Memory Networks

Federico Becattini
federico.becattini@unifi.it
University of Florence
Florence, Italy

Tiberio Uricchio
tiberio.uricchio@unifi.it
University of Florence
Florence, Italy

ABSTRACT

Memory Networks are models equipped with a storage component where information can generally be written and successively retrieved for any purpose. Simple forms of memory networks like the popular recurrent neural networks (RNN), LSTMs or GRUs, have limited storage capabilities and for specific tasks. In contrast, recent works, starting from Memory Augmented Neural Networks, overcome storage and computational limitations with the addition of a controller network with an external element-wise addressable memory. This tutorial aims at providing an overview of such memory-based techniques and their applications in multimedia. It will cover an explanation of the basic concepts behind recurrent neural networks and will then delve into the advanced details of memory augmented neural networks, their structure and how such models can be trained. We target a broad audience, from beginners to experienced researchers, offering an in-depth introduction to an important crop of literature which is starting to gain interest in the multimedia, computer vision and natural language processing communities.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**.

KEYWORDS

Memory Networks, Recurrent Neural Networks, Attention, Transformers

ACM Reference Format:
Federico Becattini and Tiberio Uricchio. 2022. Memory Networks. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3503161.3546972

1 INTRODUCTION

Memory Networks are neural networks that provide a storage component where information can be written and successively retrieved. The simplest form of memory network can be found in recurrent neural networks (RNN), like Long-Short Term Memories (LSTM) [11] or Gated Recurrent Units (GRU) [4]. Recurrent neural networks have found large use in multimedia applications, in particular to process sequential data such as text or videos [1–3, 12, 22].

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3546972

Such architectures however suffer from instability problems when processing long term dependencies and memory is a single hidden state vector that encodes all the temporal information. In recurrent neural networks, memory is addressable as a whole and the ability to address individual elements of knowledge is missing. This ability is necessary to apply algorithmic manipulation to the input data and perform complex tasks, especially over long timespans. In fact, state to state transition is unstructured and global. Being the state updated at each time-step, eventually it fails to model very long-term dependencies. Finally, the number of parameters is tied to the size of the hidden state. So, adding knowledge from the external environment, necessarily implies increasing the size of the state.

Recent works have proposed Memory Augmented Neural Networks [9, 24] to overcome the limitations of RNNs. The principal characteristic of these models is the usage of a controller network with an external elementwise addressable memory. This is used to store explicit information and access selectively relevant items. The memory controller is trained to dynamically manage memory content, optimizing predictions. Differently from RNNs, state to state transitions are obtained through read/write operations and a set of independent states is maintained. An important consideration is that in Memory Networks the number of parameters is not tied to the size of the memory, i.e., increasing the memory slots will not increase the number of parameters.

The first embodiment of a MANN has been Neural Turing Machine (NTM) [9], introduced to solve simple algorithmic tasks, demonstrating large improvements when compared to RNNs. The usage of an external memory, in fact, allows the network to store knowledge that cannot be forgotten unless deleted by the model itself. At each timestep the network can perform reasoning involving all previous observations and can perform data manipulation to emit its outputs. Follow-up works have extended and refined the formulation of the NTM. Recently, several declinations of MANNs have been proposed to tackle more complex problems such object tracking [14, 25], visual question answering [13, 15], person re-identification [20], action recognition [10], garment recommendation [5–7] and trajectory prediction [16–18].

A notable distinction between different types of MANNs can be found in its usage during training and inference. Two different approaches exist: episodic and persistent memory. An episodic memory is a working bank, where data gets manipulated across time-steps to perform active reasoning. The goal of the memory controller is to learn what to store and what to erase after the current sample has been observed. The memory bank will therefore contain a summary of an observed sequence. Similarly to RNNs, the memory is wiped out after each sequence has been observed and the output has been predicted. Examples of episodic Memory

Augmented Neural Networks are Neural Turing Machines [9], End-to-end Memory Networks [21], Key-Value Memory Networks [19] and SMEMO [18].

On the contrary, a persistent memory acts as a long-term storage where samples are gathered during the training phase. A Memory Augmented Neural Network with persistent memory need to learn what samples are important with respect to what has already been stored. Thus, the controller is trained to decide whether to add or remove observations in order to perform well on a downstream task. Examples of persistent Memory Augmented Neural Networks are MANTRA [16, 17], VQA with MANN [15], GR-MANN [5–7].

Interestingly, Memory Networks are tied with a concept that is nowadays extremely important for machine learning and multimedia applications: attention. To understand this, we can think at the definition of memory itself. Talking of memory in computer systems we refer to their storage capacity. In this sense computers have much better memory than people as they are able to store everything. Memory in humans instead is different. Human memory has a limited capacity, and thus attention determines what will be encoded. Human memory is rather the ability to select information and attend to that. Indeed, memory is attention over time. Attention and memory are important features of human cognition and they cannot operate without each other.

These intertwined concepts have been used differently in deep learning systems, with attention being at the center of the Transformer architecture [23]. Transformers are a type of Encoder-Decoder model that have been developed to solve the problem of sequence transduction, or neural machine translation.

For models to perform sequence transduction, it is necessary to have some sort of memory. Thanks to the self-attention mechanism, every element of an input sequence is matched against each other, in practice reducing a sequences to a set of tokens processed in parallel. As a direct consequence, transformers are not influenced by the distance between tokens, thus there is no long-term forgetting. This also makes transformers are more efficient than sequence-based models since matrix multiplication are performed between weights and whole sequences instead of individual tokens.

Nonetheless, limitations of transformers have been studied with respect to memory [8]. First of all, transformers are not able to track long sequences and process hierarchical inputs. If a long (potentially, unlimited) stream has to be observed, the complexity of the transformer will scale quadratically with the input length, without the possibility to update an internal state as the sequence is observed.

Another important observation is that only a fixed number of transformations can be applied to its internal states. Since both attention and feed-forward sublayer contain a fixed number of transformations, the total number of transformations between the input and output is limited by the depth of the model instead of depending on the complexity (length) of the input. Finally, at each layer, the representations for the input sequence are treated in parallel. As a consequence, a transformer does not leverage higher level representations from the past to compute the current representation. To address such limitations, a Memory Augmented Neural Network implementing a transformer with feedback memory has been proposed [8].

2 TUTORIAL OVERVIEW

The proposed tutorial aims at providing an overview of machine learning techniques exploiting memory networks and their applications in multimedia. The tutorial will cover an explanation of the basic concepts behind recurrent neural networks and will then delve into the advanced details of MANNs, their structure and how such models can be trained. We expect the event to be beneficial for the participants, offering an in-depth introduction to an important crop of literature which is starting to gain interest in the multimedia, computer vision and natural language processing communities. The tutorial targets a broad audience, from beginners to experienced researchers, providing basic concepts as well as overviews of advanced machine learning techniques.

3 BIOGRAPHIES

Federico Becattini is a Postdoctoral Researcher and Adjunct Professor at the University of Florence. His research interests focus on Autonomous Driving, Human Behavior Understanding and Fashion Recommendation, working with both international academic and industrial partners. He has co-authored more than 25 scientific papers in journals and international conferences. He also serves as reviewer for top-tier conferences and journals in multimedia and computer vision. Recently, he has also co-organized the workshops "Facial and Body Expressions" at ICPR2020, "Towards a Complete Analysis of People: From Face and Body to Clothes (T-CAP)" at ICIAP2021 and ICPR2022, "Multimedia Computing towards Fashion Recommendation (MCFR)" at ACM MM 2022 and "1st International Workshop and Challenge on People Analysis: From Face, Body and Fashion to 3D Virtual Avatars (WCPA)" at ECCV 2022. His recent work has focused on exploiting innovative deep learning architectures called Memory Augmented Neural Networks (MANN), which he has employed in several international publications including top-tier conferences and journals as CVPR, IEEE TPAMI and ACM TOMM, covering topics such as autonomous driving and fashion recommendation. He also held tutorials on the subject at ICIAP 2022 and ACM MM 2022.

Tiberio Uricchio is an adjunct professor at the University of Florence and CEO Co-Founder of Small Pixels, an academic spinoff startup on a mission to improve perceptual quality of videos. He is co-author of more than 30 publications in international scientific journals and conferences on computer vision and multimedia. He presented three tutorials on social image tagging in international conferences (ACM MM 2015, CVPR 2016 and ICIAP 2017). He was the recipient of the best demo award at the ACM MM 2019 conference, best poster award at ACM ICMR 2020 with his research on improving video quality using generative adversarial networks (GANs). He is Associate Editor of the international journal Multimedia Tools and Application, organizer of the RISS2020 workshop, and he regularly contributes as area chair and reviewer for international conferences and journals. His research interests include understanding images and videos, multimedia and deep learning.

ACKNOWLEDGMENTS

This work was supported by the European Commission under European Horizon 2020 Programme, grant number 951911 - AI4Media

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[3] Federico Becattini, Tiberio Uricchio, Lorenzo Seidenari, Lamberto Ballan, and Alberto Del Bimbo. 2020. Am I done? Predicting action progress in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 4 (2020), 1–24.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[5] Lavinia De Divitiis, Federico Becattini, Claudio Baecchi, and Alberto Del Bimbo. 2022. Disentangling Features for Fashion Recommendation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022).

[6] Lavinia De Divitiis, Federico Becattini, Claudio Baecchi, and Alberto Del Bimbo. 2021. Style-Based Outfit Recommendation. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 1–4.

[7] Lavinia De Divitiis, Federico Becattini, Claudio Baecchi, and Alberto Del Bimbo. 2020. Garment Recommendation with Memory Augmented Neural Networks. In *Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10-15, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12662)*. Springer, 282–295. https://doi.org/10.1007/978-3-030-68790-8_23

[8] Angela Fan, Thibaut Lavril, Edouard Grave, Armand Joulin, and Sainbayar Sukhbaatar. 2020. Addressing some limitations of transformers with feedback memory. *arXiv preprint arXiv:2002.09402* (2020).

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

[10] Tengda Han, Weidi Xie, and Andrew Zisserman. 2020. Memory-augmented dense predictive coding for video representation learning. *arXiv preprint arXiv:2008.01065* (2020).

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

[12] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[13] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*. 1378–1387.

[14] Zihang Lai, Erika Lu, and Weidi Xie. 2020. MAST: A memory-augmented self-supervised tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6479–6488.

[15] Chao Ma, Chunhua Shen, Anthony Dick, Qi Wu, Peng Wang, Anton van den Hengel, and Ian Reid. 2018. Visual question answering with memory-augmented networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6975–6984.

[16] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. 2020. MANTRA: Memory Augmented Networks for Multiple Trajectory Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[17] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. 2020. Multiple trajectory prediction of moving agents with memory augmented networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[18] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. 2022. SMEMO: Social Memory for Trajectory Forecasting. *arXiv preprint arXiv:2203.12446* (2022).

[19] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126* (2016).

[20] Federico Pernici, Matteo Bruni, and Alberto Del Bimbo. 2020. Self-supervised on-line cumulative learning from video streams. *Computer Vision and Image Understanding* (2020), 102983.

[21] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*. 2440–2448.

[22] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 27 (2014).

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[24] Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916* (2014).

[25] Tianyu Yang and Antoni B Chan. 2018. Learning dynamic memory networks for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 152–167.