A computational model for generative episodic memory

Zahra Fayyaz, Sen Cheng, Laurenz Wiskott

Institute for Neural Computation, Ruhr University Bochum

Introduction

Many different studies have suggested that episodic memory is a generative process, but most computational models adopt a storage view. In this work, we propose a computational model for generative episodic memory. The central hypothesis of the model is that the hippocampus stores and retrieves selected aspects of an episode, which are necessarily incomplete. Then the neocortex fills in the missing information based on general semantic information.

The proposed model consists of two parts: the visual processing network and the semantic network. At first, the images are passed through an autoencoder (AE) structure. The encoder part models the processing of episodic experiences into higher-level semantic representations. These latent semantic representations, which have already abstracted away many details, can then be reconstructed through the decoder. This structure, representing the visual pathway in the neocortex, is modeled using the Vector Quantized Variational Autoencoder (VQ-VAE) [1]. Attention is modeled by selecting parts of the latent neural representation and storing them in a model of the hippocampus. We call this stored incomplete information the gist. However, to reconstruct from this gist, first, the missing details have to be filled in. This task is handled by the semantic network, which is trained on the semantic neural representations and learns their structure and statistics. It can then generate new valid neural representations or complete the gist according to the learned semantics. This semantic network is modeled using a Pixel Convolutional Neural Network (PixelCNN) [2]. Both the VQ-VAE and the PixelCNN are state of the art machine learning generative algorithms. This allows us to use more realistic sensory inputs in contrast to the majority of the hippocampal memory models that process rather abstract and simple patterns. In our study, we have used the MNIST data set of handwritten digits. This dataset is small enough to train our model quickly but still has enough structure and details that can be exploited by the network. The decoder then forms a cortical representation of the memory in its layers. The output of the decoder is assumed to be a readout of the cortical representation of the memory.

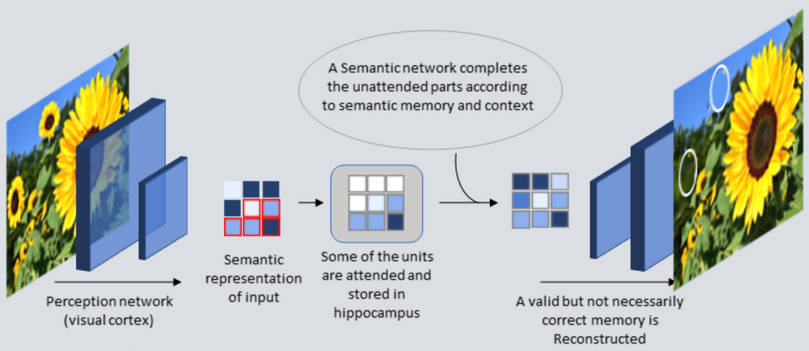

Figure 1. A general overview of the proposed model for generative episodic memory.

Methods

The visual processing network is modeled with a VQ-VAE. The VQ-VAE processes the input in three steps. The encoder, which consists of several convolutional layers, compresses the input x to generate the latent representation $z_e(x)$. This representation $z_e(x)$, is an array of $w * h$ depth feature vectors each of size d. The key innovation of the VQ-VAE is that instead of passing $z_e(x)$ directly to the decoder, it first quantizes it. A set of K shared embedding vectors $e_i \in R^d$ is introduced for that purpose. Using the vector quantization method, each of the depth feature vectors is then mapped to the closest embedding vector to form $z_q(x)$ which is the quantized version of $z_e(x)$ (see Eq. 1). The decoder, which is a deconvolutional neural network, then reconstructs the input based on the given $z_q(x)$.

$$z_q(x) = e_k, \quad \text{where} \quad k = \text{argmin}_j ||z_e(x) - e_j||_2 \quad (1)$$

During the training, VQ-VAE optimises the reconstruction loss and the quantization layer VQ loss together. Figure 2 shows a schematic of this network.

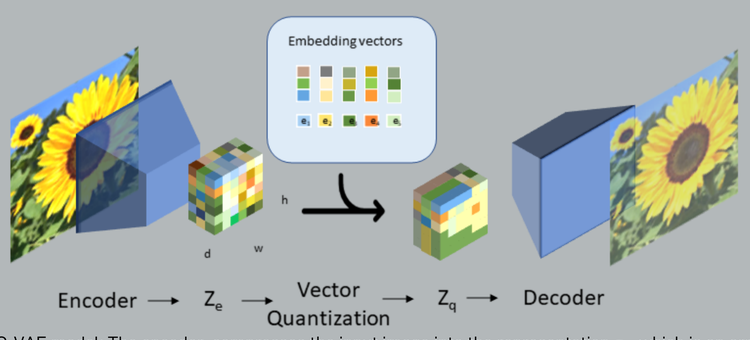

Figure 2. VQ-VAE model: The encoder, compresses the input image into the representation z_e, which is an array of $w * h$ depth feature vectors of size d. Each depth feature vector is then assigned to the closest embedding vector e_i to create the z_q vector. The decoder then reconstructs the original input based on the quantized z_q vector .

The aforementioned semantic network is modeled using a PixelCNN. PixelCNN is a famous autoregressive model that is mainly used to generate new images according to the training data distribution. After training the VQ-VAE, we can apply a PixelCNN to the $z_q(x)$ and generate new plausible latent representations. These latent representations are then passed to the decoder to generate new data . Because PixelCNN is also a good model for image completion, it can also complete the partly stored semantic representation which is then decoded.

Results

1. The interplay between semantic and episodic memory:
Episodic memory retrieval is not just a readout of a complete description of a past event, but it is a generative process, in which the episodic system provides the gist around which the semantic system reconstructs a plausible and likely scenario of the original episode. Here we have compared three different conditions. In our model, attention is modeled by storing all or some parts of the semantic representation in the hippocampus. If the episode is fully attended to, the reconstruction is faithful. However, if only some parts of the episode are attended, the reconstruction is not completely faithful but is plausible given the attended parts. Results of some fully, top 60%, and top 30% attended sample episodes and their reconstructions are displayed in Figure 3.a, 3.b, and 3.c, respectively.

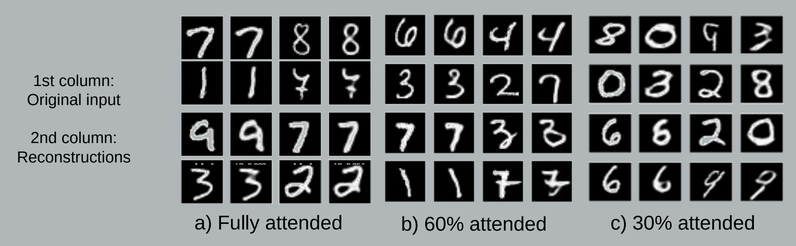

Figure 3. The left digit is the original input (i.e. experienced episode) and the right number is the reconstruction (i.e. readout of the memory recall). a) If the model pays attention to the full image the reconstruction is correct. b) If only the top 60% of the input's semantic representation is attended to the reconstruction is mostly true but the ones that are not true are still valid numbers. c) If 30% is attended there is a higher chance that the output is incorrect but it is usually still plausible (false memory).

2. The effect of context and congruency:
Experiments have shown that objects that are in a semantically congruent context are recalled better than incongruent objects, as there is no conflict between episodic and semantic memory. Also, interaction with objects (i.e. paying attention) increases memory accuracy. Moreover, it has been shown that objects that are not remembered episodically correct are more often remembered semantically than completely wrong. More information on the experimental data can be found in poster number 4: "Where is the toaster?". Here we have reproduced these experiments to validate our model. To model the effect of context, a padding has been added around the MNIST digits. The padding has horizontal bars for numbers below five and vertical bars for numbers above four to show two different contexts. The whole model is trained on this data set which we call the congruent data. Two cases are modeled: The model pays attention to the full semantic representation or only to half of it, which corresponds to the case of interaction and no interaction in the experimental data, respectively. The model is also tested with an incongruent dataset in which the context is the opposite of the congruent one (see Figure 4). As you can see in Figure 5 the trend in the results is matching the experimental data. As shown in Figure 6, when the episode is not remembered correctly it is more often dominated by the semantic memory rather than a random wrong recall.

a) Congruent data

b) Incongruent data

Figure 4. A sample from the dataset. The model is trained to semantically learn below five digits always have horizontal bars as their context and is then tested with the incongruent data which has opposite context.

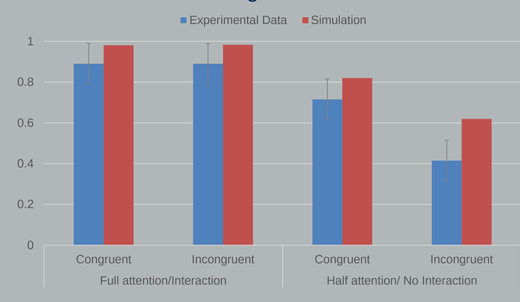

Figure 5. The simulation data matched the experimental findings that memory is better for congruent compared to incongruent objects and also that attention improves the memory accuracy.

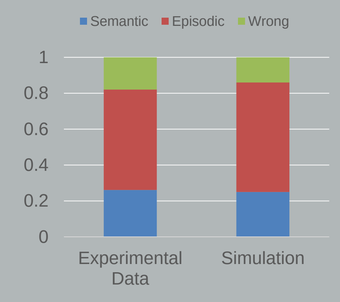

Figure 6.In the case of half attention, incongruent objects which are not remembered episodically correct are more often remembered semantically, rather than randomly wrong.

Conclusion

The results of our experiments support our hypothesis on generative episodic memory. The stored gist has far less information content than the input images; nonetheless, the input can be reconstructed from the gist with the help of a semantic network. This shows that the model is successful in capturing the complex statistics of the input images. When only parts of the latent neural representation are attended and stored, and then later recalled, the results are not necessarily faithful. Still, they are plausible and likely reconstructions of the original data.

The model also matches the experimental results about the effect of context on remembering. Memory is better for congruent compared to incongruent objects. The model simulations also confirm that objects that are not remembered episodically correctly are more often remembered semantically than completely wrong.

In conclusion, our model suggests how generative episodic memory could be implemented and provides the basis for further investigations and comparisons to neural processes.

Acknowledgements

The content of this work has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the DFG Research Group "Constructing Scenarios of the Past" (FOR 2812).
I would like to also thank Prof. Wiskott and Prof. Cheng for their supervision.

References

[1] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. "Neural discrete representation learning" (2017)
[2] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. "Conditional image generation with pixelcnn decoders." (2016)

FOR 2812 CONSTRUCTING SCENARIOS OF THE PAST

DFG Deutsche Forschungsgemeinschaft

RUHR UNIVERSITÄT BOCHUM RUB