Semantic Compression of Episodic Memories

David G. Nagy[1,2], Balázs Török [1,3], Gergő Orbán [1]
{nagy.g.david, torok.balazs, orban.gergo}@wigner.mta.hu
[1]Computational Systems Neuroscience Lab, MTA Wigner Research Centre for Physics, Budapest, Hungary
[2]Institute of Physics, Eötvös Loránd University, Budapest, Hungary
[3]Department of Cognitive Science, Budapest University of Technology and Economics, Budapest, Hungary

Abstract

Storing knowledge of an agent's environment in the form of a probabilistic generative model has been established as a crucial ingredient in a multitude of cognitive tasks. Perception has been formalised as probabilistic inference over the state of latent variables, whereas in decision making the model of the environment is used to predict likely consequences of actions. Such generative models have earlier been proposed to underlie semantic memory but it remained unclear if this model also underlies the efficient storage of experiences in episodic memory. We formalise the compression of episodes in the normative framework of information theory and argue that semantic memory provides the distortion function for compression of experiences. Recent advances and insights from machine learning allow us to approximate semantic compression in naturalistic domains and contrast the resulting deviations in compressed episodes with memory errors observed in the experimental literature on human memory. Semantic compression establishes a framework to provide a normative account for a spectrum of memory distortions in humans.

Keywords: semantic, episodic, memory errors, rate distortion

Introduction

Given the physical constraints on memory resources for the human brain, verbatim storage of all sensory experience is unfeasible. The normative framework for analysing this problem is provided by information theory, where efficient compression into memory traces hinges on the agent being able to prioritise information according to its relevance. In information theory these priorities are represented by the distortion function, which characterises the degree to which a particular form of distortion of the original experience is acceptable to the agent. While the distortion function is a critical component of efficient compression, the theory is agnostic about its specific form. Thus, efficient compression raises the question: What is the appropriate distortion function for human memory, that is, from a continuous stream of experience, how does the human brain determine what to remember and what to forget?

Storing knowledge of an agent's environment in the form of a probabilistic generative model has been established as a crucial ingredient in a multitude of cognitive tasks. Perception can be understood as probabilistic inference over the state of latent variables, where prior knowledge is integrated with noisy and ambiguous observations. In decision making, the model of the environment is used to predict likely consequences of actions, enabling the agent to find the action policy leading to maximal rewards. Learning is readily formalised as building a probabilistic model of the environment based on observations. Following previous research, we consider establishing a statistical model of the environment the domain of semantic memory and formalise it as a probabilistic generative latent variable model of the environment (Káli & Dayan, 2004; Hemmer & Steyvers, 2009; Nagy & Orban, 2016).

In addition to maintaining a probabilistic model in semantic memory, previous research has pointed out that retaining rich representations of specific experiences is also necessary (Nagy & Orban, 2016; Kumaran, Hassabis, & McClelland, 2016; Lengyel & Dayan, 2009). This form of memory, usually termed episodic memory, is an expensive representational format, which necessitates compression. We argue that in the case of memory, similar to perception, it is the inferences regarding the causes underlying sensory experience that is most relevant for the organism, which is precisely the information captured by the latent variables of semantic memory. Therefore, the information contained in the latent variables is what memory should prioritise when resources are constrained. In formal terms, we propose that semantic memory underlies the compression of episodes through providing the distortion function for episodic memory, a process we term semantic compression.

Empirically, the distortion function of an information compressing system becomes apparent in the pattern of memory errors that it produces. In the case of human memory, an extensive body of work has shown that it is indeed far from a carbon copy of sensory experience. Rather than being random noise however, these memory errors show robust and systematic biases. Such systematic biases are though to reflect rational adaptations to computational resource constraints (Schacter, Guerin, & St Jacques, 2011). Making this assumption explicit, we formalise semantic compression in the normative framework of lossy compression. This formalisation provides an opportunity for a unifying normative explanation of a wide variety of memory effects. Recent advances in machine learning yielded efficient tools to learn generative models of complex stimuli. In this study, we harness these advances to compare biases of humans in a recall task using naturalistic sketch images (Carmichael, Hogan, & Walter, 1932) with distortions introduced by semantic compression.

Rate distortion theory

The branch of information theory that deals with lossy compression is called rate distortion theory (RDT). A central insight of RDT is that while optimal compression is based on a knowledge of the statistics of the data, there is no single optimal encoding: a trade-off between the memory resources that are used for storing a given observation (rate) and the amount of distortion in the recalled memory exists. For any rate constraint, a minimal expected distortion can be established, defining the RD curve. The curve can be computed by minimising

$$L = \min\ D + \beta R.$$

Any compression method can be associated with a point on the RD plane, with optimal algorithms lying on the RD curve. Assuming the curve is strictly convex, every point on it can be identified with a single value of β, corresponding to a particular point on the rate-distortion trade-off continuum. The distortion term, the cost associated with each possible alteration of the memory trace, is defined as the expected value of the distortion, d, between the original, x, and the reconstructed observation, $\hat{x}$, so that $D = E_x[d(x,\hat{x})]$. An optimal lossy compression algorithm will selectively prioritise information such that alterations that are inconsequential according to this measure are discarded first. However, the distortion measure, $d(x,\hat{x})$, is left unspecified in RDT.

RDT was later extended to guide the choice of distortion function, in the information bottleneck (IB) method (Tishby, Fernando, & William, 1999). The IB method introduces the idea of relevant quantization: they argue that distortion should be defined so as to maximise predictive ability regarding the quantities that we are interested in. The relevant information is then defined by the mutual information between the encoding (Z) and the relevant quantities (Y) so that the loss to be minimized becomes $L_{IB} = -I(Z,Y) + \beta I(X,Z)$. The IB method is not feasible to apply to high dimensional naturalistic data, however a variational approximation to the objective called the deep variational information bottleneck (DVIB) was developed in Alemi, Fischer, Dillon, and Murphy (2016). Here we use an unsupervised version of this objective,

$$\mathcal{L}(\theta,\phi,x) = E_{z\sim q_\phi(z|x)}(\log p_\theta(x|z)) - \beta \cdot \mathrm{KL}(q_\phi(z|x)||p_\theta(z)),$$

where D corresponds to the first term (also called the reconstruction term) and R to the negative of the second term (also called the regularisation term) of the objective. Notably, Alemi et al. (2016) have shown that this also corresponds to the loss function of an approximate generative model called the β-VAE, thus providing a connection between generative models and rate distortion theory.

Semantic compression

Efficient compression is based on a knowledge of statistics of the environment. We argue that semantic memory, viewed as a probabilistic generative model, represents the best estimate the brain has of such environmental statistics. Furthermore, it provides latent variables that are shaped by stimulus statistics, rewards, tasks, and predictive success. These variables include lower level acoustic or visual features such as phonemes, or objects as well as abstract concepts such as what constitutes a good chess move or melody. In addition, they contain the information relevant for predicting future observations. As a consequence, the latent variables summarize the relevant part of sensory information for the brain and we propose that this is precisely the information that should be prioritised when memory resources are constrained.

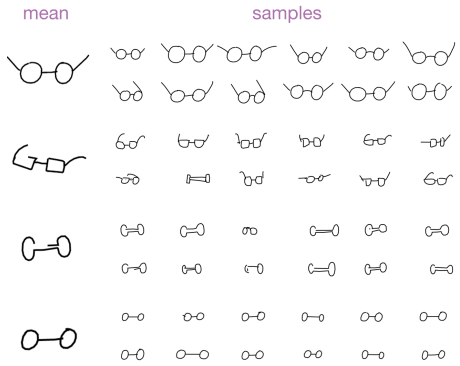

Figure 1: Learned representation. Some component means are presented along with samples. Intra-component differences are deemed smaller by the learned distortion function than inter-component differences, capturing human-like semantic distortions.

Representation of sensory experience in semantic compression occurs through inference of latent variables, z, which are then encoded as the memory trace:

$$\hat{z}(x_{obs}) = O_z[p(z \mid x = x_{obs})],$$

where O_x stands for a point estimate of the posterior distribution, such as the maximum a posteriori estimate. This formulation of the encoding process gives an opportunity to assess many of the distortions in memory associated with schematic biases and gist based errors. The encoding into a posterior over latent variables can be understood as compressing sensory experience into sufficient statistics for the latents, which is used to explain seemingly paradoxical results in the auditory cortex (McDermott, Schemitsch, & Simoncelli, 2013). Semantic compression also implies that the level of difficulty of inference affects the accuracy of the recalled memory trace. Classical memory experiments have shown that providing even a concise context which aids the interpretation of otherwise strongly ambiguous stimuli can greatly increase retention accuracy (Bower, Karlin, & Dueck, 1975; Bransford & Johnson, 1972). Finally, the statistical model of a particular stimulus set affects the efficiency with which the relevant statistics can be extracted from observations. Therefore, expertise in a cognitive domain results both in a better estimate of the observation statistics and in more efficient compression due to representations that are better suited to tasks in that domain. This explains varying recall performance for stimuli depending on how well a particular stimulus conforms the environmental statis-

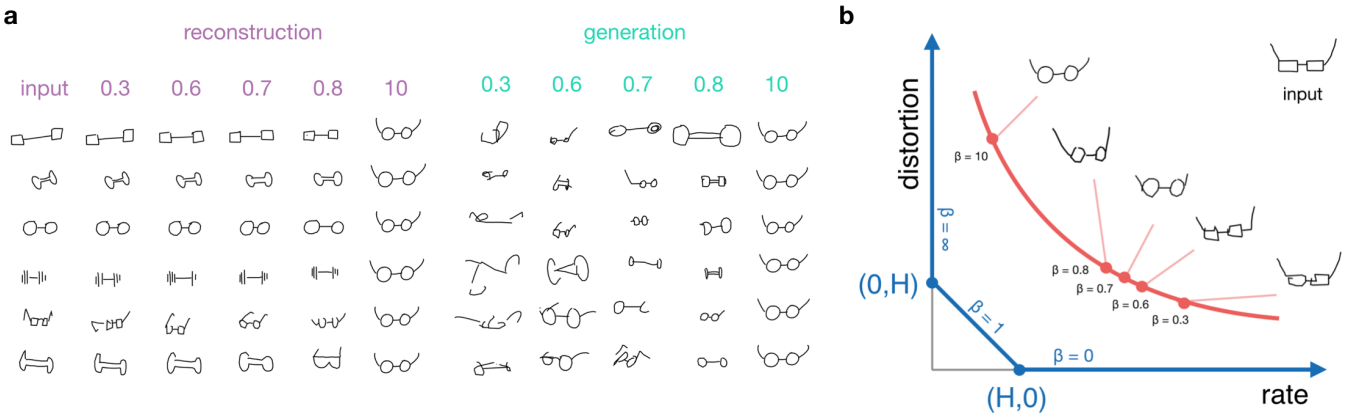

Figure 2: Reconstruction at different rates. a) Top row: value of β. Reconstruction: samples from the model with the given input in the left column. Generation: samples from the model without inputs. With increasing β, we lower the rate of compression: in reconstruction, idiosyncratic details of the input are lost. At the same time generation improves but also becomes less variable, at $\beta = 10$ producing one prototypical example. b) Blue curve: theoretical limit for best compressing models. Red curve: RD curve achievable by restricting posteriors to a parametric family such as in the sketch-rnn model. With increasing rate, compression is more faithful, while with decreasing rate, details are lost, rectangular shaped eyeglasses turn into more generic circular shaped ones.

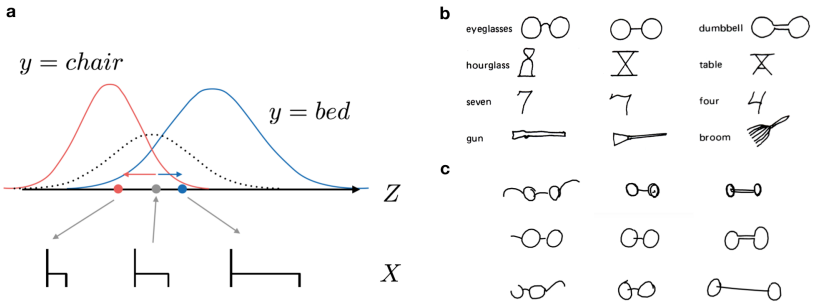

Figure 3: Memory distortions. a) Category information manifest in priors over the latent features Z (red and blue lines). Combination with the likelihood term of the observation (grey dotted line) induces biases: recall without category label (grey dot) becomes distorted (red and blue dots) b) Examples from Carmichael et al. (1932). Middle column: figures shown to subjects. Left and right columns: distorted reconstructions by participants who received the corresponding category label. c) Our reconstruction of the memory distortion.

tics (Baddeley, 1971), and performance differences between experts and non-experts (Gobet & Simon, 1996).

Recall of an observation in semantic compression is a reconstructive process. Since semantic memory is assumed to be a generative model over observed variables, it can be used to recreate an experience based on the memory trace by conditioning on the stored values for latent variables. This results in a predictive distribution over observable variables, a point estimate of which can be regarded as the representation point for the particular value of the latent:

$$\hat{x}(x_{obs}) = O_x[p(x \mid z = \hat{z}(x_{obs}))]$$

In case information about some features were lost during encoding, semantic memory can complement available information by relying on the prior distribution. This results in a gist-like reconstruction of the stimuli, where values of not retained features are substituted with what is likely to have been part of the observation. The DRM effect (Roediger & McDermott, 1995) and boundary extension (Intraub & Richardson, 1989) are good examples of such false memory effects.

Results

In order to investigate whether semantic distortion can enable efficient compression at multiple rates along the RD curve and demonstrate how it can explain systematic biases in human memory, we present a computational approximation of semantic compression in the domain of sketch-drawings under conditions where memory biases are known to emerge in human observers Carmichael et al. (1932). In this classical experiment, intentionally ambiguous hand drawn sketches of objects from common categories were presented to subjects who were asked to reproduce these images after a given amount of delay (Fig. 3b). Two separate groups of participants received different category names along with the drawings. Depending on the categorical cue, systematic biases were introduced in reproduced images.

As an approximation of the semantic model for sketch drawings we use the sketch-rnn architecture (Ha & Eck, 2017) and the β-VAE objective. We train this model on a dataset containing millions of sketch drawings of specific object categories that has recently become available in the Google QuickDraw

dataset (Ha & Eck, 2017). Since training of sketch-rnn is unsupervised, category labels can not be integrated during inference. We introduced these categories by fitting a mixture of Gaussians model on top of the latent representation. While the high dimensionality of the latent space precludes direct visualisation of the learned distortion, generating drawings from these components offers a glimpse into the kind of observations that are close in the semantic space (Fig. 1).

Fitting the model at different β values, corresponding to different trade-offs between rate and distortion, results in qualitatively different behaviours (Fig. 2). At high rates compression behaves similarly to a completely episodic system: the latents attempt to capture idiosyncratic details of the input, however there is very limited generalisation and the semantic model learned in this regime is not capable of producing realistic unconditional samples. At lower rates, recall becomes similar to a completely semantic system: it leans increasingly on reconstruction via the predictive semantic model rather than retaining details of the observation. Note, that in the extreme case of $\beta = 10$, latents become independent of the actual observation, generating a likely observation based on the marginal statistics of the data. This behaviour follows from the fact that maximum likelihood training and thus the ELBO objective does not give an explicit constraint on the latent representation, for further details see Alemi et al. (2017).

To contrast distortions introduced by semantic compression with reproduction biases revealed by the Carmichael experiment, we trained the model on sketches of specific object pairs and selected potentially ambiguous sketches. When performing inference, we incorporate the category label provided in the experiment by conditioning on the sketch being generated from the category. According to the principles of Bayesian inference the category prior introduces a bias in the encoding (Fig 3a), which will also be apparent in the generated drawing (Fig 3c).

Discussion

We gave a normative argument for compressing events in human memory using the latent variables of semantic memory formalised as a probabilistic generative model of the environment. We argued that in the framework of information theory this corresponds to using the conditional likelihood of the model as a distortion function. This correspondence enabled us to integrate recent results in machine learning with memory research to make predictions on complex, naturalistic data. Our formalisation can parsimoniously explain a variety of memory biases, and here we gave a detailed demonstration of a classic example in the domain of reproduction of sketch drawings.

Acknowledgments

The authors thank Ferenc Huszár for discussions. This work has been supported by the National Research, Development and Innovation Fund of Hungary (Grant No. K125343) and an MTA Lendület Fellowship.

References

Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2016). Deep variational information bottleneck. *arXiv preprint*, 1612.00410.

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2017). An Information-Theoretic Analysis of Deep Latent-Variable Models.

Baddeley, A. D. (1971). Language habits, acoustic confusability, and immediate memory for redundant letter sequences. *Psychonomic Science*, *22*(2), 120–121.

Bower, G. H., Karlin, M. B., & Dueck, A. (1975). Comprehension and memory for pictures. *Memory & Cognition*, *3*(2), 216–220.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *J Verbal Learning Verbal Behav*, *11*(6), 717–726.

Carmichael, L., Hogan, H. P., & Walter, A. A. (1932). An experimental study of the effect of language on the reproduction of visually perceived forms. *J Exp Psych*, *15*(1), 73–86.

Gobet, F., & Simon, H. A. (1996). Recall of rapidly presented random chess positions is a function of skill. *Psychon Bull Rev*, *3*, 159–163.

Ha, D., & Eck, D. (2017). A neural representation of sketch drawings. *arXiv preprint*, 1704.03477.

Hemmer, P., & Steyvers, M. (2009). Integrating episodic and semantic information in memory for natural scenes. *Proc 31st Ann Meeting Cogn Sci Soc*, 1557–1562.

Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *J Exp Psychol Learn Mem Cogn*, *15*, 179–87.

Káli, S., & Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nat Neurosci*, *7*(3), 286–294.

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary Learning Systems Theory Updated. *Trends in Cognitive Sciences*, *20*(7), 512–534. doi: 10.1016/j.tics.2016.05.004

Lengyel, M., & Dayan, P. (2009). Hippocampal contributions to control: The third way. , 1–8.

McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nature Neuroscience*, *16*(4), 493–498. doi: 10.1038/nn.3347

Nagy, D. G., & Orban, G. (2016). Episodic memory as a prerequisite for online updates of model structure. In *Proc 38th ann conf cog sci soc* (pp. 2699–2704).

Roediger, H. L., & McDermott, K. B. (1995). Creating False Memories: Remembering Words Not Presented in Lists. *J Exp Psychol Learn Mem Cog*, *21*, 803–814.

Schacter, D. L., Guerin, S. a., & St Jacques, P. L. (2011). Memory distortion: an adaptive perspective. *Trends Cogn Sci*, *15*, 467–74.

Tishby, N., Fernando, C. P., & William, B. (1999). The information bottleneck method. *arXiv:physics*, 0004057.