Learning Efficient Algorithms with Hierarchical Attentive Memory

Marcin Andrychowicz[*]
Google Deepmind

Karol Kurach[*]
Google / University of Warsaw

Abstract

In this paper, we propose and investigate a novel memory architecture for neural networks called Hierarchical Attentive Memory (HAM). It is based on a binary tree with leaves corresponding to memory cells. This allows HAM to perform memory access in $\Theta(\log n)$ complexity, which is a significant improvement over the standard attention mechanism that requires $\Theta(n)$ operations, where n is the size of the memory. We show that an LSTM network augmented with HAM can learn algorithms for problems like merging, sorting or binary searching from pure input-output examples. In particular, it learns to sort n numbers in time $\Theta(n \log n)$ and generalizes well to input sequences much longer than the ones seen during the training. We also show that HAM can be trained to act like classic data structures: a stack, a FIFO queue and a priority queue.

1 Intro

Deep Recurrent Neural Networks (RNNs) have recently proven to be very successful in real-word tasks, e.g. machine translation (Sutskever et al., 2014) and computer vision (Vinyals et al., 2014). However, the success has been achieved only on tasks which do not require a large memory to solve the problem, e.g. we can translate sentences using RNNs, but we cannot produce reasonable translations of really long pieces of text, like books.

A high-capacity memory is a crucial component necessary to deal with large-scale problems that contain plenty of long-range dependencies. Currently used RNNs do not scale well to larger memories, e.g. the number of parameters in an LSTM (Hochreiter & Schmidhuber, 1997) grows quadratically with the size of the network's memory. In practice, this limits the number of used memory cells to few thousands.

It would be desirable for the size of the memory to be independent of the number of model parameters. The first versatile and highly successful architecture with this property was Neural Turing Machine (NTM) proposed by Graves et al. (2014). The main idea behind the NTM is to split the network into a trainable "controller" and an "external" variable-size memory. It caused an outbreak of other neural network architectures with external memories.

However, one aspect which has been usually neglected so far is the efficiency of the memory access. Most of the proposed memory architectures have the $\Theta(n)$ access complexity, where n is the size of the memory. It means that, for instance, copying a sequence of length n requires performing $\Theta(n^2)$ operations, which is clearly unsatisfactory.

1.1 Our contribution

We propose a novel memory module for neural networks, called Hierarchical Attentive Memory (HAM). The HAM module is generic and can be used as a building block of larger neural architectures.

[*] Equal contribution.

1st Workshop on Neural Abstract Machines & Program Induction (NAMPI), @NIPS 2016, Barcelona, Spain.

Its crucial property is that it scales well with the memory size — the memory access requires only $\Theta(\log n)$ operations, where n is the size of the memory. This complexity is achieved by using a new attention mechanism based on a binary tree with leaves corresponding to memory cells. The novel attention mechanism is not only faster than the standard one used in Deep Learning (Bahdanau et al., 2014), but it also facilities learning algorithms due to a built-in bias towards operating on intervals.

We show that an LSTM augmented with HAM is able to learn algorithms for tasks like merging, sorting or binary searching. In particular, it is the first neural network, which we are aware of, that is able to learn to sort from pure input-output examples and generalizes well to input sequences much longer than the ones seen during the training. Moreover, the learned sorting algorithm runs in time $\Theta(n \log n)$. We also show that the HAM memory itself is capable of simulating different classic memory structures: a stack, a FIFO queue and a priority queue.

2 Full version

The full version of this paper can found at https://arxiv.org/abs/1602.03218.

References

Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

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

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc VV. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, Dumitru. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.