.\text{colour}) \wedge s_1.\text{x} < s_2.\text{x}\Big)$$

“The left-most shape is a blue circle.”

$$\forall s_1 \in \text{World}\colon \Big(\forall s_2 \in \text{World}\colon s_1.\text{x} \leq s_2.\text{x}\Big) \Rightarrow \text{circle}(s_1.\text{shape}) \wedge \text{blue}(s_1.\text{colour})$$

Figure 3: Two example shape world captions and their formal semantics interpretation as logical formula.

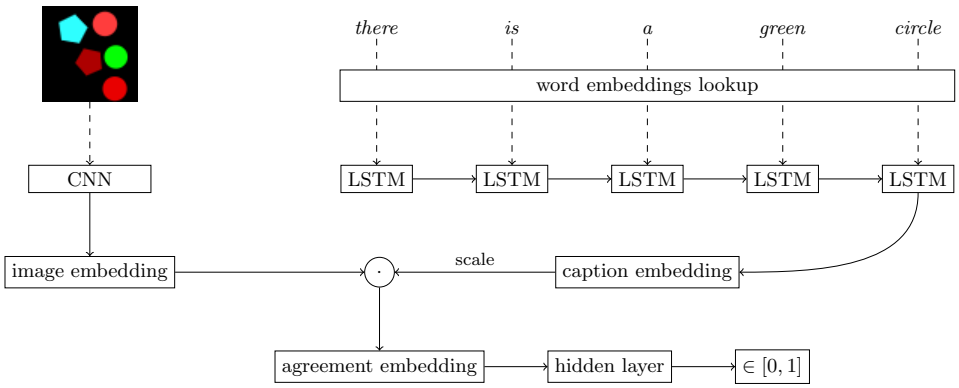


Figure 4: A schematic depiction of the multi-modal deep neural net I am currently experimenting with.

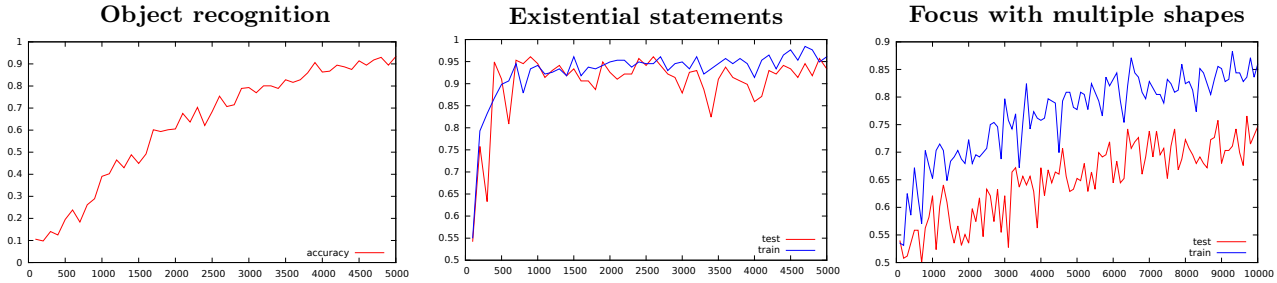


Figure 5: First results for the experiments on the object recognition task and for existential statements with images containing either one or multiple shapes (x-axis: iterations with batch size 64).

the language input. The CNN consists of three layers of convolution, max-pooling and ReLU non-linearity, producing an image embedding from the visual input matrix. The LSTM processes the embeddings associated with the caption words to produce a caption embedding. Subsequently, the smaller caption embedding is scaled to fit the image embedding, which then are both fused via pointwise multiplication. This resulting "agreement embedding" is passed on to a hidden layer with a dropout rate of 0.25 and finally produces the predicted agreement value. The entire architecture, including the word embeddings, is trained end-to-end. Figure 4 shows a schematic picture of the architecture. I plan to implement other common architectures, potentially requiring minor architecture modifications for my (untypical) plausibility score setup.

There are various possibilities for baseline systems. On the one hand, the LSTM component can be replaced by, for instance, a bag-of-words step averaging the caption word embeddings. Moreover, a separate image processing network can be trained for object recognition and, similarly, word embeddings can be trained separately, or pre-trained networks/embeddings can alternatively be used. Based on thus obtained image/caption embeddings, only the final part of the architecture is specifically trained on the task in question.

Figure 5 shows the performance curves of first experiments on:

- Object recognition (8 shapes, 7 colours, hence 56 objects) for visual network part (no caption embedding),
- Existential statements (*"There is a [colour] [shape]."*) with one withheld test combination (*"red triangle"*),
- Existential statements as before, but overall up to 4 objects present during training, or 5 in test instances.

4 What's next? – Ideas for the PhD project

Quantifier learning experiments will be the next sub-project in my PhD. I want to investigate, first, whether multi-modal deep networks can handle absolute quantification, i.e. counting, and second, whether relative quantifiers like *"most"*, *"all"*, *"no"* can be learned. It will also be interesting to see which quantifiers are more difficult to learn and to what degree the understanding is reliable and robust when presented with new situations. As part of this project, I will visit the University of Trento for two weeks in February 2017. During the ESSLLI summer school in August I talked to Raffaella Bernardi and Sandro Pezzelle, who were both involved in similar

work on quantifiers (Sorodoc et al., 2016). In a later conversation with Aurelie Herbelot I learned about their follow-up work, and we decided to combine our endeavours and collaborate on this topic.

Visual diversity can be increased by introducing pixel value noise, diversifying colour shades and shape sizes, or (potentially) applying photo filters and effects. This feature is already largely implemented.

More complex worlds can be automatically generated and captioned by the same data generation process. For instance, clipart objects can be used instead of coloured shapes similar to Zitnick et al. (2016). Another interesting option to investigate is WordsEye (`http://www.wordseye.com/`) (Coyne and Sproat, 2001), a powerful text-to-scene generator which allows to expand the visual scenario solely based on the caption.

Language noise in the form of paraphrasing can introduce linguistic variety. My earlier work on paraphrase rules as DMRS graph mapping (Copestake et al., 2016) allows to do so in the compositional framework of DMRS, which would fit in well with the current generation process.

Human captions on the images of my system (true or false) can be obtained (e.g. Amazon Mechanical Turk). On the one hand, these can be compared to the automatically generated captions and inspire paraphrasing rules for increased linguistic diversity. On the other hand, given enough variety in the training data, human captions can be used as test set for a more realistic evaluation.

Different language for captions would overcome the limitation to English, but in particular would exhibit differences in its symbolic-linguistic realisation. My system can easily switched to another language, essentially just by replacing the attribute-predicate mappings and introducing new DMRS subgraph skeletons acting as compositional components. This obviously requires a suitable underlying DELPH-IN grammar. On the annual DELPH-IN meeting at Stanford University this summer, I learned that there are suitable broad-coverage grammars for e.g. German or Japanese, and smaller ones for more exotic languages in the Grammar Matrix project (Bender et al., 2002). Note that the shape worlds do not need very broad coverage.

Multi-agent experiments can be conducted on symbolic cooperative tasks with the aim of evaluating pragmatics or language evolution. I will co-supervise a thesis project for the MPhil in Advanced Computer Science on language learning in a referential game setup along the lines of Lazaridou et al. (2016) and, following up, Lazaridou et al. (forthcoming). In contrast to their approach, however, the micro-worlds of my system are more diverse and cover a larger configuration space, so they are likely to avoid the issue of the agents developing a specialised language which exploits hidden regularities in the data.

Psycholinguistics experiments can be replicated for deep neural networks. For instance, the quantifier experiments of Pietroski et al. (2009) use images of coloured shapes to investigate how humans perceive and process the quantifier *"most"*. Such experiments can show to what degree the learning process and eventual understanding of a certain linguistic concept resembles human learning or is in fact rather different.

References

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

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

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).

Bob Coyne and Richard Sproat. 2001. WordsEye: An automatic text-to-scene conversion system. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pages 487–496, Los Angeles, CA, USA. ACM.

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.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep learning. Book in preparation for MIT Press.

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.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2016. Towards multi-agent communication-based language learning. *CoRR*, abs/1605.07133.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. forthcoming. Multi-agent cooperation and the emergence of (natural) language. *International Conference on Learning Representations (ICLR 2017) (under review)*.

Tomas Mikolov, Armand Joulin, and Marco Baroni. 2015. A roadmap towards machine intelligence. *CoRR*, abs/1511.08130.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2015, pages 427–436.

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.

Richard Sproat and Navdeep Jaitly. 2016. RNN Approaches to Text Normalization: A Challenge. *CoRR*, abs/1611.00068.

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. *CoRR*, abs/1312.6199.

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.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. forthcoming. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations (ICLR 2017) (under review)*.

C. Lawrence Zitnick, Ramakrishna Vedantam, and Devi Parikh. 2016. Adopting abstract images for semantic scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):627–638.