Rate distortion trade-off in human memory

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

From a continuous stream of experience, how does the human brain determine what parts to remember and what to forget? It is extensively documented that humans are prone to systematic biases in these decisions. Such systematic biases are often construed as byproducts of adaptive processes. We argue that the computational resource constraints on memory can be formalised in the normative framework of lossy compression and that optimal adaptation to the constraints can be achieved by exploiting a generative model of the environment for compression. Recent advances in machine learning yielded powerful tools to approximate such solutions. In this study, we harness these advances to show that generative compression can explain a wide variety of memory phenomena including the effects of domain expertise on recall, gist based distortions in recalling lists of semantically related words and the influence of contextual cues in memory for hand drawn sketches.

Keywords: semantic memory; episodic memory; memory errors; schema; rate distortion; compression

Introduction

It has long been known that human memory is far from a carbon copy of sensory experience. Rather than being random noise however, the distortions in recalled experience show robust and systematic biases. Since memory resources available for the brain are severely limited both by physical size and metabolic costs of maintaining stored information, the necessary result is loss of information. However, even forgetting can be accomplished in an optimal way: information theory provides a normative framework for analysing the problem of lossy compression, that is, making the best use of a limited amount of memory. The compression artefacts or distortions produced by widely applied lossy compression algorithms such as JPEG appear qualitatively different from those committed by human memory, which may raise questions regarding the applicability of this framework as an account of memory distortions in humans. We argue this is due mainly to two major difficulties in constructing practical compression algorithms. The first is that information theory is agnostic to how distortion is measured between original input and reconstruction. The second is the computational difficulty of constructing models of the input statistics to be compressed. In the case of the human brain, we have previously proposed that such a model is provided by semantic memory, formalised as a probabilistic generative model of the environment (Nagy, Török, & Orbán, 2018). Furthermore, we have argued that the distortion measure, that is the question of what is relevant in sensory experience, targets the same information that perception – understood as inference of latent variables – is aiming to extract from raw sensory experience. Therefore we proposed that distortion should also be defined through the latent variables of this model, terming our approach semantic compression.

Systematic biases in recalled memories are often thought to reflect rational adaptations, however, this is usually meant in the sense of being unfortunate but necessary byproducts of other adaptive processes. (Schacter, Guerin, & St Jacques, 2011; Newman & Lindsay, 2009). Here, we argue that a large variety of memory distortions can be seen as optimal adaptations to the statistics of the natural environment. Specifically, we argue that casting memory research in the normative framework of lossy compression can provide a unifying explanation for a large variety of experimental phenomena and elucidate issues in previous and future accounts of memory. Furthermore, we aim to demonstrate that various theoretical constructs of experiment driven accounts of memory errors such as gist and verbatim behaviour of memory traces arise as straightforward consequences of the normative goal of performing efficient lossy compression.

Methods

Rate distortion theory provides a normative theory for storing information under constraints on memory resources. An optimal encoding minimises the distortion while keeping utilised memory resources under some threshold:

$$D(R) = \inf_{Q}(D_Q),\ s.t.\ R_Q < R,$$

One way to achieve this, provided the $D(R)$ curve is strictly convex, is through optimising the lagrangian form of the rate distortion objective,

$$L = D + \beta R,$$

where β sets the trade-off between two terms, implicitly selecting the distortion and corresponding rate. Directly optimising this objective for naturalistic data is not feasible, however variational approximations can be made. Interestingly, a class of models called variational autoencoders (VAE) can be interpreted both in the framework of rate distortion theory, where they represent a variational approximation of the unsupervised information bottleneck problem (Alemi, Fischer, Dillon, & Murphy, 2016) but also as probabilistic generative models, where posterior inference is approximated via variational

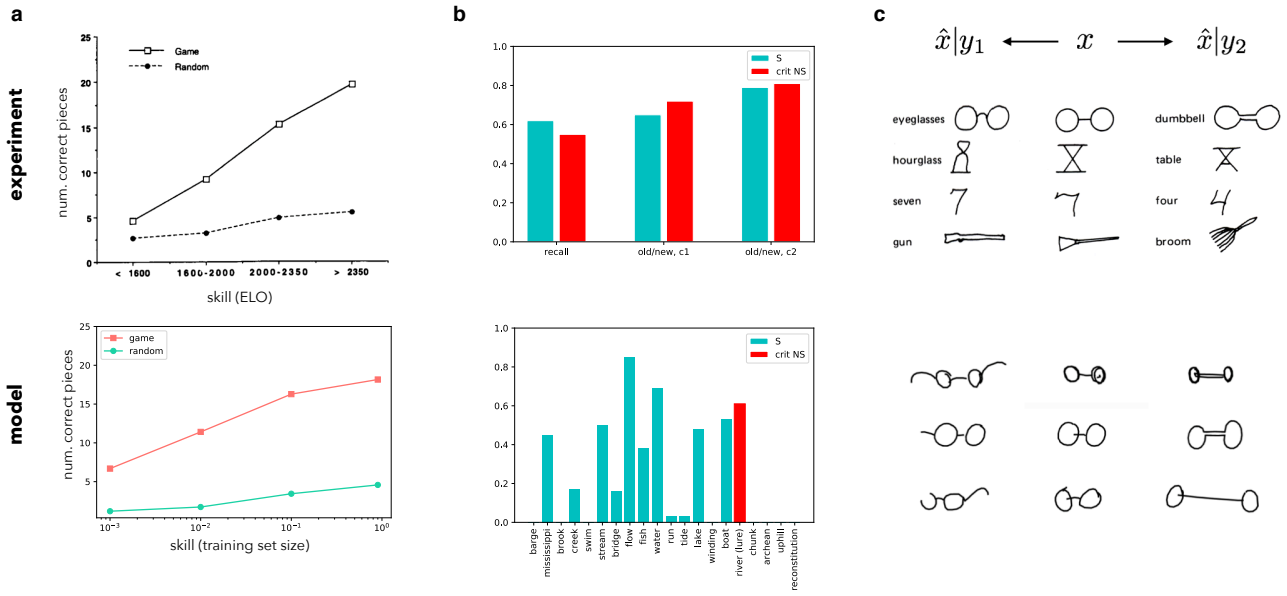

Figure 1: a) Reconstruction accuracy for game and random positions as a function of chess skill in the (Gobet & Simon, 1996) experiment (top) and in the model presented here (bottom). b) Memory for lists of related words. Recall and recognition probabilities in multiple conditions in the Roediger and McDermott (1995) experiment (top) for studied (green), critical non-studied (red) words. Frequency of recall from 100 samples for individual words in the text model (bottom). c) Memory for sketches. Label conditioned reconstructions of sketches from the original experiment of (Carmichael et al., 1932) (top) and the sketch-rnn model (bottom).

methods (Kingma & Welling, 2013). We build on this connection in order to be able to approximate semantic compression and contrast its consequences with memory experiments in naturalistic, high dimensional domains. We use the β-VAE objective (Higgins et al., 2017) to learn approximate generative models on datasets reflecting the statistics of natural domains, which at the same time can be seen as optimising a compression algorithm for these domains.

Models were trained on three domains that are prominent in the experimental literature on memory errors: chess board positions, word lists and hand-drawn sketch drawings. For chess positions, we have trained the model on games downloaded from the FICS games database [1] using a gaussian approximate posterior (or encoder) parameterised by two dense layers and a similarly parameterised generative model (or decoder). In order to create naturalistic training data in the domain of text, we have randomly subsampled the first 400 words of wikipedia articles. The model architecture was similar to the chess VAE, with the simplification in the noise model that words are generated independently, also known as the Neural Variational Document Model (NVDM) (Miao, Yu, & Blunsom, 2016) . The training data for the sketch domain consisted of pairs of object classes with 75k samples each from the Google QuickDraw dataset, and we have used the sketch-rnn VAE architecture (Ha & Eck, 2017). This model consists of a bidirectional RNN encoder and an RNN decoder which outputs the probabilities for the next pen position at each time step through a gaussian mixture model.

[1] https://www.ficsgames.org

Results

Domain expertise

Since compression hinges on accurate knowledge of environmental statistics that were learned from observations, experience in a cognitive domain increases recall accuracy for observations congruent with this statistics. This observation has widespread support in the memory literature. Chess is a particularly appealing example of this phenomenon since game statistics is widely documented and expertise can be assessed in a standardised way. Here, recall performance for chess boards was shown to strongly increase as a function of chess skill (as measured by ELO points of subjects) (Gobet & Simon, 1996). However, the difference between skilled an unskilled subjects becomes nearly nonexistent if the pieces on the board are shuffled randomly, making the board position unlikely or impossible to appear as an actual state of a chess game. In order to assess the effect of varying skills on recall performance, we varied the amount of games in the training set, ranging from 3 (unskilled) to 3000 (most skilled) games. Reconstructions of chess board positions by the model shows the characteristic increase in accuracy for 'game' boards (Fig. 1a). Data from human experiments indicate that even unskilled subjects can recall about 2.5 pieces on average, whereas our 'unskilled' model is not able to recall any. We conjecture that this is due to the fact that strategies

allowing for cross-domain transfer are available for humans (e.g. remember 'black rook at A2') whereas the entire set of observations for the model consisted only of chess boards positions. Similar results were obtained in the language domain, where recall performance for words was tested with varying levels of consistency with the statistics of English language corpus (Baddeley, 1971).

Gist-based distortions

In semantic compression, recall is interpreted as generating samples from the model conditioned on the memory trace. Lossy compression of experiences by a generative model implies that some features of the experience are retained, while others are discarded. Upon recall, discarded features are restored through samples from the model, thus favouring features that have a high probability under the memory trace-conditioned generative model. The reconstructed episode will retain features characteristic of high-level variables that can account for a large proportion of the variance in the data, the 'gist', however exact details of the episodes will be lost during lossy compression. We argue that a wide array of experimental data discussed under the rubric of gist-based distortion can be explained by generating features of an episode from a compressed memory trace.

In the case of memory performance for word lists with semantically related elements, assuming that memory constraints preclude exact recall, the recalled words will be chosen using the generative model. Since semantically related words are likely to occur together, the model will improve recall accuracy of such word lists relative to lists of randomly selected words, however the price is the intrusion of non-studied but semantically related words into the reconstructed episode (Fig. 1b). These intrusions are known as the DRM effect in the memory research literature (Roediger & McDermott, 1995) which has been found to be especially robust, even to explicit warnings. Our model trained on wikipedia articles also produces the intrusion of these 'critical' non-studied words in the recalled word lists with probability comparable to the studied words (Fig. 1b).

Since encoding of a memory trace in semantic compression is based on inference, factors that influence the inference of high level variables also affect the conditional distribution that the generative process samples from. Importantly, in the case of ambiguous stimuli, contextual information influences the inferred representation, and consequently distort lower level details in ways that better conform to the result of this inference. A paradigmatic example of contextual influence is a delayed recall experiment by where hand drawn ambiguous sketches were reproduced by participants, who – depending on the labels presented along with the drawings – produced qualitatively different reconstructions Carmichael et al. (1932). In the variational autoencoder trained on the QuickDraw dataset encoded sketches undergo similar distortions if the presentation of a label is modelled by the introduction of a conditional prior into the inference (for more details of the computational model see Nagy et al. (2018)).

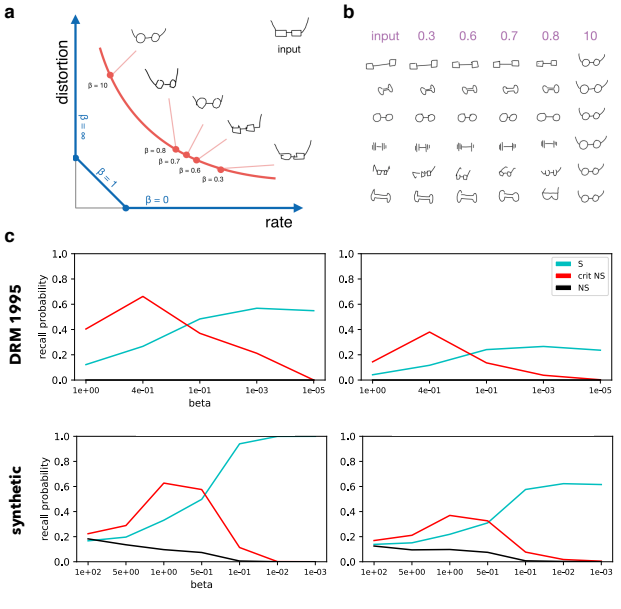

Figure 2: a) Illustration of rate distortion plane. Blue curve: theoretical limit for no restrictions on parametric model family (for details, see Alemi et al. (2017)). Red curve: RD curve achievable by restricting posteriors to a parametric family such as in the sketch-rnn model. With increasing rate, compression is more faithful, while with decreasing rate, details are lost, rectangular shaped eyeglasses turn into more generic circular shaped ones. Red R-D curve illustrative only, not actual values from model. b) Reconstructed samples from the model with the given input in the left column. c) Recall probabilities averaged for multiple word lists as a function of β (increasing amount of memory resources to the right). Top row shows results for model trained on wikipedia and word lists from article, bottom row shows results for same model trained on synthetic data and synthetic word lists.

Rate distortion trade-off

The utility of retaining information from a given episode is likely to vary with respect to a multitude of factors such as how surprising the episode is, its relevance for predicting the near future or its emotional valence. As a consequence, memory resources allocated to storing episodes are unlikely to be constant either at the time of encoding or as a function of time. If memory resources are to be distributed rationally, this memory decay should not result in random forgetting as information theory provides a principled way of discarding information so that memories degrade gracefully. Formally, optimal forgetting entails moving along the line of optimal encodings in the rate distortion plane in the direction of decreasing rate (Fig. 2a). At one extreme, where the rate distortion function intercepts the rate axis, resources are sufficient for lossless compression, meaning that verbatim recall is possible. At the other intercept, no information is retained relating to the individual episode and reconstruction is based purely on knowledge of

environmental statistics. Starting from the point corresponding to verbatim compression, the memory trace becomes increasingly gist-like, until a point where even a very high level gist of the episode is lost. In this way, the trade-off between rate and distortion results in the emergence of a continuum between gist and verbatim representations.

In order to illustrate the continuum of rate distortion tradeoffs, we have reconstructed observations using models with varying levels of compression, corresponding to different settings of the compression factor (Fig. 2a). In the word list domain, we show recall probabilities of words as in Fig. 1b, but averaging over multiple word lists, separately for studied, critical non-studied and non-studied words (Fig. 2b).

The NVDM text VAE architecture makes the simplifying assumption that words are generated independently, which means that the same word can be generated multiple times. We consider this assumption a purely computational simplification, the effects of which should be mitigated. For this reason, in addition to showing results with reconstruction through conditional sampling, we also show MAP reconstructions constrained to generate distinct words. In order to mitigate variability in the averages due to i) the low number of word lists and ii) mismatch between the statistics of the wikipedia training set and natural text, we have performed the same analysis on synthetic data where we could control these factors directly. We show results both for models trained on wikipedia with word lists from the original experiment by Roediger and McDermott (1995), and models trained on synthetic text generated from an LDA topic model with synthetic word lists generated via learned word embeddings.

In the standard DRM setting, recall probabilities of studied and critical non-studied words are similar (Fig. 1), suggesting that the amount of memory resources that humans use in the DRM experiment qualitatively corresponds to our model with $\beta = 0.1$. Increasing the rate from this level (decreasing β) results in increasing accuracy and thus the gradual disappearance of false memories. On the other hand, modeling memory decay in time by decreasing the rate results in a decrease in the recall of studied words, and an increase in false memories, similarly to what is found in a large number of experiments analysing the effect of recall delay on false memories (Toglia, Neuschatz, & Goodwin, 1999). At very high levels of compression (high β), recall for all word types is poor, as even the broad theme of the list might be forgotten.

Conclusions

In this paper we aimed to show that the principle of lossy compression can provide the basis for a unifying normative account of a wide variety of systematic memory errors. Recently developed computational approximations in machine learning enabled us to apply the framework to naturalistic, high dimensional data. We showed that the effect of domain expertise on recall accuracy in remembering chess board positions and gist based distortions in remembering semantically related word lists arise as straightforward consequences of optimising the rate distortion objective. Furthermore, we demonstrated the emergence of varying degrees of 'gistness' based purely on the statistics of observations, resulting from the rate distortion trade-off.

Acknowledgments

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.

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2017). Fixing a Broken ELBO.

Baddeley, A. D. (1971). Language Habits, Acoustic Confusability, and Immediate Memory for Redundant Letter Sequences. *Psychonomic Science*, *22*(2), 120–121. doi: 10.3758/BF03332525

Carmichael, L., Hogan, H. P., & Walter, A. A. (1932). An experimental study of the effect of language on the reproduction of visually perceived forms. *Journal of Experimental Psycology*, *15*(1), 73–86. doi: 10.1037/h0072671

Gobet, F., & Simon, H. A. (1996). Recall of rapidly presented random chess positions is a function of skill. *Psychonomic Bulletin and Review*, *3*(2), 159–163. doi: 10.3758/BF03212414

Ha, D., & Eck, D. (2017). A Neural Representation of Sketch Drawings. , 1–20.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Deepmind, G. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *Iclr*(July), 1–13.

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. (Ml), 1–14. doi: 10.1051/0004-6361/201527329

Miao, Y., Yu, L., & Blunsom, P. (2016). Neural Variational Inference for Text Processing. , *48*(Mcmc). doi: 10.1055/s-2008-1070477

Nagy, D. G., Török, B., & Orbán, G. (2018). Semantic Compression of Episodic Memories.

Newman, E. J., & Lindsay, D. S. (2009). False memories: What the hell are they for? *Applied Cognitive Psychology*, *23*(8), 1105–1121. doi: 10.1002/acp.1613

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), 803–814. doi: 10.1037/0278-7393.21.4.803

Schacter, D. L., Guerin, S. a., & St Jacques, P. L. (2011, 10). Memory distortion: an adaptive perspective. *Trends in cognitive sciences*, *15*(10), 467–74. doi: 10.1016/j.tics.2011.08.004

Toglia, M. P., Neuschatz, J. S., & Goodwin, K. A. (1999). Recall Accuracy and Illusory Memories: When More is Less. *Memory*, *7*(2), 233–256. doi: 10.1080/741944069