Examining Bio-Inspired Approaches for Continual Reinforcement Learning

by

Olya Mastikhina

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Despite the brain's inherent ability to continually learn, biological insights are rarely applied to continual reinforcement learning (RL). This thesis addresses this gap by examining four under-investigated biologically-inspired modifications within the context of continual RL: energy minimization, wire length constraints, sparse distributed memory multilayer perceptrons, and fuzzy tiling activations. We show that some of these modifications help increase plasticity and generalization as well as slightly decrease catastrophic forgetting. We additionally provide an analysis of the learned representations.

Preface

Parts of this thesis have been accepted as a workshop paper by Mastikhina, Golnaz, and White at the 2024 Reinforcement Learning Conference: Finding the Frame workshop. Newer results from the thesis may be submitted to a conference.

To my lab and to my family. To anyone aware of being aware, and wondering why

Your imagination is far more wonderful than any computer could ever be.

– Mister Rogers

Acknowledgements

I am deeply grateful for my supervisor, Dr. Martha White, as well as to my RLAI lab mates, for their invaluable guidance and support throughout this journey. In addition, I am thankful to everyone at AMII, as well as for my family and friends.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In order to be able to adapt to real world settings, reinforcement learning (RL) agents need to be able to continually learn from evolving temporally-structured information. However, unlike biological neural networks, artificial neural networks are poor at learning from sequences of data and suffer from significant challenges with memory stability and plasticity (Kudithipudi *et al.* 2022; L. Wang *et al.* 2024). These issues impact all deep continual learning sub-fields, including continual RL. However, they are also prevalent in RL as a whole as RL agents learn from data that is gradually acquired through exploration, commonly leading to training on a continuously shifting distribution of data. Plasticity loss in particular is a key RL challenge (Lyle, Zheng, Khetarpal, *et al.* 2024; Nikishin *et al.* 2022).

In this thesis, we propose to draw more inspiration from the remarkable ability of biological brains to continuously learn. Biological neural networks have evolved around learning from sequential data, and they are able to maintain stable and reusable memories while still remaining plastic enough to keep incorporating new information (Parisi *et al.* 2019). By examining already existing mechanisms that enable continuous learning in biological systems, we may be able to better understand how to build artificial continuously learning systems.

We examine four modifications to the Soft-Actor Critic (SAC) (Haarnoja, Zhou, *et al.* 2019), only one of which has been applied to RL, that can be categorized into two overarching themes derived from brain functionality: energy

constraints and memory indexing. Energy constraints impose natural limitations to brain activity, while memory indexing refers to learning constrained addresses that access whole memories. These themes are interconnected, as energy constraints have been suggested to lead to the separation of groups of neurons representing individual concepts in the brain (J. C. R. Whittington, Dorrell, *et al.* 2023).

Within machine learning itself, energy constraints have previously been shown to lead to disentanglement, or to an increase in neurons coding for individual factors of variation in the environment (J. C. R. Whittington, Dorrell, *et al.* 2023), and to the organization of neurons into clusters that correspond to areas with different functions within brains(Margalit *et al.* 2023). Indexing mechanisms, mainly through sparse activations, have been shown to improve transfer learning (H. Wang *et al.* 2024) and decrease catastrophic forgetting (Bricken *et al.* 2023).

For energy constraints, we look at non-negativity with weight and activation minimization (J. C. R. Whittington, Dorrell, *et al.* 2023) and wiring-length constraints (Margalit *et al.* 2023). We refer to the former as `small-bio`, and the latter as `wire`. `Small-bio` penalizes big weights and activations, and wire encourages the similarity of representations between nearby neurons whilst discouraging it between further away neurons. For indexing, we examine fuzzy tiling activations (FTA) (Pan *et al.* 2021) and a neural network variant of sparse distributed memory (Bricken *et al.* 2023), also called the spare distributed memory multilayer perceptron (SDMLP). Both are sparse activation methods, but SDMLP introduces additional normalizations and an excitatory to inhibitory varaint of top-k activations. Please see Chapter 3 for more background and details about the modifications.

1.1 Contributions

This thesis provides an exploration of a few biologically-inspired or aligned modifications for continual RL. The key contributions are as follows:

1. We provide a brief overview primarily of existing modifications that have

roots in brain function, but are under-investigated in RL in particular, and why we think these modifications may be helpful for continual RL. We then implement the energy constraints and indexing modifications in a SAC agent.

2. Although we did not exhaustively optimize the implementations, we show no significant benefits from the modifications in reducing catastrophic forgetting by SAC on a subset of Mujoco environments from Metaworld (Yu *et al.* 2021).

3. We also show that a few of the modifications, most particularly one of the energy constraint modifications (which we refer to as `small-bio`), already show promising benefits for plasticity and generalization, which are key desiderata of continual learning. This is shown through increased performance by SAC on Mujoco environments (Tassa *et al.* 2018; Yu *et al.* 2021), with corroborating representation metrics results, including fewer non-updated units and a lower parameter norm (Lyle, Zheng, Khetarpal, *et al.* 2024).

1.2 Thesis Structure

This thesis consists of six chapters. Chapter 2 provides a background to RL and to the Soft Actor-Critic algorithm that is used in the experimental portion of this thesis. Chapter 3 provides a brief overview of biologically inspired or aligned principles and modifications that we believe may be helpful for continual RL, and explains the modifications that we try out experimentally. The 4th chapter provides more details into the implementations of the modifications, and of our general RL setup. The experiments section shows the results of the modifications on a few setups for deconstructed continual learning problems, and we provide analyses of the properties of the resulting representations. Chapter 6, further discusses the results, their limitations, and potential future work.

Chapter 2

Reinforcement Learning Background

This chapter provides an overview of reinforcement learning to help understand the Soft Actor-Critic (SAC) algorithm, which is the underlying reinforcement learning algorithm used in this thesis.

2.1 Markov Decision Processes

Reinforcement learning is goal-directed learning where an agent interacts with and learns from its environment (Sutton and Barto 2018). The Markov Decision Process (MDP) framework (Puterman 2014) formalizes these interactions by defining them in terms of states, actions, and rewards. For each interaction step, the agent is in state s out of a state space $\mathcal{S}$ and selects an action a from an action space $\mathcal{A}$ according to policy π, receives reward r, and transitions to s', another state from the state space $\mathcal{S}$. The agent's goal is to learn a policy π, which maps the agent's actions to states, that maximizes its expected sum of future rewards.

2.2 Approximate Q-Functions

A common approach to learning a policy includes assigning the expected future sum of rewards to state-action pairs, and iteratively improving the accuracy of these values in tandem with the policy. These values are incrementally updated through temporal-difference (TD) learning. At optimality, each state-

action pair would have a Q-value, $Q^*(s, a)$, which is the expected future sum of rewards starting from state s, taking action a, and acting optimally thereafter. If the environment is deterministic, the optimal policy π^* would have the agent taking the action with the highest Q-value at each state.

When an environment has continuous states and action spaces, or when there are too many possible states and actions in general, storing Q-values for individual state-action pairs in a table can becomes infeasible. In such an event, the agent is also likely to keep encountering state-action pair combinations that it has not seen before, so it must learn to approximate Q-values by extrapolating from similar previous observations. This can be done through parameterized function approximation, now commonly done with artificial neural networks as part of deep reinforcement learning, where the agent learns the parameters $\boldsymbol{w}$ in the parameterized function $Q(s, a; \boldsymbol{\theta})$ (Sutton and Barto 2018).

2.3 Deep Reinforcement Learning Considerations

When Q-functions are learned with neural networks, target networks get employed to decrease early overgeneralization and help stabilize training (Mnih *et al.* 2015). A target network is a lagging copy of a Q-function neural network that helps update estimates of current state-action values in the main Q-network using older estimates of subsequent state-action values in the agent's trajectory.

Additionally, to decrease correlations between states and thus increase stability in neural network training, as well as to improve sample efficiency, past interactions between the agent and the environment tend to be saved in a replay buffer and are revisited as learning progresses (Mnih *et al.* 2015).

2.4 Policy Gradient and Actor-Critic Methods

While action-value functions can assist with learning policies, agents can learn policies independently with function approximation. In what is referred to as policy gradient methods, agents learn parameters $\boldsymbol{\theta}$ for policy $\pi(a|s, \boldsymbol{\theta})$, which gives the probability of taking action a in state s that would maximize the future sum of rewards. For continuous control, when the environment has a continuous action spaces, the policy network outputs the mean and standard deviation for the probability densities of the action spaces.

The policy's own estimate of the future sum of rewards is simplified compared to that of value functions, as intermediate state values are not learned, and may suffer from large variances. Actor-critic algorithms address this by instead having the policy (actor) maximize the sum of future advantages, which is the policy's estimate of the sum of future rewards minus the expected sum of future rewards obtained by a value function (critic) (Konda and Tsitsiklis 2003).

2.5 Maximum Entropy Reinforcement Learning

Instead of overfitting to only one policy that seems optimal to the agent at the time but may in fact be sub-optimal, agents need to maintain some exploration of the environment. To address this, instead of maximizing just the expected sum of rewards, maximum entropy reinforcement learning maximizes the expected sum of rewards with an entropy bonus, which improves exploration by rewarding agents for acting more randomly (Haarnoja, Tang, *et al.* 2017). The maximized objective summed over all time steps t is then $\sum_t \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))]$, where inside the summation is the expected value of the state-action pair at time step t following the current policy π, $r(s_t, a_t)$ is the reward after taking action a_t in state s_t at time step t, the entropy term $\mathcal{H}(\pi(\cdot|s_t))$ is a measure of uncertainty over the policy's distri-

bution in state s_t at time step t, and α is the temperature parameter that modules the relative importance of the entropy term.

2.6 Soft Actor-Critic

The soft actor-critic (SAC) is a maximum entropy actor-critic algorithm. It consists of a replay buffer, one actor (policy) network, two critic (Q-value) networks, and two critic target networks. There are two critics in order to avoid overestimation of state-action values; the actor uses the smallest value estimate out of the two outputted by the critics. The "soft" refers to the maximum entropy ("soft") Q-functions that now include entropy terms in their estimates of state-action values, encouraging exploration with the actor. In the most recent version of SAC, to avoid brittleness due to potential poor parameter selection, the entropy term's temperature parameter adjusts during training (Haarnoja, Zhou, *et al.* 2019). The algorithm for SAC is described in Table 2.1.

Soft Actor-Critic Algorithm	
$\phi, \theta_1, \theta_2, \mathcal{D}$	Initialize actor network, critic networks, and replay buffer
$\bar{\theta_1} \leftarrow \theta_1, \bar{\theta_2} \leftarrow \theta_2$	Initialize critic target networks
For each iteration:	
For each environment step:	
$\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t \vert \mathbf{s}_t)$	Sample action from policy
$\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1} \vert \mathbf{s}_t, \mathbf{a}_t)$	Transition to next state
$\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$	Store interactions in replay buffer
For each gradient update:	
$\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$	Update critic weights
$\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$	Update actor weights
$\alpha \leftarrow \alpha - \lambda \hat{\nabla}_\alpha J(\alpha)$	Adjust temperature for entropy
$\bar{\theta_i} \leftarrow \tau\theta_i + (1-\tau)\bar{\theta_i}$ for $i \in \{1, 2\}$	Update critic target weights

Table 2.1: SAC Algorithm as per Haarnoja, Zhou, *et al.* 2019

Chapter 3

Biologically-Inspired Approaches to Continual Learning

Living beings are arguably the only natural learning systems currently known to us. Because of that, it might be unsurprising that deep learning has its roots in studies of biological minds. As their name suggests, artificial neural networks resulted from simplified models of impulse propagation between neurons and connectionist models of the brain (McCulloch and Pitts 1943; Rosenblatt 1958). Convolutional neural networks (CNNs) were originally inspired by the visual cortex (Fukushima 1980; Fukushima and Miyake 1982; LeCun *et al.* 2015), and TD learning, central to RL, started as an analog of classical conditioning in animals (Sutton and Barto 1981; Sutton and Barto 1987). Moreover, neural and behaviour correlates keep surfacing from classical and modern RL (Subramanian *et al.* 2022).

While we believe that we will likely will not need to fully replicate the brain in order to have optimally functioning artificial learning systems, it may be beneficial to continue taking inspiration from already existing continual learning systems. However, other than for perhaps experience replay, which has correspondences to the complementary learning systems theory (Flesch *et al.* 2023; Kudithipudi *et al.* 2022; Mnih *et al.* 2015), biologically-aligned approaches to tackling continual RL are heavily under-explored.

In this section, we cover a few biologically-aligned approaches to continual

learning that we believe would be beneficial to try in continual RL, focusing on modifications that have been tried elsewhere in machine learning and would require only modest changes. We contextualize them within the framework of how the brain stores information, group the main approaches into energy constraints and indexing, and include a few separate focuses that we believe could be helpful for continual RL. We discuss how these approaches may end up being beneficial for orthogonality and generalization, which tie-in to catastrophic forgetting and plasticity.

The described energy constraints and indexing modifications are later experimentally implemented in an RL agent in the thesis.

3.1 Energy constraints

The brain is energetically expensive; in humans, despite the brain comprising on average 2% of our weight, it is responsible for approximately 20% of our resting metabolic energy (Herculano-Houzel 2011; Kety 1957). Across species, brains have a fixed energy cost per neuron, with the total metabolic budget scaling based on the number of neurons (Herculano-Houzel 2011). Because this imposes an evolutionary constraint on the brain, the activity and structure of neural networks in the brain is thought to be heavily guided by energy optimization (Herculano-Houzel 2011; Oldham *et al.* 2022; Padamsey and Rochefort 2023; Takagi 2021).

While there is likely to be a trade-off between biological constraints and other factors in the evolution of the brain (Oldham *et al.* 2022), the brain seems to be a strong example of form guiding function. For example, simply the propagation of waves guided by the geometry of the brain can be used to predict brain activity with high accuracy (Pang *et al.* 2023).

Understanding how biological constraints shape biological neural networks may help us understand how to grow artificial cognition without hard-coding specifics. For instance, in addition to broader regions of functional organization, brains contain cells that code for individual factors of variation within a task space, such as object vector cells (Høydal *et al.* 2019), or border cells

(Solstad *et al.* 2008). Additionally, variables that define a task tend to each be coded by one neuron (Flesch *et al.* 2023). Energy constraints have been proposed to play a key role in how these representations are formed (Margalit *et al.* 2023; J. C. R. Whittington, Dorrell, *et al.* 2023).

The compositional representations described above are thought to be instrumental in deriving new knowledge (Kurth-Nelson *et al.* 2023), and in more effective continual learning of tasks (Mendez and Eaton 2021; Mendez and Eaton 2022; J. C. R. Whittington, McCaffary, *et al.* 2022), particularly related ones. In the brain, this is likely done through the formation of cognitive maps, which tie sensory representations with reusable abstract spatial representations (Kurth-Nelson *et al.* 2023; Mendez and Eaton 2022; J. C. R. Whittington, McCaffary, *et al.* 2022; J. C. Whittington *et al.* 2020).

While we think that cognitive mapping is a topic that could be particularly promising to explore in RL, in this thesis we focus on the direct effects of just using energy constraints to create compositional representations. At this stage, we simply refer to them here as disentangled representations, where more individual neurons code for individual factors of variation in the input data. This form of disentanglement may create generalizable representations that then create less interference between tasks, thus being beneficial for continual learning. Here, we focus on two different types of energy constraints on SAC: biological constraints of non-negativity and energy minimisation (J. C. R. Whittington, Dorrell, *et al.* 2023), which we refer to as "small-bio", and wiring length constraints (Margalit *et al.* 2023).

3.1.1 "Small-bio"

Biological constraints of non-negativity and energy minimization in conjunction, which we refer to as `small-bio` for short, have been shown to promote disentanglement in neural networks (J. C. R. Whittington, Dorrell, *et al.* 2023). Under optimal non-negative energy minimization, particularly for linear data, neurons end up coding for at most one factor of variation in the environment, instead of potentially multiple ones (J. C. R. Whittington, Dorrell, *et al.* 2023). This energy minimization approach involves using positive activations (for ex-

ample, through the use of a ReLU activation function (Nair and Hinton 2010)), and adding additional losses for the l2 norm of the activations as well as the l2 norm of the weights:

$$L_{\text{small-bio}} = L_{default} + \beta L_{activations} + \beta L_{weights} \tag{3.1}$$

Here, $L_{default}$ is the default training loss, β modulates the strength of `small-bio`, $L_{activations}$ is the l2 norm of the activations, and $L_{weights}$ is the l2 norm of the weights.

3.1.2 Wiring Length Constraints

Long-range neuronal projections are energetically expensive to form, so wiring length may be one of the principles guiding the functional arrangement of neurons (Margalit *et al.* 2023; Oldham *et al.* 2022). Wiring length constraints, implemented within a topographic deep artificial neural network (TDANN), have been shown to recreate representations created by the brain by introducing a loss to encourage nearby neurons to have similar representations (Margalit *et al.* 2023). Wiring length constraints within artifical neural networks have not originally been shown to have a clear performance benefit, but to increase interpretability due to physical clustering; we nonetheless see improvement with `wire` in certain cases, even when applied to linear layers, as shown in Chapter 5. Within a TDANN, the following spatial loss, which we refer to as L_{wire}, is applied to convolutional layers:

$$L_{wire} = L_{default} + \alpha(1 - Corr(\vec{r}, \vec{D})) \tag{3.2}$$

Here, α is the strength of the `wire` regularization, $Corr$ is Pearson's correlation, $\vec{r}$ is a vector of pairwise similarity distances between activations, and $\vec{D}$ is a vector of the inverse distances for each pair. This loss encourages neurons that are closer to each other within a layer to have more similar activations, which leads to shorter wiring lengths between layers (Margalit *et al.* 2023).

3.2 Indexing

Although it is not an official term, we use indexing to refer to theories and empirical findings behind the brain developing neurons serve as indices into a broader concept or memory (McClelland *et al.* 1995; O'Reilly *et al.* 2014; Teyler and DiScenna 1986; Teyler and Rudy 2007). When those broad concept or memories themselves are encoded by only a few neurons, we characterize that under the category of energy constraints in the previous section.

Indexing is prevalent as a general principle in the brain, but it is particularly associated with the hippocampus and seems to greatly factor into how the brain stores knowledge - a phenomenon that is particularly relevant to continual learning.

In the hippocampal indexing theory, and the complementary learning systems (CLS) theory, the hippocampus formation learns sparse, non-overlapping representations that then index the overlapping and distributed representations in the neocortex (McClelland *et al.* 1995; O'Reilly *et al.* 2014). This leads to a high orthogonality where, for example, only a small number of the same neurons are active between multiple tasks (Flesch *et al.* 2023). As catastrophic interference during sequential task learning occurs when information is indiscriminately distributed throughout a neural network, as is the default, strategies that increase orthogonality in the representations can decrease forgetting (Flesch *et al.* 2023; Lewandowsky and Li 1995).

Below, the Sparse Distributed Memory Perceptron (SDMLP) and Fuzzy Tiling Activations (FTA) are sparse activation approaches that we see as consistent with indexing. Incidentally, both modifications additionally have ties to circuits in the cerebellum (Albus 1971; Sutton 1995; Xie *et al.* 2023).

3.2.1 FTA

FTA is an activation function that induces sparsity in neural networks by binning inputs into a larger sparse vector with a fuzzy indicator function (Pan *et al.* 2021). The fuzziness is used to avoid zero derivatives. FTA has previously been designed for RL, but effects on plasticity and catastrophic forgetting have

not yet been looked at. However, FTA has previously been shown to increase transfer learning in Deep Q-Networks (Mnih *et al.* 2015; H. Wang *et al.* 2024).

3.2.2 Sparse Distributed Memory Multilayer Perceptron

Closely related to Hopfield networks, Sparse Distributed Memory (SDM) is a mathematical associative memory model of how concepts, or patterns, are stored and retrieved in the brain (Kanerva 1988; Kanerva 1992).

SDMLP is a one hidden layer neural network implementation of SDM that treats input weights into the hidden layer as addresses, and output weights as patterns (Bricken *et al.* 2023). In supervised learning, SDMLP has been implemented with two sparse activation function variants applied to a wide hidden layer: a top-k activation function, where k gradually decreases, and a GABA switch variant of the top-k activation function (Bricken *et al.* 2023).

Within mature brains, γ-amino-butyric acid (GABA) is the main inhibitory neurotransmitter. However, early during development, it is primarily excitatory. The switch in responses from excitatory to inhibitory in neurons is referred to as the GABA switch (Ganguly *et al.* 2001). Accordingly, in the GABA switch variant of SDMLP, the GABA top-k function starts out as excitatory, and then transitions into inhibitory.

When applied together with Elastic Weight Consolidation (Kirkpatrick *et al.* 2017) to the class incremental setting without memory replay for CIFAR-10, SDMLP had been shown to obtain state-of-the art results (Bricken *et al.* 2023).

3.3 Other

We briefly include a few other biological aspects that we think could be fruitful to investigate in the context of continual RL, although related strategies are not implemented in the thesis. These aspects are plasticity within representation hierarchies, and prioritizing general representations over granular during critical learning periods.

Plasticity Within Representation Hierarchies Our brains do not have full plasticity, especially when it comes to lower-order sensory information as opposed to higher cognitive functions (Hensch 2004; Sydnor *et al.* 2021). Sensory maps become established early in our lives and stay relatively stable. In one example, kittens with vision deprivation in one eye retain a vision deficit if this vision deprivation happens early post-birth during a critical period (Wiesel and Hubel 1963). These critical periods occur progressively later during development for progressively higher order representations (Sydnor *et al.* 2021; Voss *et al.* 2017).

Critical learning periods occur in artificial neural networks as well; if an information deficit is present early in neural network training, the neural network will not be able to learn the information once it is re-added (Achille *et al.* 2019).

We propose that when it comes to the stability and plasticity trade-off in continual RL, throughout more of the agent's lifetime, it may make sense to prioritize the stability of lower order representations, or the representations in the layers closer to the input, and the plasticity of higher order representations, in the layers closer to the output.

Prioritizing General Representations over Granular Early On In the brain, the formation of representations goes from more general to more granular over time (Taylor *et al.* 2021). For example, far before the maturation of memories for specific experiences, children first develop conceptual understand and generalizable knowledge (Keresztes *et al.* 2018). Indeed, earlier in development, memory engrams are dense and imprecise (Ramsaran *et al.* 2023).

Additionally, when learning a new task, place cells for that task's memory replay begin by "hovering" between states, and information gets progressively more granular as the task is revisited (Berners-Lee *et al.* 2022).

Moreover, mice that are allowed to explore a maze before the introduction of a reward overall learn to solve the maze faster than if the reward were present from the beginning (Tolman and Honzik 1930). Similarly, pre-training

improves continual learning in artificial neural networks through the early formation of generalizable representations that then undergo less catastrophic forgetting (Mehta *et al.* 2023).

To our knowledge, similar strategies have been under-explored in continual RL. While we do not have specific approaches to recommend, as RL generally makes heavy use of replay, it may be helpful to, for example, decrease the granularity of stored experience earlier during training.

Chapter 4

Methodology

In this chapter, we describe our implementation of the energy constraints and indexing modifications introduced in the previous chapter. Three of the modifications have previously been implemented in supervised learning (Bricken *et al.* 2023; Margalit *et al.* 2023; J. C. R. Whittington, Dorrell, *et al.* 2023), and only one in reinforcement learning on DQN (Mnih *et al.* 2015; Pan *et al.* 2021). We evaluate the effects of different modifications on the Soft Actor-Critic (SAC) reinforcement learning algorithm (Haarnoja, Zhou, *et al.* 2019) on sequences of environments that are set up for investigating challenges within continual learning. While we present some ablations in this chapter, Chapter 5 has the main results.

4.1 General Setup and Environments

We use the default hyperparameter values for the base agent (Haarnoja, Zhou, *et al.* 2019), and tune additional modification-specific hyperparameters on the first environment within a sequence. We reset the replay buffer between environments, and each environment starts with $10,000$ steps of random exploration.

To evaluate the effects on negative and forward transfer, or worsened v.s. improved performance following pre-training, we use environments from the DeepMind Control Suite (Tassa *et al.* 2018). We evaluate negative transfer on `quadruped-run` following pre-training on `quadruped-walk` as SAC shows decreased performance on the second environment following the pre-training. For

effects on forward transfer, we look at `humanoid-run` following pre-training on `humanoid-walk`, where unmodified SAC shows conversely higher performance on the second environment. We trained `humanoid` for 3 million steps on walk instead of 1 million steps to ensure that the positive transfer effect is not a result of insufficient training.

We evaluate catastrophic forgetting and overfitting on robot arm tasks from Metaworld, with positions between resets kept fixed and not randomized (Yu *et al.* 2021). For catastrophic forgetting, we use separate output heads and task IDs, and train on a thematically-related but otherwise arbitrarily chosen sequence: `faucet-close-v2` → `window-close-v2`→ `faucet-close-v2`. We look at forgetting of the previous environments. For overfitting, we use one output head and train on a sequence of three `hammer-v2` tasks, also arbitrarily chosen, with separate one-hot vector input IDs for each. We reset the replay buffers between environments for all cases.

Whenever significance testing is performed, we use a p-value threshold of 0.05. As we evaluate the effects of multiple modifications, we first do a one-way analysis of variance (ANOVA) to correct for the number of modifications. For plots with ANOVA p-values below 0.05, we perform post-hoc testing with pair-wise independent student t-tests. Significance testing is done with SciPy (Virtanen *et al.* 2020).

4.2 Modifications Details

4.2.1 Energy Constraints

We find that energy constraints worked best when applied to the actor as well as the critics (not shown).

Non-negativity and Energy Minimization - "small-bio"

For `small-bio`, energy minimization is simply imposed through l2 regularization of activations as well as weights, and non-negativity can be imposed with the ReLU activation function (Nair and Hinton 2010; J. C. R. Whittington, Dorrell, *et al.* 2023), which is what we do in this thesis. Accordingly, we use

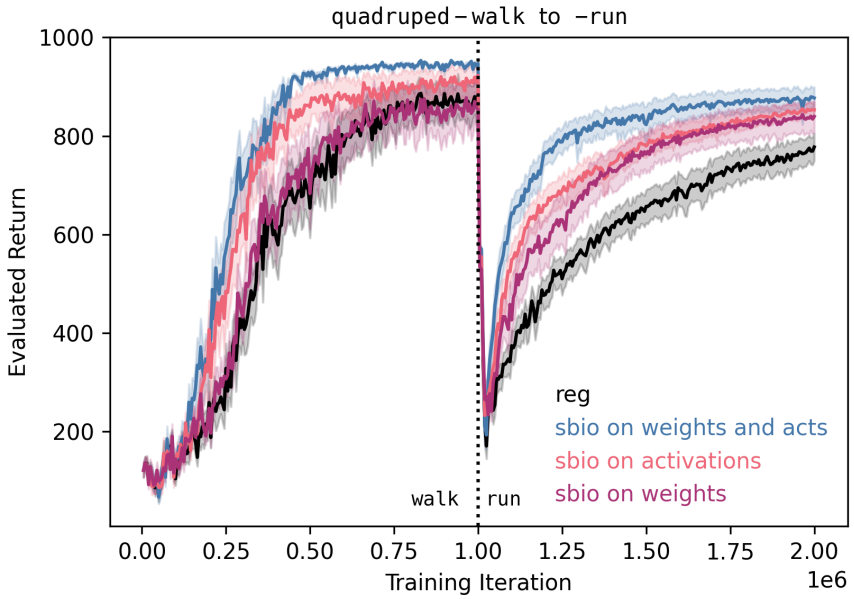

Figure 4.1: small-bio appears to work better when l2 regularization is applied to the activations as well as weights. There are 20 seeds, and the shading is standard error of the mean (SEM)

ReLU for `small-bio`, as well as for regular SAC, and the the following loss:

$$L_{sbio} = L_{default} + \beta L_{activations} + \beta L_{weights} \tag{4.1}$$

Here, $L_{default}$ is the respective regular SAC loss for the actor and the critics, β modulates the regularization strength, $L_{activations}$ is the l2 norm of the pre-ReLU activations for the respective network summed across layers, and $L_{weights}$ is the l2 norm of the weights for the respective network summed across layers. `sbio` is `small-bio`.

Consistently with the small-bio paper, we find L2 regularization of both activation and weights to appear to have better effects than L2 regularization applied to only the weights or only the activations, as seen in Figure 4.1.

Following tuning, we use a β value $1e^{-5}$ for `quadruped`, and $1e^{-6}$ for all the other environments.

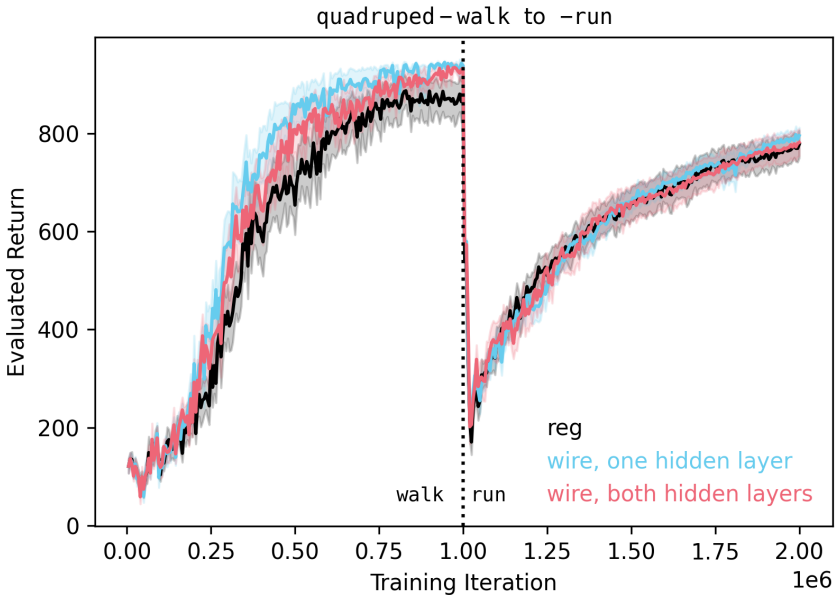

Figure 4.2: the wire modification appears to work best when applied to only one hidden layer.

Wiring Length Constraints - "wire"

We apply wiring length constraints on the second hidden layer in the actor and critic networks, and implement the Pearson's correlation spatial loss variant from the original `wire` paper (Margalit *et al.* 2023), but simply on a linear layer as opposed to a convolutional layer. The implementation consists of the following loss:

$$L_{wire} = L_{default} + \alpha(1 - Corr(\vec{r}, \vec{D})) \tag{4.2}$$

Here, α is the strength of the regularization, $Corr$ is Pearson's correlation, $\vec{r}$ is a vector of pairwise similarity distances between the pre-ReLU activations in the second hidden layer, and $\vec{D}$ is a vector of the inverse distances (in terms of position within a layer) for each pair. For pair i, $D_i = \frac{1}{d_i+1}$. This encourages neurons that are positioned close to each other within a layer to have similar activations, and neurons that are further apart to have different activations.

Figure 4.2 shows that applying `wire` on only one hidden layer appears to

sufficiently work, even thought the performance gains are minimal. We use an α value of 0.05 for `quadruped` and `humanoid`, and 0.1 for the `faucet` and `hammer` sets of environments.

4.2.2 Indexing

We find that both indexing constraints worked best when applied only to the actor, and not to the critics. This is shown below in Figure 4.3 for FTA.

Fuzzy Tiling Activations (FTA)

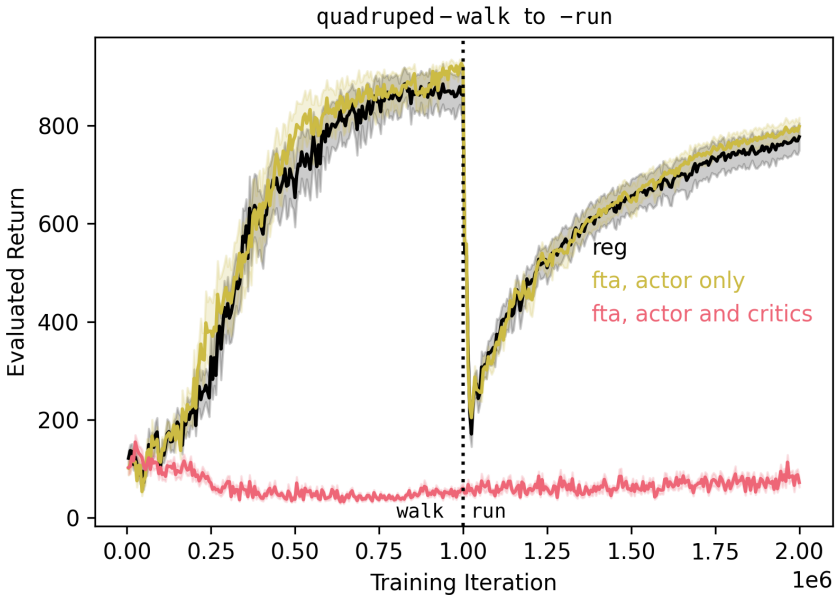

Figure 4.3: FTA did not lead to training when applied to the critics as well as the actor.

We apply FTA (Pan *et al.* 2021) to the actor's second hidden layer, and only to the actor. The remaining hidden layers have a ReLU activation function. As Figure 4.3 shows, we do not see training when FTA is applied to all of SAC's networks.

FTA is a fuzzy binning activation function. Following the original paper (Pan *et al.* 2021), the FTA activation function consists of the following:

$$\phi_\eta(z) := 1 - I_{\eta,+}(\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0)) \tag{4.3}$$

Here, δ controls the size of the FTA bins. Given an input z is a lower limit l and an upper limit u, then, the tiling vector $\mathbf{c}$ is:

$$\mathbf{c} := (l, l + \delta, l + 2\delta, ...u - \delta) \tag{4.4}$$

The fuzzy indicator function above $I_{\eta,+}(x)$, applied element-wise, is:

$$I_{\eta,+}(x) := I_+(\eta - x)x + I_+(x - \eta) \tag{4.5}$$

Here, η controls sparsity. For $x < \eta$, $I_{\eta,+}(x) = x$, and $I_{\eta,+}(x) = 1$ otherwise. Following tuning, for `quadruped` we use a lower limit of -10, an upper limit of 10, a δ of 5, and η of 5, and a hidden layer size of 1024 for the second hidden layer. For `humanoid`, these numbers are respectively -20, 20, 2, 2, and 5120. For all the other environments, we use -10, 10, 5, 1, 1024.

Sparse Distributed Memory Multilayer Perceptron (SDMLP)

SDMLP is the modification that requires the most changes to original SAC. To make SDMLP work for continual learning, the original paper's modifications include employing an activation function that is similar to a top-k function, eliminating neural network bias terms, and enforcing l2 normalization on the weights and data (Bricken *et al.* 2023).

We initially tried the non-GABA switch implementation of SDMLP, which is closer to a standard top-k activation function, but appeared to encounter the stale momentum problem reported by the paper. While using SGD without momentum is a potential fix (Bricken *et al.* 2023), we could not get training with SAC in general with SGD on `quadruped` (not shown).

For the GABA switch activation function, our implementation is close to that of the original paper's (Bricken *et al.* 2023). However, the SDMLP paper uses neural networks with only one hidden layer, while we use two, with the

GABA switch activation function applied to the second hidden layer. We additionally do not feed representations through a fixed pre-trained network, which is why we have the additional hidden layer prior to the one with the GABA switch activation function. Additionally, instead of l2 normalization of weights, we perform weight normalization (Salimans and Kingma 2016), which we found to perform better with SAC (not shown).

The GABA switch activation function is as follows:

$$\begin{aligned} z_i^* &:= \text{ReLU}(z_i - \lambda_i I) \\ I &:= \text{descending-sort}(\text{ReLU}(\mathbf{z}))_{(k+1)} \\ \lambda_i &:= \min(1, \max(-1, -1 + 2C_i/s)) \end{aligned} \tag{4.6}$$

Here, C_i counts how many times neuron i has been activated, s controls the ramp up time, and k is targeted number of neurons active at one time for when the ramp up time ends. I is activation of the $k+1$th most active neuron. As a neuron continues to get more active, λ_i progresses from -1 to 1. This means that at the very beginning of training, all neurons receive an activation boost, with the activation of the $k+1$th most active neuron added to theirs. However, as training progresses, the $k+1$th most active neuron transitions from excitatory to inhibitory, and eventually only neurons that are more active than the $k+1$th neuron remain active.

For `quadruped`, following tuning, k is 50, the size of the second hidden layer is 1024, and the number of transition steps if $750,000$. `humanoid`, these numbers are 100, 1024, $250,000$. For the `hammer` and `faucet` sets of environments, the number of k is 100, the hidden layer size is 2048, and there are $250,000$ transition steps. However, for all sets of environments, we slow the number of transition steps by 10 as a slower ramp up time helps reduce the stale momentum problem in SDMLP (Bricken *et al.* 2023), and we found that a slower ramp up time increases performance on environments subsequent to the first. Although a larger hidden layer is used in SDMLP, only 50 to 100 neurons are active at a time within the hidden layer soon after the beginning of training.

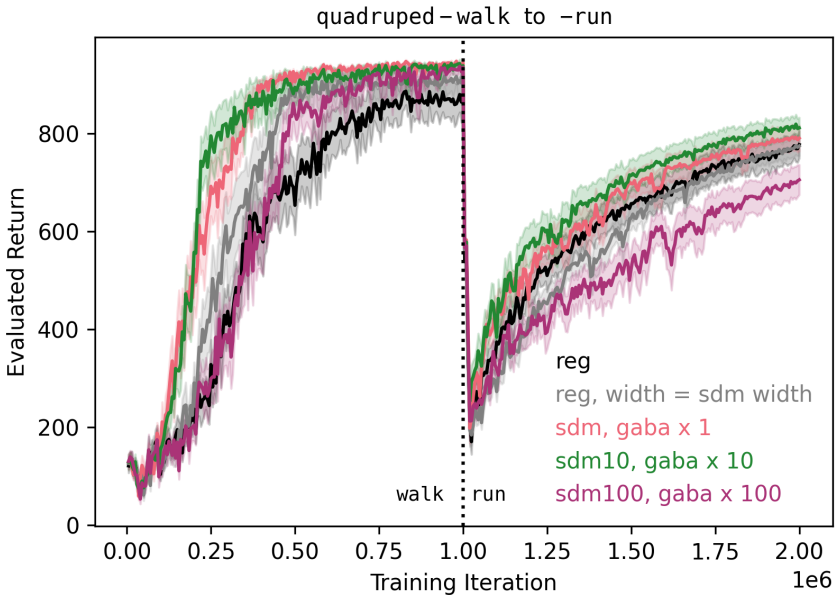

Figure 4.4: Matching hidden layer sizes does not lead to the same performance improvement with regular SAC as with the SDMLP (sdm) modified SAC. Additionally, slowing down SDMLP appears to be beneficial for performance on the second environment.

Figure 4.4 shows that increasing the second hidden layer in the actor network to the same number of neurons as in SDMLP does not lead to performance improvements on `quadruped-run` after training on `walk`.

Chapter 5

Results

In this section, we first look at the effects of the modifications on negative transfer, or losses in plasticity, forward transfer, generalization and catastrophic forgetting. Then, we analyze the resulting representations. We focus on the second hidden layer of the actor's neural network; while `small-bio` and `wire` are applied to the critics as well as the actor, FTA and SDMLP are only applied to the actor, and `wire`, FTA, and SDMLP are only applied to the second hidden layer.

5.1 Potential Mitigation Against Loss of Plasticity and Catastrophic Forgetting

5.1.1 Plasticity

As Figure 5.1 shows, `quadruped` shows seemingly negative transfer from `walk` to `run`, while `humanoid` has seemingly positive transfer that offsets any losses in plasticity that may be present. In an ideal situation in sequential task learning, we would want to see no loss in plasticity, and learning of future tasks being facilitated by the learning of similar prior tasks.

In this section, we explore the effects of the modifications on mitigating this loss in plasticity, on potentially on improving forward transfer, and also on generalization. `Small-bio` in particular is expected to help with performance past the first training environment due to it previously being shown to increase disentanglement (J. C. R. Whittington, Dorrell, *et al.* 2023), but a de-

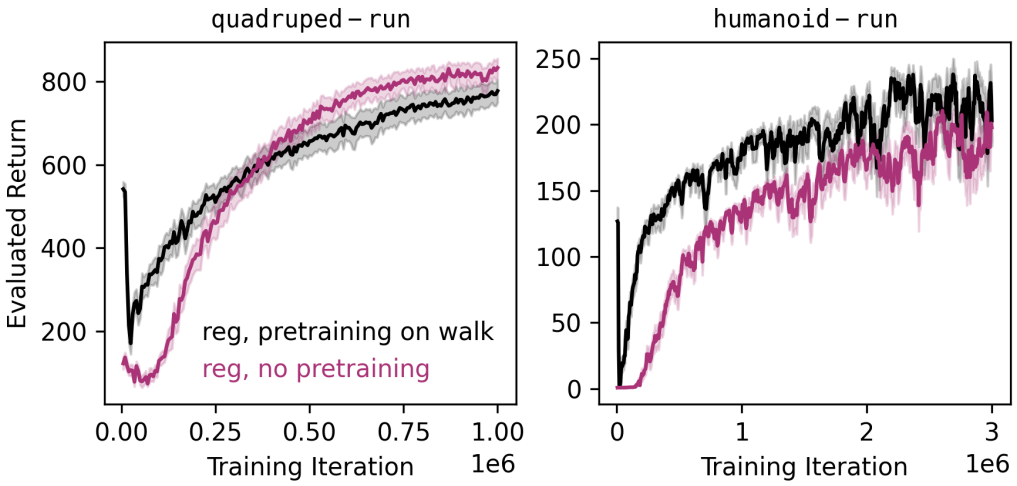

Figure 5.1: Pre-training on walk prior to training on run appears to hinder performance for quadruped, but to improve performance for humanoid.

crease in plasticity loss may also occur with the indexing modifications, FTA and SDMLP, due to potentially decreases weights overlap between environments. To evaluate significance, we perform an ANOVA on the performance plots with the modifications. For plots with ANOVA p-values below 0.05, we then perform independent student t-tests on each of the modifications versus regular SAC. We evaluate the areas under the curve (AUC) as well as the final performances when evaluating differences in plasticity.

We see a potentially mitigating effect on loss of plasticity with `small-bio`, and a reduction with overfitting (associated with plasticity and generalization), for all the modifications. However, we do not see a measurable improvement effect on positive transfer.

Forward and Negative Transfer For overall increases in performance on the first task, Figure 5.2 shows a significant performance increase from `small-bio`, `wire`, and SDMLP on `quadruped-walk`, but the ANOVA p-values are not above 0.05 for `humanoid-walk`.

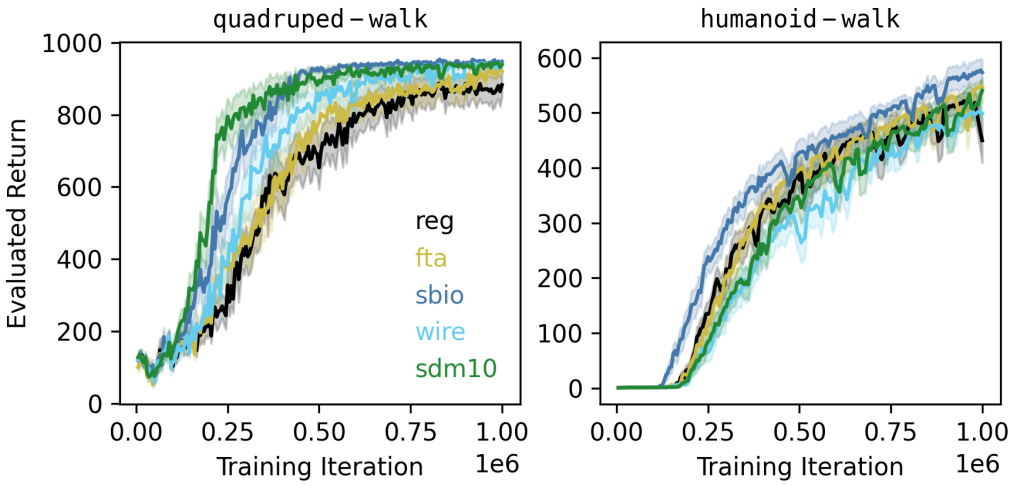

Figure 5.2: Performance on the first environment for quadruped and humanoid. sbio is small-bio, and sdm10 is SDMLP. The shading is SEM, and there are 20 seeds per run. For the area under the curve (AUC), the ANOVA p-values are 10^{-5} (left), and 0.005 (right). For final performance at the end of training, the ANOVA p-values are 0.03 (left), but also $p > 0.05$ (right). For individual t-tests relative to unmodified SAC, small-bio (0.0002), SDMLP (10^{-4}), and wire (0.046) have p-values < 0.05 for AUC for quadruped, but the final performance values are insignificant. For the AUC for humanoid, none of the p-values with individual t-tests are below 0.05.

For effects on performance on the second task, used to assess the forward and negative transfer, Figure 5.3 shows that `small-bio` significantly increases performance on `run` following pre-training on `walk` for both `quadruped` as well as `humanoid`, although only for the AUC for `humanoid`. As improvements in performance were already seen in the first set of environments in Figure 5.2, however, the increases in performance on `quadruped-run` in particular may be through an extraneous factor and not through a mitigation in plasticity loss.

Figure 5.4 attempts to isolate mitigation of loss of plasticity, or negative transfer, from possibly unrelated increases in performance; it also looks a potential increases in forward transfer. Figure 5.4 shows the following:

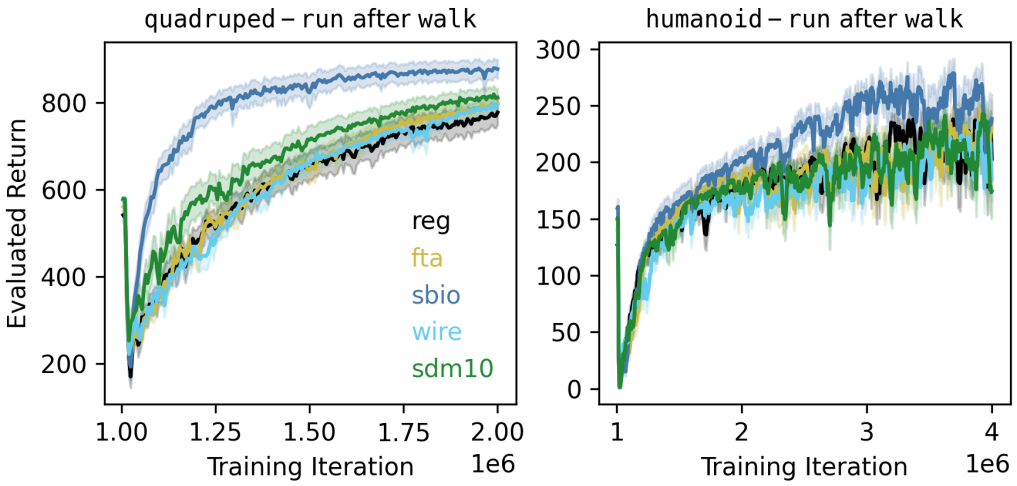

Figure 5.3: Performance on the second environment for quadruped and humanoid. The shading is SEM, and there are 20 seeds per run. For the AUC, the ANOVA p-values are $4 * 10^{-7}$ (left), and 0.0006 (right). For final performance at the end of training, the ANOVA p-values are 0.01 (left), but also $p > 0.05$ (right). For quadruped, compared to regular, small-bio has a p-value of $2 * 10^{-5}$ for the AUC, and 0.007 for the final performance. For humanoid, small-bio has a p-value of 0.003 for the AUC. The other p-values are not above 0.05

$$\begin{aligned}\text{Norm. Eval. Return} = \quad & (\text{perf. with pre-training} - \text{perf. from scratch}) \\ & - (\text{regular perf. with pre-training} \\ & - \text{regular perf. from scratch})\end{aligned} \tag{5.1}$$

That is, if pre-training causes less of a performance decrease with modified runs than with regular SAC, the normalized evaluated return will be above 0. It would also be above 0 if the performance gain for a modification due to pre-training is higher than with modified SAC. In Figure 5.4, for `quadruped`, `small-bio` is the only modification that appears to mitigate a loss in plasticity. Curiously, it also does not appear to increase positive transfer in `humanoid`, despite increasing overall performance on `humanoid-run` in Figure 5.3. This may suggest that `humanoid`, going from `walk` to `run`, does not suffer from any losses in plasticity that would reduce forward transfer in the first place.

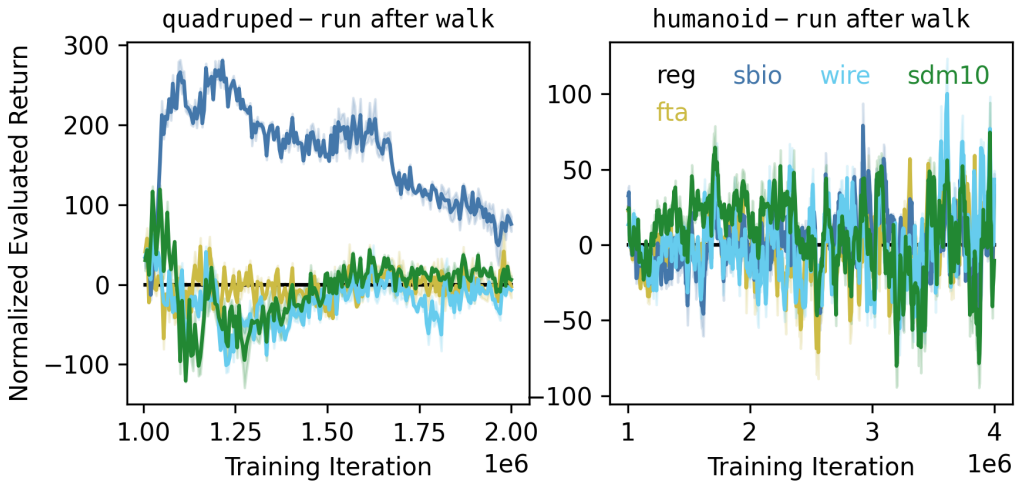

Figure 5.4: Small-bio may be the only modification that is helping prevent negative transfer (left). None of the modifications, including small-bio, appear to help increase forward transfer (right). The plots show the difference between pre-training vs from scratch performance for the modifications, minus that same difference between the unmodified runs.

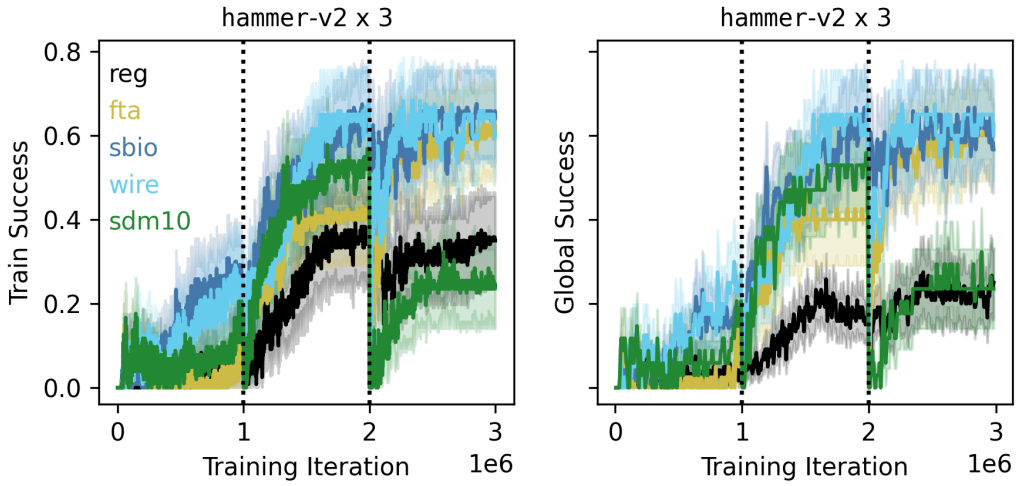

Figure 5.5: All four modifications decrease overfitting on three hammer-v2 tasks with separate task IDs. Left is training success, right is overall success on three hammer-v2s. 20 seeds per run, and shading is SEM. For final performance values, the ANOVA p-values are 0.046 (left), and 0.0005 (right). small-bio relative to regular SAC has a p-value of 0.02 (right), but all the other values relative to regular have a p-value > 0.05.

Overfitting Overfitting is associated with generalization as well as loss of plasticity (Lyle, Rowland, *et al.* 2022; Nikishin *et al.* 2022). Figure 5.5 shows that all the modifications decrease overfitting relative to regular SAC on robot arm tasks by showing comparable performance across training as well as global success on one environment with different task IDs. However, for this analysis, we are only qualitatively comparing the shapes of the curves.

We also performed significance testing for the final performances of the modifications versus regular SAC, with an ANOVA and independent student t-tests. `small-bio` shows significant performance increases in the global success plot relative to unmodified SAC.

5.1.2 Catastrophic Forgetting

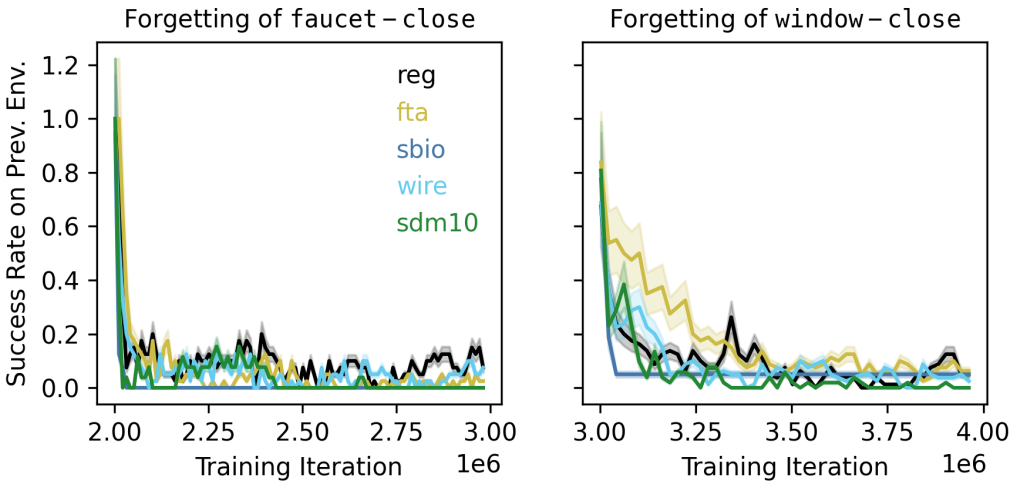

Figure 5.6: FTA may slightly slow down forgetting, although the ANOVA p-value for the AUC values in the right plot is not significant ($p = 0.051$). Forgetting of the first (left) and second (right) environment in the sequence of faucet-close → window-close → faucet-close Metaworld robot arm tasks. The shading is SEM, and there are 20 seeds per run.

In this section, we look at how slowly an environment is forgotten during the learning of a subsequent task. We primarily expect the indexing modifications, FTA and SDMLP, to slow forgetting due to the expectation that they decrease interference in weights between tasks. However, any actual improve-

ments in remembering ends up being unclear.

Figure 5.6 qualitatively shows that FTA may slightly slows down forgetting of `window-close` during the training of `faucet-close`, although the ANOVA p-value for the AUC of that plot is not significant. Conversely, `small-bio` may speed up forgetting for both `faucet-close` as well as `window-close`.

5.1.3 Summary

These data show that `small-bio`, `wire`, and SDMLP significantly increase AUC-based performance for SAC in `quadruped-walk`. Following pre-training on the respective `walk`, `small-bio` significantly increases the final performance as well as AUC-based performance in `quadruped-run` as well as `humanoid-run`, while the other modifications do not show significant improvements relative to unmodified SAC. The performance increase by `small-bio` on `quadruped-run` may be due to mitigating the loss of plasticity from the task switch for `quadruped`; however, the performance increase in `humanoid` does not appear to be through a mitigation in loss of plasticity, and possibly through some other mechanism. All modifications additionally appear to decrease overfitting. The effects of the modifications on catastrophic forgetting do not appear to be promising, although FTA may potentially slow forgetting of `window-close` during the learning of a subsequent environment.

5.2 Analysis of the Learned Representations

To help clarify the results of the previous section, this section shows an analysis of the representations produced by the actor networks. We do not see notable trends with the first three metrics, namely gradient orthogonality, gradient sparsity, and disentanglement. However, we do see trends in a few other metrics, including the non-updated units and parameter norm, that may suggest mechanisms behind increased plasticity and generalization with some of the modifications.

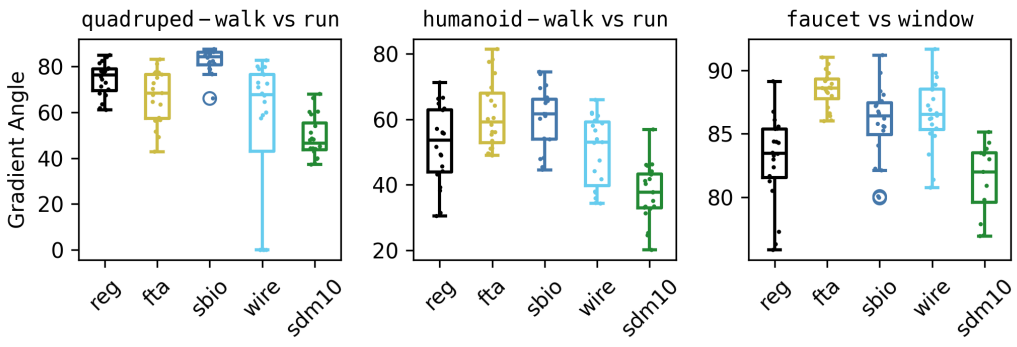

Figure 5.7: The gradient orthogonality measurements do not suggest a trend. The measurements are for the actor's final hidden layer.

5.2.1 Gradient Orthogonality

Orthogonal gradients between tasks are expected to decrease forgetting due to less interference with prior knowledge, and have been used or linked to succesful continual learning performance (Farajtabar *et al.* 2020; Mirzadeh *et al.* 2022).

However, in Figure 5.7, we do not see a meaningful trend with the angles of the gradients between tasks for the different modification; while FTA slightly slowed down forgetting of `window-close` in 5.6, and its gradient orthogonality may look a bit higher for `faucet` vs `window`, the differences appear negligible. Moreover, `small-bio`, which has the fastest forgetting in 5.6, does not appear to have the lowest gradient orthogonality. We similarly did not see a trend with the critic networks (not shown).

Increased gradient orthogonality did not appear to be a mechanism in slowing forgetting for FTA, or speeding up forgetting with `small-bio`. However, the gradient angles are compared at the end of training for each task, and it is possible that a decrease in interference between gradients for each task is happening towards the beginning of training.

5.2.2 Gradient Sparsity

An increased gradient sparsity could also decrease catastrophic forgetting, as it would imply that fewer parameters need to change between tasks (Mirzadeh *et*

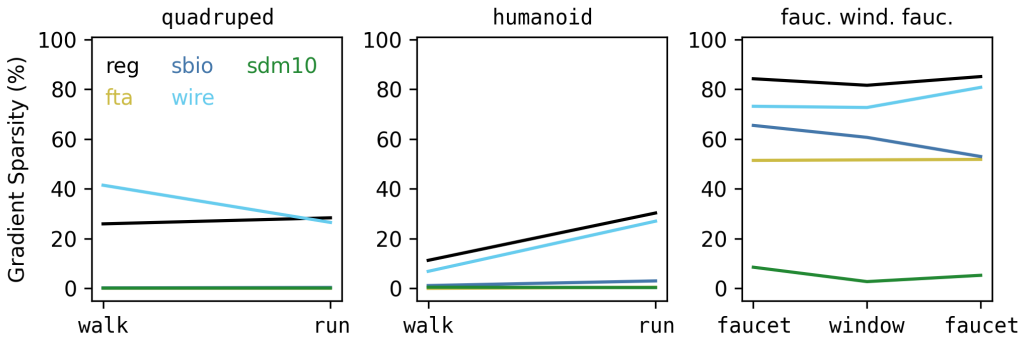

Figure 5.8: SDM maintains a high gradient sparsity, while small-bio and wire maintain a low one. The gradient orthogonality and disentanglement (MIR) measurements do not suggest a trend. The measurements are for the actor's final hidden layer.

al. 2022). A higher gradient sparsity would be particularly relevant for all tasks after the first task. However, Figure 5.8 shows regular SAC having the highest gradient sparsity, along with `wire`. Also surprisingly, the modifications with the sparse activation functions, FTA and SDMLP, have the lowest gradient sparsities.

5.2.3 Disentanglement

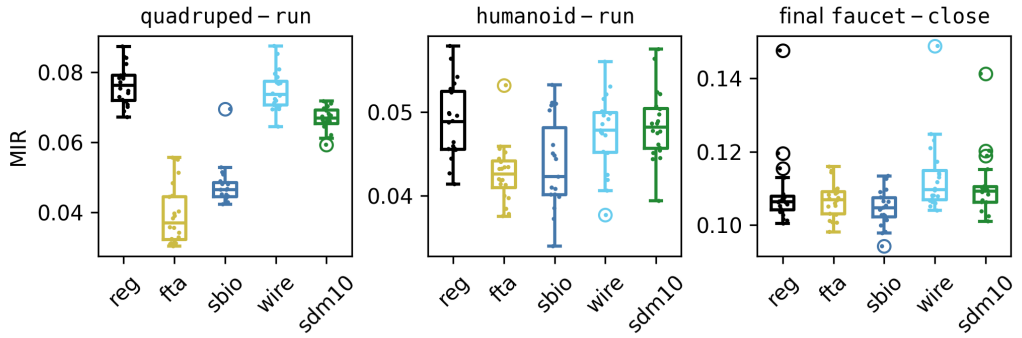

Figure 5.9: The disentanglement (MIR) measurements do not suggest a trend. The measurements are for the actor's final hidden layer.

Here, the mutual information ratio metric (MIR) attempts to measure how many neurons are coding for only one factor of variation in the data (J. C. R.

Whittington, Dorrell, *et al.* 2023).

Despite previously demonstrated increases in disentanglement with `small-bio` using MIR (J. C. R. Whittington, Dorrell, *et al.* 2023), we do not see this with `small-bio` applied to SAC. Figure 5.9 instead shows inconsistent results across environments for different modifications, which may indicate MIR being a poor metric for reinforcement learning data. However, the measurements of intrinsic dimensionality in the next section might support the measurements that we are seeing with MIR.

5.2.4 Effective and Intrinsic Dimensionality

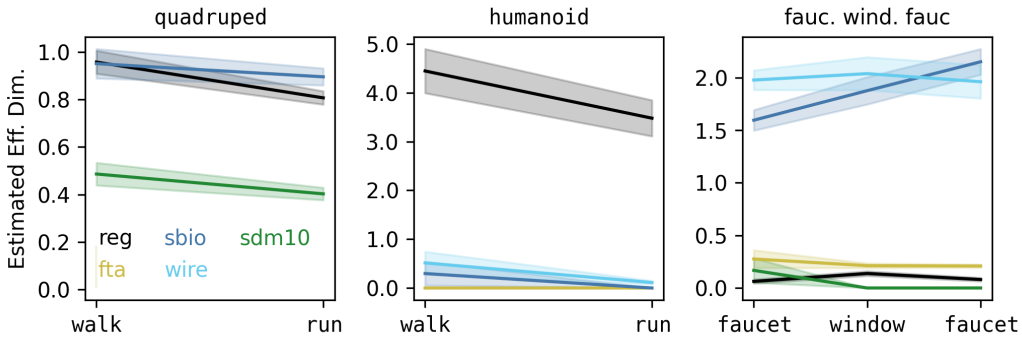

Figure 5.10: There appear to be no trends in measures of the effective dimensionality of the data within the policy network. The measurements are for the actor's final hidden layer. We use the measure with an effective dimensionality measure of n_2, as described in (Del Giudice 2021)

The effective dimensionality is a descriptive measure of how many dimensions in total are used to describe the data, while intrinsic dimensionality is an attempt to measure how many "important" features are being used (Del Giudice 2021). For instance, despite potentially many pixels used in the representation of images of nature, nature images actually have a low intrinsic dimensionality (Pope *et al.* 2021). Datasets with a lower intrinsic dimensionality are also easier to learn, and lead to better generalization within models (Pope *et al.* 2021), which can be beneficial for continual learning (Gallardo *et al.* 2021).

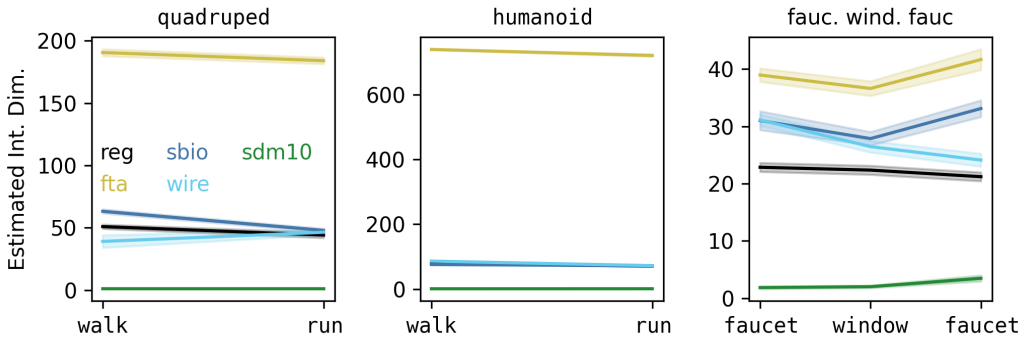

Figure 5.11: FTA greatly increases the intrinsic dimensionality of the data in the actor's final hidden layer. The plots show how many factors of variation are used to describe 90% of the data.

We do not see any trends in the effective dimensionality in Figure 5.11. However, we see that the intrinsic dimensionality is very high for FTA, suggesting that the FTA function "spreads out" the data in the neural network layer that FTA is applied to.

A higher intrinsic dimensionality does suggest more entanglement, which supports some of the MIR measurements; FTA appears to particularly lead to more entanglement compared to regular SAC for `quadruped` and `humanoid`, for which it also has the lowest disentanglement with the MIR metric in Figure 5.9. We also see `small-bio` with similar measurements to regular SAC's. However, we also see the most potential disentanglement with SDMLP in Figure 5.11, but no clear trend with it in Figure 5.9.

5.2.5 Stable Rank and Distance From Initialization

Figures 5.12 and 5.13 show that the most successful modification, `small-bio`, maintains a high stable rank and a low distance from initialization, both of which are beneficial for generalization (Nagarajan and Kolter 2019; Sanyal *et al.* 2019).

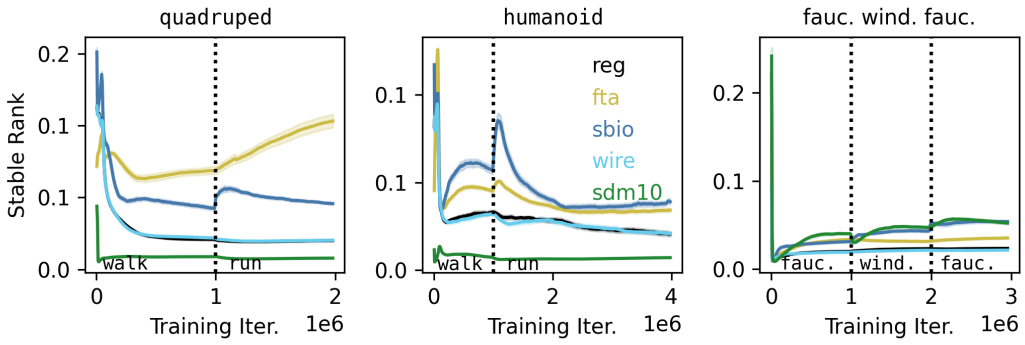

Figure 5.12: small-bio and FTA maintain an elevated stable rank across all three sets of environments. The measurements are for the actor's final hidden layer.

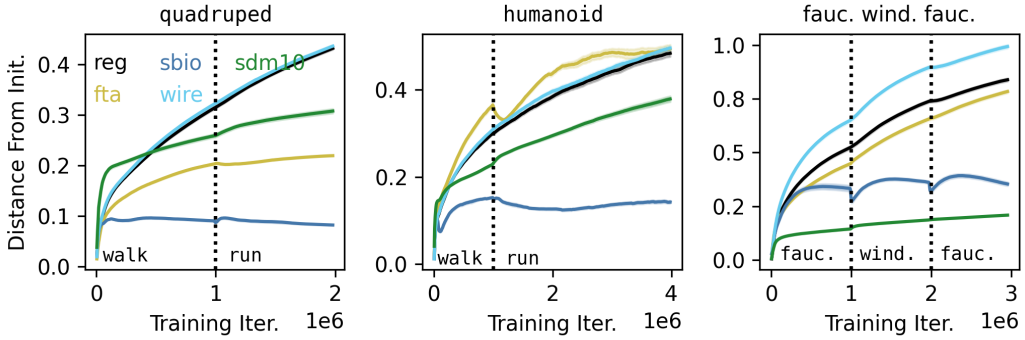

Figure 5.13: small-bio and SDMLP in particular maintain a low distance from initialization. The measurements are for the actor's final hidden layer.

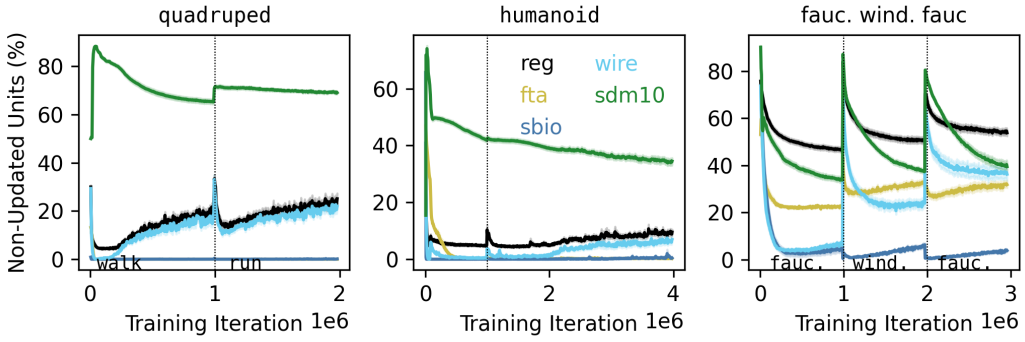

Figure 5.14: FTA and small-bio both decrease the percentage of non-updated units. The measurements are for the actor's final hidden layer.

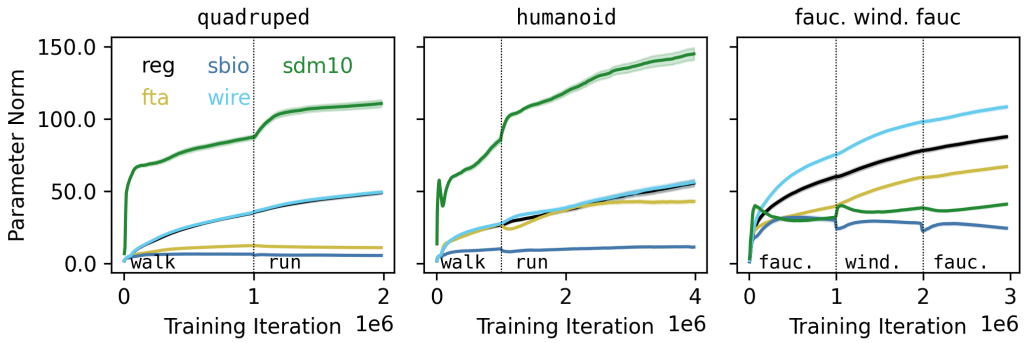

Figure 5.15: FTA and small-bio both decrease the l2 norm of the weights. SDMLP inconsistently greatly increases the l2 norm for the first two environment sets, and decreases it for the third set. The measurements are for the actor's final hidden layer.

5.2.6 Non-Updated Units and Parameter Norm

Increases in non-updated units, sometimes called dead units (Lyle, Zheng, Khetarpal, *et al.* 2024; Lyle, Zheng, Nikishin, *et al.* 2023), and in the norm of the parameters are potential mechanisms of plasticity loss (Lyle, Zheng, Khetarpal, *et al.* 2024). Consistently with this, Figures 5.14 and 5.15 show both `small-bio` and FTA, with `small-bio` in particular, maintaining the lowest percentage of non-updated units as well as parameters norms. Here, the percentage of non-updated units is calculated from mini-batches of data used for training; units are counted if they are inactive for all 256 points of data at the respective training iteration. That is, given a layer size n, the percentage of non-updated units is the following:

$$\text{Non-updated units } (\%) = 100\frac{\sum_i^n I_{non-updated}(\text{unit}_i)}{n} \tag{5.2}$$

The indicator function $I_{non-updated}$ returns 1 if the unit is non-updated, and 0 otherwise, as shown in the following equation:

$$I_{non-updated}(\text{unit}_i) = \begin{cases} 1, & \text{if unit}_i \text{ is non-updated} \\ 0, & \text{otherwise} \end{cases} \tag{5.3}$$

Each unit_i within the given layer is considered non-updated if, for all inputs j within mini-batch D, unit_i does not have an activation $h_i^j \neq 0$.

Interestingly, `small-bio` also shows as well a decrease in non-updated units and in the parameter norm with task switches in the `faucet-window-faucet` set of environments.

5.2.7 Summary

The metrics in this section primarily corroborate the increase in plasticity and decrease in overfitting potentially seen by the `small-bio` modification in the performance plots of the previous section. However, the metrics do not show an expected increase in disentanglement. Either the disentanglement metric used is unsuitable, or the improvements in plasticity and generalization are mainly through mechanisms other than disentanglement, such as a lower parameter norm.

Chapter 6

Discussion and Conclusion

This thesis presents an initial exploration of a few biologically-inspired modifications that have been understudied in continual RL. Although the modifications could benefit from more refined implementations, the modification with energy constraints and non-negativity, which we refer to as `small-bio`, has shown a particular promise for improving plasticity and generalization despite its simplicity. Fuzzy tiling activations (FTA) has also shown a potential benefit for slowing catastrophic forgetting, and all modifications suggest benefits in reducing overfitting and increasing generalization.

This study has its limitations. A key limitation is that investigating fewer methods more deeply might have yielded better insights. With modifications that are better optimized to RL in particular, we may have a better understanding of their potential. For instance, the continual learning performance of our implementation of the sparse distributed memory modification (SDMLP) on SAC did not match the continual learning performance of the original SDMLP implementation in a supervised learning environment (Bricken *et al.* 2023). Does this discrepancy show the inherent limitations of a GABA switch activation function on a slowly changing RL dataset, which we suspect may lead to the selection of less generalizable active GABA neurons compared to those selected on a supervised learning image dataset - or would certain RL-specific modifications have closed the gap? We do also note that we use a relatively large k in our implementation, while a lower k has been shown to increase remembering (Bricken *et al.* 2023).

Notably, one potential reason for this gap may have been not at least partially decoupling feature extraction from continual learning, as was done with the original SDMLP (Bricken *et al.* 2023). We overall think that this on its own is another biologically-grounded concept (Madireddy *et al.* 2023) that could be fruitful to investigate in future works.

We unfortunately were also not able to verify increased compositionality with the energy constraints. We did not see increased disentanglement with `small-bio` with the MIR metric, despite energy and positive constraints having provably been shown to lead to increased disentanglement (J. C. R. Whittington, Dorrell, *et al.* 2023). As increased disentanglement is suspected to lead to better generalization, which we did observe with `small-bio`, this may simply be an indication that the MIR metric was not suitable for the environments used in this thesis.

Additionally, we mainly investigated the representations generated by the actor, and only on one hidden layer. As all of the modifications were at minimum applied to the actor's hidden layer, this made the measurements more comparable between the different modifications. However, a further exploration of SAC's networks may also have been informative.

Furthermore, while we wanted to examine the effects of the modifications on challenging aspects of continual learning on their own, it would have been interesting to combine the modifications with traditional continual learning algorithms and observe the effect on their performance. We briefly implemented elastic weight consolidation (Kirkpatrick *et al.* 2017) (EWC - not shown), and despite not being able to observe the expected increase with remembering of prior tasks with our EWC implementation, we did note that `small-bio` in particular combined with EWC greatly exacerbated catastrophic forgetting.

We hope that this exploratory work encourages further investigations into algorithms inspired by brain function within the context of continual reinforcement learning.

In the long term, perhaps researchers will separate knowing (i.e., creating lasting representations that last following an initial external stimulus) from thinking (working with those representations) in continually learning

RL agents, and ultimately take inspiration from the default mode and theory of mind networks (Davey *et al.* 2016; Raichle *et al.* 2001; Soares *et al.* 2023) and allow machines to figure out their own place and motivations in an ever changing world.

References

[1] A. Achille, M. Rovere, and S. Soatto, "Critical learning periods in deep networks," p. 14, 2019. 14

[2] J. S. Albus, "A theory of cerebellar function," *Mathematical Biosciences*, vol. 10, no. 1, pp. 25–61, Feb. 1, 1971, ISSN: 0025-5564. (visited on 05/15/2024). 12

[3] A. Berners-Lee *et al.*, "Hippocampal replays appear after a single experience and incorporate greater detail with more experience," *Neuron*, vol. 110, no. 11, 1829–1842.e5, Jun. 1, 2022, Publisher: Elsevier, ISSN: 0896-6273. (visited on 08/02/2024). 14

[4] T. Bricken, X. Davies, D. Singh, D. Krotov, and G. Kreiman, "Sparse distributed memory is a continual learner," presented at the The Eleventh International Conference on Learning Representations, Feb. 1, 2023. (visited on 02/20/2024). 2, 13, 16, 21, 22, 38, 39

[5] C. G. Davey, J. Pujol, and B. J. Harrison, "Mapping the self in the brain's default mode network," *NeuroImage*, vol. 132, pp. 390–397, May 15, 2016, ISSN: 1053-8119. (visited on 06/28/2024). 40

[6] M. Del Giudice, "Effective dimensionality: A tutorial," *Multivariate Behavioral Research*, vol. 56, no. 3, pp. 527–542, Jul. 21, 2021, ISSN: 0027-3171. (visited on 02/21/2024). 33

[7] M. Farajtabar, N. Azizan, A. Mott, and A. Li, "Orthogonal gradient descent for continual learning," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ISSN: 2640-3498, PMLR, Jun. 3, 2020, pp. 3762–3773. (visited on 07/25/2024). 31

[8] T. Flesch, A. Saxe, and C. Summerfield, "Continual task learning in natural and artificial agents," *Trends in Neurosciences*, vol. 46, no. 3, pp. 199–210, Mar. 1, 2023, Publisher: Elsevier, ISSN: 0166-2236, 1878-108X. (visited on 06/23/2024). 8, 10, 12

[9] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, Apr. 1, 1980, ISSN: 1432-0770. (visited on 07/16/2024). 8

[10] K. Fukushima and S. Miyake, "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position," *Pattern Recognition*, vol. 15, no. 6, pp. 455–469, Jan. 1, 1982, ISSN: 0031-3203. (visited on 07/16/2024). 8

[11] J. Gallardo, T. L. Hayes, and C. Kanan, *Self-supervised training enhances online continual learning*, Oct. 22, 2021. arXiv: 2103.14010[cs]. (visited on 07/25/2024). 33

[12] K. Ganguly, A. F. Schinder, S. T. Wong, and M.-m. Poo, "GABA itself promotes the developmental switch of neuronal GABAergic responses from excitation to inhibition," *Cell*, vol. 105, no. 4, pp. 521–532, May 18, 2001, Publisher: Elsevier, ISSN: 0092-8674, 1097-4172. (visited on 08/02/2024). 13

[13] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *Proceedings of the 34th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, Jul. 17, 2017, pp. 1352–1361. (visited on 07/13/2024). 6

[14] T. Haarnoja, A. Zhou, *et al.*, *Soft actor-critic algorithms and applications*, Jan. 29, 2019. arXiv: 1812.05905[cs,stat]. (visited on 02/01/2024). 1, 7, 16

[15] T. K. Hensch, "Critical period regulation," *Annual Review of Neuroscience*, vol. 27, pp. 549–579, Volume 27, 2004 Jul. 21, 2004, Publisher: Annual Reviews, ISSN: 0147-006X, 1545-4126. (visited on 08/02/2024). 14

[16] S. Herculano-Houzel, "Scaling of brain metabolism with a fixed energy budget per neuron: Implications for neuronal activity, plasticity and evolution," *PLOS ONE*, vol. 6, no. 3, e17514, Mar. 1, 2011, Publisher: Public Library of Science, ISSN: 1932-6203. (visited on 07/31/2024). 9

[17] Ø. A. Høydal, E. R. Skytøen, S. O. Andersson, M.-B. Moser, and E. I. Moser, "Object-vector coding in the medial entorhinal cortex," *Nature*, vol. 568, no. 7752, pp. 400–404, Apr. 2019, Publisher: Nature Publishing Group, ISSN: 1476-4687. (visited on 05/14/2024). 9

[18] P. Kanerva, *Sparse Distributed Memory*. MIT Press, 1988, 194 pp., ISBN: 978-0-262-11132-4. 13

[19] P. Kanerva, "Sparse distributed memory and related models," NASA-CR-190553, Apr. 1, 1992, NTRS Author Affiliations: Research Inst. for Advanced Computer Science NTRS Document ID: 19920021480 NTRS Research Center: Legacy CDMS (CDMS). (visited on 05/10/2024). 13

[20] A. Keresztes, C. T. Ngo, U. Lindenberger, M. Werkle-Bergner, and N. S. Newcombe, "Hippocampal maturation drives memory from generalization to specificity," *Trends in Cognitive Sciences*, vol. 22, no. 8, pp. 676–686, Aug. 1, 2018, Publisher: Elsevier, ISSN: 1364-6613, 1879-307X. (visited on 08/02/2024). 14

[21] S. S. Kety, "THE GENERAL METABOLISM OF THE BRAIN *IN VIVO*," in *Metabolism of the Nervous System*, D. Richter, Ed., Pergamon, Jan. 1, 1957, pp. 221–237, ISBN: 978-0-08-009062-7. DOI: 10.1016/B978-0-08-009062-7.50026-6. (visited on 07/31/2024). 9

[22] J. Kirkpatrick *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, Mar. 28, 2017. (visited on 04/09/2024). 13, 39

[23] V. R. Konda and J. N. Tsitsiklis, "On actor-critic algorithms," *SIAM Journal on Control and Optimization*, vol. 42, no. 4, pp. 1143–1166, Jan. 2003, ISSN: 0363-0129, 1095-7138. (visited on 07/25/2024). 6

[24] D. Kudithipudi *et al.*, "Biological underpinnings for lifelong learning machines," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 196–210, Mar. 2022, Publisher: Nature Publishing Group, ISSN: 2522-5839. (visited on 06/24/2024). 1, 8

[25] Z. Kurth-Nelson *et al.*, "Replay and compositional computation," *Neuron*, vol. 111, no. 4, pp. 454–469, Feb. 2023, ISSN: 08966273. (visited on 07/20/2023). 10

[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, Publisher: Nature Publishing Group, ISSN: 1476-4687. (visited on 07/16/2024). 8

[27] S. Lewandowsky and S.-C. Li, "10 - catastrophic interference in neural networks: Causes, solutions, and data," in *Interference and Inhibition in Cognition*, F. N. Dempster, C. J. Brainerd, and C. J. Brainerd, Eds., San Diego: Academic Press, Jan. 1, 1995, pp. 329–361, ISBN: 978-0-12-208930-5. DOI: 10.1016/B978-012208930-5/50011-8. (visited on 08/01/2024). 12

[28] C. Lyle, M. Rowland, and W. Dabney, *Understanding and preventing capacity loss in reinforcement learning*, Apr. 20, 2022. (visited on 10/29/2023). 29

[29] C. Lyle, Z. Zheng, K. Khetarpal, *et al.*, *Disentangling the causes of plasticity loss in neural networks*, Feb. 28, 2024. arXiv: 2402.18762[cs]. (visited on 04/12/2024). 1, 3, 36

[30] C. Lyle, Z. Zheng, E. Nikishin, B. A. Pires, R. Pascanu, and W. Dabney, "Understanding plasticity in neural networks," in *Proceedings of the 40th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, Jul. 3, 2023, pp. 23 190–23 211. (visited on 09/01/2024). 36

[31] S. Madireddy, A. Yanguas-Gil, and P. Balaprakash, "Improving performance in continual learning tasks using bio-inspired architectures," in *Proceedings of The 2nd Conference on Lifelong Learning Agents*, ISSN: 2640-3498, PMLR, Nov. 20, 2023, pp. 992–1008. (visited on 06/23/2024). 39

[32] E. Margalit, H. Lee, D. Finzi, J. J. DiCarlo, K. Grill-Spector, and D. L. K. Yamins, *A unifying principle for the functional organization of visual cortex*, May 18, 2023. DOI: 10.1101/2023.05.18.541361. (visited on 01/25/2024). 2, 10, 11, 16, 19

[33] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory," *Psychological Review*, vol. 102, no. 3, pp. 419–457, 1995, Place: US Publisher: American Psychological Association, ISSN: 1939-1471. 12

[34] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1, 1943, ISSN: 1522-9602. (visited on 07/17/2024). 8

[35] S. V. Mehta, D. Patil, S. Chandar, and E. Strubell, "An empirical investigation of the role of pre-training in lifelong learning," *Journal of Machine Learning Research*, vol. 24, no. 214, pp. 1–50, 2023, ISSN: 1533-7928. (visited on 08/02/2024). 15

[36] J. A. Mendez and E. Eaton, *Lifelong learning of compositional structures*, Mar. 17, 2021. arXiv: 2007.07732[cs,stat]. (visited on 08/01/2024). 10

[37] J. A. Mendez and E. Eaton, *How to reuse and compose knowledge for a lifetime of tasks: A survey on continual learning and functional composition*, Jul. 15, 2022. (visited on 06/11/2023). 10

[38] S. I. Mirzadeh *et al.*, "Wide neural networks forget less catastrophically," in *Proceedings of the 39th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, Jun. 28, 2022, pp. 15 699–15 717. (visited on 06/21/2023). 31

[39] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015, Number: 7540 Publisher: Nature Publishing Group, ISSN: 1476-4687. (visited on 02/19/2024). 5, 8, 13, 16

[40] V. Nagarajan and J. Z. Kolter, *Generalization in deep networks: The role of distance from initialization*, Jan. 13, 2019. arXiv: 1901.01672[cs,stat]. (visited on 06/16/2024). 34

[41] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10, Madison, WI, USA: Omnipress, Jun. 21, 2010, pp. 807–814, ISBN: 978-1-60558-907-7. (visited on 05/10/2024). 11, 17

[42] E. Nikishin, M. Schwarzer, P. D'Oro, P.-L. Bacon, and A. Courville, "The primacy bias in deep reinforcement learning," in *Proceedings of the 39th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, Jun. 28, 2022, pp. 16 828–16 847. (visited on 10/29/2023). 1, 29

[43] R. C. O'Reilly, R. Bhattacharyya, M. D. Howard, and N. Ketz, "Complementary learning systems," *Cognitive Science*, vol. 38, no. 6, pp. 1229–1248, Aug. 2014, ISSN: 1551-6709. 12

[44] S. Oldham *et al.*, "Modeling spatial, developmental, physiological, and topological constraints on human brain connectivity," *Science Advances*, vol. 8, no. 22, eabm6127, Jun. 3, 2022, Publisher: American Association for the Advancement of Science. (visited on 07/31/2024). 9, 11

[45] Z. Padamsey and N. L. Rochefort, "Paying the brain's energy bill," *Current Opinion in Neurobiology*, vol. 78, p. 102 668, Feb. 1, 2023, ISSN: 0959-4388. (visited on 07/31/2024). 9

[46] Y. Pan, K. Banman, and M. White, "Fuzzy tiling activations: A simple approach to learning sparse representations online," presented at the International Conference on Learning Representations, Jan. 12, 2021. (visited on 02/20/2024). 2, 12, 16, 20

[47] J. C. Pang *et al.*, "Geometric constraints on human brain function," *Nature*, vol. 618, no. 7965, pp. 566–574, Jun. 2023, Number: 7965 Publisher: Nature Publishing Group, ISSN: 1476-4687. (visited on 07/25/2023). 9

[48] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, May 1, 2019, ISSN: 0893-6080. (visited on 06/26/2024). 1

[49] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein, *The intrinsic dimension of images and its impact on learning*, Apr. 18, 2021. arXiv: `2104.08894[cs,stat]`. (visited on 07/19/2024). 33

[50] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley and Sons, 2014. (visited on 12/13/2022). 4

[51] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman, "A default mode of brain function," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 676–682, Jan. 16, 2001, Publisher: Proceedings of the National Academy of Sciences. (visited on 06/28/2024). 40

[52] A. I. Ramsaran *et al.*, "A shift in the mechanisms controlling hippocampal engram formation during brain maturation," *Science*, vol. 380, no. 6644, pp. 543–551, May 5, 2023, Publisher: American Association for the Advancement of Science. (visited on 08/02/2024). 14

[53] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958, Place: US Publisher: American Psychological Association, ISSN: 1939-1471. 8

[54] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc., 2016. (visited on 02/01/2024). 22

[55] A. Sanyal, P. H. Torr, and P. K. Dokania, "Stable rank normalization for improved generalization in neural networks and GANs," presented at the International Conference on Learning Representations, Sep. 25, 2019. (visited on 06/16/2024). 34

[56] C. Soares, G. Gonzalo, J. Castelhano, and M. Castelo-Branco, "The relationship between the default mode network and the theory of mind network as revealed by psychedelics – a meta-analysis," *Neuroscience & Biobehavioral Reviews*, vol. 152, p. 105 325, Sep. 1, 2023, ISSN: 0149-7634. (visited on 06/28/2024). 40

[57] T. Solstad, C. N. Boccara, E. Kropff, M.-B. Moser, and E. I. Moser, "Representation of geometric borders in the entorhinal cortex," *Science*, vol. 322, no. 5909, pp. 1865–1868, Dec. 19, 2008, Publisher: American Association for the Advancement of Science. (visited on 05/14/2024). 10

[58] A. Subramanian, S. Chitlangia, and V. Baths, "Reinforcement learning and its connections with neuroscience and psychology," *Neural Networks*, vol. 145, pp. 271–287, Jan. 1, 2022, ISSN: 0893-6080. (visited on 07/17/2024). 8

[59] R. S. Sutton, "Generalization in reinforcement learning: Successful examples using sparse coarse coding," in *Advances in Neural Information Processing Systems*, vol. 8, MIT Press, 1995. (visited on 05/15/2024). 12

[60] R. S. Sutton and A. G. Barto, "Toward a modern theory of adaptive networks: Expectation and prediction.," *Psychological Review*, vol. 88, no. 2, pp. 135–170, 1981, ISSN: 0033-295X. DOI: 10.1037/0033-295X.88.2.135. (visited on 07/17/2024). 8

[61] R. S. Sutton and A. G. Barto, "A temporal-difference model of classical conditioning," *Proceedings of the ninth annual conference of the cognitive science society*, 1987. (visited on 07/17/2024). 8

[62] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction* (Adaptive computation and machine learning series), Second edition. Cambridge, Massachusetts: The MIT Press, 2018, 526 pp., ISBN: 978-0-262-03924-6. 4, 5

[63] V. J. Sydnor *et al.*, "Neurodevelopment of the association cortices: Patterns, mechanisms, and implications for psychopathology," *Neuron*, vol. 109, no. 18, pp. 2820–2846, Sep. 15, 2021, ISSN: 0896-6273. (visited on 08/02/2024). 14

[64] K. Takagi, "Energy constraints on brain network formation," *Scientific Reports*, vol. 11, no. 1, p. 11 745, Jun. 3, 2021, Publisher: Nature Publishing Group, ISSN: 2045-2322. (visited on 07/31/2024). 9

[65] Y. Tassa *et al.*, *DeepMind control suite*, Jan. 2, 2018. arXiv: 1801.00690[cs]. (visited on 02/01/2024). 3, 16

[66] J. E. Taylor, A. Cortese, H. C. Barron, X. Pan, M. Sakagami, and D. Zeithamova, "How do we generalize?" *Neurons, behavior, data analysis and theory*, vol. 1, p. 001c.27687, Aug. 30, 2021. (visited on 07/29/2023). 14

[67] T. J. Teyler and P. DiScenna, "The hippocampal memory indexing theory," *Behavioral Neuroscience*, vol. 100, no. 2, pp. 147–154, 1986, Place: US Publisher: American Psychological Association, ISSN: 1939-0084. 12

[68] T. J. Teyler and J. W. Rudy, "The hippocampal indexing theory and episodic memory: Updating the index," *Hippocampus*, vol. 17, no. 12, pp. 1158–1169, 2007, ISSN: 1098-1063. (visited on 05/14/2024). 12

[69] E. C. Tolman and C. H. Honzik, "Introduction and removal of reward, and maze performance in rats," *University of California Publications in Psychology*, vol. 4, pp. 257–275, 1930. 14

[70] P. Virtanen *et al.*, "SciPy 1.0: Fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, ISSN: 1548-7105. 17

[71] P. Voss, M. E. Thomas, J. M. Cisneros-Franco, and É. de Villers-Sidani, "Dynamic brains and the changing rules of neuroplasticity: Implications for learning and recovery," *Frontiers in Psychology*, vol. 8, Oct. 4, 2017, Publisher: Frontiers, ISSN: 1664-1078. (visited on 08/02/2024). 14

[72] H. Wang *et al.*, "Investigating the properties of neural network representations in reinforcement learning," *Artificial Intelligence*, vol. 330, p. 104 100, May 1, 2024, ISSN: 0004-3702. (visited on 03/11/2024). 2, 13

[73] L. Wang, X. Zhang, H. Su, and J. Zhu, *A comprehensive survey of continual learning: Theory, method and application*, Feb. 6, 2024. arXiv: 2302.00487[cs]. (visited on 06/23/2024). 1

[74] J. C. R. Whittington, W. Dorrell, S. Ganguli, and T. Behrens, "Disentanglement with biological constraints: A theory of functional cell types," presented at the The Eleventh International Conference on Learning Representations, Feb. 28, 2023. (visited on 02/20/2024). 2, 10, 16, 17, 24, 32, 33,

[75] J. C. R. Whittington, D. McCaffary, J. J. W. Bakermans, and T. E. J. Behrens, "How to build a cognitive map," *Nature Neuroscience*, vol. 25, no. 10, pp. 1257–1272, Oct. 2022, ISSN: 1097-6256, 1546-1726. (visited on 07/04/2023). 10

[76] J. C. Whittington *et al.*, "The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation," *Cell*, vol. 183, no. 5, 1249–1263.e23, Nov. 2020, ISSN: 00928674. (visited on 07/17/2023). 10

[77] T. N. Wiesel and D. H. Hubel, "Effects of visual deprivation on morphology and physiology of cells in the cat's lateral geniculate body," *Journal of Neurophysiology*, vol. 26, no. 6, pp. 978–993, Nov. 1963, Publisher: American Physiological Society, ISSN: 0022-3077. (visited on 08/02/2024). 14

[78] M. Xie, S. P. Muscinelli, K. Decker Harris, and A. Litwin-Kumar, "Task-dependent optimal representations for cerebellar learning," *eLife*, vol. 12, J. Diedrichsen, M. J. Frank, J. Diedrichsen, and H. Jörntell, Eds., e82914, Sep. 6, 2023, Publisher: eLife Sciences Publications, Ltd, ISSN: 2050-084X. (visited on 05/10/2024). 12

[79] T. Yu *et al.*, *Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning*, Number: arXiv:1910.10897, Jun. 14, 2021. arXiv: `1910.10897[cs,stat]`. (visited on 06/13/2022). 3, 17