

Embedding Learning for Declarative Memories

Volker Tresp^{1,2}, Yunpu Ma^{1,2}, Stephan Baier², Yinchong Yang^{1,2}

¹ Siemens AG, Corporate Technology, Munich, Germany

² Ludwig-Maximilians-Universität München, Munich, Germany

Abstract. The major components of the brain’s declarative or explicit memory are *semantic memory* and *episodic memory*. Whereas semantic memory stores general factual knowledge, episodic memory stores events together with their temporal and spatial contexts. We present mathematical models for declarative memories where we consider semantic memory to be represented by triples and episodes to be represented as quadruples i.e., triples in time. E.g., *(Jack, receivedDiagnosis, Diabetes, Jan1)* states that Jack was diagnosed with diabetes on January 1. Both from a cognitive and a technical perspective, an interesting research question is how declarative data can efficiently be stored and semantically be decoded. We propose that a suitable data representation for episodic event data is a 4-way tensor with dimensions subject, predicate, object, and time. We demonstrate that the 4-way tensor can be decomposed, e.g., using a 4-way Tucker model, which permits semantic decoding of an event, as well as efficient storage. We also propose that semantic memory can be derived from the episodic model by a marginalization of the time dimension, which can be performed efficiently. We argue that the storage of episodic memory typically requires models with a high rank, whereas semantic memory can be modelled with a comparably lower rank. We analyse experimentally the relationship between episodic and semantic memory models and discuss potential relationships to the corresponding brain’s cognitive memories.

1 Introduction

The main components of the brain’s declarative or explicit memory are *semantic memory* and *episodic memory*. Both are considered long-term memories and store information potentially over the life-time of an individual [9, 1, 34, 4]. The semantic memory stores general factual knowledge, i.e., information we *know*, independent of the context where this knowledge was acquired. Episodic memory concerns information we remember and includes the spatiotemporal context of events [38]. There is evidence that these main cognitive categories are partially dissociated from one another in the brain, as expressed in their differential sensitivity to brain damage [10]. However, there is also evidence indicating that the different memory functions are not mutually independent and support one another [13].

In this paper we discuss technical models for semantic and episodic memories and compare them with their biological counterparts. In particular, we consider

that a technical realization of a semantic memory is a knowledge graph (KG) which is a triple-oriented knowledge representation. Popular technical large-scale KGs are DBpedia [2], YAGO [35], Freebase [5], NELL [7], and the Google KG [32]. There exist reliable KGs with more than a hundred billion triples that support search, text understanding and question answering [32].

In our approach we model episodes as events in time that can be represented by quadruples. Thus whereas the triple $(Jack, hasDiagnosis, Diabetes)$ might reflect that Jack has diabetes, the quadruple $(Jack, receivedDiagnosis, Diabetes, Jan1)$ would represent the diagnostic event.

We propose that biologically plausible representations of both semantic and episodic memories can be achieved by a decomposition of the adjacency tensors describing the memories. The decomposition leads to a highly compressed form of the memories and exhibits a form of memory generalization or inductive learning, in form of a generalization to new triples and quadruples [26]. If each entity and each predicate has a unique latent representation, information is shared across all memory functions.

We propose that semantic memory is a long-term storage for episodic memory where the exact timing information is lost, and that both memories rely on the same latent representations. In particular we propose that semantic memory can be derived from episodic memory by marginalizing the time dimension, an operation which can be performed elegantly when nonnegative tensor decompositions are used as memory models. Whereas the storage of episodic memory typically requires decomposition models with a high rank, semantic memory can be stored with a comparably lower rank.

The paper is organized as follows. In the next section we introduce the unique representation hypothesis which postulates latent representations of generalized entities that are shared between memory functions. Section 3 covers latent representation models for semantic and episodic memory and Section 4 describes the tensor models. In Section 5, we discuss relationships between episodic and semantic memories and discuss memory querying. Section 6 discusses the biological relevance of the proposed model and Section 7 presents experimental results. Section 8 contains our conclusions.

2 Unique-Representation Hypothesis

A technical realization of a semantic memory is a knowledge graph (KG) which is a triple-oriented knowledge representation. Here we consider a slight extension to the subject-predicate-object triple form by adding the value in the form $(e_s, e_p, e_o; Value)$ where $Value$ is a function of e_s, e_p, e_o and, e.g., can be a Boolean variable ($True$ for 1, $False$ for 0) or a real number. Thus $(Jack, likes, Mary; True)$ states that Jack (the subject or head entity) likes Mary (the object or tail entity). Note that e_s and e_o represent the entities for subject index s and object index o . To simplify notation we also consider e_p to be a generalized entity associated with predicate type with index p . For the episodic memory we introduce e_t , which is a generalized entity for time t .

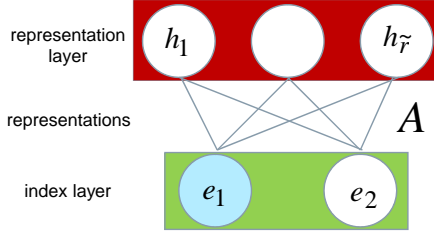


Fig. 1. A graphical view of the unique-representation hypothesis. The model can operate bottom-up and top-down. In the first case, index node e_i activates the representation layer via its latent representation, implemented as weight vector. In the figure, e_1 is activated, all other index nodes are inactive and the representation layer is activated with the pattern $\mathbf{h} = \mathbf{a}_{e_1}$. In top-down operation, a representation layer can also activate index nodes. The activation of node e_i is then equal to the inner product $\mathbf{a}_{e_i}^\top \mathbf{h}$. We consider here formalized nodes which might actually be implemented as ensembles of distributed neurons or as stable activation patterns of distributed neurons. Here and in the following we assume that the matrix A stores the latent representations of all generalized entities. The context makes it clear if we refer to the latent representations of entities, predicates, or time indices.

The *unique-representation hypothesis* assumed in this paper is that each entity or concept e_i , each predicate e_p and each time step e_t has a unique latent representation — \mathbf{a}_i , \mathbf{a}_p , or \mathbf{a}_t , respectively— in form of a set of real numbers, represented as a vector or a matrix. The assumption is that the representations are shared among all memory functions, which permits information exchange and inference between the different memories. For simplicity of discussion, we assume that the latent representations form vectors and that the dimensionalities of these latent representations for entities and predicates are \tilde{r} such that $\mathbf{a}_i \in \mathbb{R}^{\tilde{r}}$, $\mathbf{a}_p \in \mathbb{R}^{\tilde{r}}$, and for time is \tilde{r}_T such that $\mathbf{a}_t \in \mathbb{R}^{\tilde{r}_T}$. Figure 1 shows a simple network realization.

3 Semantic and Episodic Knowledge Graph Models

3.1 Semantic Knowledge Graph

We now consider an efficient representation of a KG. First, we introduce the three-way semantic adjacency tensor \mathcal{X} where the tensor element $x_{s,p,o}$ is the associated *Value* of the triple (e_s, e_p, e_o) . Here $s = 1, \dots, S$, $p = 1, \dots, P$, and $o = 1, \dots, O$. One can also define a companion tensor Θ with the same dimensionality as \mathcal{X} and with entries $\theta_{s,p,o}$. It contains the natural parameters of the model and the connection to \mathcal{X} for Boolean variables is

$$P(x_{s,p,o} | \theta_{s,p,o}) = \text{sig}(\theta_{s,p,o}) \quad (1)$$

where $\text{sig}(arg) = 1/(1+\exp(-arg))$ is the logistic function (Bernoulli likelihood), which we use in this paper. If $x_{s,p,o}$ is a real number then we can use a Gaussian distribution with $P(x_{s,p,o}|\theta_{s,p,o}) \sim \mathcal{N}(\theta_{s,p,o}, \sigma^2)$.

As mentioned, the key concept in embedding learning is that each entity and predicate e has an \tilde{r} -dimensional latent vector representation $\mathbf{a} \in \mathbb{R}^{\tilde{r}}$. In particular, the embedding approaches used for modeling a semantic KGs assume that

$$\theta_{s,p,o}^{sem} = f^{sem}(\mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o}). \quad (2)$$

Here, the function $f^{sem}(\cdot)$ predicts the value of the natural parameter $\theta_{s,p,o}^{sem}$. In the case of a KG with a Bernoulli likelihood, $\text{sig}(\theta_{s,p,o}^{sem})$ represents the confidence that the triple (e_s, e_p, e_o) is true and we call the function an *indicator mapping function*. We discuss examples in the next section.

Latent representation approaches have been used very successfully to model large KGs. It has been shown experimentally that models using latent factors perform well in these high-dimensional and highly sparse domains. Since an entity has a unique representation, independent of its role as a subject or an object, the model permits the propagation of information across the KG. For example if a writer was born in Munich, the model can infer that the writer is also born in Germany and probably writes in the German language [24, 25]. For a recent review, please consult [26].

Due to the approximation, $\text{sig}(\theta_{Jack,marriedTo,e}^{sem})$ might be smaller than one for the true spouse. The approximation also permits inductive inference: We might get a large $\text{sig}(\theta_{Jack,marriedTo,e}^{sem})$ also for persons e that are *likely* to be married to *Jack* and $\text{sig}(\theta_{s,p,o}^{sem})$ can, in general, be interpreted as a confidence value for the triple (e_s, e_p, e_o) . More complex queries on semantic models involving existential quantifiers are discussed in [19].

3.2 An Event Model for Episodic Memory

Whereas a semantic KG model reflects the state of the world, e.g., of a clinic and its patients, observations and actions describe discrete events, which, in our approach, are represented by an episodic event tensor. In a clinical setting, events might be a prescription of a medication to lower the cholesterol level, the decision to measure the cholesterol level and the measurement result of the cholesterol level; thus events can be, e.g., actions, decisions and measurements.

The episodic event tensor is a four-way tensor \mathcal{Z} where the tensor element $z_{s,p,o,t}$ is the associated *Value* of the quadruple (e_s, e_p, e_o, e_t) . The indicator mapping function then is

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

where we have added a representation for the time of an event by introducing the generalized entity e_t with latent representation $\mathbf{a}_{e_t} \in \mathbb{R}^{\tilde{r}_t}$. This latent representation compresses all events that happen at time t . As discussed, an example from a clinical setting could be *(Jack, receivedDiagnosis, Diabetes, Jan1)* which states that Jack was diagnosed with diabetes on January 1.

4 Tensor Decompositions

4.1 Tensor Decompositions

There are different options for modelling the indicator mapping functions $f^{epi}(\cdot)$ and $f^{sem}(\cdot)$. In this paper we will only consider multilinear models derived from tensor decompositions. Tensor decompositions have shown excellent performance in modelling KGs [26].

Specifically, we consider a 4-way Tucker model for episodic memory in the form

$$f^{epi}(\mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o}, \mathbf{a}_{e_t}) = \sum_{r_1=1}^{\tilde{r}} \sum_{r_2=1}^{\tilde{r}} \sum_{r_3=1}^{\tilde{r}} \sum_{r_4=1}^{\tilde{r}_T} a_{e_s, r_1} a_{e_p, r_2} a_{e_o, r_3} a_{e_t, r_4} g^{epi}(r_1, r_2, r_3, r_4). \quad (3)$$

Here, $g^{epi}(r_1, r_2, r_3, r_4) \in \mathbb{R}$ are elements of the core tensor $\mathcal{G}^{epi} \in \mathbb{R}^{\tilde{r} \times \tilde{r} \times \tilde{r} \times \tilde{r}_T}$.

4.2 Inner Product Formulation of Tensor Decompositions

Note that we can rewrite Equation 3 as

$$f^{epi}(\mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o}, \mathbf{a}_{e_t}) = \mathbf{a}_{e_o}^\top \mathbf{h}^{object}$$

where

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

Thus if we consider subject, predicate, and time as inputs, we can evaluate the likelihood for different objects by an inner product between the latent representation of the object \mathbf{a}_{e_o} with a vector \mathbf{h}^{object} derived from the latent representations of the subject, the predicate, the time and the core tensor. Similarly we can calculate likely subjects, predicates, and time instances. We propose that this formulation is biologically more plausible, since inner products are operations that are easily performed by formalized neurons [31]. Also the representation is suitable for a sampling approach in querying (see the next section).

5 Querying Memories

5.1 Probabilistic Querying

In many applications one is interested in retrieving triples with a high likelihood, conditioned on some information, thus we are essentially faced with an optimization problem. To answer a query of the form $(Jack, receivedDiagnosis, ?, Jan1)$ we need to solve

$$\arg \max_{e_o} f^{epi}(\mathbf{a}_{Jack}, \mathbf{a}_{receivedDiagnosis}, \mathbf{a}_{e_o}, \mathbf{a}_{Jan1}).$$

Of course one is often interested in a set of likely answers. In [37] it was shown that likely triples can be generated by defining a Boltzmann distribution derived from an energy function. By enforcing non-negativity of the factors and the core tensor entries, the energy function for a Tucker model becomes $\mathbb{E}(s, p, o, t) = -\log f^{epi}(\mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o}, \mathbf{a}_{e_t})$ and the quadruple probability becomes

$$P(s, p, o, t) \propto \left(\sum_{r_1=1}^{\tilde{r}} \sum_{r_2=1}^{\tilde{r}} \sum_{r_3=1}^{\tilde{r}} \sum_{r_4=1}^{\tilde{r}_T} a_{e_s, r_1} a_{e_p, r_2} a_{e_o, r_3} a_{e_t, r_4} g^{epi}(r_1, r_2, r_3, r_4) \right)^\beta \quad (4)$$

where β plays the role of an inverse temperature: A large β would put all probability mass to triples with high functional values whereas a small β would assign probability mass also to triples with lower functional values. Note that for querying, we obtain a probability distribution over s, p, o, t and we can define marginal queries like $P(s, p, o)$ and conditional queries like $P(o|s, p, t)$. It is straightforward to generate likely samples from these distributions (see Figure 2).

Since the Tucker decomposition is an instance of a sum-product network [27], conditionals and marginals can easily be computed: A conditioning means that the index nodes are simply clamped to their respective values and marginalization means that the index nodes of the marginalized variables are all active, indicated by a vector of ones. Figure 2 shows some examples.

5.2 Semantic Memory Derived From Episodic Memory

Note that we can derive semantic queries from the episodic memory. As an example, the probability for the statement *(Jack, receivedDiagnosis, Diabetes, Jan1)* can be queried from the episodic memory directly, the probability for the statement *(Jack, receivedDiagnosis, Diabetes)* can also be derived from the episodic memory if we assume that semantic memory simply aggregates episodic memory slices. In particular we get for a semantic memory (with $\beta = 1$),

$$P(s, p, o) \propto \sum_{r_1=1}^{\tilde{r}} \sum_{r_2=1}^{\tilde{r}} \sum_{r_3=1}^{\tilde{r}} a_{e_s, r_1} a_{e_p, r_2} a_{e_o, r_3} g^{sem}(r_1, r_2, r_3), \quad (5)$$

where

$$g^{sem}(r_1, r_2, r_3) = \sum_t \sum_{r_4=1}^{\tilde{r}_T} a_{e_t, r_4} g^{epi}(r_1, r_2, r_3, r_4). \quad (6)$$

Technically, semantic memory can be derived from episodic memory by setting all index nodes for time to active, as shown in Figure 2D. Thus, if we accept that the semantic memory is a long-term storage for episodic memory, *we do not need to model semantic memory separately, since it can be derived from episodic memory!*

As part of a consolidation process we propose that $g^{sem}(r_1, r_2, r_3)$ is stored explicitly. The main reason is that \tilde{r}_T is typically quite large (see discussion in Subsection 6.2) and realizing the summation in Equation 6 for each recall

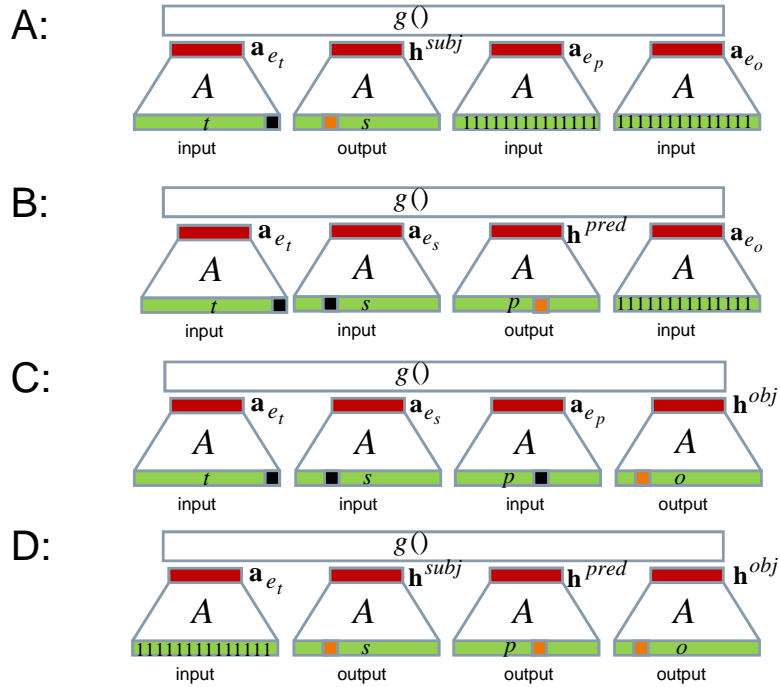


Fig. 2. The semantic decoding using a 4-dimensional Tucker tensor model for episodic memory. A: To sample a subject s given time t , we marginalize predicate p and object o . B: Here, o is marginalized and one samples a predicate p , given t, s . C: Sampling of an object o , given t, s, p . D: By marginalizing the time dimension, we obtain a semantic memory.

of semantic memory could be quite expensive. With $\tilde{r} \ll \tilde{r}_T$, the semantic memory has a small footprint and can be calculated efficiently. We also propose that the assumption that $\tilde{r} \ll \tilde{r}_T$ is quite plausible from a biological view point, as discussed in the following section.

Note that we implicitly assume that a fact that was encountered as an event is true forever. In the example above we would conclude that a diagnosed diabetes would be valid for lifetime. This would also agree with the weak expressiveness of standards like the Resource Description Framework (RDF) which do not model negations. Implementing negations and expressive constraints would require stronger ontologies. Temporal RDF graphs are discussed in [14, 15]. Some diseases, such as infections, on the other hand, can be cured. There are a number of ways of how this can be handled, for example by considering the relationships between *hasDisease* and *wasCuredFromDisease*. The latter could be implemented as an event *hasDisease* but with *Value* = -1.

6 Relationships to Human Memories

This section speculates about the relevance of the presented models to human memory functions. In particular we present several concrete hypotheses.

6.1 Unique-representation Hypothesis for Entities and Predicates

The *unique-representation hypothesis* states that each generalized entity e is represented by an index node and a unique (rather high-dimensional) latent representation \mathbf{a}_e that is stored as weight patterns connecting the index nodes with nodes in the representation layer (see Figure 1). Note that the weight vectors might be very sparse and in some models non-negative. The latent representations integrate all that is known about a generalized entity, they are the basis for episodic memory and semantic memory, and they can be instrumented for prediction and decision support in working memory. Among other advantages, a unique representation would explain why background information about an entity is seemingly effortlessly integrated into both sensor scene understanding and decision support, at least for entities familiar to the individual.

We proposed formalized nodes which might actually be implemented as ensembles of distributed neurons or as stable activation patterns of distributed neurons. Neurons which very selectively respond to specific entities and concepts have been found in the medial temporal lobe (MTL). In particular, researchers have reported on a remarkable subset of MTL neurons that are selectively activated by strikingly different pictures of given individuals, landmarks or objects and in some cases even by letter strings with their names [28]. For example, some neurons have been shown to selectively respond to prominent actors like “Jennifer Aniston” or “Halle Berry”. These are called concept cells by the authors.

In the consolidation theory of human memory it is assumed that, after some period of time, semantic memory, and possibly also episodic memory, is consolidated in cerebral cortex. Often neurons with similar receptive fields are clustered

in sensory cortices and form a topographic map [12]. Topological maps might also be the organizational form of neurons representing entities. Thus, entities with similar latent representations might be topographically close. A detailed atlas of semantic categories has been established in extensive fMRI studies showing topographically sorted local representations of semantic concepts [16].

6.2 Perception and Memory Formation

It is well established that new episodic memories are formed in the hippocampus, which is part of the MTL. We propose that the hippocampus is the region where index nodes for generalized entities are formed and that these index nodes establish a presence in the cortex during memory consolidation. The nodes in the representation layer might be in higher order sensory layers and in association cortex. The hippocampal memory index theory [36] agrees with this model and proposes that, in particular, time indices are established in the hippocampus. These are linked to the representations formed as responses to an episodic sensory input in the higher order sensory and association cortices. This model would also support the idea that \tilde{r}_T must be large since \mathbf{a}_{e_t} would need to represent all processed sensory information. The semantic decoding of \mathbf{a}_{e_t} by the episodic memory then corresponds to the semantic understanding of a sensory input, i.e., would be the essence of perception. Note that a recall of a past episode would simply mean that the corresponding node e_t is activated which then activates \mathbf{a}_{e_t} in the representation layer. \mathbf{a}_{e_t} can be semantically decoded, enabling an individual to semantically describe the past episode, and could activate the corresponding past sensory impressions, providing an individual with a sensory impression of the past episode.

6.3 Tensor Memory Hypothesis

The hypothesis states that semantic memory and episodic memory are implemented as functions applied to the latent representations involved in the generalized entities which include entities, predicates, and time. Thus neither the knowledge graphs nor the tensors ever need to be stored explicitly!

6.4 Semantic Memory and Episodic Memory

In our interpretation, semantic memory is a long-term storage for episodic memory. This is biologically attractive since no involved transfer from episodic to semantic memory is required. We propose that this is supported by cognitive studies on brain memory functions: It has been argued that semantic memory is information we have encountered repeatedly, so often that the actual learning episodes are blurred [8, 12]. Similarly, it has been speculated that a gradual transition from episodic to semantic memory can take place, in which episodic memory reduces its sensitivity and association to particular events, so

that the information can be generalized as semantic memory. Thus some theories speculate that episodic memory may be the “gateway” to semantic memory [3, 34, 33, 22, 39, 20]. [23] is a recent overview on the topic.

Our model supports inductive inference in form of a probabilistic materialization. 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. Thus both facts and inductively inferred facts about an entity are represented in its local environment. There is a certain danger in probabilistic materialization, since it might lead to overgeneralizations, reaching from national prejudice to false memories. In fact in many studies it has been shown that individuals produce false memories but are personally absolutely convinced of their truthfulness [29, 21].

7 Experiments

The goal of our experiments was to investigate the quality of the tensor decompositions for the semantic and episodic memory. The 3-way and 4-way tensors were factorized using a Tucker decomposition with unique latent representations for entities (as subjects and objects), predicates and time. We considered three model settings. The first setting was unconstrained using a binary cross-entropy (Bernoulli cost function) with additional $l2$ norm penalty on all parameters. In the second setting we constrained all parameters to be non-negative and used a mean squared error cost function and in the third setting we enforced non-negativity as well and used an $l1$ norm penalty on all parameters to encourage sparse solutions.

7.1 Data Set

Our experiments are based on the open Freebase KG, since it contains relatively many predicate types.³ Triples in the Freebase KG have been extracted from Wikipedia, WordNet, and many other web resources. In our experiments we extracted a subset which includes $10k$ entities, 285 relation types, and in total $141k$ positive triples. Most other triples were treated as unknown and only a small number of triples, generated by a corruption of observed positive triples, is treated as negative. The protocol for sampling negative triples follows Bordes *et al.* [6]: For each true triple (s, p, o) in the data set, we generated 5 negative triples by replacing the object o with corrupted entities o' drawn from the set of objects.

³ This is crucial for investigating human’s semantic memory since the system of memory includes “words and verbal symbols, their meanings and the relations between them” according to Tulving [38]. Large number of relation types contained in the knowledge graph enrich the hidden structure and enhance the complexity of the knowledge graph, and this will give a more realistic and accurate simulation of human’s semantic memory.

To generate a data set for episodic memory, we assigned a time index to each triple, in a way that all triples with the same subject obtained the same time index. Overall we used 40 different time indices.⁴ Similar to the corruption of semantic triples, the negative samples of episodic quadruples (s, p, o, t) are drawn by corrupting the objects o to o' or the temporal index t to t' , meaning that (s, p, o', t) services as a negative evidence of the episodic memory at instance t , and (s, p, o, t') is a true fact which cannot be correctly recalled at instance t' . The cost function is composed of cross-entropy and additional $L2$ or $L1$ norm and can be written for episodic quadruples as

$$\mathcal{L} = - \sum_{(s,p,o,t) \in \mathcal{T}} \log \theta_{s,p,o,t}^{epi} - \sum_{(s',p',o',t') \in \mathcal{C}} \log(1 - \theta_{s',p',o',t'}^{epi}) + \lambda \|A\|_{1 \text{ or } 2},$$

where (s, p, o, t) are true quadruples in the data set, and (s', p', o', t') are the corresponding corrupted quadruples.

7.2 Evaluation and Implementation

All the latent models were implemented using the open source libraries TensorFlow and Keras.⁵ The latent representations of all entities, predicates, time indices and the core tensor from the Tucker tensor decomposition are initialized with Glorot uniform initialization [11]. All the models are optimized using mini-batch adaptive gradient descent using the Adam update rule [18].

We split the data set, both semantic and episodic, into three subsets, where 70% were treated as the training set, 20% as the testing set, and the remaining 10% as the cross-validation set.

7.3 Experimental Results

Figure 3 shows area under precision recall (AUPRC) scores for the training and the test sets for the three settings as a function of the rank of the model. We report results for a semantic memory (“semantic”), for the episodic memory (“episodic”) and for the semantic memory derived from the episodic memory by marginalization or projection (“projection”). Note that episodic is evaluated on the episodic data (“remember”) whereas semantic and projection is evaluated on the semantic data (“know”). We see that the episodic experiment typically requires a higher rank to obtain good performance. The reason is that the episodic tensor is even sparser than the semantic tensor and contains fewer clear global patterns. The figures also show that we can obtain a semantic memory by projecting the episodic memory, confirming that episodic memory is a “gateway” to

⁴ This assignment resembles the active learning process of humans since, during the early stage of learning, we observe the true effects related to a certain subject during a fixed period of time, and store these facts in the form of episodic memory. In our experiment, we highly abstract this learning procedure.

⁵ https://en.wikipedia.org/wiki/Comparison_of_deep_learning_software. The code for the experiments is given in <https://github.com/Yunpu/Episodic-Memory>

semantic memory. In general, the unconstrained model gives better scores. But note that for the non-negative models we performed the projection as discussed in Figure 2 by entering a vector of ones for the time indices, whereas for the unconstrained setting, we first fully reconstructed the tensor entries and then did the summation on the reconstructed entries over time dimension. The latter procedure becomes infeasible for large episodic KGs.

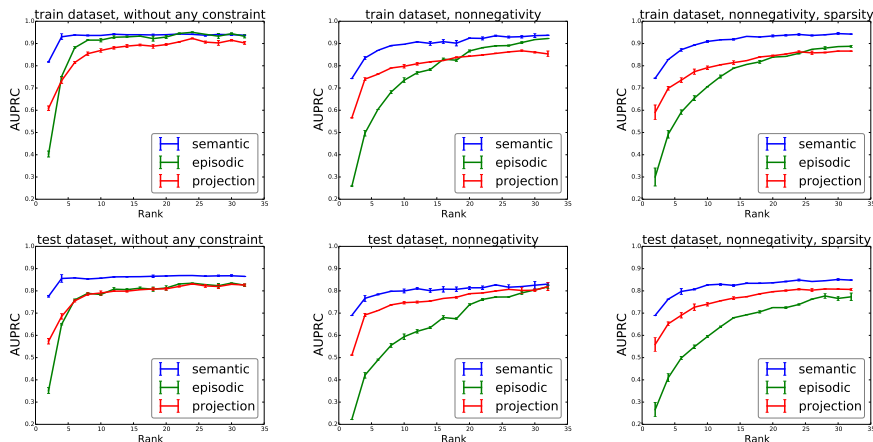


Fig. 3. AUPRC scores of the training and testing data sets for different model settings as a function of the rank.

In Figure 4 we plot recall score vs rank. In setting 3 with both non-negativity and sparsity constraints (see the third panel of Figure 4), the projection almost overlaps with the curve of semantic memory for the train and test data set. This observation explicitly indicates that the projected episodic memory function possesses the same memory capacity and quality as the semantic memory function, and this is the central result of our experiment.

Sparsity is an important feature of biological brain functions. Biological experiments indicate that the dentate gyrus (DG) and the CA3 subregion of the hippocampus sustain active neurons, which are connected by sparse projections from DG to CA3 through mossy fibers [17]. CA3 is considered to be crucial for establishing a memory trace during memory consolidation [30]. In the first unconstrained setting we obtained a sparsity of 3%, in the second setting with non-negativity constraints we obtained a sparsity of 30%, and in the third setting with non-negativity constraints and l_1 norm to encourage sparsity, we obtained 58% sparsity.

8 Conclusions

We have derived technical models for episodic memory and semantic memory based on a decomposition of the corresponding adjacency tensors. Whereas se-

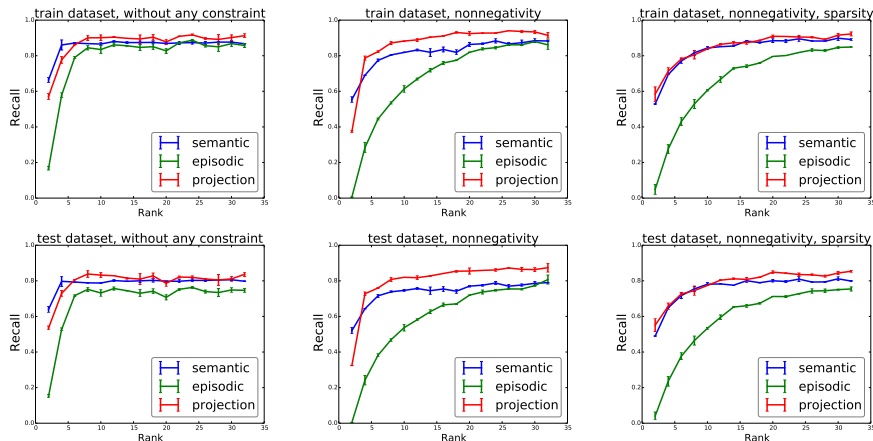


Fig. 4. Recall scores of the training and testing data set for different model settings as a function of the rank.

semantic memory only depends on the latent representations of subject, predicate and object, episodic memory also depends on the latent representation of the time of an event. We also proposed that semantic memory can be directly derived from episodic memory by marginalizing the time dimension.

As has been shown by several studies, the test set performances of tensor decomposition approaches are state-of-the-art [26]. If we want to use the models for memory recall, we are, however, mostly interested in reproducing stored memories accurately. Currently our models do not perform sufficiently well here (the training scores are around 0.9 instead of 1.0 on the training set). We attribute this to the limited ranks \tilde{r} and \tilde{r}_T of the models. As discussed, \mathbf{a}_{e_t} must encode all processed sensory information from various modalities and must be extremely high-dimensional to be able to do so, thus a large \tilde{r}_T is necessary. On the other hand, the rank for the latent representations for entities and predicates can be somewhat smaller, since the semantic memory, being formed by an integration process over the episodic memory model, requires a smaller rank for a good approximation, as confirmed in the experiment. Another issue is that the training data also contains triples and quadruples that do not follow regular patterns. It is correct to smooth over these triples and quadruples, since one cannot generalize from those, but one would want a truthful memory system to be able to recall also events that do not follow any regular patterns. Finding suitable solutions here is part of future work.

References

- [1] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. *The psychology of learning and motivation*, 1968.

- [2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. In *The Semantic Web*. 2007.
- [3] Alan D Baddeley, Graham Hitch, and GH Bower. The psychology of learning and motivation. 1974.
- [4] Frederic C Bartlett. *Remembering: A study in experimental and social psychology*, volume 14. Cambridge University Press, 1995.
- [5] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 2008.
- [6] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26*, 2013.
- [7] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. Toward an architecture for never-ending language learning. *AAAI*, 5:3, 2010.
- [8] Martin A Conway. Episodic memories. *Neuropsychologia*, 2009.
- [9] Hermann Ebbinghaus. *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot, 1885.
- [10] Michael S Gazzaniga, Richard B Ivry, and George Ronald Mangun. *Cognitive Neuroscience: The biology of the mind*. Norton, 2013.
- [11] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, volume 9, pages 249–256, 2010.
- [12] Mark A Gluck, Eduardo Mercado, and Catherine E Myers. *Learning and memory: From brain to behavior*. Palgrave Macmillan, 2013.
- [13] Daniel L Greenberg and Mieke Verfaellie. Interdependence of episodic and semantic memory: evidence from neuropsychology. *Journal of the International Neuropsychological society*, 2010.
- [14] Claudio Gutierrez, Carlos A Hurtado, and Alejandro Vaisman. Introducing time into rdf. *IEEE Transactions on Knowledge and Data Engineering*, 2007.
- [15] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard De Melo, and Gerhard Weikum. Yago2: exploring and querying world knowledge in time, space, context, and many languages. In *WWW*. ACM, 2011.
- [16] 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.
- [17] Raymond P Kesner and Edmund T Rolls. A computational theory of hippocampal function, and tests of the theory: new developments. *Neuroscience & Biobehavioral Reviews*, 2015.
- [18] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.
- [19] Denis Krompaß, Xueyan Jiang, Maximilian Nickel, and Volker Tresp. Probabilistic Latent-Factor Database Models. In *ECML PKDD*, 2014.
- [20] 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. *arXiv:1506.07285*, 2015.
- [21] Elizabeth Loftus and Katherine Ketcham. *The myth of repressed memory: False memories and allegations of sexual abuse*. Macmillan, 1996.

- [22] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex. *Psychological review*, 1995.
- [23] Neal W Morton. Interactions between episodic and semantic memory. Technical report, Vanderbilt Computational Memory Lab, 2013.
- [24] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A Three-Way Model for Collective Learning on Multi-Relational Data. In *ICML*, 2011.
- [25] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. Factorizing YAGO: scalable machine learning for linked data. In *WWW*, 2012.
- [26] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs: From multi-relational link prediction to automated knowledge graph construction. *Proceedings of the IEEE*, 2015.
- [27] Hoifung Poon and Pedro Domingos. Sum-product networks: A new deep architecture. In *ICCV*, 2011.
- [28] Rodrigo Quian Quiroga. Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 2012.
- [29] Henry L Roediger and Kathleen B McDermott. Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, 1995.
- [30] Edmund T Rolls. A computational theory of episodic memory formation in the hippocampus. *Behavioural brain research*, 2010.
- [31] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 1958.
- [32] Amit Singhal. Introducing the Knowledge Graph: things, not strings, May 2012. URL <http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html>.
- [33] Richard Socher, Samuel Gershman, Per Sederberg, Kenneth Norman, Adler J Perotte, and David M Blei. A Bayesian analysis of dynamics in free recall. In *Advances in neural information processing systems*, 2009.
- [34] Larry R Squire. *Memory and brain*. Oxford University Press, 1987.
- [35] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In *WWW*, 2007.
- [36] Timothy J Teyler and Pascal DiScenna. The hippocampal memory indexing theory. *Behavioral neuroscience*, 1986.
- [37] Volker Tresp, Cristóbal Esteban, Yinchong Yang, Stephan Baier, and Denis Krompaß. Learning with memory embeddings. *arXiv:1511.07972*, 2015.
- [38] Endel Tulving. Episodic and semantic memory 1. *Organization of Memory. London: Academic*, 1972.
- [39] Eiling Yee, Evangelia G Chrysikou, and Sharon L Thompson-Schill. *The Cognitive Neuroscience of Semantic Memory*. Oxford Handbook of Cognitive Neuroscience, Oxford University Press, 2014.