ARTICLE **Communicated by Alessandro Treves**

A Model of Semantic Completion in Generative Episodic Memory

Zahra Fayyaz
zahra.fayyaz@ini.rub.de
Aya Altamimi
aya.altamimi@ini.rub
Institute for Neural Computation, Faculty of Computer Science, Ruhr University Bochum, 44801 Bochum, Germany

Carina Zoellner
carina.zoellner@rub.de
Nicole Klein
Nicole.Klein-g5y@rub.de
Oliver T. Wolf
oliver.t.wolf@rub.de
Cognitive Psychology, Institute of Cognitive Neuroscience, Faculty of Psychology, Ruhr University Bochum, 44801 Bochum, Germany

Sen Cheng
sen.cheng@rub.de
Laurenz Wiskott
laurenz.wiskott@ini.rub
Institute for Neural Computation, Faculty of Computer Science, Ruhr University Bochum, 44801 Bochum, Germany

Many studies have suggested that episodic memory is a generative process, but most computational models adopt a storage view. In this article, we present a model of the generative aspects of episodic memory. It is based on the central hypothesis that the hippocampus stores and retrieves selected aspects of an episode as a memory trace, which is necessarily incomplete. At recall, the neocortex reasonably fills in the missing parts based on general semantic information in a process we call semantic completion. The model combines two neural network architectures known from machine learning, the vector-quantized variational autoencoder (VQ-VAE) and the pixel convolutional neural network (PixelCNN). As episodes, we use images of digits and fashion items (MNIST) augmented by different backgrounds representing context. The model is able to complete missing parts of a memory trace in a semantically plausible way up to the point where it can generate plausible images from scratch, and it generalizes well to images not trained on. Compression as well

Neural Computation **34, 1841–1870** (2022)
https://doi.org/10.1162/neco_a_01520

as semantic completion contribute to a strong reduction in memory requirements and robustness to noise. Finally, we also model an episodic memory experiment and can reproduce that semantically congruent contexts are always recalled better than incongruent ones, high attention levels improve memory accuracy in both cases, and contexts that are not remembered correctly are more often remembered semantically congruently than completely wrong. This model contributes to a deeper understanding of the interplay between episodic memory and semantic information in the generative process of recalling the past.

1 Introduction

Episodic memory enables us to remember personally experienced events and depends on the hippocampus (Clayton, Salwiczek, & Dickinson, 2007). Semantic information, on the other hand, is represented in neocortical areas and captures general facts and regularities of the world around us (Reisberg, 2013). Early concepts of episodic memory were based on the storage model, according to which the content of the memory more or less faithfully reflects the content of the experience (Tulving, 1972). This view is oversimplified since it reduces episodic recall to a mere readout process of stored complete information. However, overwhelming empirical evidence suggests that the recalled memories can be influenced by other information acquired before and after the encoding, as well as the context of encoding and recalling (Hemmer & Steyvers, 2009b). Pioneering studies suggest that semantic interpretations, rather than sensory inputs, are stored in memory (Bartlett, 1995; Sachs, 1967) and that memories are reconstructed during recall (Bartlett, 1995). In word list studies using the Deese–Roediger–McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995), participants "remember" semantically related words that were not on the study list when asked to retrieve the words studied earlier. There is also evidence that semantic and episodic memories interact and complement each other during retrieval (Greenberg & Verfaellie, 2010). For instance, Devitt, Addis, and Schacter (2017) found in a meta-analysis of eight studies based on autobiographical interviews that when participants report an episode, the internal (episodic) details and external (semantic) details they use are negatively correlated. The participants apparently use semantic information to compensate for insufficient episodic detail in their memory. Other examples are experiments by Bartlett (1995), where participants of nonmatching cultural backgrounds recalled folk tales. The recalled stories were distorted to match the participants' cultural background (semantic information). Finally, there are also paradigmatic examples of memory adjustments due to social context (Deuker et al., 2013; Hirst & Echterhoff, 2012), self-model (Axmacher, Do Lam, Kessler, & Fell, 2010), stress (Herten, Otto, & Wolf, 2017; Wolf, 2019), and many other factors (Addis, 2020; Schacter & Addis, 2020).

Few contemporary researchers would oppose the idea that episodic memory is, at least to a certain degree, generative in nature (Greenberg & Verfaellie, 2010). Nonetheless, most of the existing computational models (including some of our own: Cheng & Frank, 2011; Cheng & Werning, 2016; Neher, Cheng, & Wiskott, 2015) adopt the storage view, where memories are preserved and later retrieved faithfully (Becker, 2005; Jensen & Lisman, 1996; Rolls, 1995). Such models are usually tested with either random patterns or abstract spatial representations (Becker, 2005; Cheng & Frank, 2011; Jensen & Lisman, 1996; Neher et al., 2015; Rolls, 1995) but not with realistic sensory input. With such artificial input patterns, it is rather suggestive to think in terms of mere storage memory, since there is not much structure in the input that could be exploited. However, in natural stimuli, there is a rich hierarchical structure of features and statistical relationships, which was not exploited and not even considered in these models.

In order to model the generative process of episodic memory recall, we believe it is important (1) to use (real-world) input patterns as stimuli with enough structure that can be exploited by a semantic system for a generative process, (2) to discard some fraction of the input patterns during storage to model the inevitable loss of information in the brain due to the attentional bottleneck, and (3) to include a generative element in the model that is able to reasonably fill in the missing parts according to learned semantic information. Discarding a fraction of an input pattern can be done in at least two ways, by lossy compression and by selection (either before or after compression). The former refers to a process like mp3 encoding, a compression that tries to discard only the information that is irrelevant or trivially recoverable from what is being stored. The latter refers to a process where some part is selected for storage and another is discarded altogether; for example, from a picture of a water mill at a creek, the mill could be attended to and stored while the creek could be ignored. When recalling the mill, our semantic system probably complements it with a creek, but the creek might look very different from the original one, and we are probably not even aware of this. Such a process of scenario construction (Cheng, Werning, & Suddendorf, 2016) is able to generate a semantically plausible and consistent memory experience from an incomplete memory trace without us even noticing that the recall is not faithful. When playing the mp3 encoded song, on the other hand, there might be some noise due to the strong compression, but all in all, the song is faithfully reconstructed. We refer to the representation reduced by compression and selection as the gist, which is stored in a memory trace and from which the original episode can be reconstructed, either quite faithfully if only compression is involved or at least plausibly if also selection is involved.

Following the lossy compression approach from a perspective of rate-distortion theory and efficient coding, Bates and Jacobs (2020) and Nagy, Török, and Orbán (2020) have modeled perception and episodic memory as a generative process. Bates and Jacobs argue that a capacity-limited

perceptual system like the brain should use prior knowledge and take into account task dependencies to compress the input into an optimal representation. Nagy et al. have demonstrated that systematic distortions in memory are similar to the distortions that are characteristic of a capacity-limited generative model adapted to an environment for compression. They use a variational autoencoder (VAE) as a model for memory (Kingma & Welling, 2013). A VAE is an autoencoder architecture that maps input images to a plain gaussian model distribution and back again to images. Episodic memory can then be modeled by storing the location in the gaussian as a memory trace, which is a low-dimensional feature vector. From such a memory trace, the full image can be reconstructed. The lossy compression from input to gaussian is so extreme in this case that memory recall is largely generative. It is even possible to generate new images without any input or memory trace by sampling from the gaussian and then decoding this vector. Any such image looks similar to the images seen during training; in fact, the system can only represent such seen images or interpolations of them. These models are already generative but with a focus on optimal compression and decompression, while our focus is attentional selection and semantic completion.

Hemmer and Steyvers (2009a), among others, have suggested a Bayesian account of reconstructive memory that captures the prior knowledge interaction with episodic memory. Although their framework of generative memory is very close to ours, their work is more concerned with why generative episodic memory would be advantageous, but in this work, we suggest a model on how this process can happen.

One might think that a storage memory would be advantageous over a generative one because of its faithfulness. However, scenario construction during recall is essential to the etiological function of episodic memory because it provides far more flexibility to deal with missing data and to adjust to variable demands and constraints than a faithful reproduction of past experiences could. Moreover, the already acquired semantic knowledge can help to improve memory efficiency. Put simply, generativity is a useful feature in episodic memory, not an aberration (Schacter, Guerin, & Jacques, 2011).

2 Computational Model of Generative Episodic Memory

Based on a biologically motivated conceptual framework and using methods from machine learning, we have developed a computational model that allows us to investigate semantic completion in a generative episodic memory on real world images on a fairly abstract level but still analogous to concrete brain structures, so that predictions on a behavioral level but also about neural processes are possible (Fayyaz, Altamimi, Cheng, & Wiskott, 2021).

2.1 Conceptual Framework. We hypothesize that generative episodic memory works as follows:

1. Sensory input patterns that make up the episode are perceived by a hierarchically organized network and transformed into a hierarchical *perceptual-semantic* representation in cortical areas, such as the visual system.
2. Some elements of this representation are selected for storage in episodic memory. We call this the *episodic gist*.
3. The episodic gist is stored in hippocampal memory as a set of pointers to the corresponding perceptual-semantic elements in cortical areas, referred to as *memory trace*.
4. Triggered by some external or internal cue, or even spontaneously, the memory trace can be reactivated.
5. The pointers in the memory trace reactivate corresponding perceptual-semantic elements in cortical areas.
6. *Semantic information* in the cortical areas complements the reactivated elements by means of a recurrent dynamics to construct a plausible full representation from the incomplete gist stored in the memory trace, a process we refer to as *semantic completion*.

Several of these steps and concepts deserve closer consideration.

We speak of a *perceptual-semantic* representation, because (1) we consider the transformation from the raw input to a high-level semantic representation a gradual process, as is well known for deep neural networks (Liuzzi, Aglinskas, & Fairhall, 2020; Zhang, Han, Worth, & Liu, 2020), and (2) while we mainly remember high-level aspects, we can also remember quite low-level aspects of an episode, such as the exact color and shape of an object. So we believe that there is no clear-cut distinction between perceptual and semantic representations (Davis et al., 2021) and therefore refer to the corresponding network as the *perceptual-semantic network*. A prototypical example is the visual system, which is hierarchically organized from low-level perceptual in primary visual cortex (V1) to high-level semantic in inferior temporal cortex (IT) (Felleman & VanEssen, 1991). The very rapid recognition of high-level features in images (Thorpe, Fize, & Marlot, 1996) suggests that the generation of the semantic representation is largely done by feedforward projections. The recurrent and feedback connections in turn are instrumental for recreating a full perceptual-semantic representation from a memory trace (Takeda, 2019; Xia, Guan, & Sheinberg, 2015), although they certainly also contribute during perception.

The concept of *gist* is well known (Koutstaal & Schacter, 1997; Oliva, 2005; Sachs, 1967). The episodic gist (Cheng & Werning, 2016) contains essentials about the episode that are selected dynamically depending on attention and the context (Graham, Simons, Pratt, Patterson, & Hodges, 2000). They may be detailed in some cases and general and vague in others.

Episodic *memory traces* are pointers to perceptual-semantic elements of the sensory input rather than the representations of the input itself (Fang, Rüther, Bellebaum, Wiskott, & Cheng, 2018; Teyler & Discenna, 1986). It is not clear where the expansion from pointer to the full representational element happens. It could be on the way from hippocampus to cortical areas (Teyler & Rudy, 2007). Or the pointers in hippocampus activate pointers in cortex and only those then activate the full representational element (Reber, Stark, & Squire, 1998). Our model is more in line with the second view, because in our model, cortical semantic completion happens on the pointer level.

Semantic information is usually extracted from multiple experiences, is mostly categorical, and refers to the prototypical properties of objects or people and their relationships (Collins & Quillian, 1969; Tulving, 1972). Evidence from patients with semantic dementia suggests that semantic information is vital for episodic memory recall (Irish & Piguet, 2013). It is plausible to assume that the process of semantic completion is mainly done by recurrent connections within the perceptual-semantic network, because that is the site where the semantic information is stored (O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013) and because recurrency is well suited to perform pattern completion (Hopfield, 1982).

Because of its generative nature, we call the retrieval process *scenario construction*.

Next we describe the network architecture with which we capture the key aspects of this conceptual framework on a level abstract and efficient enough to be applicable to real world images. Since the storage and retrieval of memory patterns in the hippocampus have been modeled far more extensively than the generative aspect of episodic memory, we focus on the latter in this study.

2.2 Network Architecture. While we obtained our simulation results with specific network implementations, we do not mean to suggest that these implementations are the very ones that the brain uses. While certain properties of the networks are important for the function of the model, and we try to identify and analyze these aspects, other properties are incidental. Our computational model consists of two networks known from the field of machine learning: (1) a vector-quantized variational autoencoder (VQ-VAE) (van den Oord, Vinyals, & Kavukcuoglu, 2017), which models the perceptual-semantic network with feedforward and feedback connections, and (2) a pixel convolutional neural network (PixelCNN) (van den Oord et al., 2016), which models the recurrent connections at the semantic end of the perceptual-semantic network (see Figure 1). Due to its role, we refer to the PixelCNN also as the *semantic completion network*. We do not model the hippocampus mechanistically but assume perfect storage and retrieval, modeled simply by writing into and reading out a variable.

A VQ-VAE consists of an encoder, a decoder, and a latent representation between these two. To obtain a quantized latent representation, there is also

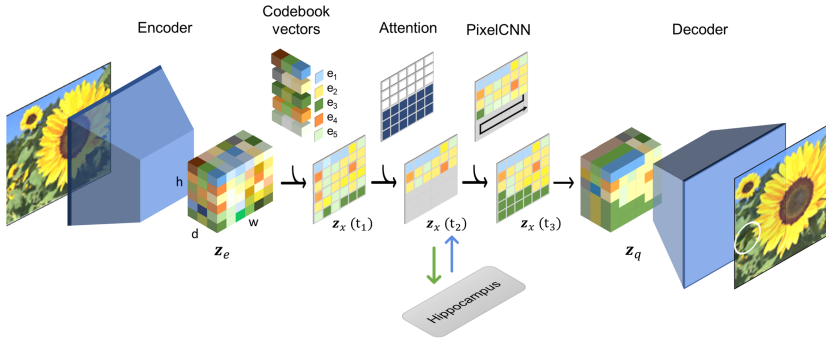

Figure 1: Network architecture. The model combines a VQ-VAE (whole pipeline from left to right) and a PixelCNN. The encoder of the VQ-VAE, which consists of several convolutional layers, converts the input image into an array $\mathbf{z}_e$ of $w \times h$ d-dimensional feature vectors. Each feature vector is then assigned to the closest codebook vector $\mathbf{e}_l$ to create the index matrix $\mathbf{z}_\mathbf{x}$ containing the indices l, from which an array $\mathbf{z}_q$ of the $w \times h$ corresponding d-dimensional codebook vectors $\mathbf{e}_l$ can be constructed. The decoder then reconstructs the original input based on the quantized array $\mathbf{z}_q$. Selective attention is modeled by discarding consecutive entries in the lower part of the index matrix (transition from t_1 to t_2.). The missing part is filled in by the PixelCNN in a recurrent process that performs semantic completion (transition from t_2 to t_3). The completion is plausible but not necessarily faithful; for example, some flowers in the background are missing here (white circle).

a set of codebook vectors, which are optimized by vector quantization but otherwise fixed. The VQ-VAE processes an input image in the following steps:

1. The VQ-VAE first converts the image of size 28×28 (or 32×32 for images with context) with a convolutional neural network (the encoder) to an array of $w \times h$ d-dimensional feature vectors ($w = h = 7$ for the original images and $w = h = 8$ for images with context; d is set to 64). The positions within the array correspond to a grid of subsampled locations in the image. Thus, this array still has a coarse spatial resolution, and the feature vectors are a description of the image around these locations.
2. This array is then converted to an array of $w \times h$ indices (called index matrix for short), each index indicating the d-dimensional codebook vector most similar to the feature vector at that position. This can be viewed as a quantization step that makes the representation more categorical (i.e., more semantic). In particular, it has been shown that each learned codebook vector in a VQ-VAE corresponds to some specific feature of the input (van den Oord et al., 2017). Up to this point,

the network has compressed the image into a more abstract and semantic representation.

3. In order to recover an image from this compressed latent representation, the array of indices is converted back to an array of $w \times h$ d-dimensional vectors by replacing the indices of the index matrix by the corresponding codebook vectors, which should be similar to the array of feature vectors of step 1.
4. The array of codebook vectors is then decoded by a deconvolution neural network (the decoder mirroring the encoder) to produce an image of original size.

A VQ-VAE alone can convert an image into a more abstract and semantic representation and back again, but it is not generative in the sense that it could produce new reasonable images from scratch or complement incomplete images. This is fine as long as the full index matrix is available. However, we also want to model attentional selection, in which case only part of the index matrix can be recalled from memory. In such cases, we need a generative component that is able to fill in the missing parts of the image. For that, we use a PixelCNN.

A PixelCNN is a probabilistic autoregressive generative model that is able to continue sequences of numbers. It can fill in missing pixel RGB values in an image in a fixed sequence—for instance, row-wise from top left to bottom right. Completing an image with a PixelCNN is a time-consuming process. We apply the PixelCNN not to the image but to the latent index matrix, which is much faster, since the index matrix is much smaller than the image. Since a PixelCNN only works in one particular order, we can model attentional selection only in a primitive form by keeping the upper rows and neglecting the lower rows of the index matrix. The level of attention determines how many rows to keep. The remaining representation of the input is what we call the episodic gist, and it is stored and retrieved in episodic memory as a memory trace.

The VQ-VAE as well as the PixelCNN are both trained on a large set of training images. First, the VQ-VAE is trained to reconstruct the input images as much as possible, despite the strong compression in the latent representation (i.e., the index matrix). The weights of the encoder and decoder are optimized as well as the codebook vectors. Once the VQ-VAE is trained, the PixelCNN can be trained on the index matrices generated by the trained VQ-VAE from the training image. See section 5 for further details on the VQ-VAE and the PixelCNN.

Our model is designed to reflect our hypotheses on generative episodic memory. That is, the stored gist has far less information content than the input images; nonetheless, the input can be reconstructed from it. The model captures complex statistics of the input and also reflects the generative nature of episodic memory that has been observed in many studies. When the attention is low (only a small part of the index matrix is stored), the

recalled memories are not necessarily faithful; still, they are valid and likely reconstructions, typical of the training data.

2.3 Analogy to the Brain. Although VQ-VAE and PixelCNN originate from the field of machine learning, we believe they can be related to aspects of neural processing in the brain, and they are an appropriate level of description for our purposes here.

The encoder network of the VQ-VAE might correspond to the feedforward processing in the visual system, which results in abstract object representations in the inferior temporal (IT) cortex. Many studies have suggested a correspondence between the hierarchy of the human visual areas and layers of CNNs (Kuzovkin et al., 2018; Lindsay, 2021; Yamins et al., 2014). The decoder has a structure symmetrical to the encoder and might be similar to the feedback connections from higher levels of the visual system to lower ones. Experimental results suggest that during retrieval, a cortical representation of the memory is formed in the lower levels of the visual pathway through feedback connections (Takeda, 2019; Xia et al., 2015). Some studies have used an autoencoder structure to model the feedback connections in the visual pathway (Al-Tahan & Mohsenzadeh, 2021). In our model, the decoder generates a cortical representation of the memory in its layers down to the image level, which we take as a readout of the cortical representation of the memory during retrieval. However, we do not mean to suggest that the brain activates sensory representations at the input level. A body of research also indicates that there is semantic learning at the level of the visual system, reflected in our model by the whole VQ-VAE network (Hu & Jacobs, 2021).

The PixelCNN learns statistical relationships between the elements of the latent representation of the VQ-VAE by repeated exposure, that is, it learns semantic information from episodes akin to how it is hypothesized also for the brain (Michaelian, 2011). It is then able to fill in missing elements in the semantic representation of an image. We hypothesize that this is akin to recurrent dynamics in the higher cortical areas that can fill in missing information in a semantically consistent and expected way (Carrillo-Reid & Yuste, 2020; Tang et al., 2018).

We do not model storage in and retrieval from the hippocampus mechanistically; we simply store and recall a perfect copy of the selected parts of the index matrix, which represents the episodic gist. Storing just the indices of the codebook vectors, and not the vectors themselves, is consistent with the indexing theory of hippocampal memory (Teyler & Discenna, 1986), although we would argue that our indices are also represented in the cortex, so that semantic completion can take place there.

3 Results

Our model is able to process real-world images, and we believe that sufficiently rich statistical structure in the input patterns is essential for a

meaningful simulation of episodic memory. However, large images require large data sets and are computationally expensive to process. As a compromise, we use the well-known MNIST data set of handwritten digits (LeCun, 1998), which is real-world and has a clear semantic structure of 10 digit classes, 0 to 9, so that simulations can be conducted efficiently. The 28×28 pixel images show white digits on black background with gray values between 0 and 256. For some experiments, we also use the Fashion MNIST data set (Xiao, Rasul, & Vollgraf, 2017), which is similar to the digit data set but shows 10 classes of fashion items rather than digits. First, we illustrate the behavior of the system, and then model a concrete memory task and compare it to experimental results (Zoellner et al., 2021).

3.1 Semantic Learning at the Level of the VQ-VAE. Semantic learning within the VQ-VAE has two aspects: (a) the semantic learning within the encoder and decoder and (b) the categorization of the feature vectors by vector quantization. Both can potentially contribute to memory efficiency in our model. Interestingly, we find that the model works well in a regime where the encoder alone actually increases the size of the representation (from $28 \times 28 = 784$ to $7 \times 7 \times 64 = 3136$ numbers), and it is the quantization that leads to a strong compression within the VQ-VAE. Since input images have 28×28 pixels with 256 gray values and index matrices z_x have 7×7 entries with 20 codebook vectors, the compression ratio as defined in van den Oord et al. (2017) is $(28 \times 28 \times \log_2(256))/(7 \times 7 \times \log_2(20)) = 29.6$ in bits.

However, semantic learning might not only help to improve memory efficiency. Here we investigate how much semantic learning helps dealing with noise and how well this generalizes from one set of images to another.

Noisy MNIST images were generated as follows. First, a noise template was generated by sampling an array of 28×28 independent and identically distributed (i.i.d.) noise from a gaussian distribution with zero mean and unit variance. This template was then added to an image (having gray values between -0.5 and 0.5) with a weighting factor between 0.01 and 1, that is, between 1% and 100%. An image with 100% noise therefore still has some original image information left. Noisy images were not clipped or normalized back to $[-0.5, 0.5]$. Using a fixed noise template for all noise levels realizes so-called frozen noise, which leads to smoother and more reliable results, because it eliminates random fluctuations between different noise levels.

To test the performance of the system on a set of images, we have to distinguish successes from failures. Although the mean squared distance between original image and reconstructed image is an obvious and frequently used measure, it is not very useful as a measure of perceptual similarity (Mathieu, Couprie, & LeCun, 2015). We have therefore trained two classifiers, each consisting of a three-layer convolutional neural network, to recognize the 10 digits (trained to a level of 98% correct classification on

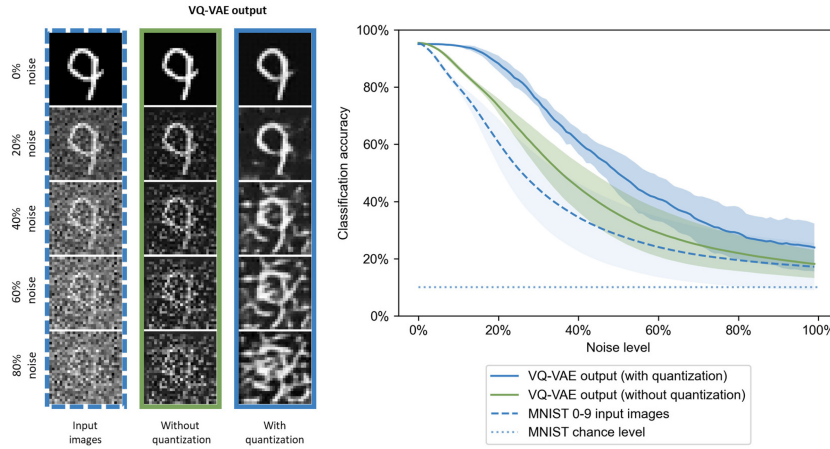

Figure 2: Effect of autoencoding and quantization on noise reduction and classification performance. Left: A test image with different noise levels (left column) is processed by a VQ-VAE without (middle column) and with (right column) vector quantization. Right: The curves show the classification accuracy determined by an MNIST digit classifier for the three settings against input noise level. The VQ-VAE network with quantization achieves the highest classification accuracy. The curves are averaged over 10 runs on 10,000 test images.

test data) or fashion item classes (91% correct classification on test data) and evaluate the performance by the classification accuracy, that is, the percentage of reconstructed images that are recognized correctly by the trained classifier.

We ran the digit or fashion MNIST classifier on the noisy images directly, on images reconstructed by a VQ-VAE without quantization (semantic learning (a)), and on images reconstructed by a VQ-VAE with quantization (semantic learning (a + b)). For the VQ-VAE without quantization, training and testing were done without quantization; thus, this corresponds to a plain autoencoder. For each noise level, we did 10 runs with a different seed for the (frozen) noise template and the VQ-VAE, each one tested on 10,000 different test images—about 1,000 per digit or class. Training was done on 60,000 images. Each of the four sets of 10 runs (with and without vector quantization $\times$ classes 0–4 and 5–9) had the same sets of 10 seeds for the noise and for the VQ-VAE to make the results more comparable.

Figure 2 shows how semantic learning contributes to robustness to noise. The VQ-VAE without quantization already reduces noise and improves performance, as is well known from autoencoders (Bhowick, Gupta, Maiti, & Shankar, 2019). Quantization in the VQ-VAE improves performance even further, presumably by dragging the internal representation toward what it

has seen before (i.e., toward the codebook vectors), thereby imposing a denoising effect. The quantization also leads to a slightly more ragged curve, presumably due to the abrupt transitions from one codebook vector to the next for increasing noise level leading to a more irregular classification behavior (see Figures S2 to S4 in the supplement). The supplementary figures also show, first, that the VQ-VAE without quantization usually switches only once from correct to wrong classification, while with quantization, it switches back and forth several times. Second, the VQ-VAE without quantization trained on digits 0 to 4 behaves almost identically to the one trained on digits 5 to 9, while there is a marked difference between these two networks with quantization in terms of when the misclassification starts and to which other digit class. Third, for the VQ-VAE without quantization, the switches to misclassification are distributed smoothly over the noise levels while the VQ-VAE with quantization has certain preferred noise levels at which a switch to misclassification happens for many test images simultaneously, visible as vertical color bands. This might be due to indices in the index matrix that initially represent black background and at a certain noise level switch to another index indicative of a wrong digit class. Since the background at that location is black and the noise pattern is identical across all test images, this switch happens at the same noise level. Fourth, there is no preference for the digit classes trained on (either 0–4 or 5–9) when tested at high noise levels, as one might have expected. This underlines the perfect generalization from digits 0–4 to 5–9 and vice versa.

The latent representation in a VQ-VAE still has some spatial resolution and can take advantage of the combinatorics in the index matrix to generalize to images with a different distribution from the one trained on. To study generalization, we tested VQ-VAEs trained on different training sets (MNIST digits 0–4, digits 5–9, fashion item classes 0–4 as well as classes 5–9) on different test sets of the same four groups, resulting overall in 16 comparisons across different noise levels. We distinguish three cases: *in sample* indicates that the training and test set were from the same group (e.g., both digits 0–4); *out of sample* indicates that the training and test set were from different groups within the same data (e.g., training on digits 0–4 and testing on digits 5–9); and *out of distribution* indicates that the training and test set were from different data sets (e.g., training on fashion item classes 0–4 and testing on digits 0–4). Within one comparison graph, we keep the test database constant, and we average the results over the two corresponding groups (e.g., digits 0–4 and digits 5–9). This eliminates confounding effects by different difficulty levels of the test sets. At high noise levels, the VQ-VAE tends to generate an output that consistently gets classified as one of the 10 classes. This seems to largely depend on the noise template with a preference for the digits 2, 3, 5, and 8 (see Figure S3). For a single run, this could lead to chance levels of either 0% or 20%, depending on whether the preferred digit is within the test set. Averaging over 10 runs and different combinations of training and test set largely eliminates this effect and

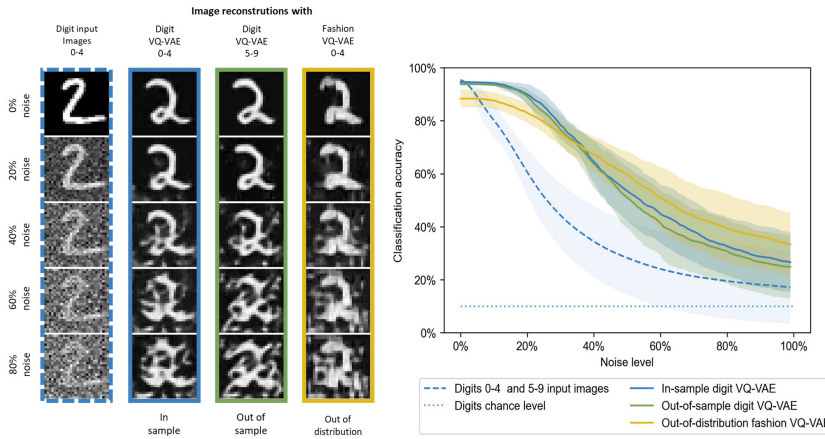

Figure 3: Generalization of the VQ-VAE to other data sets tested on digits. Left: A test image from digits 0–4 with different noise levels (left column) is processed by a VQ-VAE network trained on digits 0–4 (in sample), on digits 5–9 (out of sample), or on fashion item classes 0–4 (out of distribution). Right: Noisy images as well as their image reconstructions from the three types of VQ-VAEs are tested for classification accuracy using an MNIST classifier for digits 0–9. The curves are averaged over 10 runs on 10,000 test images, as well as different combinations of training and test set, see main text.

results in a convergence to the expected chance level of 10%. Curves are averaged over training/test combinations 0–4/0–4 and 5–9/5–9 of the same data set for in-sample, over 0–4/5–9 and 5–9/0–4 of the same data set for out-of-sample, and over 0–4/0–4, 0–4/5–9, 5–9/0–4, and 5–9/5–9 of the two different data sets for out-of-distribution results.

Figure 3 shows that the out-of-sample digit VQ-VAE performs as well as the in-sample digit VQ-VAE on digit MNIST test sets, indicating perfect generalization across digits. The fashion VQ-VAE performs slightly worse at low noise levels and better for high noise levels, which is surprising. We also performed tests on the fashion MNIST test sets, shown in Figure 4. In this case, the digit VQ-VAE performs significantly worse, indicating reduced out-of-distribution generalization. Thus, we see that generalization to different image sets is not symmetric. One possible reason is that fashion items contain strokes, which are important for digits, but digits do not contain larger white areas, which might be crucial for the fashion items.

All curves reach values higher than chance level at 100% noise level, which we attribute to the still remaining image information at that noise level.

The generalization capability demonstrated here is characteristic of a VQ-VAE. A VAE, for instance, would not be able to do that because it maps

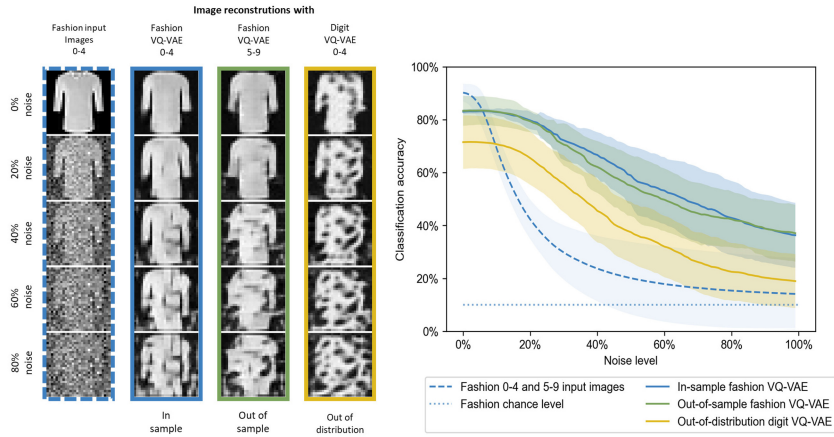

Figure 4: Generalization of the VQ-VAE to other data sets tested on fashion items. Same conventions as for Figure 3 but with digit and fashion database swapped.

the input onto just one feature vector and can therefore not take advantage of the combinatorics of feature vectors like the VQ-VAE does. We have previously tried to use a VAE in our model, but it failed to represent the out-of-sample data. Therefore, it was not suitable for modeling the experimental results of episodic-semantic confilict resolution that is described next.

3.2 Scenario Construction by Semantic Completion. At the core of our model is the concept of an episodic gist, which is incomplete but can be complemented by semantic information to reconstruct a full scenario from a partial memory trace. What is being stored in the memory trace (i.e., what makes up the episodic gist) is largely determined by attention. In our model, attentional control is somewhat constrained and only determines how many consecutive elements of the latent representation (i.e., indices of codebook vectors), are stored row-wise starting in the upper left corner. For low attention, only the upper two out of eight rows might be stored; for high attention, the upper six and a half rows might be stored. The remaining part, if needed, has to be constructed based on semantic information. It is important to note that attentional selection does not apply to the images but to the latent representation in form of the index matrix.

To illustrate the effect of semantic completion, we have compared recalled images from memory traces at several different attention levels (see Figure 5). At high attention, the reconstruction is faithful; at low attention, the reconstruction is not necessarily faithful, but it is plausible given the attended parts. Without any attention, the system can also create new images from scratch, which could be related to dreaming.

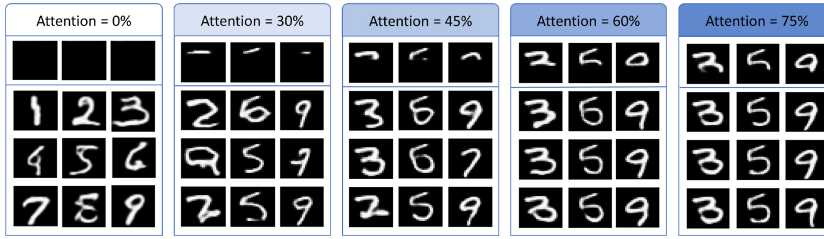

Figure 5: Semantic completion. Incomplete latent representations in the form of the index matrix are completed by the PixelCNN through a process we call semantic completion. The index matrix is already 30 times smaller than the image, and only part of that representation is selected by attention. The attention level is defined as the percentage of the index matrix that is stored in the memory trace. The partial patterns in the first row visualize the episodic memory traces. The next three rows visualize three instances of semantic completion based on the same incomplete memory traces. The first panel with 0% attention shows images generated from no information in the memory trace, one exemplar per digit from 1 to 9, which can be interpreted as dreaming. Note the higher reliability of the completion process with increased attention.

3.3 Improved Memory Efficiency by Semantic Completion. Our analysis shows that semantic learning in the VQ-VAE, in particular the quantization, leads to a compression by a factor of about 30. Here we want to investigate how much semantic completion in the PixelCNN can contribute to memory efficiency. We ran simulations with semantic completion by a fully trained PixelCNN, by a partially trained PixelCNN, and without semantic completion at all, which we refer to as *strong*, *weak*, and *none* (semantic completion), respectively. The strong network was trained for 40 epochs and had a loss of 0.65 (categorical cross-entropy), the weak network was trained for 3 epochs and had a loss of 0.75, and for the none case, the pixels in the nonattended parts were set to black. Figure 6 shows how the three different training levels of the PixelCNN affect the semantic completion of memory traces with different attention levels. The recall performance was measured by a simple convolutional digit classifier with an accuracy of 99% on original test images. Semantic completion can significantly contribute to recall quality measured by classification accuracy and thereby save memory, although the saving is not nearly as large as for the compression by the VQ-VAE, maybe a factor of two.

3.4 Modeling Episodic-Semantic Conflict Resolution in Humans. An important goal of our modeling effort is to reproduce experimental results from episodic memory research and eventually make suggestions and predictions for new experiments. Here we relate to a recent experiment by Zoellner et al. (2021) on the conflict resolution between episodic memory and semantic information in humans.

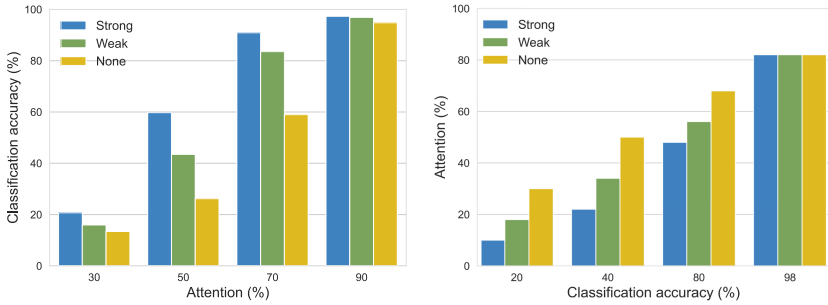

Figure 6: Improved memory efficiency by semantic completion. Left: A classifier trained on MNIST is used to evaluate the classification accuracy on recalled patterns of different attention levels. The stronger the PixelCNN with its semantic information, the better the accuracy for any given attention level. Right: For any expected classification accuracy, the stronger PixelCNN requires less attention.

The experiment took place on three days. On day 1, participants were asked to explore an apartment in a virtual environment. The apartment had three main rooms: bedroom, kitchen, and bathroom, each room containing eight household objects. Half of the objects were in a room where one would expect such an object (e.g., an alarm clock in the bedroom), which is referred to as *congruent context*. The other half of the objects were placed in an unexpected room (e.g., a toaster in a bathroom), which is referred to as *incongruent context*. Participants first familiarized themselves with the apartment and were then instructed to perform some tasks to interact with half of the objects, for example, to make a sandwich using the toaster. Such objects are called *task relevant*, the others *task irrelevant*. This design provided some control over the level of attention with which the different objects were perceived (see Figure 7, left).

Participants' memory was then tested on the next day with a *recognition task*. In this task, participants ranked on a 6-scale from -3 (surely not seen) to 3 (surely seen) how confident they were that they had seen a specific household object and, if they think they had seen it, decided which room it was in (see Figure 7, right). In addition to the 24 household objects from the apartment, 24 similar-looking distractor objects were presented as well, to avoid random guessing. Each object was presented once. Confidence level was highly correlated with task relevance (mean confidence 2.4) versus task irrelevance (mean confidence 0.9). The same task was repeated after seven days to check how memory changes over time. Since the results show no significant difference between day 2 and day 8, and we are not modeling memory accuracy over time, we pooled the data from the two days.

In the recognition task, there are three possible outcomes in the incongruent cases if the object has been remembered: the semantically incongruent

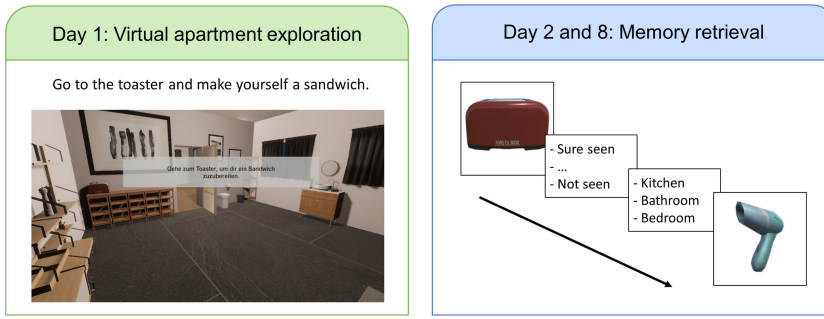

Figure 7: The recognition memory task. Participants explored an apartment in a virtual environment containing a number of household objects, half of which were placed in an incongruent room. Later they were asked to rate their confidence having seen an object in the apartment and to indicate which room the object was in.

but episodically correct room is remembered (episodic recall), the semantically congruent but episodically incorrect room is remembered (semantic recall), or the semantically incongruent and episodically incorrect room is remembered (wrong recall). In the congruent cases, the semantically congruent recall is also episodically correct (correct recall); the other two rooms are both wrong recalls. The episodic recalls are sometimes also called correct recalls for convenience.

To reproduce the experiment with our model, we padded the images to size 32×32 and augmented them with three different backgrounds in the bottom half: a background with triangles for digits 0 to 3, squares for digits 4 to 6, and circles for digits 7 to 9. The backgrounds provide a context for the digits that can be congruent, such as a 3 in front of triangles, or incongruent, such as a 5 in front of circles (see Figure 8). The background covers only the bottom half of the images, so that it is possible to store images that show the object without background, that is, only the upper half. With a more flexible attention mechanism, we could also use backgrounds across the whole image. The model implicitly learns the association between a digit and its congruent background by repeated exposure, for example, 2 is always shown together with triangles. The model is trained only on congruent images and then tested on congruent as well as incongruent images. The different levels of attention are modeled by selecting various fractions of the latent representation, the index matrix. The attention levels were not directly measured in the experiment; however, we believe that the reported confidence levels are a good proxy for attention.

For the model simulation, we proceed as follows: First, the perceptual-semantic network, VQ-VAE and PixelCNN, is trained on the congruent data set. Then a number of congruent and incongruent images are shown to the

Figure 8: Sample images showing combinations of digits and backgrounds and the reconstructions of the model. Top: Congruent and incongruent images for training and testing. Bottom: Recalled images from memory traces with medium and high attention. Green indicates correct/episodic recalls; yellow indicates semantic recalls; red indicates wrong recalls; gray indicates cases where the reconstructed digit has not been recognized correctly. We have dropped the latter from our analysis as they count as not remembered. Right: Reconstructions from low (top left) to high (bottom right) attention levels. For each image the 8×8 index matrix up to the position of the image in the array were attended to, and the rest were reconstructed with the semantic completion network. Odd columns were dropped. Thus, the third image in the first row was reconstructed from only six known index values in the top row of the index matrix. See Figure S1 for more examples.

system and stored in memory traces with varying levels of attention: 5% (low attention), 52% (medium attention), or 63% (high attention) of the index matrix is stored. The stored memory traces are then recalled by the network and semantically completed by the PixelCNN. A trained classifier for digits and another one for backgrounds are used to model the responses of the participants. If the digit classifier recognizes the digit from the recalled image correctly, this counts as if the participant remembers having seen the object. Only then is the background classifier applied to determine the type of background. The digit classifier network is a basic CNN with three convolutional layers that was trained on digits in both congruent and incongruent contexts so that it is not biased by the background. This network has an accuracy of 99% on test data. The context classifier is a simple pattern matching algorithm that assigns the pattern class based on mean squared error.

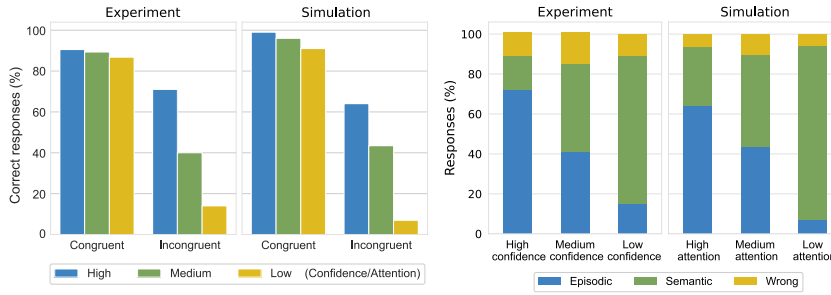

Figure 9: The effect of context and attention in experiment and simulation. Left: The fraction of correctly recalled contexts for congruent and incongruent cases depending on confidence/attention level. Right: A more detailed result for the incongruent cases.

Since the PixelCNN has been trained only on congruent examples, it usually fills in the semantically congruent background in the bottom half of the image if it has stored a particular digit at attention level 50% or less (here 5%), because it has no background information. If it fails to do so, it should fill in one of the other two backgrounds with equal probability. However, if the attention level is higher (here 52% or 63%), the PixelCNN should infer the (possibly incongruent) background from the bits that are preserved about it in the memory trace and complete it. The more information it has, the more reliably it recovers the correct background. Thus, the results should be trivial for congruent images because the congruent background is always recalled correctly, but for incongruent images, the outcome depends on attention level. That is, for low attention levels, the model plausibly constructs the congruent background (semantic recall), and for high attention levels, the model correctly recalls the incongruent background (episodic recall). It should usually not recall a background that is both incongruent and incorrect (wrong recall).

Experimental as well as simulation results are shown in Figure 9 in a direct comparison and can be summarized:

- Congruent contexts are recalled better than incongruent ones, even for high attention levels, as there is no conflict between episodic memory and semantic information.
- High confidence of having seen an object, modeled by high attention levels, increases memory accuracy in both congruent and incongruent cases, but much more so in the latter case, because in the former, performance is at a high level throughout.
- Contexts that are not remembered episodically correctly are more often remembered semantically congruently than completely wrong.

- Episodic and wrong recalls are equally likely in incongruent cases if the confidence/attention level is low, since there is (presumably) no information about the episodically correct room in the memory trace. This is expected for symmetry reasons if there is no particular prior toward one of the rooms.

To match the experimental results we tuned the three different attention levels, which are not well quantified in the experiment. However, the fact that there are wrong recalls of the context in incongruent cases and the good match of their proportion to the semantic recalls are emergent properties of the model.

4 Conclusion

With this work, we present a model of generative episodic memory at a rather abstract level with a network architecture combining known methods from machine learning: VQ-VAE and PixelCNN. It can process real images, includes the potential for spatial attentional selection (although still in a primitive form), can represent images that are quite different from those it was trained on, and models usage of semantic information for encoding, which relates to abstraction; for quantization, which relates to categorization; and for semantic completion to complement parts neglected by spatial attention. The term *semantic* might seem overly ambitious here, but we believe that the semantic information these generative models capture shares essential characteristics with what we would normally refer to as semantic, namely, general regularities of the world that hold beyond and are represented independent of particular episodes. If one would scale up the model in size and complexity, the semantic information would gradually be of a more high-level nature.

4.1 The Six Steps of Generative Episodic Memory. Our model shows how generative episodic memory can work in principle. It supports our conceptual framework for human episodic memory:

1. Sensory input (an image in our case) is processed by a multilayer perceptual-semantic network, for example, the visual system, to generate more abstract representation.
2. Some aspects of this representation, the episodic gist, are selected, presumably by attention depending on many factors.
3. Pointers to the selected perceptual-semantic elements in the hierarchical representation are then stored in the form of a memory trace in the hippocampus.
4. During recall, a memory trace is reactivated in the hippocampus.
5. The pointers in the memory trace in turn reactivate the perceptual-semantic elements.

6. Perceptual-semantic information is finally used to fill in missing parts in a dynamic process.

The last step makes episodic memory generative, and we call this process scenario construction.

4.2 Memory Efficiency. Three factors in the model contribute potentially to its memory efficiency: encoding and decoding; quantization, in both the VQ-VAE; and semantic completion in the PixelCNN. Interestingly, we find that the model works well in a regime where the encoding actually expands the representation by about a factor of four and only the quantization performs compression, so that combined, we have a compression by a factor of about 30 in the VQ-VAE. From Figure 6 one can infer that the semantic completion contributes a factor of only up to 2 to the overall compression of input images into the memory traces. We expect that for richer data sets and when the temporal dimension is taken into account, the contribution of semantic completion to the memory efficiency becomes much larger. However, we hypothesize that semantic completion, also serves other purposes.

4.3 Semantic Completion by the PixelCNN. The model shows good semantic completion capabilities. It is remarkable how well the PixelCNN completes the index matrix to generate plausible complete images even from small fragments, where faithful reconstruction is not possible (see Figure 6). Semantic completion can help to generalize better. It has been hypothesized that the main purpose of episodic memory is not to remember the past but to help us make decisions for the future (La Corte & Piolino, 2016; Schacter & Addis, 2007a, 2007b). Thus, if our knowledge about the world changes, maybe our memories of the past should also change to be maximally useful to deal with the future. Semantic completion can do exactly that.

4.4 Advantages of Using a VQ-VAE. We are not the first to model the generative nature of episodic memory. Two recent studies have used variational autoencoders (VAE) to reconstruct images from memory traces (Bates & Jacobs, 2020; Nagy et al., 2020). In contrast to a VAE, the latent representation in the VQ-VAE that we use here maintains some spatial resolution. This has two advantages. First, we can model not only compression but also spatial selection by attention. We do that by discarding some fraction of the feature vectors in the array and keeping the rest. Second, the model can also store and recall input patterns that are quite different from those seen during training, because the known feature vectors can be combined in many different new spatial constellations. For instance, a VQ-VAE trained on digits 0 to 4 can equally well represent digits 5 to 9 (see Figure 3), something a VAE cannot do. This generalization capability extends to the PixelCNN

in a remarkable way. Figure 8 shows that the PixelCNN is able to disentangle digit and background in the training data and put them together in an unseen way. We see another advantage of the VQ-VAE for our purposes in that the feature vectors are quantized, which is in analogy to semantic categorization in the brain (Persaud, Hemmer, Kidd, & Piantadosi, 2017).

4.5 Is Machine Learning the Right Level of Abstraction? It could be questioned whether machine learning methods are a good basis for modeling the brain like we do here. There is always a trade-off between the scale of the model and the biological details that it can account for. It is currently impossible to keep all the biological details when modeling an extensive system as we do here. Therefore, we believe that the artificial neural networks we use here offer a good level of abstraction with biological relevance. The convolutional neural networks, as well as recurrent neural networks on which our model is based, were inspired by the brain and are remarkably successful also in computational neuroscience (Kuzovkin et al., 2018; Lindsay, 2021; Papadimitriou, Vempala, Mitropolsky, Collins, & Maass, 2020; Savage, 2019; Yamins et al., 2014). Furthermore, they offer the advantage of efficiency, so that real world images can be processed. This is an important factor for two reasons. First, the frequently used artificial random stimuli in earlier memory modeling studies lack the statistical structure and regularities that are essential in studying the interplay between episodic memory and semantic information. Without statistical regularities that can be exploited, episodic memory cannot be generative by design. Second, being able to process images that are closer to real world images is an important step toward closing the gap between model simulations and experimental studies with human participants.

4.6 Modeling Episodic Semantic Conflict Resolution in Humans. We have successfully modeled the episodic memory experiment by Zoellner et al. (2021). Both experiments and simulations show that congruent contexts are recalled better than incongruent ones (van Kesteren, Rignanese, Gianferrara, Krabbendam, & Meeter, 2019), that attention improves correct recall in both cases, and that incorrectly recalled contexts in incongruent cases are more often remembered semantically correct than completely wrong. Figure 9 shows that we have achieved good agreement with the experimental results.

4.7 Future Research. Overall we believe this model advances our understanding and sharpens our concepts of generative episodic memory. However, the model can and should be developed in future work:

- The current attentional selection is rather restrictive and needs to be more flexible, so that any location could be selected. One option

would be to replace the PixelCNN by a more flexible transformer network (Chen et al., 2020; Parmar et al., 2018; Sanh, Debut, Chaumond, & Wolf, 2019).
- Even though the encoder is hierarchical, consistent with our conceptual framework, the index matrix on which the selection and semantic completion are done is not. It is possible to employ a hierarchical version of the VQ-VAE (Razavi, van den Oord, & Vinyals, 2019) to allow for semantic completion in a truly hierarchical representation.
- The storage process in the hippocampus is currently not modeled. This could be addressed, for instance, by adding a model of one-shot storage of pattern sequences in hippocampal memory (Melchior, Bayati, Azizi, Cheng, & Wiskott, 2019). This would also allow for the investigation of sequential episodes, not just snapshots. Sequentiality has been claimed to be one essential characteristic of episodic memory (Cheng, 2013; Cheng & Werning, 2016).
- Besides developing the model further, it needs to be compared to more experiments on human episodic memory to further constrain the model and contribute to the design of new experiments. For example, our model predicts that if a person's semantic information changes considerably, that person will probably recall old memories biased toward the newly learned semantics or just less accurately; this might partially explain the phenomenon of infantile amnesia (Robinson-Riegler & Robinson-Riegler, 2012). One other possible prediction is that if we assume that stress during retrieval temporarily blocks the access to episodic traces, we would see more semantic construction when a person is under stress during retrieval compared to unstressed retrieval (Wolf, 2019).

Overall, we believe the model we present here makes a significant step toward understanding how generative episodic memory might work, and it opens numerous options for future research to investigate more aspects of scenario construction.

5 Methods

Vector-quantized variational autoencoders (VQ-VAE) are autoencoders with a discrete latent representation that process input in three steps. First, the encoder, a convolutional neural network (CNN), transforms the input $\mathbf{x}$ to generate the encoding $\mathbf{z}_e$, an array of $w \times h$ d-dimensional feature vectors. Second, these feature vectors are quantized with the help of k codebook vectors $\mathbf{e}_l \in \mathbb{R}^d$ according to the vector quantization (VQ) framework in equation 5.1, that is, each one is mapped to the closest codebook vector $\mathbf{e}_l$, and the index matrix $\mathbf{z}_x$ contains the corresponding indices:

$$\mathbf{z}_x^{(i,j)} = \text{argmin}_l \|\mathbf{z}_e^{(i,j)} - \mathbf{e}_l\|_2$$
$$\mathbf{z}_q^{(i,j)} = \mathbf{e}_{\mathbf{z}_x^{(i,j)}}$$
$$\text{with } i \in \{1, \ldots, w\},\ j \in \{1, \ldots, h\}, l \in \{1, \ldots, k\}. \tag{5.1}$$

The codebook vectors are initialized randomly, but they get optimized during training together with the encoder and decoder. The quantized version of $\mathbf{z}_e$ is denoted $\mathbf{z}_q$ and is a $w \times h$ array of codebook vectors. Third, the decoder, a deconvolutional neural network, then generates a reconstruction $\mathbf{y}$ of the input $\mathbf{x}$ based on the given $\mathbf{z}_q$.

A VQ-VAE can be trained end to end based on the loss function

$$L = \log p\left(\mathbf{y} = \mathbf{x} \mid \mathbf{z}_q(\mathbf{x})\right) + \|\text{sg}[\mathbf{z}_e(\mathbf{x})] - \mathbf{e}\|_2^2 + \beta\|\mathbf{z}_e(\mathbf{x}) - \text{sg}[\mathbf{e}]\|_2^2. \tag{5.2}$$

The first term is the reconstruction loss, which is the negative log likelihood of the decoder output $\mathbf{y}$ being equal to the input $\mathbf{x}$, given the quantized latent representation $\mathbf{z}_q$; this term optimizes the decoder and the encoder. The second term is the VQ objective, which optimizes the codebook vectors. This term uses the l_2 norm to push the codebook vectors toward $\mathbf{z}_e$ and minimizes the quantization error. Since the codebook vectors $\mathbf{e}_l$ may not train as fast as the encoder, it might happen that the encoder outputs grow arbitrarily large. To make sure that the encoder commits to the codebook vectors, the third term is added. Essentially it pushes the encoder outputs toward the codebook vectors. This third term is called the commitment loss and constrains the scale of the encoder output (van den Oord et al., 2017).

Since quantization is a discrete operation, it is not possible to calculate its gradient for backpropagation. Therefore, the stop gradient (sg) operator is introduced here. During the forward pass, it works like an identity operator. During the backward pass (backpropagation), the gradient $\nabla_z L$ is passed directly from $\mathbf{z}_q$ to $\mathbf{z}_e$. The second and the third terms have identical values; the second one updates the codebook $\mathbf{e}$ via quantization (i.e., due to nonzero $\nabla_z L$), and the third one affects only the encoder.

In our simulations, we trained the VQ-VAE with 20 codebook vectors of size 64. The weight for the commitment loss, β, was set to one. The batch size was 128, and the learning rate was 3×10^{-4}. The encoder consists of two convolutional layers with a kernel size of three and stride two with 16 and 32 filters, respectively, followed by another layer with stride one and 64 filters. The decoder has an architecture symmetrical to the encoder but with transposed convolutional layers. Figures S5 and S6 in the appendix provide a visualization of the network.

A VQ-VAE by itself is not a generative model. However, its quantized index matrix ($\mathbf{z}_x$) makes it possible to sample new data with the help of a PixelCNN (van den Oord et al., 2016). A PixelCNN is a well-known autoregressive model, mainly used to generate new images given a training data

distribution. The basic principle of this model is that each pixel x_i in an image $\mathbf{x}$ has a probability distribution that depends on all the pixels that came before:

$$p(\mathbf{x}) = \prod_{i=1}^{D} p(x_i \mid \mathbf{x}_{<i}) \text{ where } \mathbf{x}_{<i} = (x_1, \ldots, x_{i-1}). \tag{5.3}$$

PixelCNNs generate images pixel by pixel and in a sequence (e.g., from top left to bottom right) conditioned on all previously sampled pixels. This process is slow for large images; in our case, we use the PixelCNN only on the index matrix $\mathbf{z}_x$, which is much smaller. After training, the model can complete a partial index matrix and even generate a new one from scratch based on the semantic information learned from the training data. This is then converted to an array $\mathbf{z}_q$ of codebook vectors and passed to the decoder to generate the output.

In this work, we use a gated PixelCNN (van den Oord et al., 2016) with the activation function

$$\mathbf{y} = \tanh(W_{k,f} * \mathbf{z}_x) \odot \sigma(W_{k,g} * \mathbf{z}_x) \tag{5.4}$$

instead of the more common rectified linear activation function. σ is the sigmoid function, k is the number of the layer, $\odot$ is the element-wise product, and $*$ represents the convolution operator. These multiplicative units help the network to model more complex functions (van den Oord et al., 2016).

The PixelCNN has 12 gated blocks and was also trained with the Adam optimizer with a batch size of 128 and a learning rate of 3×10^{-4}. The convolutional layers in the PixelCNN have 32 feature maps. Both the VQ-VAE and PixelCNN implementation were adopted from Royer (2019). A visualization of the PixelCNN network is included in Figure S7 in the appendix.

Alternatively one could have used other more advanced autoregressive models like image GPT (Chen et al., 2020) or transformer-based generative models (Parmar et al., 2018) and expect similar results. Since the latent representation was quite small in our case, the PixelCNN was sufficient.

Code

The codes related to this paper can be found at the following link: https://github.com/ZahraFayyaz/Generative-episodic-memory.

Acknowledgments

This work was supported by a grant from the German Research Foundation (DFG), "Constructing scenarios of the past: A new framework in episodic memory," FOR 2812, project number 419039588, P5 (L.W.) and 419039274,

P4 (O.T.W.). We thank Dr. Anand Subramoney for his helpful comments on the manuscript.

References

Addis, D. R. (2020). Mental time travel? A neurocognitive model of event simulation. *Review of Philosophy and Psychology*, *11*(2), 233–259. 10.1007/s13164-020-00470-0

Al-Tahan, H., & Mohsenzadeh, Y. (2021). Reconstructing feedback representations in the ventral visual pathway with a generative adversarial autoencoder. *PLOS Computational Biology*, *17*(3), e1008775. 10.1371/journal.pcbi.1008775, PubMed: 33760819

Axmacher, N., Do Lam, A. T., Kessler, H., & Fell, J. (2010). Natural memory beyond the storage model: Repression, trauma, and the construction of a personal past. *Frontiers in Human Neuroscience*, *4*, 211. 10.3389/fnhum.2010.00211, PubMed: 21151366

Bartlett, F. C. (1995). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press. 10.1017/CBO9780511759185

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review*, *127*(5), 891. 10.1037/rev0000197, PubMed: 32324016

Becker, S. (2005). A computational principle for hippocampal learning and neurogenesis. *Hippocampus*, *15*(6), 722–738. 10.1002/hipo.20095, PubMed: 15986407

Bhowick, D., Gupta, D. K., Maiti, S., & Shankar, U. (2019). *Stacked autoencoders based machine learning for noise reduction and signal reconstruction in geophysical data.* arXiv:1907.03278.

Carrillo-Reid, L., & Yuste, R. (2020). Playing the piano with the cortex: Role of neuronal ensembles and pattern completion in perception and behavior. *Current Opinion in Neurobiology*, *64*, 89–95. 10.1016/j.conb.2020.03.014, PubMed: 32320944

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. In *Proceedings of the International Conference on Machine Learning* (pp. 1691–1703).

Cheng, S. (2013). The CRISP theory of hippocampal function in episodic memory. *Frontiers in Neural Circuits*, *7*, 88. 10.3389/fncir.2013.00088

Cheng, S., & Frank, L. M. (2011). The structure of networks that produce the transformation from grid cells to place cells. *Neuroscience*, *197*, 293–306. 10.1016/j.neuroscience.2011.09.002, PubMed: 21963867

Cheng, S., & Werning, M. (2016). What is episodic memory if it is a natural kind? *Synthese*, *193*(5), 1345–1385. 10.1007/s11229-014-0628-6

Cheng, S., Werning, M., & Suddendorf, T. (2016). Dissociating memory traces and scenario construction in mental time travel. *Neuroscience and Biobehavioral Reviews*, *60*, 82–89. 10.1016/j.neubiorev.2015.11.011, PubMed: 26627866

Clayton, N. S., Salwiczek, L. H., & Dickinson, A. (2007). Episodic memory. *Current Biology*, *17*(6), 189–191. 10.1016/j.cub.2007.01.011

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247. 10.1016/S0022-5371(69)80069-1

Davis, S. W., Geib, B. R., Wing, E. A., Wang, W.-C., Hovhannisyan, M., Monge, Z. A., & Cabeza, R. (2021). Visual and semantic representations predict subsequent memory in perceptual and conceptual memory tests. *Cerebral Cortex*, *31*(2), 974–992. 10.1093/cercor/bhaa269, PubMed: 32935833

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17. 10.1037/h0046671, PubMed: 13664879

Deuker, L., Müller, A. R., Montag, C., Markett, S., Reuter, M., Fell, J., . . . Axmacher, N. (2013). Playing nice: A multi-methodological study on the effects of social conformity on memory. *Frontiers in Human Neuroscience*, *7*, 79. 10.3389/fnhum.2013.00079, PubMed: 23515273

Devitt, A. L., Addis, D. R., & Schacter, D. L. (2017). Episodic and semantic content of memory and imagination: A multilevel analysis. *Memory and Cognition*, *45*(7), 1078–1094. 10.3758/s13421-017-0716-1, PubMed: 28547677

Fang, J., Rüther, N., Bellebaum, C., Wiskott, L., & Cheng, S. (2018). The interaction between semantic representation and episodic memory. *Neural Computation*, *30*(2), 293–332. 10.1162/neco_a_01044, PubMed: 29220304

Fayyaz, Z., Altamimi, A., Cheng, S., & Wiskott, L. (2021). *A model of semantic completion in generative episodic memory.* arXiv:2111.13537.

Felleman, D. J., & VanEssen, D. C. (1991, January). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47. 10.1093/cercor/1.1.1, PubMed: 1822724

Graham, K. S., Simons, J. S., Pratt, K. H., Patterson, K., & Hodges, J. R. (2000). Insights from semantic dementia on the relationship between episodic and semantic memory. *Neuropsychologia*, *38*(3), 313–324. 10.1016/S0028-3932(99)00073-1

Greenberg, D. L., & Verfaellie, M. (2010). Interdependence of episodic and semantic memory: Evidence from neuropsychology. *Journal of the International Neuropsychological Society*, *16*(5), 748–753. 10.1017/S1355617710000676, PubMed: 20561378

Hemmer, P., & Steyvers, M. (2009a). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189–202. 10.1111/j.1756-8765.2008.01010.x

Hemmer, P., & Steyvers, M. (2009b). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin and Review*, *16*(1), 80–87. 10.3758/PBR.16.1.80

Herten, N., Otto, T., & Wolf, O. T. (2017). The role of eye fixation in memory enhancement under stress–an eye tracking study. *Neurobiology of Learning and Memory*, *140*, 134–144. 10.1016/j.nlm.2017.02.016, PubMed: 28267591

Hirst, W., & Echterhoff, G. (2012). Remembering in conversations: The social sharing and reshaping of memories. *Annual Review of Psychology*, *63*, 55–79. 10.1146/annurev-psych-120710-100340, PubMed: 21961946

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8), 2554–2558. 10.1073/pnas.79.8.2554

Hu, R., & Jacobs, R. A. (2021). Semantic influence on visual working memory of object identity and location. *Cognition*, *217*. 10.1016/j.cognition.2021.104891, PubMed: 34481197

Irish, M., & Piguet, O. (2013). The pivotal role of semantic memory in remembering the past and imagining the future. *Frontiers in Behavioral Neuroscience*, *7*, 27. 10.3389/fnbeh.2013.00027, PubMed: 23565081

Jensen, O., & Lisman, J. E. (1996). Novel lists of 7 ± 2 known items can be reliably stored in an oscillatory short-term memory network: Interaction with long-term memory. *Learning and Memory*, *3*(2–3), 257–263. 10.1101/lm.3.2-3.257, PubMed: 10456095

Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational Bayes.* arXiv:1312.6114.

Koutstaal, W., & Schacter, D. L. (1997). Gist-based false recognition of pictures in older and younger adults. *Journal of Memory and Language*, *37*(4), 555–583. 10.1006/jmla.1997.2529

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J.-P., Baciu, M., Kahane, P., . . . Aru, J. (2018). Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications biology*, *1*(1), 1–12. 10.1038/s42003-018-0110-y, PubMed: 29809203

La Corte, V., & Piolino, P. (2016). On the role of personal semantic memory and temporal distance in episodic future thinking: The TEDIFT model. *Frontiers in Human Neuroscience*, *10*, 385. 10.3389/fnhum.2016.00385, PubMed: 27524964

LeCun, Y. (1998). *The MNIST database of handwritten digits.* http://yann.lecun.com/exdb/mnist/

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031. 10.1162/jocn_a_01544, PubMed: 32027584

Liuzzi, A. G., Aglinskas, A., & Fairhall, S. L. (2020). General and feature-based semantic representations in the semantic network. *Scientific Reports*, *10*(1), 1–12. 10.1038/s41598-020-65906-0, PubMed: 31913322

Mathieu, M., Couprie, C., & LeCun, Y. (2015). *Deep multi-scale video prediction beyond mean square error*. arXiv:1511.05440.

Melchior, J., Bayati, M., Azizi, A., Cheng, S., & Wiskott, L. (2019). *A hippocampus model for online one-shot storage of pattern sequences.* arXiv:1905.12937.

Michaelian, K. (2011). Generative memory. *Philosophical Psychology*, *24*(3), 323–342. 10.1080/09515089.2011.559623

Nagy, D. G., Török, B., & Orbán, G. (2020). Optimal forgetting: Semantic compression of episodic memories. *PLOS Computational Biology*, *16*(10), e1008367. 10.1371/journal.pcbi.1008367, PubMed: 33057380

Neher, T., Cheng, S., & Wiskott, L. (2015). Memory storage fidelity in the hippocampal circuit: The role of subregions and input statistics. *PLOS Computational Biology*, *11*(5), e1004250. 10.1371/journal.pcbi.1004250, PubMed: 25954996

Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 251–256). Amsterdam: Elsevier. 10.1016/B978-012375731-9/50045-8

O'Reilly, R., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. (2013). Recurrent processing during object recognition. *Frontiers in Psychology*, *4*. https://www.frontiersin.org/article/10.3389/fpsyg.2013.00124. 10.3389/fpsyg.2013.00124

Papadimitriou, C. H., Vempala, S. S., Mitropolsky, D., Collins, M., & Maass, W. (2020). Brain computation by assemblies of neurons. In *Proceedings of the National Academy of Sciences*, *117*(25), 14464–14472. 10.1073/pnas.2001893117

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (vol. 80, pp. 4055–4064). https://proceedings.mlr.press/v80/parmar18a.html

Persaud, K., Hemmer, P., Kidd, C., & Piantadosi, S. (2017). Seeing colors: Cultural and environmental influences on episodic memory. *i-Perception*, *8*(6), 2041669517750161.

Razavi, A., van den Oord, A., & Vinyals, O. (2019). *Generating diverse high-fidelity images with VQ-VAE-2.* arXiv:1906.00446.

Reber, P. J., Stark, C. E. L., & Squire, L. R. (1998). Cortical areas supporting category learning identified using functional MRI. In *Proceedings of the National Academy of Sciences*, *95*(2), 747–750. 10.1073/pnas.95.2.747

Reisberg, D. (2013). *The Oxford handbook of cognitive psychology*. Oxford: Oxford University Press. 10.1093/oxfordhb/9780195376746.001.0001

Robinson-Riegler, B., & Robinson-Riegler, G. (2012). *Cognitive psychology: Applying the science of the mind*. London: Pearson.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803. 10.1037/0278-7393.21.4.803

Rolls, E. T. (1995). A model of the operation of the hippocampus and entorhinal cortex in memory. *International Journal of Neural Systems*, *6*, 51–70.

Royer, A. (2019). *VQ-VAE for image generation.* https://ameroyer.github.io/portfolio/2019-08-15-VQVAE/

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics*, *2*(9), 437–442. 10.3758/BF03208784

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter.* arXiv:1910.01108.

Savage, N. (2019). How AI and neuroscience drive each other forwards. *Nature*, *571*(7766), S15–S15. 10.1038/d41586-019-02212-4, PubMed: 31341311

Schacter, D. L., & Addis, D. R. (2007a). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 773–786. 10.1098/rstb.2007.2087

Schacter, D. L., & Addis, D. R. (2007b). The ghosts of past and future. *Nature*, *445*(7123), 27–27. 10.1038/445027a

Schacter, D. L., & Addis, D. R. (2020). Memory and imagination: Perspectives on constructive episodic simulation. In A. Abraham (Ed.), *The Cambridge handbook of the imagination*. Cambridge: Cambridge University Press. 10.1017/9781108580298.008

Schacter, D. L., Guerin, S. A., & Jacques, P. L. S. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences*, *15*(10), 467–474. 10.1016/j.tics.2011.08.004, PubMed: 21908231

Takeda, M. (2019). Brain mechanisms of visual long-term memory retrieval in primates. *Neuroscience Research*, *142*, 7–15. 10.1016/j.neures.2018.06.005, PubMed: 29964078

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., . . . Kreiman, G. (2018). Recurrent computations for visual pattern completion. In *Proceedings of the National Academy of Sciences*, *115*(35), 8835–8840. 10.1073/pnas.1719397115

Teyler, T. J., & Discenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience*, *100*(2), 147–154. 10.1037/0735-7044.100.2.147, PubMed: 3008780

Teyler, T. J., & Rudy, J. W. (2007). The hippocampal indexing theory and episodic memory: Updating the index. *Hippocampus*, *17*(12), 1158–1169. 10.1002/hipo.20350, PubMed: 17696170

Thorpe, S., Fize, F., & Marlot, C. (1996, June). Speed of processing in the human visual system. *Nature*, *381*, 520–522. 10.1038/381520a0, PubMed: 8632824

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). Cambridge, MA: Academic Press. https://psycnet.apa.org/record/1973-08477-000

van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., & Kavukcuoglu, K. (2016). *Conditional image generation with PixelCNN decoders.* arXiv:1606.05328.

van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). *Neural discrete representation learning.* arXiv:1711.00937. https://arxiv.org/abs/1711.00937

van Kesteren, M. T., Rignanese, P., Gianferrara, P. G., Krabbendam, L., & Meeter, M. (2019). *Integrating memories: Congruency and reactivation aid memory integration through reinstatement of prior knowledge*. bioRxiv:716076.

Wolf, O. T. (2019). Memories of and influenced by the Trier social stress test. *Psychoneuroendocrinology*, *105*, 98–104. 10.1016/j.psyneuen.2018.10.031, PubMed: 30409385

Xia, R., Guan, S., & Sheinberg, D. L. (2015). A multilayered story of memory retrieval. *Neuron*, *86*(3), 610–612. 10.1016/j.neuron.2015.04.017, PubMed: 25950629

Xiao, H., Rasul, K., & Vollgraf, R. (2017). *Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms.* CoRR, abs/1708.07747.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. In *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. 10.1073/pnas.1403112111

Zhang, Y., Han, K., Worth, R., & Liu, Z. (2020). Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature Communications*, *11*(1), 1–13. 10.1038/s41467-019-13993-7, PubMed: 31911652

Zoellner, C., Klein, N., Cheng, S., Schubotz, R., Axmacher, N., & Wolf, O. T. (2021, 11). *Where was the toaster? Interplay of episodic memory traces and semantic knowledge during scenario construction.* PsyArXiv. https://psyarxiv.com/2kmwy

Received February 28, 2022; accepted May 3, 2022.