Tutorial : Memory Augmented Neural Networks for Natural Language Processing

Caglar Gulcehre, Sarath Chandar
Montreal Institute for Learning Algorithms
Université de Montréal, Canada.
{ca9lar,apsarathchandar}@gmail.com

Abstract

Memory Augmented Neural Networks (MANNs) can store and read information from an external memory. While the traditional machine learning algorithms (including neural networks) accepts an input and process it to perform a prediction, MANNs can use the explicit memory to store necessary information during the execution of the task and retrieve information from the memory when needed. This can be helpful for complex tasks like reasoning, planning, question answering, and dialogue systems. The aim of this tutorial is to introduce this paradigm of memory augmented neural networks to the NLP community since this has large scope in several complex NLP tasks like question answering, reading comprehension, dialogue systems, and summarization.

Description of Tutorial Content

Designing of general-purpose learning algorithms is a long-standing goal of artificial intelligence. A general purpose AI agent should be able to have a memory that it can store and retrieve information from. Despite the success of deep learning in particular with the introduction of LSTMs and GRUs to this area, there are still a set of complex tasks that can be challenging for conventional neural networks. Those tasks often require a neural network to be equipped with an explicit, external memory in which a larger, potentially unbounded, set of facts need to be stored. They include but are not limited to, reasoning, planning, episodic question-answering and learning compact algorithms. Recently two promising approaches based on neural networks to this type of tasks have been proposed: Memory Networks and Neural Turing Machines.

In this tutorial, we will give an overview of this new paradigm of "neural networks with memory". We will present a unified architecture for Memory Augmented Neural Networks (MANN) and discuss the ways in which one can address the external memory and hence read/write from it. Then we will introduce Neural Turing Machines and Memory Networks as specific instantiations of this general architecture. In the second half of the tutorial, we will focus on recent advances in MANN which focus on the following questions: How can we read/write from an extremely large memory in a scalable way? How can we design efficient non-linear addressing schemes? How can we do efficient reasoning using large scale memory and an episodic memory? The answer to any one of these questions introduces a variant of MANN. We will conclude the tutorial with several open challenges in MANN and its applications to NLP.

We will introduce several applications of MANN in NLP throughout the tutorial. Few examples include language modeling, question answering, visual question answering, and dialogue systems.

For updated information and material, please refer to our tutorial website `https://sites.google.com/view/mann-emnlp2017/`.

Outline of the tutorial

1. Introduction and Motivation [10 mins]

2. Basics [30 mins]
   (a) Neural Networks and Backpropagation
   (b) Recurrent Neural Networks
   (c) Long Short Term Memory (LSTM) Networks
3. Memory Augmented Neural Networks (MANN) [60 mins]
   (a) Why Memory Augmented Neural Networks?
   (b) General paradigm: Neural Networks with Memory
   (c) Addressing Mechanisms for accessing memory
   (d) Neural Turing Machines [3]
   (e) Memory Networks [10, 9]
4. Advances in MANN [70 mins]
   (a) Dynamic Neural Turing Machines with soft and hard addressing schemes. [6]
   (b) Hierarchical Attentive Memory [1]
   (c) Dynamic Memory Networks with episodic Memory [7, 11]
   (d) Scalable memory access for MANNs with extremely large memory. [2, 8]
   (e) Differentiable data structures. [5]
   (f) Differentiable Neural Computers. [4]
5. Challenges and Open Questions [10 minutes]

Instructors

Speaker 1:
Caglar Gulcehre,
Ph.D. candidate, Université de Montréal,
ca9lar@gmail.com

Caglar Gulcehre is a final year PhD student in University of Montreal under the supervision of Yoshua Bengio. His work mainly focuses on applications of neural networks, in particular recurrent architectures such as GRU and LSTMs on NLP and sequence to sequence learning tasks. His research also investigates different optimization approaches and architectures which are easier to optimize for neural networks. His recent research focuses on building neural network models that have external memory structures. He has done research internships at IBM Watson Research Center, Google Deep Mind. He was a PC Member at ECML and IJCAI 2016 Deep Reinforcement Learning Workshop. Prior to joining MILA as a PhD student, he finished his master degree in Middle East Technical University in Cognitive Science department. The complete list of his publications can be found at: https://scholar.google.ca/citations?user=7hwJ2ckAAAAJ&hl=en&oi=ao

Speaker 2:
Sarath Chandar,
Ph.D. student, Université de Montréal,
apsarathchandar@gmail.com

Sarath Chandar is currently a PhD student in University of Montreal under the supervision of Yoshua Bengio and Hugo Larochelle. His work mainly focuses on Deep Learning for complex NLP tasks like question answering and dialog systems. He also investigates scalable training procedure and memory access mechanisms for memory network architectures. In the past, he has worked on multilingual representation learning and transfer learning across multiple languages. His research interests includes Machine Learning, Natural

Language Processing, Deep Learning, and Reinforcement Learning. Before joining University of Montreal, he was a Research Scholar in IBM Research India for a year. He was a co-organizer of Deep Reinforcement Learning Workshop at IJCAI 2016. He has previously given a tutorial on "Multilingual Multimodal Language Processing using Neural Networks" at NAACL 2016. To view the complete publication list and presenter profile, please visit: `http://sarathchandar.in/`

References

[1] Marcin Andrychowicz and Karol Kurach. Learning efficient algorithms with hierarchical attentive memory. *CoRR*, abs/1602.03218, 2016.

[2] Sarath Chandar, Sungjin Ahn, Hugo Larochelle, Pascal Vincent, Gerald Tesauro, and Yoshua Bengio. Hierarchical memory networks. *CoRR*, abs/1605.07427, 2016.

[3] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[4] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio G. Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià P. Badia, Karl M. Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis. Hybrid computing using a neural network with dynamic external memory. *Nature*, advance online publication, October 2016. ISSN 0028-0836. doi: 10.1038/nature20101. URL `http://dx.doi.org/10.1038/nature20101`.

[5] Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. Learning to transduce with unbounded memory. In *Advances in Neural Information Processing Systems*, pages 1828–1836, 2015.

[6] Caglar Gulcehre, Sarath Chandar, Kyunghyun Cho, and Yoshua Bengio. Dynamic neural turing machine with soft and hard addressing schemes. *CoRR*, abs/1607.00036, 2016.

[7] Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. *CoRR*, abs/1506.07285, 2015.

[8] Jack W. Rae, Jonathan J. Hunt, Tim Harley, Ivo Danihelka, Andrew W. Senior, Greg Wayne, Alex Graves, and Timothy P. Lillicrap. Scaling memory-augmented neural networks with sparse reads and writes. *CoRR*, abs/1610.09027, 2016.

[9] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. *In Proceedings of NIPS*, 2015.

[10] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *In Proceedings Of The International Conference on Representation Learning (ICLR 2015)*, 2015. In Press.

[11] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. *CoRR*, abs/1603.01417, 2016.