The Tensor Memory Hypothesis

Volker Tresp and Yunpu Ma
Siemens AG and Ludwig Maximilian University of Munich

Abstract. We discuss memory models which are based on tensor decompositions using latent representations of entities and events. We show how episodic memory and semantic memory can be realized and discuss how new memory traces can be generated from sensory input: Existing memories are the basis for perception and new memories are generated via perception. We relate our mathematical approach to the hippocampal memory indexing theory. We describe the first mathematical memory models that are truly declarative by generating explicit semantic triples describing both memory content and sensory inputs. Our main hypothesis is that perception includes an active semantic decoding process, which relies on latent representations of entities and predicates, and that episodic and semantic memories depend on the same decoding process.

1 Introduction

It still is a great puzzle that the brain can easily deal with continuous high-dimensional signals, such as images and sounds, but at the same time relies on very discrete concepts like "Jennifer Aniston" [40]. The latter faculty is most apparent in language, where we clearly make statements about discrete entities: "Jennifer Aniston is an American actress, who became famous through the television sitcom *Friends*". We argue that a key concept might be representation learning, i.e., that each discrete entity is represented as a vector of real numbers. Based on the latent representations, different declarative memory functions can be implemented, in particular episodic memory and semantic memory. In this paper, we use the framework of tensor decompositions to mathematically describe the declarative nature of memory and perception. The effectiveness of tensor models for the realization of technical memory functions is well established [33, 31] and here we explore their relevance in modelling human memories. Representation learning might also be the basis for perception: New memories are formed by mapping sensory inputs to latent event representations which can be stored as episodic memories; these can then also be semantically decoded, using the tensor framework. Thus sensory impressions can be interpreted by the brain, become declarative and thus can verbally be described. Perception, in form of a semantic decoding of sensor input, as well as episodic and semantic memory all depend on the same latent representations, since the brain must know about entities and entity classes to understand and interpret new sensory inputs. We describe the first mathematical memory models that are truly declarative by generating explicit semantic triples describing both memory content and sensory inputs.

In contrast to a previous paper [53], here we relate our mathematical approach to the hippocampal memory indexing theory, which is one of the main theories for forming episodic memories [51, 52]. In summary, our main hypothesis is that perception includes an active semantic decoding process which relies on latent representations of entities and predicates and that episodic and semantic memories depend on the same decoding process.

The paper is organized as follows. In the next section we describe the tensor memory models and in Section 3 we discuss different memory operations and illustrate how sensory inputs can generate new episodic memories. In Section 4 we discuss the potential relevance of the model for human perception and human memories. Section 5 contains our conclusions where we make the point that the brain might use the latent representations in many additional functions, such as prediction, planning, and decision making.

2 Tensor Memories

Let e_i be a symbol that represents entity i. We associate with e_i a latent vector $\mathbf{a}_{e_i}$.[1] Similarly, we assume that a predicate p is represented by a symbol e_p with latent representation $\mathbf{a}_{e_p}$. We now propose that the probability that a triple statement[2] (s, p, o) is part of semantic memory ("facts we know") can be modelled as $P((s, p, o)) = sig(\theta_{s,p,o})$, where $sig(\cdot)$ is the logistic function and $\theta_{s,p,o}$ is calculated as a function of the latent representations of the involved predicate and entities,

$$\theta_{s,p,o} = f^s(\mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o}).$$

The function $f^s(\cdot)$ can be derived from tensor factorization or can be realized as a neural network. An overview of models can be found in [31]. To simplify the discussion we focus on the Tucker model with

$$f^s(\mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o}) = \sum_{r_1=1}^{\tilde{r}} \sum_{r_2=1}^{\tilde{r}} \sum_{r_3=1}^{\tilde{r}} a_{e_s,r_1}\ a_{e_p,r_2}\ a_{e_o,r_3}\ g(r_1, r_2, r_3).$$

Here, $g(\cdot)$ is an element of the core tensor. The effectiveness of tensor models for the realization of semantic memory functions is well established [31].

For an episodic memory ("facts we remember") we add a representation for time. Let e_t be the symbol for time instance t. Its latent representation then is $\mathbf{a}_{e_t}$. The probability of the observation of triple (s, p, o) at time t is $P((s, p, o, t)) = sig(\theta_{s,p,o,t})$ with $\theta_{s,p,o,t} = f^e(\mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o}, \mathbf{a}_{e_t})$ and Tucker decomposition

$$f^e(\mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o}, \mathbf{a}_{e_t}) = \sum_{r_1=1}^{\tilde{r}} \sum_{r_2=1}^{\tilde{r}} \sum_{r_3=1}^{\tilde{r}} \sum_{r_4=1}^{\tilde{r}} a_{e_s,r_1}\ a_{e_p,r_2}\ a_{e_o,r_3}\ a_{e_t,r_4}\ g(r_1, r_2, r_3, r_4). \tag{1}$$

[1] An entity can, e.g., be a particular person or a location like "Munich"

[2] An example would be *(Jack, knows, Mary)*

Whereas the episodic memory would be able to retrieve the fact that *(Jack, diagnosed, Diabetes, Yesterday)*, the semantic memory would represent *(Jack, diagnosed, Diabetes)*. Note that the tensor models can reconstruct known memories by assigning a high probability to facts known to be true but they also assign high probabilities to facts which follow the patterns found in the memory systems; thus they realize a form of probabilistic inductive reasoning [32]. As an example, consider that we know that Max lives in Munich. The probabilistic materialization that happens in the factorization should already predict that Max also lives in Bavaria and in Germany. Generalization from existing facts by probabilistic inductive reasoning is of particular importance in perception, where the predicted triple probability might serve as a prior for information extraction [9, 3]. There is a certain danger in probabilistic materialization, since it might lead to overgeneralizations, reaching from prejudice to false memories [41, 25].

3 Memory Operations and Sensory Inputs

3.1 Querying and Association

The models so far calculated the probability that a triple or quadruple is true. For querying, we re-normalize the model and $\exp \beta\theta_{s,p,o}$ becomes proportional to the probability that the memory *produces the triple* (s, p, o) as a generative model. If the inverse temperature β is high, only highly likely triples are produced.[3][4] Thus query answering can be implemented as a sampling process [53]. The set of generated triples then represents likely query answers. As an example, by fixing $s = $ *Jack* and $p = $ *likes*, querying would generate likely entities as objects that *Jack* is fond of.

A recall of a past previous events at time t' simply means that triples are generated from episodic memory using Equation 1 with $\mathbf{a}_{e_{t'}}$ fixed. The pattern $\mathbf{a}_{e_{t'}}$ might have been generated by the activation of $e_{t'}$ or by some associative process.

To recall what is known about an entity i, one applies $\mathbf{a}_{e_i}$ and generates triples $(s = i, p, o)$ from semantic memory, describing facts about i. Alternatively, one can generate quadruples $(s = i, p, o, t)$ from episodic memory, describing events that i participated. The set of triples could be used by the language modules (in the brain this would be Broca's area) to generate language, i.e. to verbally describe an episodic or semantic memory. It is also straightforward to generate an association, e.g., to retrieve entities semantically similar to, let's say, $\mathbf{a}_{e_i}$.

[3] For nonnegative models we can replace $\exp \beta\theta_{s,p,o}$ with $\theta_{s,p,o}$.

[4] Note that in associative memories a similar concentration on the most likely interpretations can be achieved by exponentiation [16] or by using polynomial energy functions [23].

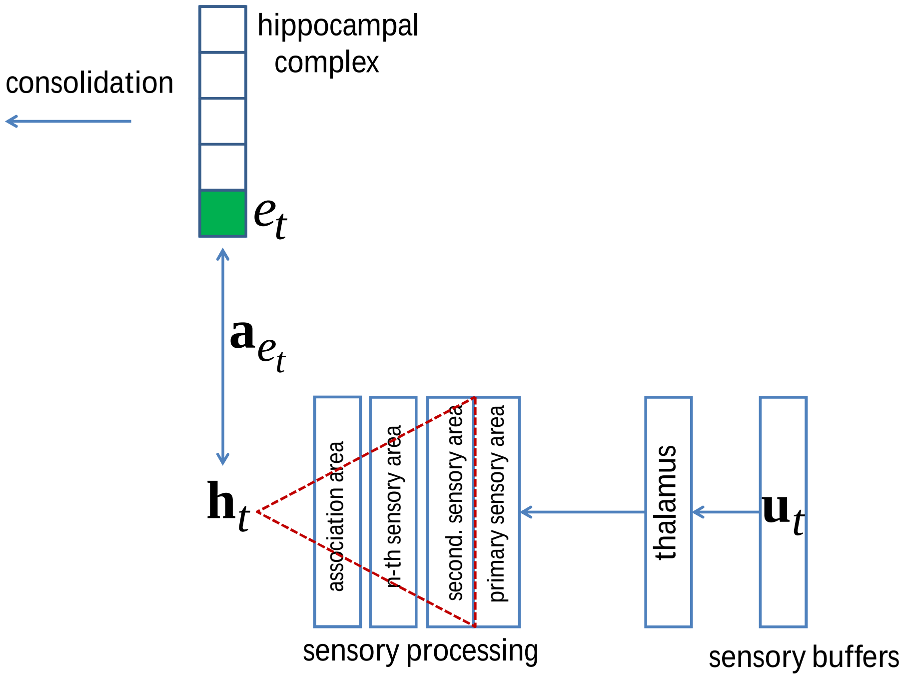

Fig. 1. The figure sketches the hippocampal memory indexing theory. Here, $\mathbf{u}_t$ represents the sensory input which is relayed through the thalamus (except olfaction), a part of the forebrain, and then forms the input to the different sensory processing paths. Based on sensory processing, the latent representation for the time instance $\mathbf{h}_t$ is generated. As indicated, in the earlier processing layers most nodes are activated, while at the higher layers, only a smaller number of scene-specific nodes are active. In the mathematical model, this mapping is described as $\mathbf{h}_t = \mathbf{f}^M(\mathbf{u}_t)$. For sensory inputs to be stored as episodic memory —because the event is e.g., significant, novel, or attached with emotion— an index e_t is formed in the hippocampus. $\mathbf{a}_{e_t} = \mathbf{h}_t$ is the sparse weight pattern between the index and the sensory layers which binds the sensory activation patterns with the index e_t. In recalling a memory from a past time t', the corresponding index $e_{t'}$ is activated, which then reactivates an approximate $\mathbf{h}_{t'}$ via weight patterns $\mathbf{a}_{e_{t'}}$. In memory consolidation, the index and its connection pattern find a representation in neocortex, a process which is supposed to occur during sleep [50]. In neocortex, indices are topologically ordered.

3.2 New Memories and Semantic Decoding of Sensory Inputs

In some applications, information enters the memory system via sensory inputs, like vision and audition. The sensory input generates a latent representation

$$\mathbf{h}_t = \mathbf{f}^M(\mathbf{u}_t) \tag{2}$$

where $\mathbf{u}_t$ represents the sensory memory (e.g., the image) at time t, and where the multivariate function $\mathbf{f}^M(\cdot)$ might be realized by a deep convolutional neural network (CNN). This latent representation is stored as $\mathbf{a}_{e_t} \leftarrow \mathbf{h}_t$ and can be semantically decoded by substituting this $\mathbf{a}_{e_t}$ into the episodic tensor model in Equation 1.

4 A Cognitive Perspective

4.1 Forming Episodic Memories

We relate our mathematical approach to the hippocampal memory indexing theory, which is one of the main theories for forming episodic memories [51, 52] (Figure 1). As in Equation 2, a sensory input $\mathbf{u}_t$ activates a hierarchical multi-layered activation pattern $\mathbf{h}_t$ in the sensory layers of the cortex. As indicated in Figure 1, this pattern is more specific towards the higher sensory layers. Thus, as also known from deep convolutional neural networks (CNNs), for a given sensory input, only few (but varying) nodes are activated on top of the hierarchy. Following the theory, an index e_t is formed in the hippocampus for sensory impressions important to be remembered by forming a map from the higher layers in the sensory hierarchy to the index. Since the representation of a sensory input becomes more sparse toward higher order sensor processing layers, the representations for different sensory inputs become also more orthogonal towards higher layers. In a sense, the index is simply the top of the hierarchy: the complete index might involve a large number of neurons but a given sensory input only activates a small number of them, realizing a sparse (pseudo-) orthogonal distributed representation. In machine learning terms, e_t is a new class trained by one-shot learning where the activation pattern $\mathbf{h}_t$ is reflected as a typically sparse weight pattern $\mathbf{a}_{e_t}$. A feature typically not present in CNNs is that the connection is assumed to be bidirectional and the activation of index e_t can reactivate the pattern $\mathbf{h}_t = \mathbf{a}_{e_t}$ in memory recall by back projection, retrieving a sensory pattern of the past episode. The index performs a binding of memory patterns. Instead of the hippocampus indexing all of neocortex, it is probable that it participates in a hierarchical indexing scheme whereby it indexes the association cortex, which then indexes the rest of neocortex [52]. An important biological function would be to recall previous episodic memories that are similar to $\mathbf{h}_t$ and to associate past emotional impressions, predictions (what happened next), past actions, and past outcomes of those actions.

The formation of the index in the hippocampus is highly complex [52, 43] and one should interpret the index and the associated weight patterns as functional descriptions, and not as biological implementations of indices or explicit synaptic weights. Details on the biological aspects of the index formation in the hippocampus can be found in the Appendix.

We can look at Equation 2 as an encoder for sensory inputs, followed by a semantic decoder with

$$P(s, p, o|t) \propto \exp \beta f^e(\mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o}, \mathbf{a}_{e_t} = \mathbf{h}_t) \quad (3)$$

which assumes the form of a generalized nonlinear model. Encoder-decoder networks are the basis for neural machine translation systems [7]. [3] is an example where semantic decoding is applied to real-world images.

Evidence for time cells in the hippocampus (CA1) has recently been found [11]. In fact, it has been observed that the adult macaque monkey forms a few thousand new neurons daily [14], possibly to encode new information. These time cells might be related to the indices e_t.

4.2 Memory Consolidation

In a consolidation process, the index and its connection pattern might also build representations in neocortex, where memories form topological maps. This is an involved process, likely happening during sleep and involving specific oscillation patterns such as sharp waves and ripples [49, 13, 46, 26]. The relevance of neural oscillations, in particular the coupling of theta and gamma rhythms, has been discussed in [34].

A simple mechanism would be that the indices reactivate past memory activation patterns which then trigger new index formations in neocortex by replay. Memory consolidation is a controversial issue and there exist several theories concerning the system consolidation of memory. In the standard theory, memories from the hippocampal region, where memories are first encoded, are moved to the neo-cortex in a more permanent form of storage [49]. In this process the hippocampus is "teaching" the cortex, eventually making the memory hippocampus-independent: The neocortex then can support memory indefinitely. An alternative view is given by the multiple trace theory (MTT), which clearly distinguishes between episodic and semantic memory [28, 29]. MTT argues that the hippocampus is always involved in the retrieval and storage of *episodic* memories but that it is less necessary during the encoding and use of *semantic* memories. It is thought that semantic memories, including basic information encoded during the storage of episodic memories, can be established in structures apart from the hippocampal system such as the neo-cortex in the process of consolidation. Hence, while proper hippocampal functioning is necessary for the retention and retrieval of episodic memories, it is less necessary during the encoding and use of semantic memories.

Replay of episodic memories might not only transfer episodic memories but might also contribute to other forms of training of brain structures, as suggested in the complementary learning systems (CLS) theory, which also relies on memory replay [27, 24]. According to that theory, effective learning requires two complementary systems: one, located in the neocortex, serves as the basis for the gradual acquisition of structured knowledge about the environment, while the other, centered in the hippocampus, allows for rapid learning of the specifics of individual items and experiences.

4.3 Representations for Entities

We would suggest that indices for entities e_i are formed in the hippocampal area, as well. [39] identified concept cells with focussed responses to individuals like the actresses Jennifer Aniston or Halle Berry. These concept cells might also eventually be consolidated in neocortex where they become part of topological maps [19]. The latent representations $\mathbf{a}_{e_i}$ associated with the indices would be distributed in the brain. Thus the representation pattern for the concept "hammer" might activate brain regions associated with the typical appearance of a hammer (visual cortex), but also with the sound of hammering (auditory cortex) and with the activity of hammering (motor cortex) [20]. As another example, if a subject recalls a person, all sensory impressions of that person are restored. Note that each entity has a distributed representation since an index e_i and its representation $\mathbf{a}_{e_i}$ are always jointly activated. The assumption we are making is that an index, which, as concept cells, might involve the activation of one or a small number of neurons [40], uniquely identifies an entity (see Subsection 4.1).

4.4 Episodic and Semantic Memories

At a certain abstraction level the brain has been modelled as a graph of computing nodes [4].

Here, we discuss distributed graphical implementations of the mathematical models. Figure 2 shows a graphical structure with five layers. In this example, $s,\ p,\ t$ are given and the goal is to find an o that leads to likely quadruples. In the second layer the latent representations a_{e_s}, a_{e_p}, and a_{e_t} are activated. The latent representation for the unknown object is set to ones as indicated. The third layer calculates $a_{e_s,r_1}\ a_{e_p,r_2}\ a_{e_t,r_4}\ g(r_1, r_2, r_3, r_4)$ and the fourth layer (second decoder from the right) calculates the sum

$$\mathbf{h}_o = \sum_{r_1=1}^{\tilde{r}} \sum_{r_2=1}^{\tilde{r}} \sum_{r_4=1}^{\tilde{r}} a_{e_s,r_1}\ a_{e_p,r_2}\ a_{e_t,r_4}\ g(r_1, r_2, r_3, r_4).$$

The fifth layer samples one or several objects based on

$$P(o|s, p, t) \propto \exp \beta \mathbf{a}_{e_o}^T \mathbf{h}_o. \tag{4}$$

We exploit here that a multi-linear expression as in Equation 1, can be written as $\mathbf{a}_{e_o}^T \mathbf{h}_o$.[5] $\mathbf{h}_o$ can be interpreted as an activation pattern generated in neocortex, which is then matched with existing representations of entities $\mathbf{a}_{e_o}$.

[5] We can write the Tucker model as $f^e(\mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o}, \mathbf{a}_{e_t}) = \mathcal{G} \bullet_1 \mathbf{a}_{e_s} \bullet_2 \mathbf{a}_{e_p} \bullet_3 \mathbf{a}_{e_o} \bullet_4 \mathbf{a}_{e_t}$. $\mathcal{G}$ is the core tensor and the term in the bracket is a vector. $\bullet_n$ is the n-mode vector product [21].

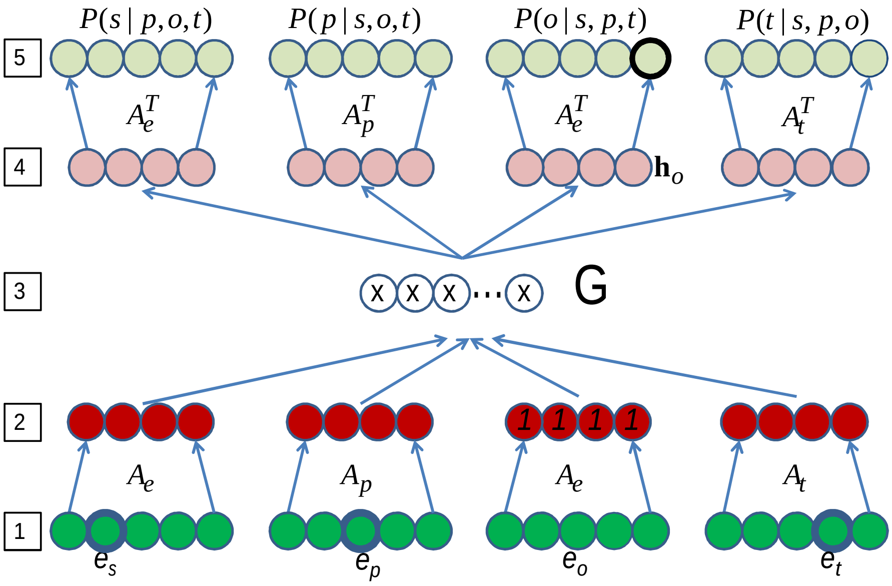

Fig. 2. An architecture implementing the episodic memory. As concrete example, we assume that in the layer 1 the nodes for the corresponding subject $s = 2$ and predicate $p = 3$ and time $t = 4$ are activated. Thus the query would be: Which objects are likely associated with subject $s = 2$ and predicate $p = 3$ at the time $t = 4$. Layer 2 is the representation layer and the corresponding latent representations $\mathbf{a}_{e_s}$, $\mathbf{a}_{e_p}$, $\mathbf{a}_{e_t}$ are represented. Nodes in the product layer calculate $a_{e_s,r_1}\ a_{e_p,r_2}\ a_{e_t,r_4}\ g(r_1, r_2, r_3, r_4)$. Since in the example we query an object, the latent representation for the object is replaced by ones in layer 2. In the second to right decoder of layer 4 we form $\mathbf{h}_o = \sum_{r_1=1}^{\tilde{r}} \sum_{r_2=1}^{\tilde{r}} \sum_{r_4=1}^{\tilde{r}} a_{e_s,r_1}\ a_{e_p,r_2}\ a_{e_t,r_4}\ g(r_1, r_2, r_3, r_4)$. The layer 5 then samples an object from $P(o|s, p, t) \propto \exp \beta \mathbf{a}_{e_o}^T \mathbf{h}_o$. The columns of A_e, A_p, A_t contain the latent representations of the entities, predicates and time steps. By following different paths through the graph, we can generate a sample from $P(s|p, o, t)$, from $P(p|s, o, t)$, or from $P(t|s, p, o)$. We can also sample from marginalized distributions. Since nonnegative tensor models belong to the class of sum-product model [38], marginalization means to enter a vector of ones for quantities to be marginalized. For example to marginalize out p, we would enter a vector of ones $\mathbf{1}$ to the index layer for predicates in layer 1. Biologically, this can be interpreted as a “neutral” input. The figure shows nicely that an entity only communicates with the rest of the network via its latent representation. We assume that the processing steps in layers 1-5 are executed sequentially.

By entering the vectors of ones at the latent representations in layer 2 for $\mathbf{a}_{e_s}$, $\mathbf{a}_{e_p}$, or $\mathbf{a}_{e_t}$ and by taking different paths through the models, we can calculate the conditional probabilities $P(s|p, o, t)$, $P(p|s, o, t)$, $P(t|s, p, o)$, respectively. With nonnegative tensor models, we can also marginalize: Since the (nonnegative) tensor models belong to the class of sum-product networks, a marginalization simply means the application of all ones for the variable to be marginalized in the index layer 1. This permits us to, e.g., calculate $P(o|s, t)$, marginalizing out p.[6]

In [54], we argue that a semantic memory can be generated from an episodic memory by marginalizing time. Thus, as an example, to estimate $P(o|s)$ we would activate the index for e_s in layer 1, apply vectors of ones at layer 1 for predicate and time (to marginalize out both) and apply a vector of ones at layer 2 for the object (the quantity we want to predict). The generation of semantic memory from episodic memory would biologically be very plausible.[7] This relationship between both memories is supported by cognitive studies on brain memory functions: It has been argued that semantic memory is information an individual has encountered repeatedly, so often that the actual learning episodes are blurred [8]. Thus episodic memory may be the "gateway" to semantic memory [1, 48].

It appears from Figure 2 that some representations are redundant. For example A_e appears several times in the figure. Also layers 2 and 4 both represent latent representations and layers 1 and 5 both represent indices. Figure 3 shows a computational path without redundancies. It also reflects that the entity sets for subject and object are really identical.

Tensor models have been used previously as memory models but the main focus was the learning of simple (auto-) associations [16, 18, 35] and compositional structures [47, 37, 36, 15]. In the tensor product approach (Smolensky [47]) encoding or binding is realized by a tensor product (generalized outer product), and composition by tensor addition. In the STAR model [15], predicates are represented as tensor products of the components of the predicates; in the triples considered here, these would be subject, predicate and object. None of these approaches uses tensor *decompositions*.

---

[6] Note that the final operation as described in Equation 4 is the dot product of a vector generated from the latent representations of s, p, t, i.e., $\mathbf{h}_o$, with the latent representations of the object, i.e., $\mathbf{a}_{e_o}$. The general structure in Figure 2 is the same for different kinds of tensor models, whose implementations vary in layer 5. For example, a RESCAL model [32] calculates $\mathbf{h}_o = G^p \mathbf{a}_{e_s}$ which can be related to many classical associative memory models. The predicate matrix G^p is a slice in RESCAL's core tensor. Main differences to the classical associative memory models are that here the factors are latent and that the system is trained end-to-end, whereas classical systems rely on some form of Hebbian learning [17, 30]. Also classical memory models are often auto-associative, i.e., their main concern is to restore a noisy pattern, whereas the models considered here are associative, in that they predict an object, given subject and predicate.

[7] This form of marginalization only works for nonnegative models.

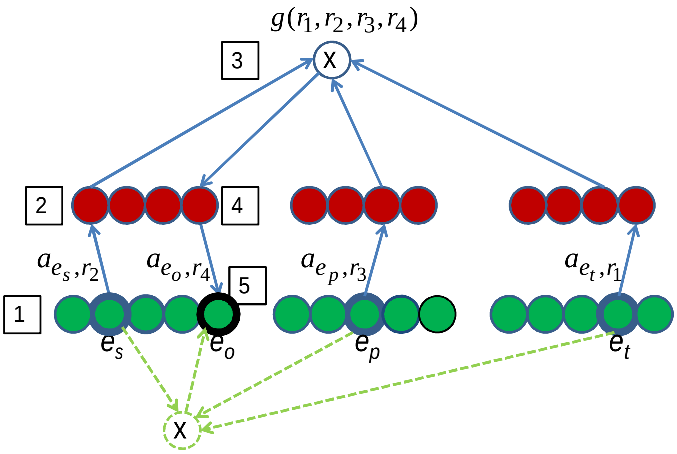

Fig. 3. Here we show the information flow that involves a specific product node (one out of $\tilde{r}^4$ of such nodes) of the graphical representation in Figure 2. The figure also shows that entities have unique representations (left), independent of their roles as subject or object. Also layers 2 and 4 identical in Figure 2, as are layers 1 and 5, and really and only need single representations, as shown here. The nodes on the activation layer 2 are connected to the product node as shown which has a multiplicative bias $g(r_1, r_2, r_3, r_4)$. As discussed in the caption of Figure 2, the product node calculates $a_{e_s,r_1}\ a_{e_p,r_2}\ a_{e_t,r_4}\ g(r_1, r_2, r_3, r_4)$. In layer 4 (which is identical to layer 2), the sum over all product nodes is calculated to form $\mathbf{h}_o$, and an object is sampled from the resulting probabilities in layer 5 (which is identical to layer 1). Although the complexity of the network is approximately $\mathcal{O}(\tilde{r}^4)$, this number is fixed and does not grow with the number of entities, predicates or time steps. In contrast, in an explicit representations of facts (lower part of the figure in light green), which is the basis of other tensor-based memories [15], the number of product nodes grows proportional to the number of episodic facts to be stored. This explicit fact representation might be useful for "unusual" facts that cannot easily be represented by factorization models.

4.5 Future Work: Perception as an Active Process

Equation 3 describes a basic encoder-decoder system. It can be made more powerful by integrating attention mechanisms, serial processing and recurrent structures [7, 6, 2]. As part of future work we will explore recent developments in encoder-decoder approaches for sensory processing and semantic decoding. In [3] a semantic prior is combined with a triple extraction process that involves visual attention based on extracted bounding boxes. This is another line of research worth exploring.

5 Conclusions

The work presented in this paper is an extension to the hippocampal memory indexing theory, which is one of the main detailed theories about the forming of episodic memory. We have extended the model by also considering semantic decoding into explicit triples and by providing explicit models for episodic and semantic memory. Our approach is built upon latent representations of generalized entities representing, e.g., objects, persons, locations, predicates, and time instances. As in the hippocampal memory indexing theory, an activation of an index for a past memory would activate its representation in association cortex which might reconstruct sensory impressions of past memories. A very useful function of an episodic memory would be to recall what happened after a past episode that is similar to the current situation and what action was applied and with what consequences. As a new contribution, we propose that past representations are also decoded semantically, producing triples of past memories. Semantic decoding might be an important intermediate step to generate language, i.e., to explicitly report about perceived sensory inputs, past episodes and semantic knowledge. Language, of course, is a faculty specific to humans.

The activations patterns $\mathbf{a}$ might represent an entity, a predicate, or a time instance on the subsymbolic level, whereas the corresponding index e represents a concept on a discrete symbolic level.

We have shown how semantic memory can be generated from the latent representations used by episodic memory. Thus we provide an explicit model to support the hypothesis that semantic memory is derived from episodic memory and that both rely on the same representations. In this line of thought, semantic memory is a long-term storage for episodic memory, where the exact timing information is lost.

We have demonstrated that episodic and semantic memories can be modelled using tensor decompositions. In a way this is an existence proof, showing that there are biologically plausible architectures that can implement episodic and semantic memory. The brain might use different mathematical structures, although the marginalization of episodic memory, which is the basis of semantic memory in our approach, is strictly only possible in the way described for sum-product models, such as tensor decompositions.

In this paper, we have focussed on memory models. We propose that other functions like prediction, planning, reasoning and decision making would use the indices and their latent representations as well [12, 56]. In particular short-term memory might exploit both semantically decoded indices and their latent representations. Cognitive control functions and working memory functions are typically associated with prefrontal cortex.

References

1. Alan D Baddeley, Graham Hitch, et al. Working memory. *The psychology of learning and motivation*, 8:47–89, 1974.
2. Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.
3. Stephan Baier, Yunpu Ma, and Volker Tresp. Improving visual relationship detection using semantic modeling of scene descriptions. *ISWC*, 2017.
4. Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, 10(3):186, 2009.
5. Erika Cerasti and Alessandro Treves. How informative are spatial ca3 representations established by the dentate gyrus? *PLoS computational biology*, 6(4), 2010.
6. Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886, 2015.
7. Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
8. Martin A Conway. Episodic memories. *Neuropsychologia*, 47(11):2305–2313, 2009.
9. Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.
10. Howard Eichenbaum. Memory on time. *Trends in cognitive sciences*, 17(2):81–88, 2013.
11. Howard Eichenbaum. Time cells in the hippocampus: a new dimension for mapping memories. *Nature Reviews Neuroscience*, 15(11):732–744, 2014.
12. Cristóbal Esteban, Volker Tresp, Yinchong Yang, Stephan Baier, and Denis Krompaß. Predicting the co-evolution of event and knowledge graphs. In *Information Fusion (FUSION), 2016 19th International Conference on*, pages 98–105. IEEE, 2016.
13. Paul W Frankland and Bruno Bontempi. The organization of recent and remote memories. *Nature Reviews Neuroscience*, 6(2):119–130, 2005.
14. Mark A Gluck, Eduardo Mercado, and Catherine E Myers. *Learning and memory: From brain to behavior*. Palgrave Macmillan, 2013.
15. Graeme S Halford, William H Wilson, and Steven Phillips. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, 21(6):803–831, 1998.
16. Douglas L Hintzman. Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, 16(2):96–101, 1984.

17. John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
18. Michael S Humphreys, John D Bain, and Ray Pike. Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, 96(2):208, 1989.
19. Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 2016.
20. Markus Kiefer and Friedemann Pulvermüller. Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex*, 48(7):805–825, 2012.
21. Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 2009.
22. Benjamin J Kraus, Robert J Robinson, John A White, Howard Eichenbaum, and Michael E Hasselmo. Hippocampal time cells: time versus path integration. *Neuron*, 78(6):1090–1101, 2013.
23. Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems*, pages 1172–1180, 2016.
24. Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016.
25. Elizabeth Loftus and Katherine Ketcham. *The myth of repressed memory: False memories and allegations of sexual abuse*. Macmillan, 1996.
26. Lisa Marshall and Jan Born. The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in cognitive sciences*, 11(10):442–450, 2007.
27. James L McClelland, Bruce L McNaughton, and Randall C O'reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.
28. Lynn Nadel and Morris Moscovitch. Memory consolidation, retrograde amnesia and the hippocampal complex. *Current opinion in neurobiology*, 7(2):217–227, 1997.
29. Lynn Nadel, A Samsonovich, Lee Ryan, and Morris Moscovitch. Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. *Hippocampus*, 10(4):352–368, 2000.
30. Maximilian Nickel. Private communication, 2017.
31. Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.
32. Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A Three-Way Model for Collective Learning on Multi-Relational Data. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.
33. Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. Factorizing yago: scalable machine learning for linked data. In *Proceedings of the 21st international conference on World Wide Web*, pages 271–280. ACM, 2012.
34. Erika Nyhus and Tim Curran. Functional role of gamma and theta oscillations in episodic memory. *Neuroscience & Biobehavioral Reviews*, 34(7):1023–1035, 2010.
35. Adam F Osth and Simon Dennis. Sources of interference in item and associative recognition memory. *Psychological review*, 122(2):260, 2015.

36. Tony Plate. A common framework for distributed representation schemes for compositional structure. *Connectionist systems for knowledge representation and deduction*, pages 15–34, 1997.
37. Jordan B Pollack. Recursive distributed representations. *Artificial Intelligence*, 46(1):77–105, 1990.
38. Hoifung Poon and Pedro Domingos. Sum-product networks: A new deep architecture. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 689–690. IEEE, 2011.
39. Rodrigo Quian Quiroga. Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 13(8):587–597, 2012.
40. Rodrigo Quian Quiroga, Itzhak Fried, and Christof Koch. Brain cells for grandmother. *Scientific American*, 308(2):30–35, 2013.
41. Henry L Roediger and Kathleen B McDermott. Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, 21(4):803, 1995.
42. Edmund T Rolls. Memory systems in the brain. *Annual review of psychology*, 51(1):599–630, 2000.
43. Edmund T Rolls. A computational theory of episodic memory formation in the hippocampus. *Behavioural brain research*, 215(2):180–196, 2010.
44. Edmund T Rolls. Advantages of dilution in the connectivity of attractor networks in the brain. *Biologically Inspired Cognitive Architectures*, 1:44–54, 2012.
45. Edmund T Rolls and Raymond P Kesner. A computational theory of hippocampal function, and empirical tests of the theory. *Progress in neurobiology*, 79(1):1–48, 2006.
46. C Daniela Schwindel and Bruce L McNaughton. Hippocampal-cortical interactions and the dynamics of memory trace reactivation. *Prog. Brain Res*, 193:163–177, 2011.
47. Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1):159–216, 1990.
48. Larry R Squire. *Memory and brain.* Oxford University Press, 1987.
49. Larry R Squire and Pablo Alvarez. Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current opinion in neurobiology*, 5(2):169–177, 1995.
50. Robert Stickgold. Sleep-dependent memory consolidation. *Nature*, 437(7063):1272, 2005.
51. Timothy J Teyler and Pascal DiScenna. The hippocampal memory indexing theory. *Behavioral neuroscience*, 1986.
52. Timothy J Teyler and Jerry W Rudy. The hippocampal indexing theory and episodic memory. *Hippocampus*, 2007.
53. Volker Tresp, Cristóbal Esteban, Yinchong Yang, Stephan Baier, and Denis Krompaß. Learning with memory embeddings. *NIPS 2015 Workshop on Nonparametric Methods for Large Scale Representation Learning (extended TR)*, 2015.
54. Volker Tresp, Yunpu Ma, Stephan Baier, and Yinchong Yang. Embedding learning for declarative memories. In *European Semantic Web Conference*, pages 202–216. Springer, Cham, 2017.
55. Alessandro Treves and Edmund T. Rolls. Computational constraints suggest the need for two distinct input systems to the hippocampal ca3 network. *Hippocampus*, 2(2):189–199, 1992.
56. Yinchong Yang, Peter A Fasching, Markus Wallwiener, Tanja N Fehm, Sara Y Brucker, and Volker Tresp. Predictive clinical decision support system with rnn encoding and tensor decoding. *arXiv preprint arXiv:1612.00611*, 2016.

Appendix: More Details on Hippocampal Memory

Evidence suggests that dentate gyrus (DG) and CA3 subregions of hippocampus sustain the sparsest active neurons. The DG subregion processes the incoming inputs prior to the associative learning in the CA3 region; it performs pattern separation and pattern orthogonalization in a sense that the overlapping of neurons firing patterns of similar inputs becomes reduced. It is biologically realized by the sparse firing activity of DG cells as well as sparse connections between DG and CA3 cells via mossy fiber. Sparse projection from DG to CA3 further reduces the degree of correlation inside CA3 and increases the storing capacity [55, 5]. The concept of an index in our model is inspired by the sparse representation in the DG/CA3 subregions of the hippocampus.

It is predicted that the retrieval of hippocampus-dependent information from a retrieval clue depends on the CA3 subregion, which can be modeled as an attractor neural network with recurrent collateral connections. For each neuron the number of synapses connecting with the recurrent collaterals is supposed to be much smaller than the number of CA3 neurons; this results in a diluted connectivity of the CA3 recurrent network. Diluted connectivity increases the number of attractor states (or stable states) of recurrent network and enhances the memory capacity [44, 45].

We assume that a time index e_t is formed if the sensory perception is novel and emotionally significant. Considerable recent works have argued that the hippocampus is responsible for the temporal organization of memories [10, 11]; it becomes activated when a sequence of serial events are being processed. To be explicit, the CA1 subregion of the hippocampus is involved in the sequential memory for temporal ordered events. The CA1 subregion has direct connections with the CA3 subregion as well as the medial entorhinal cortex. There is an interesting hypothesis that the medial entorhinal cortex might supervise the formation of sequence memories since the grid cells in the medial entorhinal cortex might provide both spatial information and temporal information required for sequence encoding [22]. A direct connection from medial entorhinal cortex to CA1 specializes the firing pattern of neurons in CA1 at different time instances. This process is implemented by competitive networks, which are essential for the item association operation in CA3 [42]. In our model this process is abstracted as follows: entorhinal cortex organizes time indices in a sequence, e_t, e_{t-1}, $\cdots$, they can further trigger indices detecting and binding in the CA3 subregion, and outputs a series of time-ordered triples $(e_s, e_p, e_o)_t$ that together form a sequential episodic memory.