A Maximum Entropy Deep Reinforcement Learning Neural Tracker

Shafa Balaram, Kai Arulkumaran, Tianhong Dai, and Anil Anthony Bharath

Imperial College London, London SW7 2AZ, UK
{shafa.balaram15,kailash.arulkumaran13,tianhong.dai15,a.bharath}
@imperial.ac.uk

Abstract. Tracking of anatomical structures has multiple applications in the field of biomedical imaging, including screening, diagnosing and monitoring the evolution of pathologies. Semi-automated tracking of elongated structures has been previously formulated as a task for deep reinforcement learning (DRL), albeit it remains a challenge. We introduce a maximum entropy continuous-action DRL neural tracker capable of training from scratch in a complex environment in the presence of high noise levels, Gaussian blurring and cell detractors. The trained model is evaluated on mouse cortical two-photon microscopy images. At the expense of slightly worse robustness compared to a previously applied DRL tracker, we reach significantly higher accuracy, approaching the performance of the standard hand-engineered algorithm used for neuron tracing. The higher sample efficiency of our maximum entropy DRL tracker indicates its potential of being applied directly to small biomedical datasets in the absence of artificial models.

Keywords: Tracking · Tracing · Neuron · Axon · Reinforcement learning · Maximum entropy.

1 Introduction

In the field of image analysis, tracking can be defined as the process of locating an object through time and/or space [17], introducing an ordering in the observation points. While segmentation can be a precursor, tracking is used to provide additional information about structures. Additionally, tracking is often used to address partial loss of visibility of the structure caused by apparent gaps, low contrast or occlusion, a common problem in video analysis [20], through inference. In addition, morphological attributes of spatial structures, such as branching patterns [5], can be obtained about biological structures, which can aid the detection and treatment of ophthalmologic and cardiovascular pathologies [10].

Hand-engineered trackers have been applied to obtain measurements about thin, elongated structures in biomedical imaging [5,17]. To address the dependence of trackers on specific biomedical datasets, Dai *et al.* [4] extended prior work on the application of deep reinforcement learning (DRL) to tracking in biomedical images [23], performing subpixel neural tracking while coping with a limited number of labelled images [3] through the use of a synthetic dataset, eventually performing zero-shot transfer learning on axonal images from two-photon imaging of mouse cortex.

In order to improve the sample efficiency, and hence applicability, of the DRL neural tracker to other biomedical datasets, we opt for the off-policy soft actor-critic (SAC) algorithm [8]. SAC also uses a maximum entropy formulation of reinforcement learning (RL), which leads to better exploration. In comparison to Dai *et al.* [4], our tracker could be trained and validated on artificially-simulated images modelled with a higher degree of complexity and detractors, including background structure mimicking cell debris and high noise levels.

We show that not only does SAC benefit from vastly improved sample efficiency, but it also achieves far greater accuracy than Dai *et al.*'s tracker—approaching that of the standard hand-engineered algorithm used [14]—with only a slight drop in robustness. Furthermore, we show that the ability to train on more complex synthetic environments increases the tracker's generalisation to real data. Together, this makes our approach more appealing for application to other biomedical image datasets.

2 Background

2.1 Tracking

Segmentation was used in previous two-photon axon image analysis by Li *et al.* [11] as a prior step to improve their neuron tracing algorithm. However, by combining both techniques, the performance of the tracking algorithm becomes dependent on the accuracy of the segmentation process. Instead, our semi-automatic tracking algorithm employs a local exploration strategy, where a seed starting point of each neuron is specified explicitly. This is similar to the Vaa3D algorithm [14], an ImageJ plugin which is currently the standard for neuron tracing. Through the availability of seed points, the algorithm can adapt locally to changing image quality and contrast conditions, which becomes important in neuron images with non-uniform backgrounds, such as varying noise levels, inhomogeneous microscopic blurring and presence of cell debris [19].

2.2 Maximum Entropy Reinforcement Learning

RL is a branch of machine learning which provides a mathematical framework for an agent to learn independently by interacting with its environment with the aim of maximising its return (sum of rewards) [21]. In conventional RL, the environment produces a state $\mathbf{s}_t$ at every timestep t after which the agent samples an action $\mathbf{a}_t$ from a policy π, a probability distribution which maps states to actions. Consequently, the agent receives a successor state $\mathbf{s}_{t+1}$ from the environment together with a scalar reward r_{t+1} as feedback to the decision taken. This closed loop mechanism ends when a terminal state is reached.

In our scenario, a trained agent should be able to trace the centreline of neurons in the presence of varying imaging conditions and detractors by learning an optimal sequence of decisions related to displacements in a 2D Cartesian coordinate system. Previously, Dai *et al.* [4] used the on-policy proximal policy optimisation (PPO) algorithm [18], which utilises an actor-critic (policy π and state value function $V(\mathbf{s}_t)$) setup [21], where both the policy and state value function are parameterised by neural networks. In contrast, we use the off-policy maximum entropy SAC algorithm [8].

SAC is also an actor-critic algorithm, but utilises two (soft) state-action value functions $Q(\mathbf{s}_t, \mathbf{a}_t)$ in place of the single state value function [9]. Unlike on-policy algorithms, off-policy algorithms can learn from past trajectories, improving their sample efficiency over on-policy algorithms. Then, the RL formulation used by SAC has been extended with a maximum entropy term to improve exploration and robustness [24,7]. The objective is to learn an optimal stochastic policy π^* which maximises both the expected discounted return and its expected entropy [8]. Finally, SAC uses a temperature parameter which determines the relative influence of the reward and entropy terms on the policy and thus balances the exploration-exploitation trade-off [8]. The most significant change we made to apply this to medical images was to use a different neural network architecture that makes use of privileged information during training [15].

2.3 Deep Reinforcement Learning in Biomedical Imaging

Deep neural networks have been used successfully as function approximators of the policy and the state-action value function in multiple applications, including real-world visual navigation tasks where RL agents can learn directly from raw pixel values [2]. Their applications in biomedical imaging include landmark detection [6], view planning [1] and vascular centreline tracing [23], which all make use of deep Q-network algorithm [12], constraining them to using discrete action spaces. In order to predict the centreline observation points to subpixel accuracy, a continuous action space is required. This issue was addressed by Dai *et al.* [4] through the use of PPO.

Biomedical datasets manually-labelled by experts can be both limited and expensive to acquire. With a small dataset of 20 annotated microscopy images available [3], Dai *et al.* [4] simulated synthetic images of single neurons based on two-dimensional splines for training and tuning of hyperparameters during validation. Since they had access to the ground truth locations of synthetic centrelines during training, they used an asymmetric actor-critic architecture which improves value function learning [15]. This is achieved by providing the critic (value) network with extra information only during training, i.e., the binary maps containing the neuron centrelines. The trained tracker was then tested directly on microscopy data—which can be considered transductive or “zero-shot” transfer [13]. Similarly, we also employ the asymmetric actor-critic architecture, whereby the two soft-Q functions (critics) of SAC are each provided with the binary ground truth maps.

3 Entropy-based Deep Reinforcement Learning for Tracking

We now define the key entities of RL in the context of tracking neural centrelines.
Environment: Owing to the availability of only 20 greyscale expert-annotated axonal mouse cortex images [3], we simulate artificial images of single neurons "on the fly" in a pseudo-random manner with controlled degrees of complexity during training. We introduce a few modifications to the images generated by Dai *et al.* [4] in terms of the cosinusoidal axon intensity profiles and presence of detractors with the aim of imitating the complex natural environment of neurons (refer to Subsection 4.1 for further details). Figure 1 shows examples of synthetic (both ours and Dai *et al.*'s [4]) and microscopy images for visual comparison; matching the rough structure of the real data appears to be enough for some generalisation to real data. The terminal state definition is described in the supplementary material.
State Space: Our state space is the same as that of Dai *et al.* [4]. We refer the reader to the supplementary material for further details.
Action Space: Actions in a continuous control space are sampled by the agent directly from its stochastic policy. We parameterise the policy as a Gaussian "squashed" by a tanh function; the actions are then floating point numbers $\in (-1, 1)$, and represent subpixel displacements in the image's coordinate system without requiring any further processing.
Reward Function: The reward function has to be defined in such a way so as to achieve the aim of tracing the neural centreline with subpixel accuracy. We use a simplified version of the original reward function formulated by Dai *et al.* [4] and include a full description in the supplementary material.
Agent: The agent's policy and two soft Q-functions are modelled using convolutional neural networks, as shown in Figure 2 of the supplementary material along with their input states. The actor network outputs the parameters of two independent Gaussian distributions, namely the means $\boldsymbol{\mu}$ and logarithm of the standard deviations $\log \boldsymbol{\sigma}$. We constrain $\log \boldsymbol{\sigma} \in [-20, 2]$ to prevent highly deterministic or stochastic policies. The support for the distributions is bounded with a tanh squashing function and the sampled actions represent displacements in the $x - y$ coordinate system of the images. The two critic networks each output a scalar soft Q-function, $Q(\mathbf{s}_t, \mathbf{a}_t)$. We provide more explanation behind the choice of this SAC variant [9] as well as the final training algorithm for subpixel neural tracking in the supplementary material.

4 Experiments

4.1 Datasets & Performance Evaluation

There are two different datasets used throughout our experiments: the synthetic and microscopy images of neurons (see Figure 1 for examples). For reasons discussed in Section 2.1, the starting points of the neurons of each dataset are provided to all trackers as seed points.

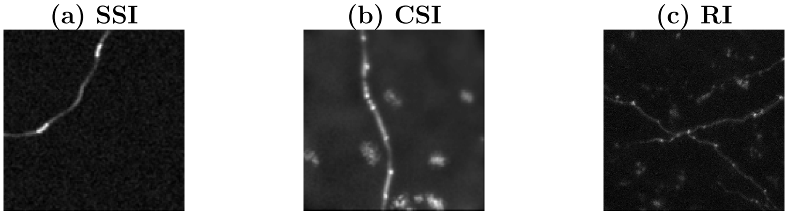

Fig. 1. Comparison of synthetic and microscopy datasets: (a) a simple artificially-simulated image (SSI) used during training by Dai *et al.* [4] with relatively low levels of background noise, no blurring and no cells, (b) a complex synthetic image (CSI) used in our training, generated with Gaussian blurring, "cell debris" and higher levels of background noise, and (c) a real image (RI) obtained from the somatosensory cortex of a mouse using two-photon microscopy [3].

Synthetic Dataset We first train and validate our tracker on synthetic single-neuron images. The ability to train the agent in the synthetic dataset is also an implicit part of performance evaluation. We increased the complexity of images simulated by Dai *et al.* [4] in several ways. We choose cosine intensity profiles for axons to imitate regions in different stages of synaptic transmission. Highly-illuminated blob-like structures are also present to simulate synaptic boutons. We add background image structure mimicking cell debris as well as Gaussian and Poisson noise as detractors from the centreline to be tracked. Finally, a common artefact in real datasets is an out-of-focus microscope, which we try to capture using the Gaussian blurring operation.

Microscopy Dataset We evaluate the performance of our best performing maximum-entropy DRL trained tracker on a mouse cortical axon dataset [3][1]. There are 20 greyscale images, maximum-intensity projected from 3D stacks, with their corresponding binary ground truth images annotated by an expert.

Metrics In order to quantify and compare the performance of our tracker, two measures are used: the root mean squared error (RMSE) and coverage [4]. The RMSE quantifies the perpendicular error between the predicted and target centrelines and thus, represents the accuracy of the tracker. To measure the robustness of the tracker, we utilise the coverage, which is the proportion of the neuron tracked by predicted points that lie within a margin of 3 pixels of the target centreline.

4.2 Training & Validation

We tune the original SAC model's hyperparameters on a held-out synthetic dataset of 20 images for a 10:1 training/validation split. During the validation

[1] Dataset available at: https://www.zenodo.org/record/1182487#.XP2UBS2ZMxc.

process, we investigated whether the absence of detractors, such as cells and boutons, in the environment has an impact on the agent's transfer learning ability. In addition, we looked into whether adding a local response normalisation layer [16] after each ReLU activation function of the feature learning stage could tackle the lower contrast of the microscopy dataset and improve generalisation to the real data. Finally, for each variant mentioned, we considered automatic entropy tuning [9]. We observed that models trained in the absence of detractors (simulated boutons or cell debris) performed worse on the microscopy dataset (details in the supplementary material), indicating the importance of being able to train DRL agents on more challenging image data. In contrast, we were unable to train a PPO tracker [4] on this more complex synthetic image data. Our best performing model is the original SAC trained for 3.5×10^5 timesteps using a fixed temperature parameter α of 1.0 in the presence of all detractors. Its training requires approximately 2000 synthetic images and takes around 5 hours on K80 GPUs. We refer the reader to the supplementary material for the variance across 5 random seeds during training. Our code is available at https://bitbucket.org/bicv/maximum-entropy-drl-tracker.

4.3 Testing

Our best SAC model was tested on the microscopy dataset, without further tuning of parameters or hyperparameters. As in Dai *et al.* [4], we take into consideration the higher resolution of the microscopy images by extracting larger windows of sizes 15×15 pixels and 31×31 pixels before downscaling all views to 11×11 pixels. We increase the maximum episode length from 200 to 350 to account for the larger dimensions of the microscopy images. The starting point of each axon is provided to the agent separately in multi-axon images.

Table 1 compares the performance of our best-performing tracker against Dai *et al.*'s PPO tracker [4] and the Vaa3D algorithm [14]. The RMSE and coverage values of the individual microscopy images are shown in the supplementary material. Despite the slightly lower coverage of our maximum entropy tracker in comparison to PPO, it achieves much higher accuracy, approaching that of the Vaa3D algorithm.

Table 1. Test performance of 20 microscopy images: mean and ±1 standard deviation of the root mean squared error (RMSE) in pixels and coverage of the SAC and PPO DRL trackers, and the Vaa3D algorithm. Note that although the maximum-entropy DRL tracker is less robust than the PPO DRL tracker, its accuracy approaches that of the Vaa3D algorithm.

SAC		PPO		Vaa3D	
RMSE	Coverage	RMSE	Coverage	RMSE	Coverage
4.36 ± 3.63	0.808 ± 0.242	27.62 ± 27.96	0.841 ± 0.130	1.75 ± 1.73	0.923 ± 0.089

5 Conclusion

We proposed a maximum entropy DRL tracker trained in a complex environment simulated to mimic an axon microscopy dataset. Our training algorithm combines the state-of-the-art SAC algorithm [9] with the asymmetric actor-critic architecture [15]. Our improvements on prior work [4] include the requirement for only a small training dataset owing to the high sample efficiency of the chosen off-policy training algorithm. We also demonstrate the ability to track neural centrelines to subpixel accuracy in the presence of background image structure mimicking cell debris and higher noise levels. Finally, while our maximum entropy DRL tracker is less robust with its slightly lower coverage value, it has an accuracy 6-fold higher than the PPO tracker, with its RMSE approaching that of the standard algorithm for neuron tracing [14].

The maximum entropy DRL tracker can be combined with active contour methods to track boundaries of other structures, such as walls of blood vessels, after redefining the reward function based on the nature of the structure of interest. Furthermore, the need for a smaller labelled dataset of only 2000 images increases the likelihood of training the tracker directly on biomedical datasets in cases where artificial models cannot be built easily. Future work could include training in synthetic multi-axon images and accounting for branching of neurons, potentially by combining the network with an automatic junction detection algorithm [22], or extending to subvoxel tracking through extending the environment to 3D, and introducing highly anisotropic spatial sampling into the environment—a common challenge in confocal imaging.

References

1. Alansary, A., Le Folgoc, L., Vaillant, G., Oktay, O., Li, Y., Bai, W., Passerat-Palmbach, J., Guerrero, R., Kamnitsas, K., Hou, B., et al.: Automatic view planning with multi-scale deep reinforcement learning agents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 277–285. Springer (2018)
2. Arulkumaran, K., Deisenroth, M.P., Brundage, M., Bharath, A.A.: Deep Reinforcement Learning: A Brief Survey. IEEE Signal Processing Magazine **34**(6), 26–38 (2017)
3. Bass, C., Helkkula, P., De Paola, V., Clopath, C., Bharath, A.A.: Detection of axonal synapses in 3D two-photon images. PloS one **12**(9), 1–18 (2017)
4. Dai, T., Dubois, M., Arulkumaran, K., Campbell, J., Bass, C., Billot, B., Uslu, F., de Paola, V., Clopath, C., Bharath, A.A.: Deep reinforcement learning for subpixel neural tracking. In: Proceedings of the International Conference on Medical Imaging with Deep Learning. pp. 130–150 (2019)
5. Fraz, M.M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A.R., Owen, C.G., Barman, S.A.: Blood vessel segmentation methodologies in retinal images–a survey. Computer methods and programs in biomedicine **108**(1), 407–433 (2012)
6. Ghesu, F.C., Georgescu, B., Zheng, Y., Grbic, S., Maier, A., Hornegger, J., Comaniciu, D.: Multi-scale deep reinforcement learning for real-time 3d-landmark detection in CT scans. IEEE trans. on pattern analysis and machine intelligence **41**(1), 176–189 (2017)

7. Haarnoja, T., Tang, H., Abbeel, P., Levine, S.: Reinforcement learning with deep energy-based policies. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1352–1361. JMLR.org (2017)
8. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. arXiv preprint arXiv:1801.01290 (2018)
9. Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al.: Soft actor-critic algorithms and applications. arXiv preprint arXiv:1812.05905 (2018)
10. Kanski, J.J., Bowling, B.: Clinical ophthalmology: a systematic approach. Elsevier Health Sciences (2011)
11. Li, R., Zeng, T., Peng, H., Ji, S.: Deep learning segmentation of optical microscopy images improves 3-d neuron reconstruction. IEEE trans. on medical imaging **36**(7), 1533–1541 (2017)
12. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529 (2015)
13. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE trans. on knowledge and data engineering **22**(10), 1345–1359 (2009)
14. Peng, H., Ruan, Z., Long, F., Simpson, J.H., Myers, E.W.: V3d enables real-time 3d visualization and quantitative analysis of large-scale biological image data sets. Nature biotechnology **28**(4), 348 (2010)
15. Pinto, L., Andrychowicz, M., Welinder, P., Zaremba, W., Abbeel, P.: Asymmetric actor critic for image-based robot learning. arXiv preprint arXiv:1710.06542 (2017)
16. Pinto, N., Cox, D.D., DiCarlo, J.J.: Why is real-world visual object recognition hard? PLoS computational biology **4**(1), e27 (2008)
17. Poulin, P., Cote, M.A., Houde, J.C., Petit, L., Neher, P.F., Maier-Hein, K.H., Larochelle, H., Descoteaux, M.: Learn to track: Deep learning for tractography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 540–547. Springer (2017)
18. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
19. Skibbe, H., Reisert, M., Nakae, K., Watakabe, A., Hata, J., Mizukami, H., Okano, H., Yamamori, T., Ishii, S.: Pat—probabilistic axon tracking for densely labeled neurons in large 3-d micrographs. IEEE trans. on medical imaging **38**(1), 69–78 (2018)
20. Smeulders, A.W., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: An experimental survey. IEEE trans. on pattern analysis and machine intelligence **36**(7), 1442–1468 (2013)
21. Sutton, R.S., Barto, A.G.: Reinforcement learning : an introduction. MIT Press Ltd, Cambridge, Massachusetts (2018), https://mitpress.mit.edu/books/reinforcement-learning-second-edition
22. Uslu, F., Bharath, A.A.: A multi-task network to detect junctions in retinal vasculature. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 92–100. Springer (2018)
23. Zhang, P., Wang, F., Zheng, Y.: Deep reinforcement learning for vessel centerline tracing in multi-modality 3d volumes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 755–763. Springer (2018)
24. Ziebart, B.: Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy (2010), http://search.proquest.com/docview/845728212/