Transfer Learning in Reinforcement Learning:

Differentiable plasticity: training plastic neural networks with backpropagation

&

A Model-based Approach for Sample-efficient Multi-task Reinforcement Learning

João Valério - joao.agostinho@estudiantat.upc.edu

Universitat Politècnica de Catalunya

Master in Artificial Intelligence

Barcelona, Spain

Abstract

This paper delves into the topic of **transfer learning** in **artificial intelligence**'s field of **reinforcement learning**. **Transfer learning** is a technique that leverages 'external expertise from other domains to benefit the learning process of the target task' [1]. Its significance has grown in recent years across various domains, including **reinforcement learning**.

In this study, we focus on two notable scientific works: "Differentiable Plasticity: training plastic neural networks with Backpropagation" [2] published in 2018, and "A Model-based Approach for Sample-efficient Multi-task Reinforcement Learning" [3] published in 2019. By examining these approaches, the paper aims to provide a comprehensive understanding of the importance of **transfer learning** in **reinforcement learning**, as well as the scientific advancements made in each work.

Keywords: transfer learning, artificial intelligence, reinforcement learning, reinforcement learning

1. Introduction

'Reinforcement Learning (RL) is an effective framework to solve sequential decision-making tasks, where a learning agent interacts with the environment to improve its performance through trial and error' [4]. Diverging from other artificial intelligence fields, RL stands out due to its distinctive approach by focusing on 'how to map situations to actions' [4], which has propelled significant advancements within the field. However, despite 'its remarkable advancement, RL still faces intriguing difficulties induced by the exploration-exploitation dilemma' [4].

'Specifically, for practical RL problems, the environment dynamics are usually unknown, and the agent cannot exploit knowledge about the environment to improve its performance until enough interaction experiences are collected via exploration. Due to the partial observability, sparse feedback, and the high complexity of state and action spaces, acquiring sufficient interaction samples can be prohibitive, which may even incur safety concerns [...], where the consequences of wrong decisions can be too high to take. The abovementioned challenges have motivated various efforts to improve the current RL procedure. As a result, transfer learning [...], which is a technique to utilize external expertise from other domains to benefit the learning process of the target task, becomes a crucial topic in RL' [1].

In addressing these challenges, several approaches can be employed, each with its own specificities, advantages, and drawbacks. Examples include learning one task and fine-tuning policy and values for the subsequent task, input randomization to prepare for different scenarios or the use of entropy in the policy, multitasking and meta-learning, and transfer of information between tasks, among others.

This study focuses on two scientific papers: "Differentiable Plasticity: training plastic neural networks with Backpropagation" [2] published in 2018, and "A Model-based Approach for Sample-efficient Multi-task Reinforcement Learning" [3] published in 2019. These papers present distinct approaches in reinforcement learning. The former [2] draws inspiration from the synaptic plasticity of the brain, 'carefully tuned by evolution to produce efficient lifelong learning' [2]. The latter [3] proposes 'to learn a dynamical model during the training process and use this model to perform sample-efficient adaptation to new tasks at test time' [3].

Although these two approaches may belong to different domains, they share a common objective of constructing a flexible model capable of adapting swiftly and efficiently to new environments.

2. Background & Proposal

To facilitate a comprehensive understanding of the proposals presented in each work and enable a fair comparison, it is crucial to delve into the backgrounds that underpin their development. In this regard, Chapter 2.1 provides an exploration of the background knowledge incorporated in [2], while Chapter 2.2 delves into the background knowledge upon which [3] was built. By examining these backgrounds, we can establish the foundations and contextual framework essential for comprehending the contributions and significance of each respective work.

2.1. Differentiable Plasticity - Paper [2]

Machine Learning techniques have garnered significant attention in the field of artificial intelligence due to their remarkable ability to learn from extensive training on vast amounts of data (Krizhevsky et al., 2012; Mnih et al., 2015; Silver et al., 2016). These techniques have led to groundbreaking advancements in various domains, including reinforcement learning. However, 'after learning is complete, the agent's knowledge is fixed and unchanging; if the agent is to be applied to a different task, it must be re-trained (fully or partially), again requiring a very large number of new training examples' [2]. Consequently, the adaptation process between domains becomes expensive and impractical.

By recognizing that 'biological agents exhibit a remarkable ability to learn quickly and efficiently from ongoing experience' [2], the authors sought to reproduce this capacity of autonomous learning on neural networks. However, at the time, the majority of autonomous learning approaches did not recognize that 'in biological brains, long-term learning [...] occurs primarily through synaptic plasticity – the strengthening and weakening of connections between neurons as a result of neural activity [...]' [2]. Considering this, Hebb's rule plays a vital role:

"*Neurons that fire together, wire together.*"

'Which is to say that neural pathways consistently activated together become physiologically modified to facilitate future signal transductions' [8].

Therefore, building upon this knowledge, the authors propose to 'expand backpropagation training to networks with plastic connections – optimizing through gradient descent not only the base weights but also the amount of plasticity in each connection' [2], a concept referred to as differentiable plasticity. This approach introduces a flexible domain adaptation mechanism, distinct from most works on neural networks that primarily focus on non-plasticity. Thus, by drawing inspiration from a previous work demonstrating the feasibility of this concept, [9], the authors aim to illustrate that 'this approach can train large (millions of parameters) networks for non-trivial tasks' [2], yielding remarkable results.

To demonstrate the potential benefits of this vision, the authors apply their approach to the following tasks:

- Complex pattern memorization (including natural images).
- One-shot classification (on the Omniglot dataset).
- Reinforcement Learning (in a maze exploration problem).

Evidence that the implications of differentiable plasticity extend beyond the realm of reinforcement learning, showcasing its potential in diverse fields.

2.2. Multi-task - Paper [3]

In contrast to the approach discussed in Chapter 2.1, Paper [3] introduces a different perspective by focusing on the technique of multi-tasking to enable flexible adaptation between domains. The objective of multi-task reinforcement learning can be summarized in two main goals:

1. 'Efficiently learn by training against multiple tasks' [3].
2. 'Quickly adapt, using limited samples, to a variety of new tasks' [3].

In the context of this work, 'the tasks correspond to reward functions for environments with the same (or similar) dynamical models' [3].

In addition to the fundamental concepts, the authors also consider the Model Agnostic Meta-Learning (MAML) algorithm. This algorithm has demonstrated success in both supervised and reinforcement learning settings for multi-task learning (Finn et al., 2017). MAML 'learns a shared policy initialization across tasks' [3], which is then adapted at test time when encountering new tasks through policy gradient updates.

However, this approach has two main constraints when it comes to policy transfer

1. Firstly, 'assumimg the existence of a policy initialization from which many task-specific policies may be found via a few local updates' [3], may fail when policies differ significantly across tasks.
2. Secondly, it 'affects the time and sample efficiency of the adaptation phase' [3], even with exhaustive training.

To address these points, the present work intends 'to address these limitations by revisiting the model-based approach to multi-task reinforcement learning' [3]. Therefore, the authors propose two key modifications:

1. 'Learn and adapt with a shared dynamical model, rather than a policy' [3].
2. 'A "warm-up" phase of adaptation' [3] during which a policy is trained on the learned dynamical model.

However, these modifications are based on the assumption that the tasks being tackled have similar associated dynamical models, allowing 'the tasks, and the policies required to solve them, to vary arbitrarily' [3]. Thus, while training a separate policy for each task may result in a computationally expensive adaptation phase, it offers the advantage of requiring few, if any, additional samples. The authors summarize this idea by stating:

> ‘*Consider the extreme case in which we acquire a perfect dynamical model of the environment during training: adapting a policy to a new task may be computationally challenging, but will require no new samples*’ [3].

With this scientific perspective, the authors aim to make three primary contributions:

1. Introduce a new model-based multi-task algorithm in reinforcement learning.
2. Conduct numerical comparisons with the MAML algorithm ‘on several continuous control MuJoCo reinforcement learning benchmarks of varying difficulty’ [3].
3. Provide numerical evidence showcasing the algorithm's ability to handle out-of-distribution tasks, shifts in dynamical models, and active task selection.

3. Technical Considerations

Building upon the theoretical considerations discussed in Chapter 2 for both papers, a set of technical considerations have been formulated. This chapter is dedicated to summarizing the key technical considerations implemented by the authors, divided into two subtopics: 3.1 for paper [2] and 3.2 for paper [3]. It is important to note that while this chapter provides a summary, for a detailed understanding of the engineering processes, referring to the original papers is essential.

3.1. Differentiable Plasticity - Paper [2]

In order ‘to train plastic networks with backpropagation, a plasticity rule must be specified’ [2]. The formulation of this rule can vary depending on the specific work being developed. In the case of [2], the authors adopt a flexible formulation that separates the components distinguishing the plastic and non-plastic baselines in each connection. This design choice enables easy implementation of multiple Hebbian rules within the framework.

Drawing from the foundational knowledge of neural networks[1], the non-plastic component corresponds to the traditional connection weight, as on the most common neural network models. However, what sets apart the non-plastic neural networks from the approach presented in [2] is the addition of the Hebbian trace, which is a plastic component that ‘varies during a lifetime[2] according to ongoing inputs and outputs’ [2]. Moreover, the ‘relative importance of plastic and fixed components in the connection is structurally determined by the plasticity coefficient’ [2]. Consequently, depending on the weight and the plasticity coefficient, a connection can assume three states: fully fixed (or non-plastic), fully plastic (no fixed component), or a balance between both components.

In each lifetime, the Hebbian trace is initialized to 0, while the weights and the plastic coefficients ‘are [...] structural parameters of the network that are conserved across lifetimes, and optimized by gradient descent between lifetimes (descending the gradient of the error computed during episodes), to maximize expected performance over a lifetime/episode’ [2]. As a last note, the learning rate is an optimized parameter on the network, that assumes the same value for all the connections.

3.2. Multi-task - Paper [3]

The main areas addressed by paper [3] regard the ‘sample complexity in the training time and adaptation time, and [...] zero-shot adaptation to similar tasks’ [3]. Thus, in order to achieve this, the following points are considered:

1. ‘Transfer the parameters of the learned dynamical model, rather than the policy’ [3].

[1] It is assumed that the reader holds the fundamental knowledge of neural networks.
[2] The authors claim that the terms “lifetime” and “episode” are used interchangeably.

2. '"Warm-up" a task's policy by training on learned "virtual" dynamics prior to interaction' [3].

Based on those assumptions, the sequential multi-task training overview is divided into three main topics:

1. Task Sampling: Sampling tasks from the distribution.
2. Policy Warm-Up: Initializing 'a random policy and warm it up on the new task by training on learned dynamical models' [3], except for the first task, where the warm-up is skipped. This process is referred as VirtualTraining.
3. Data Collection: 'Alternately collect new data, fit a dynamics model and then perform VirtualTraining for n iterations' [3].

Therefore, this training steps integrate a set of key design choices, such as:

- In steps 2. and 3., in order to prevent 'the policy from over-fitting to a particular dynamical model' [3], the policy is trained against several learned dynamical models, with an intentional over-parameterization of the neural network, which is an improved approach empirically proven in [11].
- The policy (step 2.) is warmed-up on the trained model, allowing adaptation without any new samples, and, ultimately, on the test phase, the ' adaptation will be exactly this warm-up phase' [3].
- The training process is sequential, indicating that the task is selected, the trained is performed and the data is collected, before passing into the next task.

Furthermore, during the adaptation or test phase, when a new task is presented, the learned model is utilized to train a policy from the ground up. This involves randomly initializing the policy parameters, performing a warm-up phase, and subsequently continuously adapting the policy during training.

Moreover, the paper also delves into the subject of active task selection. This feature allows 'to select a particularly difficult or diverse sequence of tasks, which may be desirable if it speeds training' [3], according to a defined function. The selection of tasks is guided by a defined function, which aids in determining the optimal sequence to enhance learning efficiency.

4. Experiments

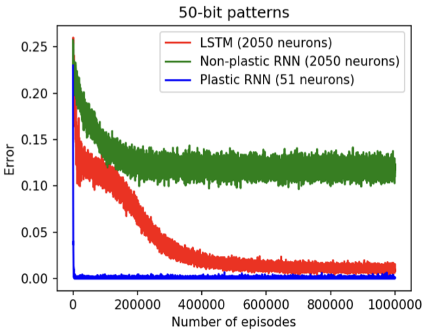
Building upon the theoretical and technical considerations discussed earlier, the authors proceeded to conduct a series of experiments, yielding significant scientific contributions. In this chapter, we aim to provide a summary of the key results obtained in each paper, namely Chapter 4.1 for paper [2] and Chapter 4.2 for paper [3]. These experiments shed light on the effectiveness and potential of the proposed approaches, showcasing their impact and offering valuable insights into their performance.

4.1. Differentiable Plasticity - Paper [2]

Pattern memorization: Binary patterns

The first experiment conducted regards 'memorizing sets of arbitrary high-dimensional patterns (including novel patterns never seen during training), and reconstructing these patterns when exposed to partial, degraded versions of them' [2].

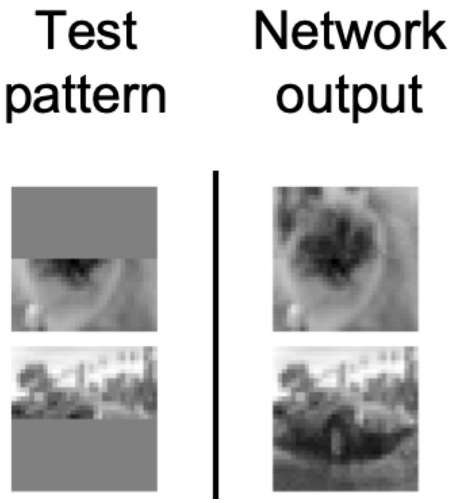
Through a comparative analysis involving an LSTM, a non-plastic RNN, and a plastic RNN, the authors effectively demonstrate the superiority of the plastic RNN. Notably, the plastic RNN achieves remarkably low error rates (below 0.01) coupled with an exceptionally rapid convergence rate, outperforming the LSTM by a factor of 250. Furthermore, it is worth noting that the plastic RNN achieves these results with a less complex network architecture. Consequently, in the context of the specific task examined, the authors conclude that 'plastic recurrent networks seem considerably more powerful than LSTMs' [2] and non-plastic RNNs.



Pattern memorization: Natural Images

Next, the method is applied to address the challenging task of 'memorizing natural images with graded pixel values, which contain much more information per element' [2] compared to the previous task. The CIFAR-10 dataset is utilized, providing a diverse range of images for evaluation.

Through rigorous testing, the model demonstrates its remarkable capability to successfully tackle the 'non-trivial task of memorizing and reconstructing previously unseen natural images' [2]. This achievement can be attributed to the enhanced flexibility introduced by the non-plastic term. Furthermore, the authors highlighted that 'the main outcome is that independent plasticity coefficients for each connection improve performance for this task' [2].

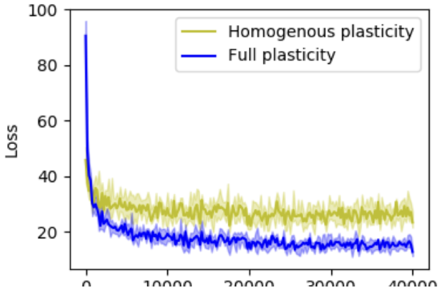


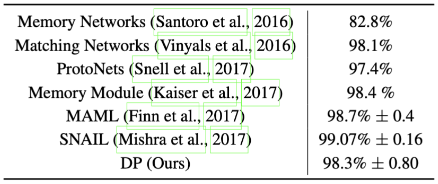
One-shot pattern classification: Omniglot task

After testing pattern memorization, the authors proceed to apply the model 'to the standard task for one-shot and few-shot learning, namely, the Omniglot task' [2] proposed in [12].

The authors' first significant finding is that employing independent plasticity coefficients for each connection leads to improved performance of the model.

Moreover, by conducting a comparative analysis against various state-of-the-art models encompassing diverse approaches, the developed model consistently delivers commendable results. Notably, the authors demonstrated 'that [...] adding a few plastic connections in the output of the network [...] allows for competitive one-shot learning over arbitrary man-made visual symbols' [2].

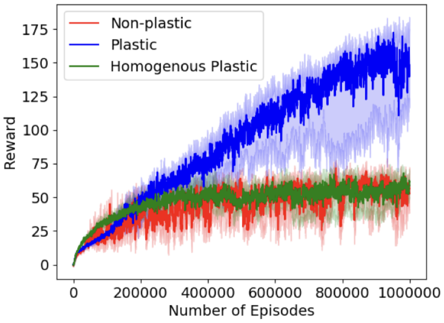




Memory Networks (Santoro et al., 2016)	82.8%
Matching Networks (Vinyals et al., 2016)	98.1%
ProtoNets (Snell et al., 2017)	97.4%
Memory Module (Kaiser et al., 2017)	98.4 %
MAML (Finn et al., 2017)	98.7% ± 0.4
SNAIL (Mishra et al., 2017)	99.07% ± 0.16
DP (Ours)	98.3% ± 0.80

Reinforcement learning: Maze exploration task

In their final experiment, the authors apply the developed model to a maze exploration task to assess the potential of differentiable plasticity in enhancing the learning capabilities of the network in such tasks.



By running 'the experiments under three conditions: full differentiable plasticity, no plasticity at all, and homogenous plasticity in which all connections share the same (learnable) [...] parameter' [2], the authors prove that the differentiable plasticity technique improves the performance significantly. Moreover, they emphasize that in this particular type of tasks, 'RNNs get "stuck" on a sub-optimal strategy' [2] and 'individually sculpting the plasticity of each connection is crucial in reaping the benefits of plasticity for this task' [2].

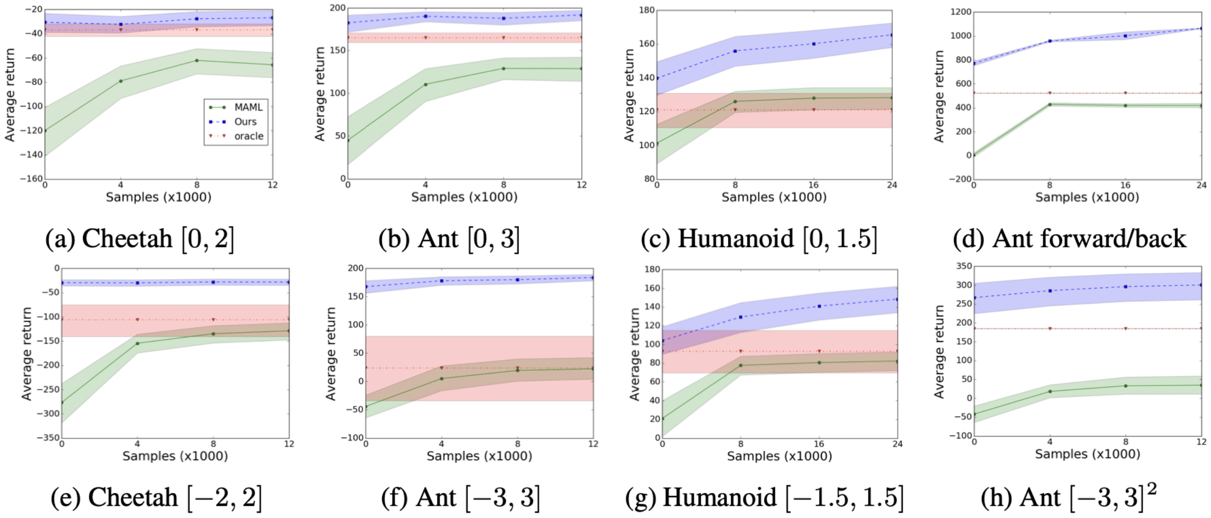
4.2. Multi-task - Paper [3]

In this paper, the evaluation is conducted on several continuous control tasks utilizing environments from the rllab benchmark (Duan et al., 2016), which relies on the MuJoCo physics simulator (Todorov et al., 2012).

Comparison to MAML:

To assess the performance of the proposed algorithm, a comparison is made with MAML and an oracle policy that 'is trained jointly across the task distribution' [3]. The evaluation is based on two metrics: the average return and the number of samples. The comparison results are illustrated below.

Figure 4.2.1 - Developed model vs MAML vs oracle policy [3].



(a) Cheetah $[0, 2]$ (b) Ant $[0, 3]$ (c) Humanoid $[0, 1.5]$ (d) Ant forward/back

(e) Cheetah $[-2, 2]$ (f) Ant $[-3, 3]$ (g) Humanoid $[-1.5, 1.5]$ (h) Ant $[-3, 3]^2$

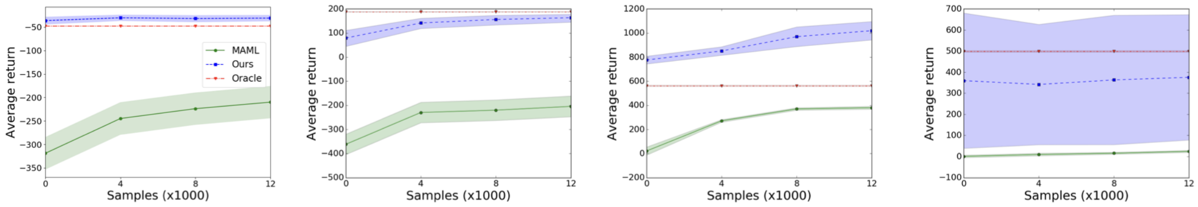
The results clearly demonstrate that the developed algorithm surpasses MAML in performance across all tasks, even with a reduced number of samples. As previously mentioned, the limitation of the MAML algorithm is that it cannot guarantee the existence of an initialization that enables near-optimal policies for all tasks within a task family with just a few gradient steps. In contrast, the proposed approach achieves significantly superior outcomes.

Additionally, it is worth noting that the proposed model exhibits the ability to outperform the oracle policy even without any samples. This is attributed to the fact that 'the policy produced by the warm-up stage (before any samples are collected from the test environment) may already be high-performing' [3].

Task Distribution and Dynamical Model Shift:

The model is further assessed in a scenario where 'the distribution of tasks at test time differs from that at train time' [3], commonly known as domain adaptation in supervised learning (Ben-David et al., 2010).

Figure 4.2.2 - Developed model vs MAML vs oracle policy in domain adaptation [3].



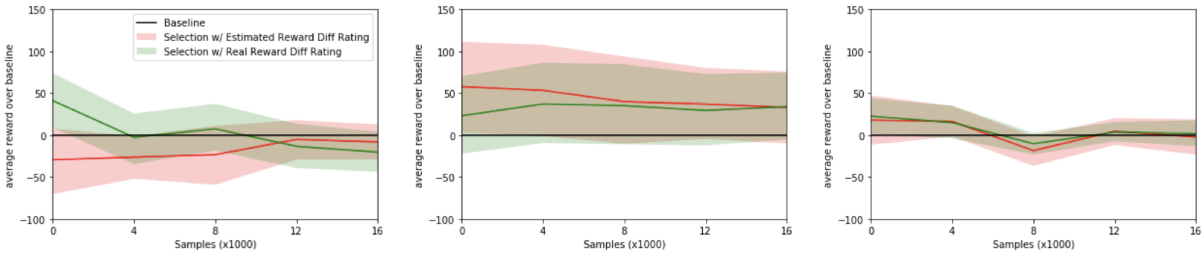
(a) Cheetah Pos-to-Neg (b) Ant Pos-to-Neg (c) Ant low friction (d) Ant crippled

The obtained results in (a) and (b) demonstrate that the developed model maintains its capability 'to achieve some gains over MAML as a result of the warm-up phase' [3], even in scenarios with changing reward distributions. However, the effectiveness of the warm-up phase decreases (figure (c)) when there are changes in the underlying dynamical model, particularly when the ant is crippled (figure (d)).

Active Task Selection:

Lastly, the authors conducted a test regarding the active task selection mentioned earlier. The results of this test are as follows:

Figure 4.2.3 - Active Task Sampling [3].



(a) 5 tasks sampled (b) 20 tasks sampled (c) 30 tasks sampled

While some potential benefits of active sampling can be observed in figures (a) and (b), these results are preliminary and not sufficient to fully explore the advantages of active sampling. Further investigation is required to thoroughly understand the benefits it may offer. These initial findings provide a glimpse into the potential advantages, but the authors acknowledge that a more detailed study is necessary. They leave the exploration of active sampling for future research endeavors.

5. Key Limitations

Like any scientific work, the approaches presented in this study have their limitations. This chapter aims to summarize the key limitations that should be considered when interpreting the findings.

5.1. Differentiable Plasticity - Paper [2]

In summary, the main limitation are:

- At the moment, there is a limited 'understanding of the implications of this new tool' [2].
- Due to the insertion of more parameter, more fine-tuning is needed.

5.2. Multi-task - Paper [3]

On paper [3], the set of key limitations are:

- The approach assumes 'similar dynamical models across tasks' [3].
- Virtual training ' is compute-intensive and the warm-up phase proves the longest part [...]' [3].

6. Conclusions & Future Directions

The two approaches presented in this paper represent significant contributions to the field of transfer learning in reinforcement learning, each with its unique development premises.

The incorporation of optimized plasticity as a key component in learning, as explored in the first approach, showcases the natural potential of this neurologically inspired feature. The promising results obtained demonstrate the effectiveness of considering plasticity in the learning process. Furthermore, the broader implications of this work extend beyond reinforcement learning, as the concept of plasticity holds great potential for various domains and applications of neural networks. The possibilities for future research in this area are extensive, given the wide range of fields that can benefit from the integration of plasticity into learning algorithms.

In the case of the novel multi-tasking approach, the demonstrated benefits in specific tasks highlight the potential of this method when applied appropriately. While the current tasks used in the evaluation were relatively simple, the authors suggest future work involving more challenging tasks to further explore the capabilities of the model. Additionally, the exploration of active task selection and the analysis of online settings and adversarial tasks present exciting avenues for future research.

In summary, these works contribute to the growing body of evidence that diverse solutions can be found to address the challenges of transfer learning. They provide valuable insights and pave the way for further advancements in the field. The future of transfer learning holds immense possibilities, and these studies have laid a solid foundation for ongoing progress and exploration.

7. References

[1] ZHU, Z.; LIN, K.; *et al.* (2022). *Transfer Learning in Deep Reinforcement Learning: A Survey*. USA: Cornell University.

[2] MOCONI, T.; CLUNE, J.; STANLEY, K. (2018). *Differentiable plasticity: training plastic neural networks with backpropagation*. USA: Cornell University.

[3] LANDOLFI, N.; THOMAS, G.; MA, T. (2019). *A Model-based Approach for Sample-efficient Multi-task Reinforcement Learning*. USA: Stanford University.

[4] SUTTON, R.; BARTO, A. (2018). *Reinforcement learning: An introduction*. USA: MIT press.

[5] KRISHEVSKY, A; SUTSKEVER, I; HINTON, G (2012). *ImageNet Classification with Deep Convolutional Neural Networks*. Canada: University of Toronto.

[6] KAVUKCUOGLU, M.; SILVER, K.; *et al.* (2015). *Human-level control through deep reinforcement learning*. USA: Springer Nature.

[7] SILVER, D.; HUANG, A.; *et al.* (2016). *Mastering the game of go with deep neural networks and tree search*. USA: Springer Nature.

[8] BROWN, R.; BLIGH, T.; GARDEN, J. (2021). *The Hebb Synapse Before Hebb: Theories of Synaptic Function in Learning and Memory Before Hebb (1949), With a Discussion of the Long-Lost Synaptic Theory of William McDougall*. Canada: Dalhousie University.

[9] MOCONI, T. (2016). *Backpropagation of hebbian plasticity for continual learning*. USA: NIPS.

[10] FINN; C.; ABBEEL, P.; LEVINE, S. (2017). *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks*. USA: Cornell University.

[11] LUO, Y; XU, H, *et al.* (2019). *Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees*. USA: Cornell University.

[12] LAKE, B.; SALAKHUTDINOV, R.; TENENBAUM, J. (2015). *Human-level concept learning through probabilistic program induction*. USA: Science.

[13] DUAN, Y; CHEN, X.; *et al.* (2016). RL: *Fast Reinforcement Learning via Slow Reinforcement Learning*. USA: Cornell University.

[14] TODOROV, E; EREZ, T; TASSA, Y. (2012). *Mujoco: A physics engine for model-based control*. USA: IEEE/RSJ.

[15] BEN-DAVID; S.; PEREIRA, F.; *et al.* (2010). *A theory of learning from different domains*. USA: Springer.