A Model for Perception and Memory

Volker Tresp, Sahand Sharifzadeh and Dario Konopatzki (firstname.lastname@lmu.de)
LMU Informatik, Oettingenstraße 67, 80538 München, Germany

Abstract

We analyze the close link between perception and memory. Our main hypothesis is that some of the main memory systems of the human brain, e.g., the episodic memory, the semantic memory, and to some degree also the working memory, are by-products of the need for humans to gradually extract more meaningful and more complex information from sensory inputs. Our model is an extension to the tensor memory approach. The key notions are index representations for entities, concepts, relationships and time instances, embeddings associated with the indices, a working memory layer, and a sensory memory layer. Perception and memory are realized as an interplay between the different layers. Our model is both competitive to other technical solutions and, as we argue, biologically plausible. Our experiments demonstrate that semantic memory can evolve from perception as a distinguishable functional module.

Introduction

Perception has evolved from simple stimulus-reaction in lower animals to the ability of a deep analysis of sensory input in humans. An important capability, for example, is the comparison to previous experiences: if a certain event is very similar to a past event, and that past event triggered a certain action, it makes sense that the current event should trigger the same action. Another important function is the identification of concepts and their relationships: "a child, located on a swing" will trigger very different actions than "a child, running in front of a car". Clearly a more refined perception is tightly linked to an improved understanding of the world, its schema, objects and their relationships, or as Goethe put it: "you only see what you know". In this paper we argue that episodic memory, i.e., the faculty to recall and restore past events, and semantic memory, i.e., knowledge about the world, are by-products of an evolving perceptual system which developed to deal with an increasingly complex world: our hypothesis is that episodic memory and semantic memory did not initially evolve as separate memory functions but instead repurposed faculties developed in perception for a semantic decoding of sensor stimuli. Furthermore, working memory might have evolved out of the need to store information to improve perceptual decoding.

The work in this paper is based on the tensor memory approach (Tresp et al., 2015; Tresp & Ma, 2016) which is an extension to the hippocampal memory indexing theory (Teyler & DiScenna, 1986). The key concepts of that approach are sparse index representations for entities, relationships and time instances. Each index has an associated distributed embedding, and memory and perception are based on an interplay between both. Perception, episodic memory and semantic memory might evoke sub-symbolic associations, but they are also declarative, indicated by the abilities of humans to report verbally about perception and memory contents. The semantic decoding in the tensor memory has exactly that declarative nature!

Here we significantly modify and extend that model. In the tensor memory model, the calculations of conditional probabilities required for decoding require marginalization operations which are costly and might be difficult to realize with biological wetware. Also, several indices and their embeddings needed to be active at the same time, which might not be biologically plausible (binding problem) and the approach required units to implement multiplication. Here, we propose a layered approach, where the sensory information is processed by a working memory layer, a representation layer and an index layer. The operations can be described as a single recurrent neural network where semantic memory evolves as an identifiable functional module.

The remaining parts of the paper are organized as follows. After we provide a brief review of the tensor memory approach in the next section, we present our model and mathematical operations performed by the model. Then follows a discussion on the neural substrate and a presentation of experimental results. The last section contains our conclusions.

Tensor Memories

Triple-based graphs have evolved into major data structures for representing semantic information. Concrete examples are knowledge graphs which store world facts (e.g., *(Munich, partOf, Bavaria)*) and scene graphs for describing image content (e.g., in the actual image, *(Dog, bites, Person)*).[1] The graphs are based on *(s, p, o)*-triples where the subject *s* and the object *o* are entities represented as the nodes in the graph, and where a directed link, labeled by *p*, represents a predicate. In the tensor memory approach, a graph was represented as a 3-way tensor, which was approximated by tensor factorization involving latent embeddings as vectors of real numbers: $\mathbf{a}_{e_s}$ is the embedding associated with the subject, $\mathbf{a}_{e_o}$ is the embedding associated with the object, $\mathbf{a}_p$ is the embedding associated with the predicate, and $\mathbf{a}_t$ is the embedding associated with the time instance, or image, t. Note that an entity has a unique representation, independent of its role as a subject or object. The factorized models deliver estimates for the probability that a triple is true at time t, given image information at time t, i.e., $P(s,p,o|t)$, and $P(s,p,o)$, which is the

[1] The nodes in the graph represent entities. In a knowledge graph, the nodes are labeled by identifiers (*Jack*), in scene graphs by concept labels (*Person*).

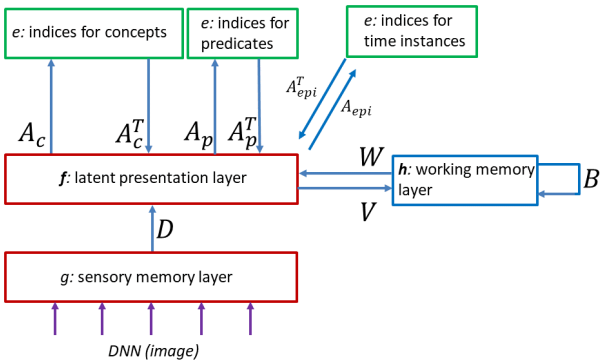

Figure 1: Our model architecture consists of four layers. Extracted representations from images are represented at the bottom layer (sensory memory, $\mathbf{g}$) which is connected to the representation layer $\mathbf{f}$. The top layer $\mathbf{e}$ contains the indices for concepts, predicates, and time instances. The working memory $\mathbf{h}$ is an integration layer and $\mathbf{g}$ is the sensory layer.

prior probability for observing the triples (s,p,o).[2]

The tensor memory model has some technical shortcomings when used in perception. For example, the semantic memory was derived from a marginalization over time, which is a computationally expensive operation that might not easily be implemented in biological wetware and can only be executed efficiently for some models (Tresp & Ma, 2016). Other problems are the polynomial scaling with the rank of the tensor model and the need for units that can perform multiplications.

A Model for Perception and Memory

A Layered Architecture: Figure 1 shows our model architecture. As in the tensor memory model, we assume an **index representation layer** $\mathbf{e}$ for entities, predicates and time instances, which is shown at the top of the figure. The indices can activate the **representation layer** $\mathbf{f}$ via connection matrix A_c^T for the concepts, A_p^T for the predicates, and A_{epi}^T for time instances. The embedding of concept e_i is the vector $\mathbf{a}_{e_i}$, which is the transpose of the i-th row of A_c. Similar for the predicates and the time instances. When index e_i is active and all other indices are inactive, then $\mathbf{f} = \mathbf{a}_i$. We introduce the **working memory layer** $\mathbf{h}$. This layer has some internal dynamics and receives inputs from the representation layer $\mathbf{f}$. In the following, we assume that we want to retrieve two concepts and their relationships at time, or image, t. Let t be the time constant of perception (on the order of hundreds of milliseconds). The micro time-step τ is the time constant for the decoding of the sensory input ($\tau << 100ms$). We now discuss the individual processing steps.

[2]More explicitly, $P(s,p,o|t)$ stands for the probability of observing a subject entity and an object entity at time t, where the subject belongs to concept s, the object belongs to concept o, and both are related by predicate p.

Decoding the Subject: Consider that $\mathbf{g}(t)$ is the embedding of the sensory input at time t. The activations of the working memory become, with $\mathbf{h}_{in}(t) = 0$,

$$\mathbf{h}(t) = sig(\mathbf{h}_{in}(t) + VD\mathbf{g}(t)).$$

The activations in the representation layer and the index layer are calculated as

$$\mathbf{f}(t) = D\mathbf{g}(t) + W\mathbf{h}(t) \quad and \quad \mathbf{e}(t) = sig(A_c\mathbf{f}(t)).$$

Thus the activations of the indices are determined by the inner product of their embeddings with the activation of the representation layer. In training, $\mathbf{e}(t)$ is set to be a one-hot vector indicating the index of the true subject. In testing, we proceed with $\mathbf{e}(t)$.[3] Finally, we set,

$$\mathbf{f}(t) \leftarrow A_c^T\mathbf{e}(t) = \mathbf{a}_{e_s} \quad and \quad \mathbf{h}_{in}(t+\tau) = B\mathbf{h}(t) + V\mathbf{a}_{e_s}.$$

In training, $\mathbf{f}(t)$ is now set to be the embedding of the true subject e_s, and in testing, it is an average, weighted by $\mathbf{e}(t)$; $\mathbf{h}_{in}(t+\tau)$ is the input activation for the working memory in the next time step. All weight matrices D, V, W, B and the matrices containing the embeddings A_c, A_p, A_{epi} are learned in training. Note that here, and in the following, there is a direct short cut, not involving the potentially slower working memory, in the form of $\mathbf{e}(t) = sig(A_cD\mathbf{g}(t))$.

Decoding the Object: The object decoding is identical to the subject decoding, if we replace t with $t+\tau$, $t+\tau$ by $t+2\tau$, and $\mathbf{a}_{e_s}$ by $\mathbf{a}_{e_o}$.

Decoding the Predicate: The predicate decoding is identical to the subject decoding, if we replace t with $t+2\tau$, $t+\tau$ by $t+3\tau$, $\mathbf{a}_{e_s}$ by $\mathbf{a}_p$, and A_c by A_p. Note that the decoding is asymmetrical and can distinguish between *(Dog, bites, Person)* and *(Person, bites, Dog)*. For a given image, the decoding can generate a large number of triples, which, in their entirety, present a visual input as an ensemble scene graph.

Discussion

Sensory Memory Layer: $\mathbf{g}$ is the visual sensory memory, maintaining visual information to be processed and analyzed. $\mathbf{g}$ represents properties of the respective focus of attention (in technical systems, these would be the bounding boxes). We assume that sensor processing involves an attention mechanism, such that $\mathbf{g}(t)$ represents the subject bounding box, $\mathbf{g}(t+\tau)$ represents the object bounding box, and $\mathbf{g}(t+2\tau)$ represents the predicate bounding box. The latter includes the two previous bounding boxes and some surrounding image area. In the brain, it is assumed that the sensory memory layer involves the visuospatial sketchpad of the working memory, associated with the parietal-occipital region.

[3]In testing we could perform a sampling from a normalized version of $\mathbf{e}(t)$; but this sampling introduces noise and would have to be repeated many times; proceeding with $\mathbf{e}(t)$ can be considered an approximation to the sampling.

Index Layer: The index layer $\mathbf{e}$ consists of indices for concepts, like *Cat*, and predicates like *nextTo*, and time instances. Generally it is assumed that indices are formed in the hippocampus and their long-term representation might involve the pole of the temporal lobe. An index might be realized by a small number of interacting neurons (Teyler & DiScenna, 1986; Quiroga, 2012). Over the path $\mathbf{e} \rightarrow \mathbf{f} \rightarrow \mathbf{g}$, an index can also excite a sensory impression. The indices (including the indices for time instances) have a relational memory function in the sense that they bind together different dimensions in the representation layer.

Representation Layer: The representation layer is important for the information path from $\mathbf{g}$ to $\mathbf{e}$ and it interacts with the working memory $\mathbf{h}$. If index e_i is activated, the activation of layer $\mathbf{f}$ reflects $\mathbf{a}_i$. Thus, whereas the sensory layer is primarily visually grounded, the representation layer is primarily concept grounded. If the concept "cat" is active in the index layer, the representation layer would contain abstract representations of the concept cat, without a reference to the actual cat in the sensory input. In the brain, these representations might involve the parietal lobe and the posterior region of the temporal lobe.

Working Memory Layer: The working memory layer integrates information from visual input and the decoding process (*subject, predicate, object*), and eventually the complete scene with its visual representations and decoded concepts and predicates. Working memory might have initially been developed biologically to support a more complex scene understanding and event processing. Its integrative functions are typically associated with the prefrontal cortex (PFC) in the frontal lobe and its interaction with the representation layer might reflect the event-specific relational memory functions in perception and memory recall. The PFC is profusely and reciprocally connected with the hippocampus, and cortices of association of the temporal and parietal lobes. Note that this layer is the "intelligence on top", since a simpler decoding $\mathbf{g} \rightarrow \mathbf{f} \rightarrow \mathbf{e}$ would not involve the working memory layer.

Semantic Decoding, Schema, and Semantic Memory: Whereas the restoration of an episodic memory trace is mostly sub-symbolic and might lead to an autonoetic experience, our model also contains a semantic decoding for perception and episodic memory. It produces a set of triples on a symbolic level involving indices for concepts and predicates and their embeddings, which are encoded as connection patterns (Tresp et al., 2015). In the cognitive sciences, representations for concepts form what is called a *schema*, which aids in the interpretation of events. Studies have shown that individuals can analyze perceptual information significantly more easily when this information is related to an acquired schema. According to our model, an improvement in the schema would go hand in hand with a refined perception. (Moscovitch et al., 2016) defines a schema as "adaptable associative networks of knowledge extracted over multiple similar experiences", which is in agreement with our model. The same paper states that "memories for recent events draw on interactions between schemas, semantics, and perceptual aspects of an experience, mediated in part by different regions in the anterior and posterior neocortex", which we would interpret as the multilevel processing in our model.

Early in evolution, it was important for individuals to recognize particular classes of objects (e.g., "tigers", "snakes"); object recognition then became the basis for a more meaningful information extraction in form of semantic triples. Our model requires a storage layer which maintains information about already extracted concepts; as proposed already, this storage might have been the initial motivation for the brain to evolutionary develop a working memory in the PFC.

Another by-product in our approach is semantic memory. In our model, semantic memory uses the same layered structure, ignoring the sensory input, and models the prior probability for observing a triple. Thus semantic memory involves only the top three layers and is independent of the context provided by the sensory input. Assume the index for *Cat* is activated in the index layer by some internal or external cue. Then, without any perceptual input, the decoding process might generate, with some probability, that *(Cat, sitsOn, Stove)*. Mathematically, the semantic memory here models $P(p = \textit{sitsOn}, o = \textit{Stove} | s = \textit{Cat})$. In our model, the semantic memory is implemented as the connection pattern between the index layer and the representation layer. In the brain, semantic memory involves the anterior temporal cortex (Moscovitch et al., 2016).

A scene graph describes entities and their relationships. So far we focused on the concept attributes of the entities: *(Dog, bites, Person)* and not identifier attributes as in *(Sparky, bites, Jack)*. Humans have an enormous capacity to represent a large number of entities; but consider a less complex mammal which needs to have only knowledge about a smaller number of specific entities, such as the leader hierarchy in a pack. We propose that, for significant entities, indices are formed as well. So in the previous example, there would be indices for *Jack* and *Sparky*, in addition to the indices for *Person* and *Dog*.

Our model does not explicitly consider properties like *large, red, threatening*. These can be treated as concepts in conjunction with the predicate *hasAttribute* where the visual information for subject and object originate from an identical image region. Also the representations in the sensory layer and in the representation layer might convey attribute information.

Episodic Memory: Most researchers consider temporal coding to be a core function of the hippocampus and not a derived property (Teyler & DiScenna, 1986; Eichenbaum, 2014; Moscovitch et al., 2016). Our model agrees with this view and we assume that an index for a time instance is formed for a sensory input that is associated with an emotion or with novelty (Figure 1). In its simplest form, the t-th row of the matrix A_{epi} copies $\mathbf{f}$. Biologically, time indices might involve a small network of interacting neurons (Quiroga, 2012); together with their connection patterns (in our model A_{epi}) they form mem-

ory traces or engrams. It is assumed that the original purpose of this index was to be able to compare the current event to previously encountered events (familiarity) and their associated actions, supporting the individual in decision making. In the course of evolution, this decision oriented process was repurposed and various cues were able to activate the indices which, using bidirectional connections, are then able to restore a past memory as a personal experience. Subsequently, this function became more elaborate and enabled future-oriented mental time travel to evaluate future consequences of actions. Humans became able to mentally place themselves in the past, in the future, or in counterfactual situations, a process called autonoetic consciousness. Episodic memory traces can also be used to train implicit memories in perceptual and procedural memories or even train complex action patterns (Kumaran, Hassabis, & McClelland, 2016). An episodic memory experience is an active process that involves details of the event and its location (Moscovitch et al., 2016). Sometimes the reconstruction is considered a Bayesian process of reconstructing the past as accurately as possible based on the engram information. According to the standard consolidation theory, indices are consolidated in neocortex, whereas the multiple trace theory proposes that the hippocampal representation maintains its function over long periods and a memory trace is only partially consolidated in neocortex (Moscovitch et al., 2016). In our model, consolidation would involve a reimplementation of an index and its connection pattern.

Experiments

We use the Stanford Visual Relationship data set, which is the basis for many works on scene analysis, e.g., (Baier, Ma, & Tresp, 2017). We used 100 concepts and 70 predicates with 4000 images for training and 1000 for testing. The results of our model are comparable to highly optimized models in other works (Table 1). We also see that the working memory is essential for obtaining good results. The dimensions for the layers are $\mathbf{g}/4096$, $\mathbf{f}/4096$, $\mathbf{h}/500$. For comparison, we report results from (Baier et al., 2017).

We also did experiments where we removed the visual inputs and our model performed as a semantic memory. The table shows that the performance of this derived model is worse than a model optimized on semantic data (Baier) but much better than random. The table also shows that by starting with a perception model (trained on 10 epochs) and then adding (1 or 9) epochs, where we only use the semantic triples without perceptual input, significantly improves the extracted semantic model with only a small performance drop in perception.

Conclusion

We have presented a mathematical model for perception, episodic memory and semantic memory, and related it to cognitive models of the human brain. Our main hypothesis is that episodic memory, semantic memory, and to some degree also working memory, are by-products of the need for humans to extract more meaningful and more complex information from sensory inputs. We could show experimentally that semantic memory can evolve as a by-product of perception. The semantic memory represents prior probabilities, which might be an interesting basis for a Bayesian brain interpretation. We propose that the model we presented is in a sense minimalist, containing necessary perceptual components.

Table 1: ph stands for phrase detection and pr stands for predicate detection. In phrase detection, a triple with its corresponding bounding boxes is considered a success, if both the triple and the bounding boxes are correctly detected. In predicate detection, subject concept and object concept are given and the task is to predict the predicate (Baier et al., 2017). For z-s-ph/z-s-ph (zero shot), we only evaluate the test set performance on triples that did not occur in training. The first row (Model) shows results for our model. The fourth row (Baier) shows the results from literature. Dir are results where we removed the working memory. Our model gives better results for the zero-shot experiments. The last two columns report recall results for only the semantic memory. The first row shows results where the semantic memory was extracted from our perceptual model. The result (82.46 and 53.53) are worse than the result for Baier, where the latter was trained directly on the triple data. S1 and S9 show results where we added 1 and 9 epochs of pure semantic training to the perception model. We see that the semantic model improves significantly with almost no degradation on perception.

Method	ph	z-s-ph	pr	z-s-pr	@10	@1
Model	23.45	**10.95**	93.32	78.79	82.46	53.53
S1	23.32	10.44	93.17	**80.07**	93.46	67.55
S9	22.61	9.24	92.77	79.47	94.77	68.68
Baier	**25.11**	7.96	**93.81**	76.05	**95.86**	**70.50**
Dir	11.13	7.87	77.19	65.61	-	-
Rand	0.01	0.00	18.53	16.51	0.08	0.01

References

Baier, S., Ma, Y., & Tresp, V. (2017). Improving visual relationship detection using semantic modeling. In *ISWC.*

Eichenbaum, H. (2014). Time cells in the hippocampus. *Nature Reviews Neuroscience*, *15*(11).

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? *Trends in Cognitive Sciences*, *20*(7), 512–534.

Moscovitch, M., et al. (2016). Episodic memory and beyond. *Annual review of psychology*.

Quiroga, R. Q. (2012). Concept cells: the building blocks of declarative memory functions. *Nat Rev Neurosci*, *13*(8).

Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory. *Behavioral neuroscience*.

Tresp, V., & Ma, Y. (2016). The tensor memory hypothesis. In *Nips workshop on representation learning).*

Tresp, V., et al. (2015). Learning with memory embeddings. *NIPS Workshop on Representation Learning*.