Chunking as policy compression in capacity-limited recurrent neural networks

Matthieu B. Le Cauchois
École polytechnique fédérale de Lausanne (EPFL)
Department of Cognitive Science
University of California, San Diego
matthieu.lecauchois@epfl.ch

Alexander Mathis
École polytechnique fédérale de Lausanne (EPFL)
alexander.mathis@epfl.ch

Jonathon R. Howlett*
VA San Diego Healthcare System
Department of Psychiatry
University of California, San Diego
jhowlett@ucsd.edu

Marcelo G. Mattar*
Department of Cognitive Science
University of California, San Diego
mmattar@ucsd.edu

Abstract

Any physical system operating with limited capacity must represent data efficiently. The brain, an example of a capacity-constrained system, must balance the goal of maximizing reward against the costs of representing complex behavioral responses to each situation. While previous work has characterized the informational complexity of neural representations in perceptual and memory systems, much less is known about the constraints in representations of behavioral policies. Here, we employ the normative framework of rate-distortion theory to examine the effect of policy compression in reinforcement learning. To induce a compressed policy representation, we introduced a structural bottleneck to a recurrent neural network trained to encode a policy. We hypothesized that tighter bottlenecks would give rise to chunking, whereby a behavioral policy is compressed by grouping separate actions into holistic sequences. To test this hypothesis, we trained recurrent agents to map each of 16 inputs to different action outputs. A subset of inputs appeared frequently in the same order (e.g. 0, 1, 2, 3, 4, 5, 6, 7). We found that our agents displayed various signatures of optimal compression through chunking. Activity at different stages of the networks revealed compressed and dynamic representations leveraging the temporal statistics of inputs. These findings were not observed in unconstrained networks, suggesting that information bottlenecks encouraged chunk learning. Interestingly, constrained agents recovered faster in domain adaptation tasks. In sum, our results show that networks with limited representational capacity learn compressed chunking policies tuned to the statistics of the environment. Our findings also invite an information-theoretic interpretation for the bottleneck architecture of the basal ganglia, a brain structure crucially involved in representing behavioral policies.

Keywords: Information Bottleneck, Rate Distortion Theory, Chunking, Reinforcement Learning

Acknowledgements

We would like to thank Andrew Zimnik and Fred Callaway for helpful feedback.

*Joint senior authors.

1 Introduction

If optimal decision making is hard, optimal decision making under resource limitations is even harder. Confronted with a formidable amount of situations and eligible responses, an agent must learn to efficiently store or compute appropriate behaviors. While the framework of Reinforcement Learning (RL) formalizes how an agent might learn the optimal behavior in response to each situation (i.e., the optimal policy), it rarely considers the agent's limitations in storing and/or computing such a policy. To address this limitation, the field of information theory, and more specifically the branch called Rate–Distortion Theory (RDT), provides the theoretical foundations for identifying policies that are both efficient and rewarding. RDT captures the notion that the representation of certain distinctions between states can be omitted in the policy if the associated cost of doing so is low.

Counter-intuitively, information bottlenecks can confer benefits to the agent by increasing generalization and robustness. These phenomena have been studied in the context of unsupervised learning [1] and more recently, RL [2, 3]. Given the presence of structural bottlenecks in the brain, in the basal ganglia for instance [4], it is essential to consider whether and how such bottlenecks affect a network's representational capacity, and –crucially– how such constraints ultimately affect an agent's behavior.

Here we investigate the impact of bottlenecks in deep RL through the lens of RDT, similarly to previous work with supervised learning for the visual system [5]. We hypothesized that, in the presence of such capacity constraints, the agent might learn to identify temporal patterns in the data to learn a compact yet rewarding policy, *chunking* separate actions into holistic sequences. To test this hypothesis, we trained several stochastic funneled recurrent neural networks on a sequential decision-making task designed to highlight chunking. By varying bottleneck dimension and noise, we first found that chunking scales with capacity constraints. Capacity allocation is tuned to the statistics of the environment, similarly to RDT. Furthermore, constrained agents show behavioral patterns of inflexibility reminiscent of habits. Second, by probing our models we found compressed representations that balanced state-dependency and recurrent retrieval. Finally, constrained agents responded faster to domain adaptation. Altogether our work calls for a novel look into the impact of information bottlenecks in decision-making. Further experiments should examine the interplay between chunking and the large compression factor in the basal ganglia, a brain structure involved in action selection [4].

2 Approach

Task To study how policy compression in capacity-limited agents leads to chunking, we trained a recurrent neural network on a sequential decision-making task [6]. Inputs sampled from 16 states were presented sequentially for eight time-steps. Each state was associated with a single rewarded action out of 16 possible actions. States were represented as one-hot vectors. Different modalities such as 100×100 pixel images resulted in conclusions equivalent to those shown here. Importantly, a specific set of eight states was frequently presented in a fixed order (Figure 1a). We call this set of states the *frequent sequence*, while *chunks* refers to actions executed holistically. Our intuition was that a capacity-limited agent will group these states into a chunk, reducing the amount of information required to store an effective policy.

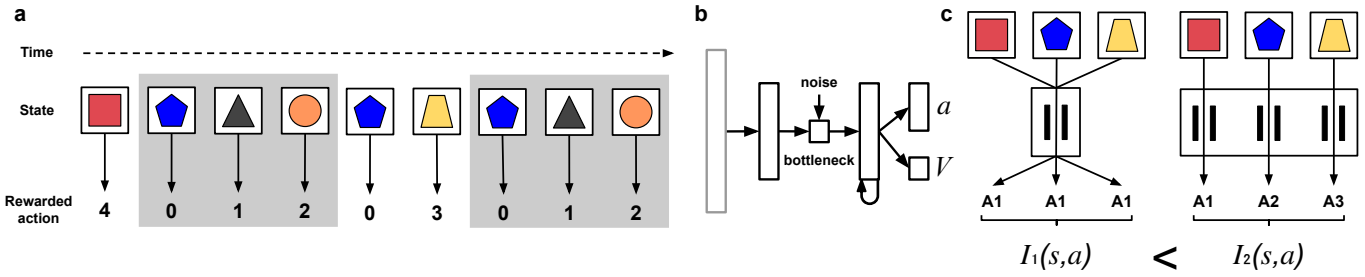

Figure 1: **(a)** Task structure. Grey rectangles highlight boundaries of the frequent sequence. **(b)** Architecture of our model. Number of units in the bottleneck layer and noise magnitude are varied across experiments. The grey module represents the input signal. **(c)** Impact of bottlenecks on information flow: mutual information between states and actions is reduced. This loss must be resolved using sequential information.

Modeling A Markov Decision Process (MDP) is a tuple $\mathcal{M} = \langle S, A, r, p, \gamma \rangle$ which consists of finite sets of states S and actions A, a reward function $r : S \times A \to \mathbb{R}$, transition dynamics $p(s'|s, a)$ and a discount factor γ. When r and p are unknown, a set of algorithms known as model-free RL can be used to learn the policy $\pi^*(a|s)$ that maximizes the expected discounted returns. In line with previous work [2, 6], we study capacity-limited RL with the normative framework of RDT. In RDT, the following variational problem must be solved:

$$\begin{aligned} \underset{\pi}{\text{argmax}} \quad & V^{\pi} \\ \text{subject to} \quad & I^{\pi}(s,a) = C \end{aligned}$$

where $V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_t \gamma^t r_t |\, s_0 = s\right]$ denotes the value function and $I(s,a) = H(s) - H(s|a)$ denotes the mutual information (H is the Shannon entropy). In short an agent must optimize for reward under a capacity constraint, measured by the mutual information between its actions and the states encountered. RDT provides an upper-bound on the reward achievable for a fixed capacity. We posit that the high compression factors found in the brain impose such a constraint. Instead of specifying the constraint in the optimization problem, we employ a neural network model with a stochastic bottleneck (Figure 1b) in which inputs are hypothesized to be mapped in a cluttered space such that outputs are harder to resolve, thereby limiting mutual information (Figure 1c). We model funneled cortical loops using an encoder and a Long Short-Term Memory layer (LSTM) linked by a bottleneck layer of adjustable dimension (Figure 1b). The input to the encoder can be of any modality; we experimented with a single feedforward encoding layer for one-hot inputs and a convolutional neural network for 100×100 pixel inputs, yielding equivalent results. From encoder output to LSTM output, the number of units across layers is 64-bottleneck-64. The model is trained using the proximal policy optimization algorithm [7] for an increasing number of bottleneck units (we only display 2, 4, 6 and 64 for clarity) and three different seeds.

Stochasticity Importantly, we add Gaussian noise to the activity of the bottleneck layer (Figure 1b), much like real communication channels in RDT and similarly to previous work on noise injection in RL [3]. Since our model has real-valued units at the bottleneck and our task consists of discrete inputs, noise is crucial to impose capacity constraints.

3 Experiments

Chunking as compression To examine the emergence of chunking, we looked at the network outputs (with fixed weights) and the average return obtained for different inputs (Figure 2a and 2b). We compared the average return obtained for tasks with (i) the frequent sequence in original order; (ii) the frequent sequence in reverse order; (iii) no frequent sequence (none). We found that capacity-limited networks obtained lower returns for the reversed sequence than for the original sequence. A similar pattern was observed in the absence of a frequent sequence. The same was not true for unconstrained networks. This is consistent with our hypothesis that chunking is a form of policy compression: chunks require less state-dependent behavior, and are retrieved holistically. The efficacy of chunking as a compression strategy is made apparent when looking at the poor performance of an equivalently-sized feedforward agent. The pressure to compress should increase with the difficulty of the task. Chunking is thus expected to be weaker for smaller frequent sequence lengths, when capacity can accommodate all states independently. To evaluate this hypothesis, we measured chunking by computing the difference between episode returns for the original and reversed sequence. This difference is expected to decrease with frequent sequence size. For fair comparison, we scaled it to the largest frequent sequence tested. A *set-size* effect is indeed observed for a bottleneck with four units (Figure 2c), and is modulated by the noise magnitude (high noise levels limit capacity and enforce chunking). Another insight into this compression process can be obtained by tuning the appearance frequencies or reward rates of the states outside the frequent sequence (Figure 2d). Without tuning, the success of a particular state is noisy and depends on seeds. If a state is more likely to appear or has a higher reward however, the constrained agent will prioritize storage of that state similarly to RDT.

Inflexibility Previous work [8] postulated that conceptualizing habits as chunks allows model-based RL to account for phenomena that model-free RL alone cannot, such as inflexibility in contingency change tasks. We hypothesized that a hybrid stance could be adopted for our model where compression-induced model-free chunking leads to contingency change inflexibility. To test this, we randomly switched the rewarded action of a state located at the middle of the frequent sequence (state 4) halfway through training and reported the average success rate for that state (Figure 2e). Performance for that state is near-perfect in the first part of the experiment for all bottleneck dimensions. Following the action switch, the unconstrained agent immediately recovers previous performance levels, whereas constrained agents never fully recover in the time imparted.

Open-loop control We then examined the agent's open-loop behavior. We hypothesized that, when presented with starting states of the frequent sequence and with randomly picked subsequent states, capacity-limited agents would retrieve the entire action sequence in a state-independent fashion. We indeed observed such behavior in constrained agents (Figure 2f), with actions corresponding to the frequent sequence chosen at a higher rate and in the right order even if presented with different states. The same was not observed for unconstrained agents (Figure 2g). Importantly, these *action slips* occur only for large-enough initiations.

Compressed representations To better understand the neural mechanisms underlying these behavioral evidences of chunking, we studied the learned representations. We applied Principal Component Analysis (PCA) to the LSTM hidden unit activity for two bottleneck dimensions (4 and 64) and for different levels of initiation of the frequent sequence. Plotting the mean trajectories onto the first two principal components revealed attractor dynamics for the constrained

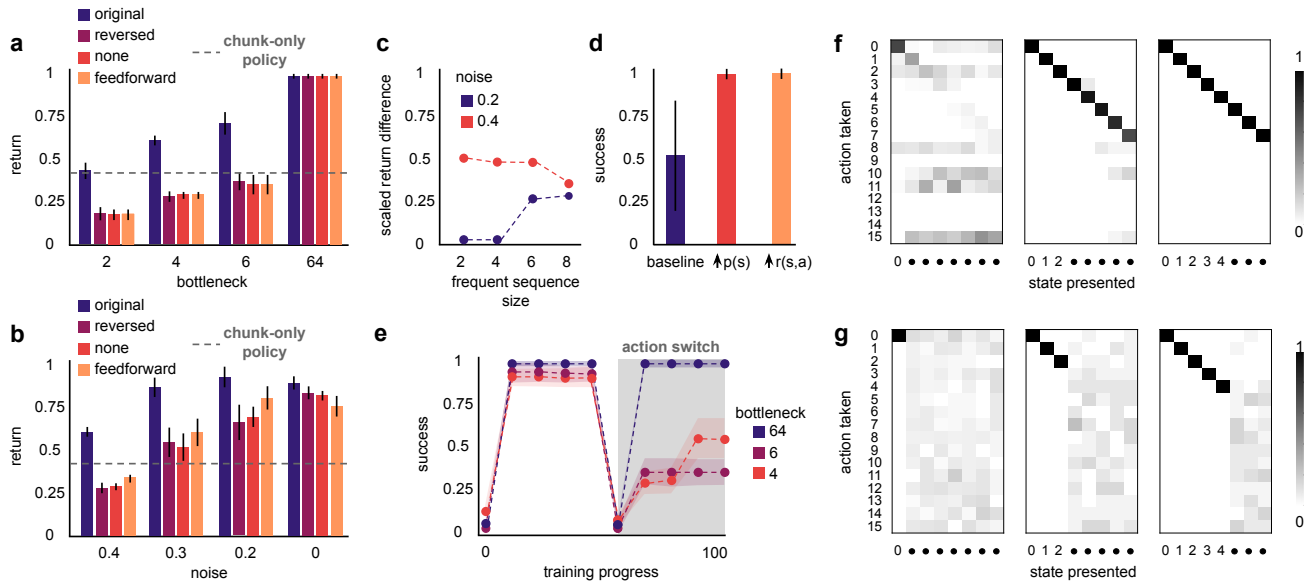

Figure 2: **(a)** Episode return for four bottleneck dimensions. First three hue colors represent modifications on the task's frequent sequence. Last hue is for a feedforward network (its performance is equivalent in each task modification). Dashed grey line is the return obtained by only executing the chunk. **(b)** Same as (a) for four noise standard deviations and a fixed bottleneck of 4 units. **(c)** Normalized chunking metric for four different frequent sequence sizes. Hue colors represent different noise levels. **(d)** Success rate for states outside the frequent sequence. Left bar is baseline, middle bar is for targeted increased state frequency and right bar is for targeted increased reward rate. **(e)** Success rate for state 4 throughout training for three different bottleneck dimensions. Halfway through training, the rewarded action for that state is switched (grey background). **(f)** Action distribution for three different initiations of the task's frequent sequence. The • represent any state. Model used has a bottleneck of 4 units. **(g)** Same as (f) for a bottleneck of 64 units.

agent (Figure 3a). Importantly, the attractor pull grows with the size of the frequent sequence initiation. The attractor dynamics are absent for the unconstrained agent (Figure 3b). Next, we investigated the decodability of the input state at different stages of the network and for different bottleneck dimensions using logistic regression (Figure 3c and 3d). The F1 score is maximum everywhere for the unconstrained agent. For the constrained agent, the F1 score in the bottleneck is low for most states. Initial states of the frequent sequence, however, present higher decodability reminiscent of the previously documented bracketing phenomenon in the striatum [9]. A couple of states outside the frequent sequence also show high decodability. These states have high variance: the capacity in the bottleneck is allocated to random states as reported in Figure 2c. In the recurrent layer of the constrained agent, states of the frequent sequence are highly separable: loss in the bottleneck is resolved using sequential information. Following previous work [10], we measured the information loss and analyzed the state-specificity of layers by computing an upper bound [11] to the mutual information between unit activities and states (Figure 3e). The mutual information drop between the encoder and the bottleneck is symptomatic of an information bottleneck at small bottleneck dimensions. The state-specificity is partially recovered in the recurrent layer, indicating that mutual information is an effective tool for studying chunking in neural circuits.

Adaptation Previous work on capacity constraints in RL [2, 3] reported enhanced generalization in constrained systems. We hypothesized that in addition to being an effective compression strategy, chunking is beneficial for generalization. To do so, we switched the input features from one-hot to randomly sampled patterns halfway through training and froze the recurrent layer (where chunks are implemented). We found that constrained agents recovered almost immediately (albeit not to previous performance) whereas unconstrained agents were much slower (Figure 3f).

4 Discussion

We found that training compresses policies in funneled recurrent neural networks by chunking actions according to the temporal statistics of the environment. We identified several key behavioral patterns resembling habits, and reverse-engineered the corresponding neural implementation by dissecting the learned representations. In future work, we will iterate on our model of chunking and evaluate neural and behavioral connections with the basal ganglia. Our preliminary results also extend on the idea that compression can benefit biological and artificial agents [1, 2, 3]. Given the close relationship between chunking and *options*, future work should study the interplay of bottlenecks and recurrence in hierarchical RL and options-learning.

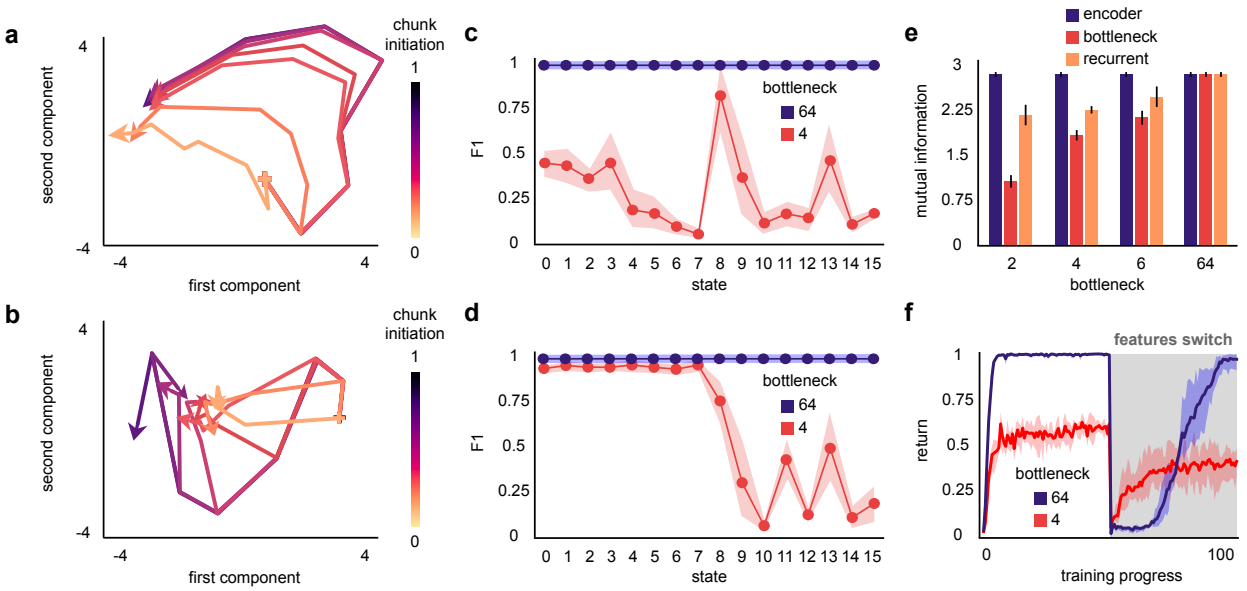

Figure 3: **(a)** PCA on the LSTM hidden unit activity: projection onto the top two principal components for different completion levels of the original frequent sequence. Model used has a bottleneck of 4 units. **(b)** Same as (a) for a bottleneck of 64 units. **(c)** F1 score for multi-class classification of the input state using logistic regression on the bottleneck unit activity for two bottleneck dimensions. **(d)** Same as (c) for the LSTM hidden unit activity. **(e)** Estimate of the mutual information between unit activities of each layer and input state for four bottleneck dimensions. **(f)** Episode return for two bottleneck dimensions. Halfway through training, the input features to the model are switched (grey background).

References

[1] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[2] R. Lerch and C. Sims, "Rate-distortion theory and computationally rational reinforcement learning," *Proceedings of Reinforcement Learning and Decision Making (RLDM) 2019*, pp. 7–10, 2019.

[3] M. Igl, K. Ciosek, Y. Li, S. Tschiatschek, C. Zhang, S. Devlin, and K. Hofmann, "Generalization in reinforcement learning with selective noise injection and information bottleneck," *CoRR*, vol. abs/1910.12911, 2019.

[4] I. Bar-Gad, G. Morris, and H. Bergman, "Information processing, dimensionality reduction and reinforcement learning in the basal ganglia," *Progress in Neurobiology*, vol. 71, no. 6, pp. 439–473, 2003.

[5] J. Lindsey, S. A. Ocko, S. Ganguli, and S. Deny, "A unified theory of early visual representations from retina to cortex through anatomically constrained deep cnns," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019.

[6] L. Lai and S. J. Gershman, "Policy compression: An information bottleneck in action selection," in *Psychology of Learning and Motivation - Advances in Research and Theory*, vol. 74, pp. 195–232, Academic Press Inc., jan 2021.

[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017.

[8] A. Dezfouli and B. W. Balleine, "Habits, action sequences and reinforcement learning," *European Journal of Neuroscience*, vol. 35, no. 7, pp. 1036–1051, 2012.

[9] X. Jin and R. M. Costa, "Start/stop signals emerge in nigrostriatal circuits during sequence learning," *Nature*, vol. 466, no. 7305, pp. 457–462, 2010.

[10] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017.

[11] A. Kolchinsky and B. D. Tracey, "Estimating mixture entropy with pairwise distances," *Entropy*, vol. 19, no. 7, p. 361, 2017.