One bottleneck is not enough

Mani Hamidi (mani.hamidi@uni-tuebingen.edu)
Human and Machine Cognition Lab
University of Tübingen
Tübingen, 72076 DE

Mihály Bányai (mihaly.banyai@tuebingen.mpg.de)
Department of Computational Neuroscience
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Charley M. Wu (charley.wu@uni-tuebingen.de)
Human and Machine Cognition Lab
University of Tübingen
Tübingen, 72076 DE

Abstract

Rate-distortion theory has been used to model the relationship between the capacity of a resource-limited agent and learning performance. However, most of these models define a single bottleneck on either the representational or policy complexity of the agent, but not both. Here we explicitly model representational capacity and policy capacity separately, and show that they make independent and non-interchangeable impacts on learning performance and efficiency of learned representations. This preliminary work has the potential to provide normative guidance about how to design more efficient RL agents, while also informing better descriptive models of human behavior by capturing different forms of cognitive constraints.

Keywords: rate-distortion theory, compression, contextual bandits, representation learning, explore-exploit

Introduction

RDT was first introduced to describe lossy compression in capacity-limited communication channels (Shannon, 1959). Compression is cast as a constrained optimization problem, where the capacity (or "rate") of the channel is measured by the mutual information between the input and output, and bounded by a maximum limit. The goal is then to minimize distortion in the signal, which is measured by a choice of loss function, such as a mean squared error.

More recently, RDT has been applied to reinforcement-learning (RL), where the agent is treated as a capacity-limited channel striving to maximize reward (Malloy & Sims, 2022; Gershman, 2020; Genewein, Leibfried, Grau-Moya, & Braun, 2015). Some have focused on the *representational capacity* of the agent, for encoding and remembering features in the input signal that are relevant for reward acquisition (Malloy & Sims, 2022; Bates, Lerch, Sims, & Jacobs, 2019). Others, have applied constraints to the *policy capacity*, constraining how much an agent's policy can diverge from a stimulus-independent prior (Gershman, 2020; Lai & Gershman, 2021). In both cases a single parameter (typically denoted with β) is used to capture the effects of capacity limitations.

A tale of two bottlenecks

Yet, representation and policy bottlenecks correspond to different types of constraints, with conceivably different impacts on behavior. For example, the resolution of one's representations impact how well stimuli are mapped onto rewards (representational capacity), while more fine-grained distinctions in the mapping of actions to value (policy capacity) impacts the explore-exploit trade-off. Inspired by past work suggesting the potential for non-trivial interaction between multiple bottlenecks (Genewein et al., 2015; Tishby, Pereira, & Bialek, 2000), we address this gap by exploring the impact of independently constrained representation and policy capacities on the performance of an RL agent.

In this work, we use a value-learning architecture based on the β-variational autoencoder (β-VAE; Alemi et al., 2018) to learn latent representations for each arm in a contextual bandit task. We systematically manipulate both the capacity to learn these latent representations, and the capacity to deploy actions w.r.t. these representations, in order to elucidate the interplay between constraints on both channels. We present preliminary results showing that the two capacities independently impact the performance of the agent and the efficiency of learned representations.

Methods

We use a contextual multi-armed bandit task, where each arm a_i corresponds to a different binary vector $\mathbf{x}_i$ of length N, with 2^N arms in total. By analogy, each action can be thought of choosing a fish lure (Fig. 1a) with N binary features (e.g., hook type, color, etc...), which have different contributions to the probability of catching a fish $R(\mathbf{x})$. The problem is to learn which arms (e.g., lures) maximize reward. We define the reward function $R(\mathbf{x})$ such that each of the N features have a different contribution to the desirability of the lure, $R(\mathbf{x}) = \mathbf{w} \cdot \tilde{\mathbf{x}}$, where feature weights $\mathbf{w}$ are sampled from an exponential distribution $w_j \sim \exp(\lambda)$ with $\lambda = 1$, and $\tilde{\mathbf{x}}$ providing a transformation of $\mathbf{x}$ to randomize whether the rewarding option corresponds to $x_i = 1$ or $x_i = 0$.

We define a RL agent using a neural network that receives

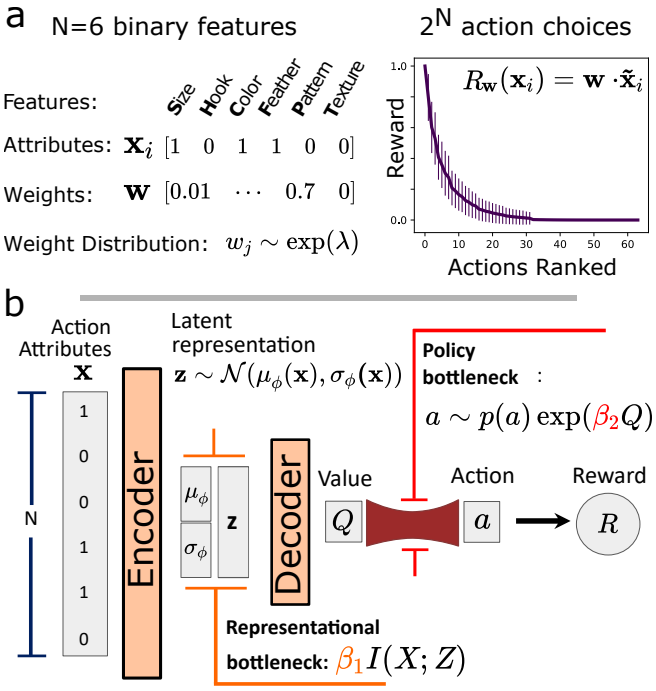

Figure 1: Task and model. **a**) Contextual bandit task with 2^N arms, corresponding to $N = 6$ binary features, each with a different weighted contribution to reward. **b**) Agent architecture with representational and policy bottlenecks.

the feature vector $\mathbf{x}_i$ as input for a given action a_i, and encodes it into a latent vector $\mathbf{z}$. The latent vector is then decoded into an estimated action-value Q_i. The objective function combines reward prediction error (mean squared error; MSE) between observed R and predicted rewards Q_i, with mutual information regularization, controlled by β_1 to capture the representation learning capacity of the agent:

$$\mathcal{L} = \sum_i^{2^N} (R_i - Q_i)^2 - \frac{1}{\beta_1} I(\mathbf{z};\mathbf{x}) \quad (1)$$

The latent vector $\mathbf{z}$ is sampled from a multivariate normal distribution, parametrized by the encoder: $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_\phi(\mathbf{x}), \sigma_\phi(\mathbf{x}))$. The representational bandwidth, or $I(\mathbf{z};\mathbf{x})$, is then calculated as the KL-divergence of this distribution from the prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ (Kingma & Welling, 2013)[1].

The policy module of the agent is simply a softmax distribution over Q-values:

$$\pi(a_i|\mathbf{x}) \propto p(a_i) \exp[\beta_2 Q(\mathbf{x}_i)] \quad (2)$$

where the prior policy $p(a)$ is assumed to be uniform. It follows that β_2 captures the policy capacity of the agent, who in order to exploit the value-relevant information about the feature attributes of a given arm, $Q(\mathbf{x})$, must reduce the exploration rate (increase β_2), in order to deviate from the uniform prior policy $p(a)$.

[1] This can be computed in closed form using: $I(\mathbf{z};\mathbf{x}) = \frac{1}{2} \sum_k^{|z|} (\mu_k^2 + \sigma_k^2 - 1 - \log(\sigma_k^2))$

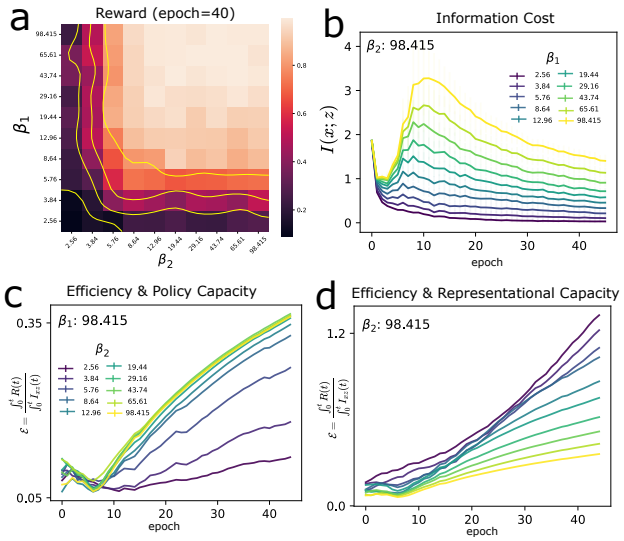

Figure 2: Results. Each epoch corresponds to 32 trials. **a**) Average reward (colors), with contour lines showing areas of equivalence. **b**) Information cost (in terms of mutual information) in the representation channel across different values of β_1. **c**) Cost-reward efficiency as a function of policy capacity. **d**) Efficiency as a function of representational capacity.

Results

We performed 10,000 simulations, using each combination of β_1 and β_2 values from a set of 10 evenly log-spaced values from 2.56 to 98.42. Each combination of parameters was simulated over 100 different randomly generated reward landscape (Fig. 1a), where we set $N = 6$, and the encoder and decoder consisted of a single layer of 10 and 20 neurons, respectively (Fig. 1b).

We first examined whether β_1 and β_2 are *interchangeable*. We test this by examining whether changes in performance as a result of modifying one capacity limit can be counteracted by appropriate modifications to the other. Figure 2a visualizes average performance across variations in β_1 and β_2, where the contour lines indicate regions of equivalence. These contours are largely orthogonal, suggesting that no increase in representational or policy capacity can compensate for limited capacity in the other channel. Note that improved performance with higher βs is expected, as they reflect increase in representational and policy capacities of the agent.

However, improved performance comes at a cost, and the dynamics of these costs over time are differently shaped by β_1 and β_2 . In Figure 2b, we first plot the direct relationship between β_1 on the representational costs of the agent $I(\mathbf{z};\mathbf{x})$ over time, using an illustrative case where policy capacity is large ($\beta_2 = 98.42$) and representational capacity serves as the main bottleneck for the agent. As expected, larger representational capacities allow for more complex representations to be learned (higher $I(\mathbf{z};\mathbf{x})$), highlighted by distinct maxima in early phases of learning. Notably, the actual capacity does

not always match maximal capacity limits, as representations become more compressed with further training.

We then analyze the dynamics of reward-cost efficiency using $\mathcal{E}(T) = \sum_{t=0}^{T} R(t) / \sum_{t=0}^{T} I_{xz}(t)$, and describe how this is differently impacted by the two bottlenecks. Figure 2c shows how when policy capacity is the limiting bottleneck on the agent (β_1 set to 98.42), higher policy capacity trivially yields more efficient agents regardless of the elapsed time. On the other hand, Figure 2c shows that when representational capacity accounts for the main bottleneck in the agent (β_2 set to 98.42), lower capacity corresponds to higher efficiency, but with some notable convergences during intermediate phases of learning.

In future work, we intend to model constraints on policy capacity within the same loss function as representational capacity (Eq. 1), allowing us to more symmetrically characterize how the two bottlenecks impact one another and to describe their interplay in meta-learning across multiple RL tasks.

Conclusion

We present preliminary work seeking to extend the RDT paradigm in RL by independently manipulating representational and policy capacities. Our early results suggest that performance and the efficiency of learned representations cannot be captured by a single bottleneck, with both independently influencing performance. This work has the potential to provide normative guidance about how to design more efficient RL agents, while also informing better descriptive models of human behavior by capturing different forms of cognitive constraints (Zenon, Solopchuk, & Pezzulo, 2019).

Acknowledgments

This work is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645 and the Marie Skłodowska-Curie Action RELEARN-DLV-897042.

References

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., & Murphy, K. (2018). Fixing a broken elbo. In *International conference on machine learning* (pp. 159–168).

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019, February). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *J. Vis.*, *19*(2), 11.

Genewein, T., Leibfried, F., Grau-Moya, J., & Braun, D. A. (2015). Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, *2*, 27.

Gershman, S. J. (2020, November). Origin of perseveration in the trade-off between reward and complexity. *Cognition*, *204*, 104394.

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

Lai, L., & Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. In *Psychology of learning and motivation* (Vol. 74, pp. 195–232). Elsevier.

Malloy, T., & Sims, C. R. (2022). A beta-variational auto-encoder model of human visual representation formation in utility-based learning. *Journal of Vision*, *22*(14), 3747–3747.

Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, *4*(142-163), 1.

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Zenon, A., Solopchuk, O., & Pezzulo, G. (2019). An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, *123*, 5–18.