Abstract

Re-detecting landmarks in pre- and post-operative brain tumor images is a hard identification problem, due to drastic changes in the anatomical landscape caused by resection induced tissue displacement. Due to this massive change in structure, classical image registration methods oftentimes fail in the vicinity of the tumor. However, locations situated near the tumor are most interesting for further radiomic use and alignment of subsequent scans. We successfully expand the reinforcement learning method with anatomical guidance of [Wal+20] to a multi agent one like [Ala+19] has shown for a different problem. We use multi-scaling as introduced by [Ghe+17] for single agent applications and compare three different methods of choosing the best location when an agent starts to oscillate in the search process. We contribute 3D attention augmented convolutions, an expansion of [Bel+19b] 2D method to capture importance in images, and compare it with standard attention [Vas+17] for our problem. By training and testing multiple agents simultaneously we achieve a speed up of about an order of magnitude, when utilising the most effective method to choose the best location when oscillating, while slightly improving the accuracy to lower than 3mm.

Contents

1 Introduction

1.1 Landmark detection in pre- and post-operative MR brain tumor scans

The most effective treatment for aggressive human brain tumors is tumor resection, typically followed by chemo or radiation therapy [DeA01]. When evaluating the surgical outcome and deciding on further treatment, areas of tumor re-growth are compared to the corresponding areas before surgery. Tumor resection and re-growth almost always lead to shift in brain tissue in the surrounding area, which is large enough for classical image registration techniques to fail in exactly those areas of highest interest.

As [Wal+20] we aim to re-detect landmarks to aid the use of quantitative radiomic approaches [Lam+12] and to simplify correct image alignment in future scans for a given patient. In figure 1.1, a sample re-detection of two different landmarks is shown.

The goal of this work is to improve the speed and accuracy of landmark redetection in pre- and post-operative brain tumor magnetic resonance (MR) scans beyond the performance of existing state-of-the-art approaches [Wal+20].

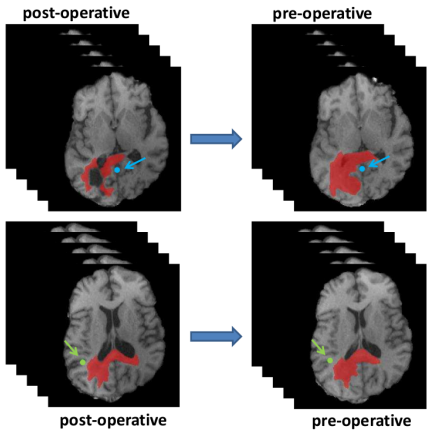

Figure 1.1: Sample redetection of two different landmarks (top, bottom) in the same patient with anatomical guidance mask (red)

1.2 Related Work

1.2.1 Landmark Detection in MR brain tumor scans using single agent deep reinforcement learning

[Wal+20] uses single agent deep reinforcement learning (RL) and an anatomical guidance mask. The anatomical guidance mask helps the agent to avoid the tumor in the pre- operative and the resection cavity in the post-operative scan.

A Dueling Deep Q-Network is used to learn the policy of the agent.

As [Wal+20] is the only paper exploring the exact same problem, we refer to only similar and other applications for the expansion of our architecture to achieve better performance.

1.2.2 Multi agent Reinforcement Learning

[Vlo+19] uses multi agent reinforcement learning (MARL) to improve on previously by [Ala+19] compared architectures for landmark detection in brains without tumors.

MARL improves speed because multiple landmarks can be detected simultaneous as shown by [Vlo+19]

It is assumed that the improved accuracy of MARL is achieved through the sharing of knowledge between agents which is hypothesised to be helpful due to the fact that anatomical landmarks are non-randomly disseminated throughout the human body [Vlo+19].

1.2.3 Attention

Attention can be considered as a tool to bias allocation of available processing resources towards the most informative parts of an input signal.

Attention has become the *de facto* standard tool for sequence-based tasks because of its ability to capture long distance interactions [BCB14; Bel+16; Bel+19a; VFJ15].

Most notably, attention was proposed by [BCB14] for alignment in Machine Translation in combination with a Recurrent Neural Network (RNN) [HS97].

[Vas+17] achieved state-of-the-art results in Machine Translation by introducing the Transformer, which uses multi headed self attention to map from (embedded-) sequence to sequence. This method has resulted in multiple recent advances in Natural Language Processing (NLP) [Dev+18; Rad+19; Bro+20].

Attention has been successfully used in combination with convolution for NLP tasks [Yan+18; Yu+18; SLL19] and Reinforcement Learning [Zam+19].

Multiple attention variants addressing the shortcomings of convolutions for image tasks have been proposed [Hu+17; Hu+18; Che+18; Woo+18; Wan+17; Zha+18; Bel+19b].

CBAM [Woo+18] independently refines convolutional features in the spatial and channel dimension. Non-local neural networks [Wan+17] use non-local residual blocks that employ self-attention in convolutional architectures. [Bel+19b] propose augmented self attention combined with convolution and show improvements for image classification on CIFAR-100 [Kri12], ImageNet [Den+09] and object detection on COCO [Lin+14].

1.3 Theoretical Background

1.3.1 Reinforcement Learning

In RL, a so-called agent is in a state of an environment where he takes an action for which he receives a reward [SB18]. The mapping of actions to states of an environment is a policy. The goal is to find the optimal policy which leads the agent from any state of the environment to the target by maximising accumulated rewards.

The agent in state $s \in S$ taking action $a \in A$ receives a reward signal $r \in R$ for each time step t. For weighting future against immediate rewards a discount rate $\lambda \in [0, 1]$ is introduced, yielding an accumulated discounted reward after τ time steps of

$$R_\tau = \sum_{\tau=0}^{\infty} \gamma^\tau r_{t+\tau+1} \tag{1.1}$$

The sequential decision process of taking actions can be viewed as a Markov Decision Process (MDP) where all a_t and s_t are conditionally independent of all previous actions and states [OW12]. Because the MDP of our problem is incomplete, the optimal policy can not be directly computed, and we thus must resort to iteratively sample states and actions and learn from experience. Due to the recent success [Ala+19; Vlo+19] of employing Q-learning in medical image analysis we subsequently limit our scope to reinforcement learning using Q-learning.

In Q-learning the action-value function $Q(s, a)$ assigns a quality score to every state-action pair. The Q-function $Q(s, a)$ which is optimised during training can be solved recursively with the Bellman optimality equation [SB98]:

$$Q(s, a) = \mathbb{E}\left[r + \gamma \max_{a'} Q(s', a')\right] \tag{1.2}$$

[Mni+15] proposed a Deep Q-Network (DQN) where a deep convolutional neural network approximates

$$Q(s, a) \approx Q(s, a; \theta), \tag{1.3}$$

achieving human-level performance in multiple Atari games. Although naive implementations suffer from divergence and instability, these problems were overcome by

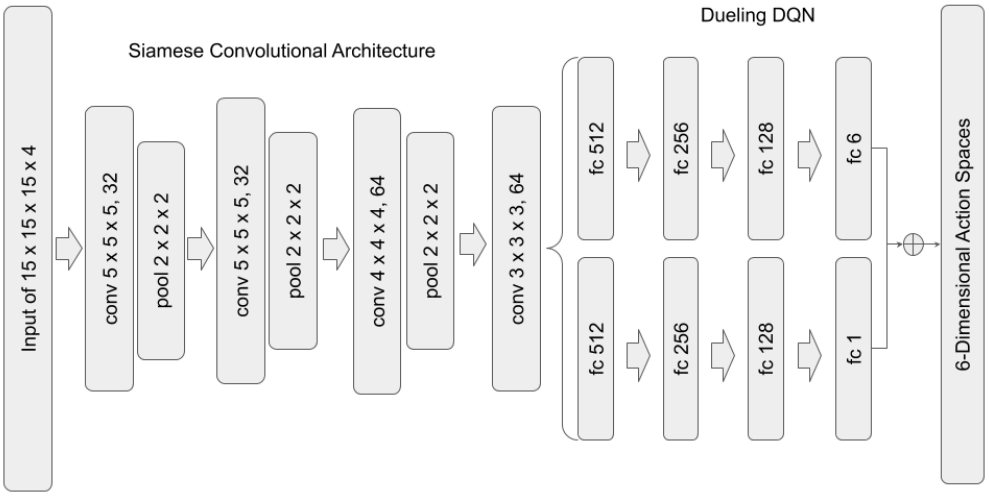

Figure 1.2: Dueling DQN architecture

only periodically updating the target network $Q(w^-)$ every n iterations, the use of experience replay [Lin92] and gradient clipping. Experience replay stores transitions $[s, a, r, s']$ in memory from which mini-batches are randomly sampled to avoid successive data sampling. Using stochastic gradient descent (SDG) on the derivate of the DQN loss

$$L_n(\theta_n) = \mathbb{E}_{s,a,r,s'}\left[(r + \gamma \max_{a'} Q(s', a'; \theta_n^-) - Q(s, a; \theta_n))^2\right] \tag{1.4}$$

an approximation of the optimal parameters w^* can be learned.

Dueling DQN [Wan+16] further improves the DQN architecture by decomposing the action-state value function $Q(s, a)$ into a state-value function $V(s)$ and action-advantage function $A(s, a)$.

$$Q(s, a) = V(s) + A(s, a) \tag{1.5}$$

The state-value function $V(s)$ aims to estimate the value a state provides independent of the possible actions the agent has in that state. The action advantage function $A(s, a)$ aims to approximate the value of taking any of the possible actions given that the agent is in state $s \in S$. As visible in Figure 1.2, the fully connected layers of the DQN architecture are split to independently compute the advantage function and state-value function. Subsequently the advantage function and state-value function are combined by an aggregation layer to provide a single Q-function.

2 Main

2.1 Methods

We explored

1. MARL Architecture
   a) Best location according to [Wal+20]
   b) Best location chosen as the min of mean of surrounding Q-values
   c) Best location chosen as the most frequented point in location history
2. MARL + Attention Architecture
   a) MARL + standard attention
   b) MARL + 3D attention augmented convolutions

In detail as follows:

2.1.1 Multi Agent Reinforcement Learning

We formulate the re-detection of multiple landmarks as a multi agent reinforcement learning problem. Similar to single agent reinforcement learning being viewed as a MDP one can also view the n agents taking actions in the environment as a series of n concurrent MDPs, see figure 2.1. Concurrently, each agent learns its individual policy toward its landmark during training.

We define the environment as a 3D MR scan of a human brain. We subsequently limit our analysis on the case that all agents are located in the same environment. To every agent only the surrounding 15x15x15 voxel centred on the agents ongoing position is observable. Because the problem space is not fully observable to each agent, our formulation of multi agent reinforced re-detection of landmarks is a concurrent Partially Observable Markov Decision Process (co-POMDP) [GE15].

At every step in the decision process each agent has three dimensions of freedom: x, y, z, creating six possible actions per agent. Each agent chooses the action with the highest associated reward received from the environment. The reward function is defined as the relative improvement in euclidean distance from state s_t to s_{t+1}.

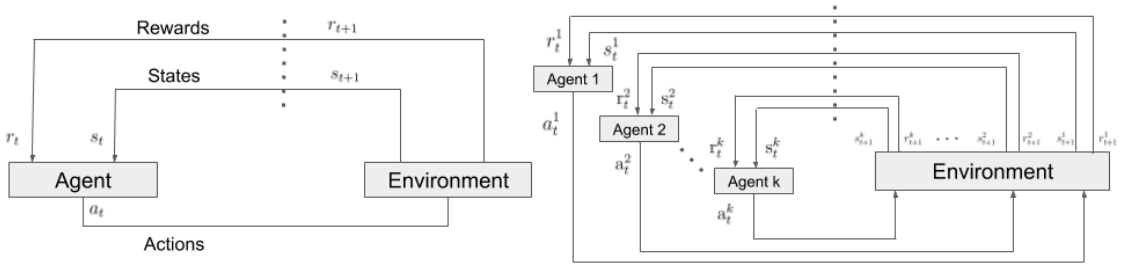

Figure 2.1: Action reward loop for single and multi agent RL

We combine the anatomical guidance from the single agent architecture of [Wal+20] with the multi agent Collab-DQN architecture from [Vas+17]. The architecture without attention is detailed in the following:

As in the Collab-DQN from [Vlo+19] we use a Siamese convolutional architecture to generate the deep features for all agents, see figure 2.2. From these deep features the action-advantage and the state-value are predicted for each agent individually.

Anatomical Guidance

Like [Wal+20] we use anatomical guidance to aid the agents when stepping into the tumor in pre-operative and the resection in post-operative scans. We implement this by returning a negative reward when the agent steps into the mask, identical to when the agent tries to leave the image. The masks are hand drawn from the clinical expert when setting the ground truth landmarks. See 2.2.1 for more information on the dataset.

Multi Resolution

We use multi resolution to aid the agent in faster converging to the goal state, as demonstrated by [Ghe+19a; Ghe+17; Ghe+19b] for single agent RL. Contrary to [Vlo+19]

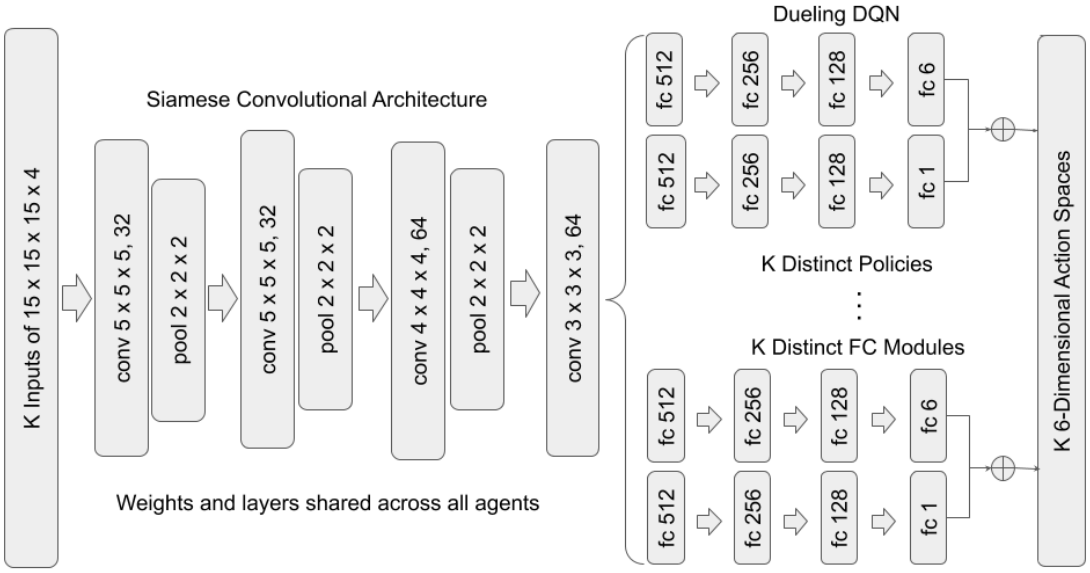

Figure 2.2: Multi agent dueling DQN

who seem to not use multi scaling coupled with multi agent RL we combine both methods.

At the beginning the agent starts with a step size of 9mm. When oscillations are detected, the agent reduces the step size by 3mm and continues from the best position in the location history. This is repeated until the step size is one and oscillations occur, the best location at this point is the result.

The different possibilities of choosing the best location when oscillation occurs are detailed in the following section 2.1.1.

For our purpose we define oscillation as the occurrence of a location n times in the location history. We use $n = 4$ and a location history length of 20.

Best Location

We choose the best location in three different ways: a. Best location according to [Wal+20; Vlo+19; Ala+19]: minimum of the maximum of the Q-values of the surrounding locations b. Best location using min of mean of surrounding Q-values c. Best location using most frequented point in location history

Choosing the best location is important because when using multi-scaling it seeds the "rounds" after decreasing the step size and determines the final position and thus

the distance while testing and training.

The two newly proposed methods to choose the best location have shown to be the most effective among various methods tried on a single patient.

2.1.2 Multi Agent Reinforcement Learning with Attention

Standard Attention

We flatten a given input tensor (H, W, D, F_{in}) to $X \in \mathbb{R}^{HWD \times F_{in}}$ and subsequently perform multi headed self attention as proposed by [Vas+17]. H, W, D and F_{in} denote height, width, depth and number of input filters of an importance map. d_v, d_k and N_h denote the depth of values, the depth of keys and queries and the number of attention heads. We assume that N_h evenly divides by d_v and d_k. The output of a single head h can be written as:

$$O_h = \text{Softmax}\left(\frac{(XW_q)(XW_k)^T}{\sqrt{d_k^h}}\right)(XW_v) \tag{2.1}$$

where $W_q, W_k \in \mathbb{R}^{F_{in} \times d_k^h}$ and $W_v \in \mathbb{R}^{F_{in} \times d_v^h}$ are learned linear transformations that map X to queries $Q = XW_q$, keys $K = XW_k$ and values $V = XW_v$. d_k^h denotes the depth of keys per head h. The individual ouputs of all heads are subsequently concatenated and a linear transformation is applied:

$$\text{MHA}(X) = \text{Concat}\left[O_1, \cdots, O_{Nh}\right] W^O. \tag{2.2}$$

W^O denotes the weights of the learned linear transformation. Finally $\text{MHA}(X)$ is reshaped to its original dimensions (H, W, D, F_{in}). In this naive form attention has a runtime of $\mathcal{O}((HWD)^2 d_k)$ and a memory cost of $\mathcal{O}((HWD))^2 N_h)$ because individual attention maps are stored for each head.

We limit the application of standard attention in our architecture to the output of the last convolution where all spatial dimensions are one. The standard attention layer is wrapped in a residual connection. The outputs after applying the residual are normalised to [0,1] as visualised in figure 2.3. We use anatomical guidance as well as multi-resolution and choose the best location by the minimum of the mean of the surrounding Q-values.

3D Attention Augmented Convolutions

We propose 3D Attention Augmented Convolutions to address the shortcomings of the attention mechanism and convolutions in visual applications. Namely without embedding position information attention is permutation equivariant, thus loosing

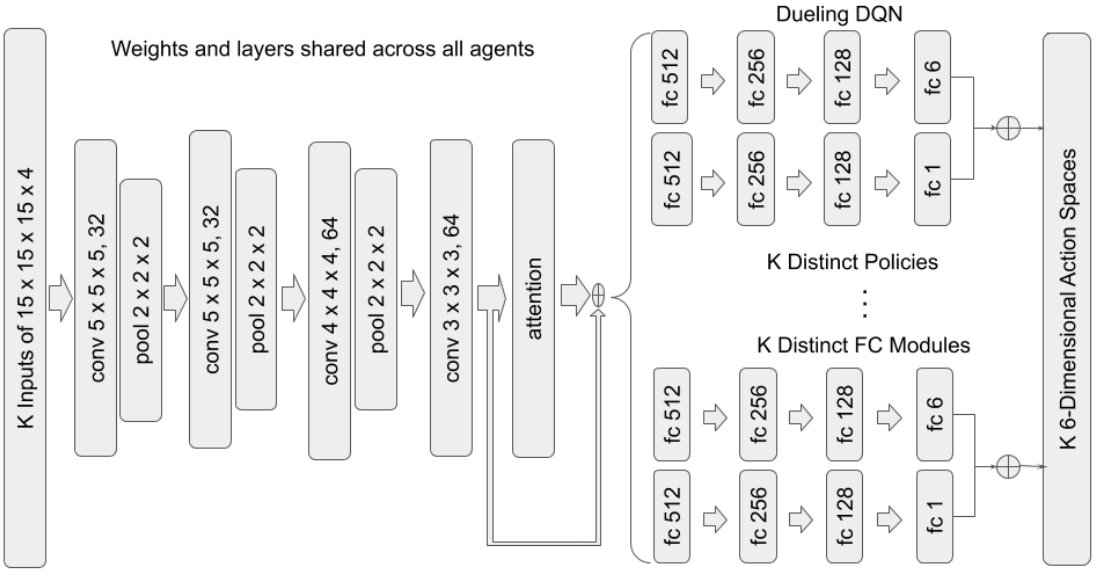

Figure 2.3: Multi agent dueling DQN with standard attention

structural relationships when modelling image data. Convolutions are limited due to locality and lack of global context understanding. Although explicit spatial encodings have been proposed [Par+18; Liu+18], they lack translational equivariance, which is a desired property for visual tasks. [Bel+19b] proposes an efficient extension to 2D of relative position encodings [SUV18] based on the Music Transformer [Hua+18].

Subsequently, we will formally outline Attention Augmented Convolutions [Bel+19b] extended to three dimensions. The attention logit representing how much pixel $i = (i_x, i_y, i_z)$ attends to pixel $j = (j_x, j_y, j_z)$ is computed as:

$$l_{i,j} = \frac{q_i^T}{\sqrt{d_k^h}}(k_j + r^W_{j_x - i_x} + r^H_{j_y - i_y} + r^D_{j_z - i_z}) \tag{2.3}$$

where q_i is the query vector for pixel i, k_j the key vector for pixel j. $r^W_{j_x - i_x}, r^H_{j_y - i_y}$ and$r^D_{j_z - i_z}$ are learned embeddings for relative width $j_x - i_x$, height $j_y - i_y$ and depth $j_z - i_z$. The ouput of head h becomes:

$$O_h = \text{Softmax}\left(\frac{(QK^T + S^{rel}_H + S^{rel}_W + S^{rel}_D}{\sqrt{d_k^h}}\right) V \tag{2.4}$$

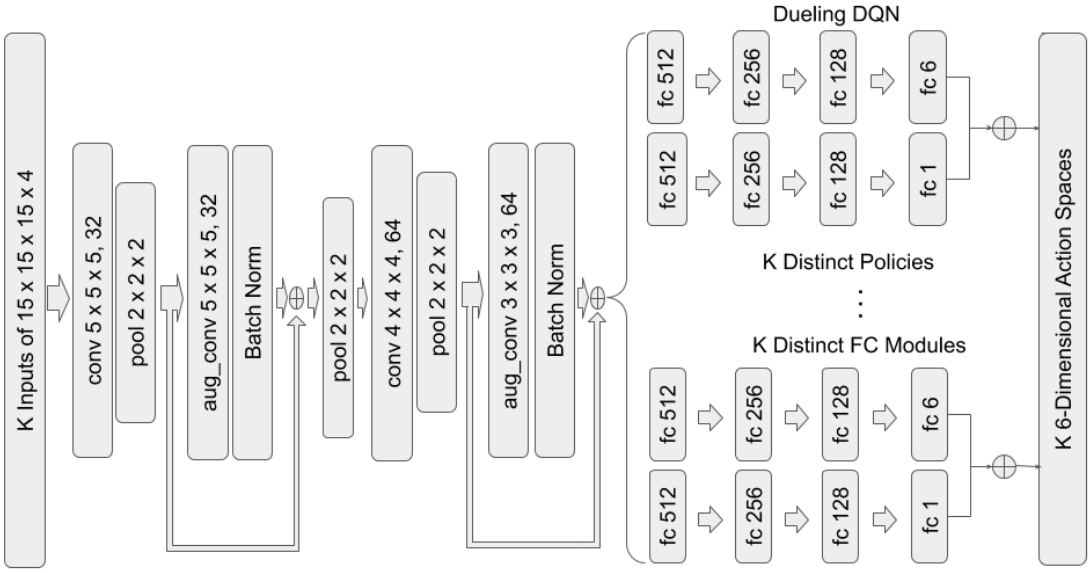

Figure 2.4: Multi agent dueling DQN with 3D attention augmented convolutions

where $S_H^{rel}, S_W^{rel}, S_D^{rel} \in \mathbb{R}^{HWD \times HWD}$ are matrices of relative position logits along height, width and depth dimensions which satisfy $S_H^{rel} = q_i^T r_{j_y - i_y}^H$, $S_W^{rel} = q_i^T r_{j_x - i_x}^H$ and $S_D^{rel} = q_i^T r_{j_z - i_z}^H$.

We further extend the memory efficient implementation by [Bel+19b], which is based on the algorithm presented in [Hua+18], to yield a final memory cost of $\mathcal{O}(HWDd_k^h)$ instead of $\mathcal{O}((HWD))^2 d_k^h)$ which would prove to be prohibitive due to $d_k^h > N_h$.

The final output of a 3D attention augmented convolution is the concatenation of a standard 3D convolution with the output of a relative multi headed attention:

$$\text{AAConv}(X) = \text{Concat}\left[\text{Conv}(\text{X}), \text{MHA}_{\text{rel}}(X)\right] \tag{2.5}$$

Analogue to [Bel+19b], we apply batch normalisation to the attention augmented convolution and allow for a residual connection. We exchange all depth preserving convolutional layers with this method. The complete architecture can be seen in figure 2.4.

2.2 Experiments and Results

2.2.1 Dataset

We evaluate our architectures on the same dataset as [Wal+20], provided by the BraTS challenge [Men+15] and TCIA [Cla+13]. The dataset compromises of 10 scans before and 10 scans after tumor resection for 10 individual patients. All images are skull-stripped, rigidly co-registered and interpolated to a common resolution of $1mm^3$. The inital resolution ranges between 3mm and 8mm for most sequences.

Additionally to the images the dataset includes segmentation masks for the tumor in the pre-operation scan and the resection cavity in the post-operation scan.

The dataset further includes initialisation boxes of size 50x50x50 around the target.

Three Landmarks are set by a clinical expert per patient in the post-operative scan in varying distances up to 4cm around the resection-affected region. To generate ground truth annotations, the same landmarks are then re-detected by the same expert in the corresponding pre-operative image.

2.2.2 Training

All architectures described in 2.1 Methods where trained with batch size 48, 3750 steps per epoch for 20 epochs, update frequency of 4, activation function PReLU for convolutional layers, ReLU for fully conected layers, adam optimiser with $\gamma = 1e-3$ and $\epsilon = 1e-3$. He initialisation [He+15] was used for all kernels of convolutional layers.

An experience memory size of 1e5 was used. The discount factor was set to 0.9.

The agents follow an ϵ-greedy exploration/exploitation strategy. At every step the agents choose the actions with the highest associated Q-values with probability $(1-\epsilon)$. Consequential a random action is taken with probability ϵ. The initial exploration rate ϵ is set to 1, decaying linear to 0.1 in the first 10 epochs.

During training we freeze the network parameters and disable updates as soon as one agent reaches his goal landmark. All other agents may still individually continue to explore the environment until they either reach their landmark as well or a pre-defined maximum number of movements are made by the agent.

Every attempt to find the landmark is named an *episode*. We limit the maximum number of frames i.e. movements an agent makes in the environment per episode to 1500.

Although all initialisation seeds where fixed, full reproducibility was not achieved. This is most likely due to the highly parallelized execution on GPUs.

	mean distance [mm]	median distance [mm]
expert	2.18	2.0
AM	2.82	1.53
MinMax	3.09	1.41
MinMean	2.79	1.41
Most	3.18	1.73
Attn	56.61	22.56
AugAttn	155.86	157.58

Table 2.1: Comparison of distance errors for AM, MinMax, MinMean, Most, Attn, AugAttn against an human expert annotation

2.2.3 Testing

Testing is done with 20 random initialisations sampled from the initialisation space around the tumor but not inside the mask for the tumor or resection cavity.

During testing no exploration is done, the agents always select the action with the highest Q-value.

2.2.4 Computational Performance

Training for one patient with three Landmarks takes $\sim$ 3.5 hours on a single NVIDIA Titan-Xp, 12GB, with approximately 1GB of VRAM usage for the models without attention and 4GB with attention.

Single threaded testing with 20 random initialisations takes around three minutes per patient.

2.2.5 Results

We compare the results of the baseline method (AM) [Wal+20] with anatomical guidance to all architectures described in section 2.1 i.e. multi-scaled multi agent reinforcement learning with and without attention. Concretely we test our multi agent approach without attention with the best location chosen as [Wal+20] by the minimum of the maximum surrounding Q-values (MinMax), the minimum of the mean of the surrounding Q-values (MinMean) and the most frequented position in the location history (Most). Further, we compare our multi agent approach with standard attention (Attn) and with 3D attention augment convolutions (AugAttn).

As visible in figure 2.5 Attn does not converge in training for the 2. landmark for all patients while the agents for the 1. and 3. landmark do converge. In comparison

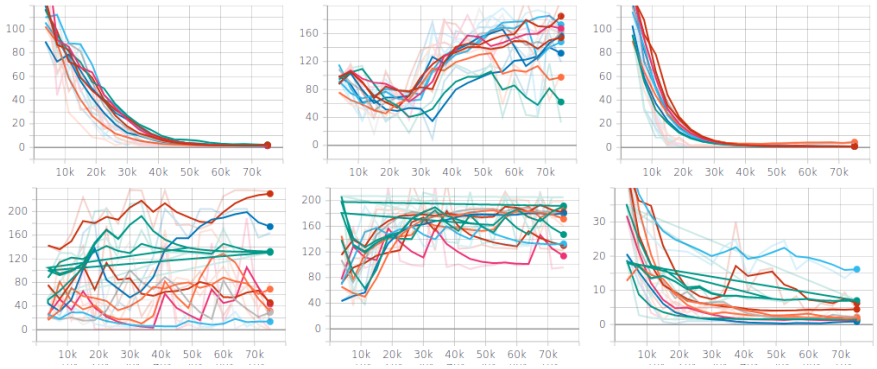

Figure 2.5: Smoothed training (top) and testing (bottom) mean distance per agent for Attn in mm against number of training steps

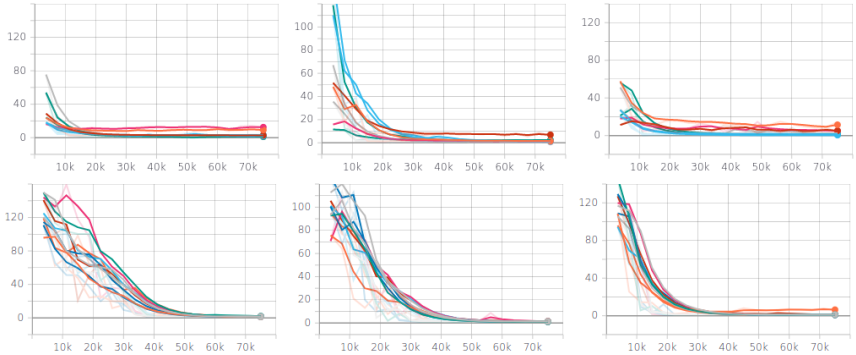

Figure 2.6: Smoothed training (top) and testing (bottom) mean distance per agent for MinMean in mm against number of training steps

to agent 3 the learning of agent 1 does not generalise as well to the testing data. This is evident by the minimum average distance achieved across all patients for agent 1 which is 51.28mm whereas the same value for agent 3 is 4.27mm.

We achieve a slight increase in accuracy with MinMean in comparison to the baseline with anatomical guidance AM [Wal+20] while decreasing computational effort about an order of magnitude.

In figure 2.6 we show the convergence of both training and test accuracy of MinMean.

3 Conclusion

In this work we consider the use of multi agent reinforcement learning for landmark detection in pre and post operative MR brain tumor scans. We introduce a novel 3D augmented attention mechanism and perform multi-scaling in multi agent reinforcement learning. We demonstrate that the combination of multi-scaling, multi agent reinforcement learning and the correct choice of the best location can drastically improve the computational performance for this problem while slightly improving the accuracy.

Possible future work can be broadly categorised into improvement of the dataset, the implementation and investigations into expansions of the proposed architectures and experiments.

We would like to improve the data quantity and quality for the problem. Improving the quality of the dataset can be achieved by collecting landmarks from several clinical experts and assuming the average to be the ground truth. Additionally, more patients are needed so that a proper training, test and evaluation split can be done.

The implementation can be improved by a rewrite in PyTorch, which would most likely improve the usability and performance. Fixing the visualisation of the agents movements inside of the scans would facilitate communication and help investigate results. The performance might be further improved by an investigation into bottlenecks at execution time. We suspect that improved data loading holds the greatest potential to increase resource utilisation (CPU/GPU). Inference may be sped up by using the same multi threaded evaluation method as done at training time. Training speed might be further improved by a multi threaded async multi agent implementation as demonstrated by [Mni+16].

For our problem the architecture may be expanded by prepending a segmentation network (e.g. [RFB15]) to generate the segmentation mask. The non convergence of both attention based methods should be further analysed as attention based methods have shown the best training accuracies. A starting point might be the pretraining of convolutional layers and limiting the use of attention to later epochs in the training process.

In the future the experiments may be expanded by an investigation into cross patient training, the differences of landmarks set by different clinical experts and the robustness of our approaches to these differences.

List of Figures

List of Tables

Bibliography

[Ala+19] A. Alansary, O. Oktay, Y. Li, L. Folgoc, B. Hou, G. Vaillant, K. Kamnitsas, A. Vlontzos, B. Glocker, B. Kainz, and D. Rueckert. "Evaluating Reinforcement Learning Agents for Anatomical Landmark Detection." In: *Medical Image Analysis* 53 (Feb. 2019). DOI: 10.1016/j.media.2019.02.007.

[BCB14] D. Bahdanau, K. Cho, and Y. Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." In: (Sept. 1, 2014). arXiv: 1409.0473v7 [cs.CL].

[Bel+16] I. Bello, H. Pham, Q. V. Le, M. Norouzi, and S. Bengio. "Neural Combinatorial Optimization with Reinforcement Learning." In: (Nov. 29, 2016). arXiv: 1611.09940v3 [cs.AI].

[Bel+19a] I. Bello, S. Kulkarni, S. Jain, C. Boutilier, E. Chi, E. Eban, X. Luo, A. Mackey, and O. Meshi. *Seq2Slate: Re-ranking and Slate Optimization with RNNs*. 2019.

[Bel+19b] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le. "Attention Augmented Convolutional Networks." In: (Apr. 22, 2019). arXiv: 1904.09925v4 [cs.CV].

[Bro+20] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. "Language Models are Few-Shot Learners." In: (May 28, 2020). arXiv: 2005.14165v3 [cs.CL].

[Che+18] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng. "A^2-Nets: Double Attention Networks." In: (Oct. 27, 2018). arXiv: 1810.11579v1 [cs.CV].

[Cla+13] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository." In: *Journal of Digital Imaging* 26.6 (July 2013), pp. 1045–1057. DOI: 10.1007/s10278-013-9622-7.

[DeA01] L. M. DeAngelis. "Brain Tumors." In: *New England Journal of Medicine* 344.2 (Jan. 2001), pp. 114–123. DOI: 10.1056/nejm200101113440207.

[Den+09] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database." In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[Dev+18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: (Oct. 11, 2018). arXiv: 1810.04805v2 [cs.CL].

[GE15] J. Girard and M. R. Emami. "Concurrent Markov decision processes for robot team learning." In: *Engineering Applications of Artificial Intelligence* 39 (Mar. 2015), pp. 223–234. DOI: 10.1016/j.engappai.2014.12.007.

[Ghe+17] F. C. Ghesu, B. Georgescu, S. Grbic, A. K. Maier, J. Hornegger, and D. Comaniciu. "Robust Multi-scale Anatomical Landmark Detection in Incomplete 3D-CT Data." In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Springer International Publishing, 2017, pp. 194–202. DOI: 10.1007/978-3-319-66182-7_23.

[Ghe+19a] F. Ghesu, B. Georgescu, Y. Zheng, S. Grbic, A. Maier, J. Hornegger, and D. Comaniciu. "Multi-Scale Deep Reinforcement Learning for Real-Time 3D-Landmark Detection in CT Scans." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.1 (2019), pp. 176–189.

[Ghe+19b] F.-C. Ghesu, B. Georgescu, Y. Zheng, S. Grbic, A. Maier, J. Hornegger, and D. Comaniciu. "Multi-Scale Deep Reinforcement Learning for Real-Time 3D-Landmark Detection in CT Scans." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.1 (Jan. 2019), pp. 176–189. DOI: 10.1109/tpami.2017.2782687.

[He+15] K. He, X. Zhang, S. Ren, and J. Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1026–1034.

[HS97] S. Hochreiter and J. Schmidhuber. "Long short-term memory." In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[Hu+17] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. "Squeeze-and-Excitation Networks." In: (Sept. 5, 2017). arXiv: 1709.01507v4 [cs.CV].

[Hu+18] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi. "Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks." In: (Oct. 29, 2018). arXiv: 1810.12348v3 [cs.CV].

[Hua+18] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck. "Music Transformer." In: (Sept. 12, 2018). arXiv: 1809.04281v3 [cs.LG].

[Kri12] A. Krizhevsky. "Learning Multiple Layers of Features from Tiny Images." In: *University of Toronto* (May 2012).

[Lam+12] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, et al. "Radiomics: extracting more information from medical images using advanced feature analysis." In: *European journal of cancer* 48.4 (2012), pp. 441–446.

[Lin+14] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. "Microsoft COCO: Common Objects in Context." In: (May 1, 2014). arXiv: 1405.0312v3 [cs.CV].

[Lin92] L.-J. Lin. "Reinforcement Learning for Robots Using Neural Networks." PhD thesis. USA: Carnegie Mellon University, Schenley Park Pittsburgh, PA, United States, 1992.

[Liu+18] R. Liu, J. Lehman, P. Molino, F. Petroski Such, E. Frank, A. Sergeev, and J. Yosinski. "An intriguing failing of convolutional neural networks and the CoordConv solution." In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018, pp. 9605–9616.

[Men+15] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, C. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. V. Leemput. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)." In: *IEEE Transactions on Medical Imaging* 34.10 (Oct. 2015), pp. 1993–2024. DOI: 10.1109/tmi.2014.2377694.

[Mni+15] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S.

Legg, and D. Hassabis. "Human-level control through deep reinforcement learning." In: *Nature* 518.7540 (Feb. 2015), pp. 529–533. DOI: 10.1038/nature14236.

[Mni+16] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. "Asynchronous Methods for Deep Reinforcement Learning." In: *ICML 2016* (Feb. 4, 2016). arXiv: 1602.01783v2 [cs.LG].

[OW12] M. van Otterlo and M. Wiering. "Reinforcement Learning and Markov Decision Processes." In: *Adaptation, Learning, and Optimization*. Springer Berlin Heidelberg, 2012, pp. 3–42. DOI: 10.1007/978-3-642-27645-3_1.

[Par+18] N. Parmar, A. Vaswani, J. Uszkoreit, Ł. Kaiser, N. Shazeer, A. Ku, and D. Tran. "Image Transformer." In: (Feb. 15, 2018). arXiv: 1802.05751v3 [cs.CV].

[Rad+19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. "Language models are unsupervised multitask learners." In: *OpenAI Blog* 1.8 (2019), p. 9.

[RFB15] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In: (May 18, 2015). arXiv: 1505.04597v1 [cs.CV].

[SB18] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[SB98] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. 1st. Cambridge, MA, USA: MIT Press, 1998. ISBN: 0262193981.

[SLL19] D. R. So, C. Liang, and Q. V. Le. "The Evolved Transformer." In: (Jan. 30, 2019). arXiv: 1901.11117v4 [cs.LG].

[SUV18] P. Shaw, J. Uszkoreit, and A. Vaswani. "Self-Attention with Relative Position Representations." In: (Mar. 6, 2018). arXiv: 1803.02155v2 [cs.CL].

[Vas+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need." In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[VFJ15] O. Vinyals, M. Fortunato, and N. Jaitly. "Pointer Networks." In: (June 9, 2015). arXiv: 1506.03134v2 [stat.ML].

[Vlo+19] A. Vlontzos, A. Alansary, K. Kamnitsas, D. Rueckert, and B. Kainz. "Multiple Landmark Detection using Multi-Agent Reinforcement Learning." In: *Lecture Notes in Computer Science, vol 11767. Springer, Cham*. Oct. 2, 2019. DOI: 10.1007/978-3-030-32251-9_29. arXiv: 1907.00318v2 [cs.CV].

[Wal+20] D. Waldmannstetter, F. Navarro, B. Wiestler, J. S. Kirschke, A. Sekuboyina, E. Molero, and B. H. Menze. "Reinforced Redetection of Landmark in Pre- and Post-operative Brain Scan Using Anatomical Guidance for Image Alignment." In: *Biomedical Image Registration*. Springer International Publishing, 2020, pp. 81–90. DOI: 10.1007/978-3-030-50120-4_8.

[Wan+16] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas. "Dueling Network Architectures for Deep Reinforcement Learning." In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 1995–2003.

[Wan+17] X. Wang, R. Girshick, A. Gupta, and K. He. "Non-local Neural Networks." In: (Nov. 21, 2017). arXiv: 1711.07971v3 [cs.CV].

[Woo+18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. "CBAM: Convolutional Block Attention Module." In: (July 17, 2018). arXiv: 1807.06521v2 [cs.CV].

[Yan+18] B. Yang, L. Wang, D. F. Wong, L. S. Chao, and Z. Tu. "Convolutional Self-Attention Network." In: (Oct. 31, 2018). arXiv: 1810.13320v2 [cs.CL].

[Yu+18] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. "QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension." In: (Apr. 23, 2018). arXiv: 1804.09541v1 [cs.CL].

[Zam+19] V. Zambaldi, D. Raposo, A. Santoro, V. Bapst, Y. Li, I. Babuschkin, K. Tuyls, D. Reichert, T. Lillicrap, E. Lockhart, M. Shanahan, V. Langston, R. Pascanu, M. Botvinick, O. Vinyals, and P. Battaglia. "Deep reinforcement learning with relational inductive biases." In: *International Conference on Learning Representations*. 2019.

[Zha+18] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. "Self-Attention Generative Adversarial Networks." In: (May 21, 2018). arXiv: 1805.08318v2 [stat.ML].