ON MEMORY IN HUMAN AND ARTIFICIAL LANGUAGE PROCESSING SYSTEMS

Aida Nematzadeh, Sebastian Ruder & Dani Yogatama*
{nematzadeh,ruder,dyogatama}@google.com
DeepMind

ABSTRACT

Memory in humans and artificial intelligence (AI) systems has similar functions—both are responsible for encoding, retrieving, and storing of information. While memory in humans has specialized systems for different functions (e.g., working memory, semantic memory, episodic memory), memory in AI systems is often implicitly represented in the weights of parametric neural networks. Focusing on language processing systems, we argue that this property makes it hard for AI systems to generalize across complex linguistic tasks. We consider the separation of computation and storage as necessary, suggest desired properties of the storage system, and discuss the benefit of integrating different types of human memory (as separate modules) into next-generation language processing systems.

1 INTRODUCTION

The uniquely human ability to acquire, comprehend, and produce language is highly dependent on our memory: it stores our knowledge of various aspects of language and enables us to follow a conversation (Carroll, 2007). We can broadly categorize human memory processes into two categories: *computation* corresponding to encoding and retrieval of information and *storage* responsible for maintaining information over time. Interestingly, human memory consists of specialized systems for different functions; *e.g.*, the episodic memory is responsible for the memory of personal experiences and events, the semantic memory stores information about a language sound system, and the working memory is involved when comprehending sentences. Moreover, the storage of information in human memory is imperfect—we forget as we learn.

Similarly, AI systems require processes for both computation and storage of information. Although previous work has explored separating computation and storage (*e.g.*, Graves et al., 2014; Weston et al., 2015; Sprechmann et al., 2018), the dominant approach focuses on building large neural models where both computation and storage are modeled implicitly in model parameters (i.e., neural network weights). Such an AI system optimizes an objective function for a specific task to determine what to encode (write into memory) and requires additional supervision to retrieve knowledge (Bosselut et al., 2019), which limits their ability to generalize to new tasks.

We argue that achieving human-like linguistic behavior requires models that consider computation and storage as separate processes. Furthermore, the storage should be a dynamic module that is structured to facilitate faster search and is capable of managing its size complexity by forgetting or compressing (abstracting away). We consider such a modular design to be a necessary structural bias and a more viable alternative to training an ever larger neural network.

In addition, none of the existing work implements an integrated model of specialized memory systems that are involved in human language processing (*e.g.*, working memory, semantic memory, episodic memory). We conjecture that language processing models should incorporate different memory systems with specific functions as independent modules. An intuitive reason is that it is simpler to train a memory module to perform one function (*e.g.*, store events) as opposed to training an end-to-end model that needs to decide what functions are needed to perform given a task.

*All authors contributed equally.

Memory	Functions in human language processing	AI tasks	AI models
Episodic memory	Tracking verbs, events, and ongoing discourse	Story comprehension, discourse analysis, event detection	GEM, A-GEM, MbPA, MbPA++, Matching Networks
Semantic memory	Storing words, sounds, and pragmatic aspects of language	Knowledge base construction, open-domain question answering, common sense reasoning	NELL, LSC, Progress & Compress
Working memory	Sentence comprehension, understanding references to pronouns	Reading comprehension, coreference resolution, word sense disambiguation, entity linking, parsing	LSTM, DNC, Memory Networks

Table 1: Memory types, their function in human language processing, AI tasks where such functions might be necessary, and AI models that implement analogous functions. Model references (left to right, top to bottom): Lopez-Paz & Ranzato (2017); Chaudhry et al. (2019); Sprechmann et al. (2018); Masson et al. (2019); Vinyals et al. (2016); Mitchell et al. (2015); Chen et al. (2015); Schwarz et al. (2018); Hochreiter & Schmidhuber (1997); Graves et al. (2014); Weston et al. (2015).

In general, we believe that focusing on improving memory in AI systems is timely since it mostly relies on algorithmic advances—compared to the hardware or data requirements that impede making progress in language grounding, interactions, and others.

2 Human Memory

Before the 1970s, experimental research primarily considered memory as a unified system with a single function. However, recent neuroscience research shows that the human brain consists of memory systems with distinct functions: One brain pathway encodes our personal experience and another one keeps our knowledge of words (for a review, see Eichenbaum, 2012). Here, we focus on memory systems that are highly involved in language acquisition and processing, *i.e.*, declarative and working memory.

Declarative memory is long-term and consists of two different memory systems of *knowledge* and *personal experience*: (1) *Semantic memory* refers to our knowledge of facts. For example, our knowledge of sounds, words, and concepts is encoded in semantic memory. (2) *Episodic memory* encodes our individual experiences of the world and gives us the capacity to replay such experiences and imagine future ones. Moreover, the episodic memory stores information about *when*, *where*, and in which *order* events occur. Finally, while the hippocampus is responsible for forming declarative memories, long-term memories are not stored in the hippocampus (Rolls, 2000).

Working memory is the temporary storage used for processing information (*e.g.*, Baddeley, 1986). For example, to follow a conversation, we need to store some information about what our conversation partner has shared.

Forgetting is a crucial aspect of human memory; paradoxically, not remembering everything we encounter facilitates the learning and retrieval of information (Bjork, 1994). For example, forgetting enables us to form abstractions—knowing what a cat is without needing to remember the details of all cats we have seen (Vlach et al., 2008).

3 Memory in AI

State-of-the-art AI models (*i.e.*, neural networks) work well on a single dataset, but they often struggle with long-term dependencies and fail to reuse previously acquired knowledge. A main approach to mitigate this problem is to augment neural networks with a memory module. However, the term "memory" itself is used with multiple different interpretations. We classify memory-augmented neural networks into two broad categories based on their connections to the human memory system and summarize how each model implements the read and write mechanisms for its memory module. Table 1 lists examples of human memory functions, AI tasks that require similar memory, and models that implement similar memory modules.

Models with declarative memory. *Episodic memory* models have been largely used to prevent catastrophic forgetting (McCloskey & Cohen, 1989; Ratcliff, 1990) in a lifelong learning setup on multiple tasks. The episodic memory module is a key-value module that stores (a subset of) training examples from each task that have been seen by the model. In lifelong language learning, the key is typically a learned vector representation of the example and the value is its original representation (*e.g.*, a string for a natural language example). The strategy to select which examples to write into the memory is often simple (*e.g.*, the last N examples seen for a task, random decision whether to write or not), and more sophisticated methods seem to not perform as well due to a coverage issue (i.e., the need to ensure the distribution of stored examples is representative of the distribution of true examples; Isele & Cosgun 2018; Masson et al. 2019). The episodic memory is used to either constraint the gradient updates (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019), locally adapt the base model to a new test example (Vinyals et al., 2016; Sprechmann et al., 2018), or for experience replay (Wang et al., 2019; Masson et al., 2019).

Semantic memory is often represented as knowledge bases in artificial intelligence. Work on using knowledge bases to improve AI systems has progressed from the early decades of AI with the design of rule-based expert systems (Jackson, 1990) to the modern era prior to the deep learning revolution (Banko & Etzioni, 2007; Mitchell et al., 2015). A knowledge base is seen as a component of many lifelong learning algorithms (Chen et al., 2015), some of which represent it with a neural network (Rusu et al., 2016; Kaiser et al., 2017; Schwarz et al., 2018). In addition, many researchers (Mikolov et al., 2013b; Kulkarni et al., 2015; Hamilton et al., 2016) have remarked upon the ability of word embeddings (Mikolov et al., 2013a) to represent concepts and their relations, which resemble information stored in a semantic memory.

Models with working memory. We consider a model to be augmented with a working memory mechanism if the memory module is mainly used to store local context (e.g., a conversation thread in a dialog system, an article context in language modeling). A classic model is the Long Short-Term Memory (LSTM; Hochreiter & Schmidhuber, 1997) unit, which has been a driving force behind advances in sequence modeling (including language processing). An LSTM network has a memory cell (*i.e.*, a vector) that is modulated by three gates: input, forget, and output. These gates regulate what is written and read from the memory cell, and the entire network is trained end to end on a particular task. This memory cell serves as a working memory to store information across multiple timesteps when processing a sequence.

Another influential model is the Differentiable Neural Computer (DNC; Graves et al., 2014), which augments a neural network with a memory module in the form of an external memory matrix. Access to the memory is regulated by a neural network controller. The controller has three main types of heads that are implemented as differentiable attentions: content lookup (read head), memory allocation (write head), and an extra head that tracks transitions between written memory locations (temporal transition head). These heads have been suggested to operate in a way similar to the hippocampus (e.g., fast memory modification, sparse weightings to increase representational capacity, the formation of temporal links). DNC works well on a toy question answering task that requires reasoning across multiple sentences. Many other related variants of DNC are used in natural language processing, *e.g.*, Memory Networks (Weston et al., 2015; Sukhbaatar et al., 2015), stack-augmented neural networks (Joulin & Mikolov, 2015; Grefenstette et al., 2015; Yogatama et al., 2018), and cache-based sequence models (Graves et al., 2014; Merity et al., 2017; Khandelwal et al., 2020). Nevertheless, learning to use the controller remains a challenging task and is currently outperformed by simply writing at uniformly spaced intervals (Le et al., 2019).

Forgetting. Catastrophic forgetting—where a model performs poorly on examples from a dataset that is encountered in a distant past (after seeing multiple other datasets)—is a common problem in AI systems (Yogatama et al., 2019; Greco et al., 2019). On the other hand, the ability to forget is a crucial part of the human memory system. A recent study shows that there exist forgettable examples—defined as examples that transition from being predicted correctly to incorrectly—when training a model on a particular dataset (Toneva et al., 2019). Moreover, identifying such forgettable examples appears key to train a model that efficiently generalizes with fewer examples.

4 The Gap between Human Memory and AI

In this section, we discuss how we might bridge the gap between human memory systems and memory-augmented neural networks as a possible way to improve AI models. We focus on models of computation and storage as well as improvements to core operations in the memory.

As discussed in §1, a recent line of research (Devlin et al., 2019; Radford et al., 2019) suggests that increasing the number of parameters in a neural network is sufficient for improving AI systems and an explicit memory module is not needed. In these models, the computation and storage systems are fundamentally intertwined as neural network parameters (weights). It has been shown that relatively low-level knowledge (*e.g.*, word similarity)—which would be stored in a semantic memory—can be extracted from these models in an unsupervised way (Petroni et al., 2019; Bouraoui et al., 2020). However, the extraction of more complex ones (*e.g.*, common sense) requires explicit supervision (Bosselut et al., 2019), which indicates a limitation of such a model.

In contrast to the above approach, we believe that **the separation of computation and storage** is necessary to incorporate structural bias into AI systems. In memory-augmented neural networks (§3), the separation between computation and storage enables the computation module to focus on processing examples and the storage module to focus on learning to represent a persistent storage efficiently. Lillicrap & Santoro (2019) argue that a modular treatment of computation and storage is also useful from an optimization perspective—particularly for long-term credit assignments—to propagate gradient information from the present with high fidelity. Separating neural network components based on their functions has been found useful in other contexts, *e.g.*, for attention with a query, key, and value component (Daniluk et al., 2017). In this position paper, we make no claim about the form of the storage module. For example, it is possible to implement it as a neural network such as in recent work (Schwarz et al., 2018) to facilitate efficient compression of information.

The three main operations that are introduced by the separation of computation and storage are encoding (writing), retrieval (reading and searching), and storing (compression and forgetting). We argue that **the storage system should be implemented as a module that has a structured space** to facilitate faster search, similar to human memory retrieval. Furthermore, the storage needs to be able to **dynamically manage its size complexity by forgetting or compressing to form abstractions**. Compression is needed from a practical standpoint, since it is unrealistic to have a neural network that keeps growing in its storage size. Forgetting with an explicit memory module may also enable fine-grained control of what is stored, which is important for privacy purposes (Carlini et al., 2019). Moreover, our environment imposes a specific schedule on what we observe and what we can learn. The interaction of this schedule and our memory not only prevents the notion of catastrophic forgetting, but also improves our learning and retrieval of information (Anderson & Milson, 1989; Bjork, 1994). Automatically discovering a task-independent schedule (*e.g.*, via meta learning, curriculum learning) could be important to mitigate catastrophic forgetting in AI systems.

Finally, existing systems only implement a particular type of human memory systems—either a working memory module to capture long-term dependencies or a declarative memory module to remember facts or examples of a given task. It has been argued that intelligent behaviors rely on multiple memory systems (Tulving, 1985; Rolls, 2000). Next generation models should seek to **integrate different memory types (as separate modules) in a single language processing model** with functions that closely resemble memory in human language processing. We see this integration to be crucial for an AI system to be able to combine information from a long-term local context (e.g., in a Wikipedia article) and persistent global context (e.g., other Wikipedia articles). Such an integration is needed for tasks such as answering factual questions, conversing about a particular topic, and others.

5 Conclusion

We considered similarities and differences of memory implementations in human and artificial language processing systems. We argued for the separation of computation and storage and suggested necessary properties of the storage system in the form of a structured space and the ability to forget and compress (form abstractions). We conjectured that next-generation language processing systems would integrate different types of memory as separate modules with functions that resemble memory in human language processing.

REFERENCES

John R Anderson and Robert Milson. Human memory: An adaptive perspective. *Psychological Review*, 96(4):703, 1989.

Alan Baddeley. Oxford psychology series, no. 11. working memory. new york, ny, us, 1986.

Michele Banko and Oren Etzioni. Strategies for lifelong knowledge extraction from the web. In *Proceedings of the 4th international conference on Knowledge capture*, pp. 95–102, 2007.

Robert A Bjork. Memory and metamemory considerations in the training of human beings. In Janet Ed Metcalfe and Arthur P Shimamura (eds.), *Metacognition: Knowing about knowing.*, pp. 185–205. The MIT Press, 1994.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli elikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. Inducing Relational Knowledge from BERT. In *Proceedings of AAAI 2020*, 2020.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In *28th USENIX Security Symposium (USENIX Security 19)*, 2019.

David Carroll. *Psychology of language*. Nelson Education, 2007.

Arslan Chaudhry, Marc Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient Lifelong Learning with A-GEM. In *Proceedings of ICLR 2019*, 2019.

Zhiyuan Chen, Nianzu Ma, and Bing Liu. Lifelong Learning for Sentiment Classification. In *Proceedings of ACL*, pp. 750–756, 2015.

Michal Daniluk, Tim Rocktäschel, Johannes Weibl, and Sebastian Riedel. Frustratingly Short Attention Spans in Neural Language Modeling. In *Proceedings of ICLR 2017*, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL 2019*, 2019.

Howard Eichenbaum. Memory systems. *Handbook of Psychology, Second Edition*, 3, 2012.

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

Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. Psycholinguistics meets Continual Learning : Measuring Catastrophic Forgetting in Visual Question Answering. In *Proceedings of ACL 2019*, pp. 3601–3605, 2019.

Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. Learning to transduce with unbounded memory. In *Proc. of NIPS*, 2015.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1489–1501, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1141. URL `https://www.aclweb.org/anthology/P16-1141`.

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

David Isele and Akansel Cosgun. Selective Experience Replay for Lifelong Learning. In *Proceedings of AAAI 2018*, 2018.

Peter Jackson. Introduction to expert systems. 1990.

Armand Joulin and Tomas Mikolov. Inferring algorithmic patterns with stack-augmented recurrent nets. *arXiv preprint*, 2015.

Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to Remember Rare Events. In *Proceedings of ICLR 2017*, 2017.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *Proc. of ICLR*, 2020.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 625–635, 2015.

Hung Le, Truyen Tran, and Svetha Venkatesh. Learning to Remember More with Less Memorization. In *Proceedings of ICLR 2019*, 2019.

Timothy P Lillicrap and Adam Santoro. Backpropagation through time and the brain. *Current opinion in neurobiology*, 55:82–89, 2019.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient Episodic Memory for Continuum Learning. In *Proceedings of NIPS 2017*, 2017.

Cyprien De Masson, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. Episodic Memory in Lifelong Language Learning. In *Proceedings of NeurIPS 2019*, 2019.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *Proc. of ICLR*, 2017.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013a.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013b. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/N13-1090`.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language Models as Knowledge Bases? In *Proceedings of EMNLP 2019*, 2019. URL `http://arxiv.org/abs/1909.01066`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review*, 97(2):285, 1990.

Edmund T. Rolls. Memory systems in the brain. *Annual Review of Psychology*, 51(1):599–630, 2000.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive Neural Networks. *arXiv preprint arXiv:1606.04671*, 2016.

Jonathan Schwarz, Jelena Luketina, Wojciech M. Czarnecki, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & Compress : A scalable framework for continual learning. In *Proceedings of ICML 2018*, 2018.

Pablo Sprechmann, Siddhant M Jayakumar, Jack W Rae, Alexander Pritzel, Benigno Uria, and Oriol Vinyals. Memory-based Parameter Adaptation. In *Proceedings of ICLR 2018*, 2018.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Proceedings of 29th Conference on Neural Information Processing Systems*, pp. 2440–2448, 2015.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An Empirical Study of Example Forgetting during Deep Neural Network Learning. In *Proceedings of ICLR 2019*, 2019.

E. Tulving. How many memory systems are there? *American Psychologist*, 40:385–398, 1985.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. In *Proceedings of NIPS 2016*, 2016.

Haley A Vlach, Catherine M Sandhofer, and Nate Kornell. The spacing effect in childrens memory and category induction. *Cognition*, 109(1):163–167, 2008.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, William Yang Wang, and Santa Barbara. Sentence Embedding Alignment for Lifelong Relation Extraction. In *Proceedings of NAACL 2019*, 2019.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory Networks. In *Proceedings of ICLR 2015*, 2015.

Dani Yogatama, Yishu Miao, Gabor Melis, Wang Ling, Adhiguna Kuncoro, Chris Dyer, and Phil Blunsom. Memory architectures in recurrent neural network language models. In *Proc. of ICLR*, 2018.

Dani Yogatama, Cyprien de Masson D'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. Learning and Evaluating General Linguistic Intelligence. *arXiv preprint arXiv:1901.11373*, 2019.