Multi-task Deep Reinforcement Learning for Cognitive Spectrum-agile Communications

Mohamed A. Aref
Communications and Information Sciences Laboratory
ECE Department, University of New Mexico
Albuquerque, NM, USA
maref@unm.edu

Sudharman K. Jayaweera
Communications and Information Sciences Laboratory
ECE Department, University of New Mexico
Albuquerque, NM, USA
jayaweera@unm.edu

***Abstract*—This paper introduces a cognitive engine design to achieve spectrum-agile communications over a heterogeneous wideband spectrum. The proposed cognitive approach has the ability to learn and avoid interference signals and other harmful signals. The targeted spectrum in this work is much wider than the ones proposed in the literature, most likely covering several hundreds of MHz. The proposed approach is based on deep reinforcement learning (DRL), more specifically on a double deep Q-network (DDQN) made of a convolutional neural network (CNN). The wideband spectrum is divided into a number of sub-bands and each sub-band consists of a number of channels. The problem is modeled as a multi-task DRL, where each sub-band represents a single task. Transfer learning is used between tasks to speed up the learning process. It is shown, through simulations, that the proposed technique can efficiently learn an effective strategy to avoid harmful signals in a noncontiguous wideband spectrum. Furthermore, it outperforms other DRL-based approaches in the literature while operating in a much wider spectrum and maintaining low computational complexity.**

***Index Terms*—Double deep Q-network, multi-task deep reinforcement learning, spectrum agility, transfer learning, wideband autonomous cognitive radios.**

I. Introduction

The rapid growth of mobile broadband traffic, driven by smart phones and new wireless communications networks that support high data rates such as 5th generation (5G) cellular communication systems, has lead to an increase in the demand for the RF spectrum. On the other hand, many studies have reported that the localized temporal and geographic spectrum utilization is extremely low [1], [2]. This has motivated the spectrum regulatory organizations to develop new spectrum policies that will allow secondary users (SUs) to opportunistically access a licensed band when the primary user (PU) is absent. The cognitive radio (CR) was introduced as a solution to improve the spectrum utilization in these scenarios [3], [4].

Since cognitive radios are considered secondary users in accessing the licensed spectrum, they should be able to independently detect spectrum opportunities without any assistance from the PUs. This process is called spectrum sensing, which is considered one of the most critical components in cognitive radio networks (CRNs) that distinguishes a cognitive radio from a legacy radio. Several spectrum sensing techniques have been studied in literature including matched filtering, energy detection, cyclostationary feature detection and compressive sensing [5]–[7]. Most of these techniques, however, limited to a few tens, or a few hundreds at most, of MHz wide spectrum. Furthermore, the proposed cognitive radios are assumed to be operating in a single mode of operation in which the spectrum sensing is limited to the spectrum occupied by a specific primary system.

In this paper we discuss the design of future cognitive radios that is capable of operating over several noncontiguous bands spread over a wide range of frequencies. The spectrum of interest in this work is much wider than the ones proposed in the literature, most likely covering several hundreds of MHz or few GHz. The spectrum of interest can be accessed by several systems that can introduce different types of signals at different frequencies creating a heterogeneous environment. In order for the CR to operate in the best mode, it should have intelligence capabilities to find the spectrum opportunities in such heterogeneous widenand spectrum. Furthermore, it should be spectrum-agile to avoid interference and other harmful signals, such as jamming. Therefore, the problem can be considered as an interference avoidance problem or as an anti-jamming problem in the presence of jamming attacks.

The literature is rich with several work that have studied the interference avoidance and anti-jamming problems for cognitive radios [8]–[10]. Most of these consider either Markov decision processes (MDPs) or stochastic games for system modeling. Moreover, they explore different reinforcement learning (RL) techniques to achieve their ultimate goal including, Q-learning, minimax Q-learning, or win-or-learn-fast (WoLF) policy hill climbing (PHC). In a simple RL scenario, an agent would first observe the surrounding environment to identify its current state. Then, it executes an action that move it to a new state. The RL is based on delayed-reward principle in which the agent receives a reward from the environment after executing each action [11]. The value of the reward indicates how good or bad the action is. The objective of the agent is then to choose actions that maximize the rewards.

In this paper we consider a heterogeneous wideband spectrum that might experience different scenarios at each time instant. Thus, the possible number of states and actions could be large. RL, however, has limitations when dealing with systems with a large state-action space. To tackle this, several work has recently proposed using deep reinforcement learning (DRL) [12]. The DRL relies on the power of deep neural networks to improve training and learning processes of traditional RL making it suitable for systems with a large state-action space [12]. Most existing DRL techniques are based on deep Q-network (DQN) algorithm that extends the Q-leaning by using a convolutional neural network (CNN) to learn an approximate Q-function [11], [12].

The authors in [13] proposed a two dimensional anti-jamming technique to enable a SU in a CRN to avoid the PU's

978-1-7281-3706-3/19/$31.00

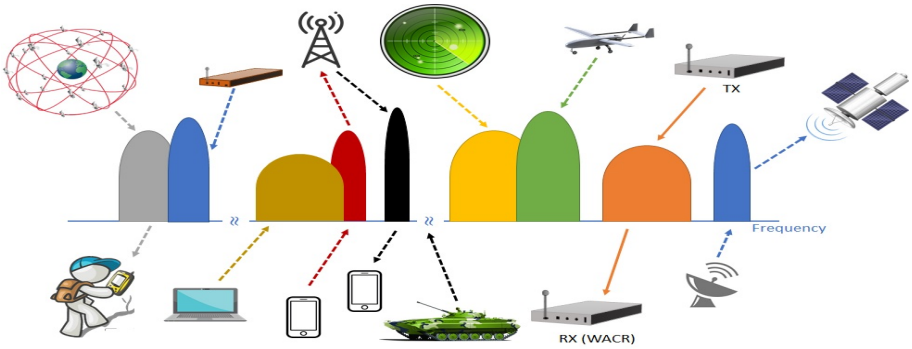

Figure 1. System model including heterogeneous RF environment.

transmission as well as jamming attacks using DQN with a CNN. At any time instant, the SU performs two actions: (1) Stay in or leave the current cell (2) Select a frequency channel for communications. The authors in [14] extended the model in [13] for an underwater acoustic network (UAN). In [15], the authors adopted DQN with a recursive convolutional neural network (RCNN) to enable the SU to learn an optimal frequency hopping strategy to avoid jamming. A cognitive approach for interference avoidance and anti-jamming was proposed in [16] that is suitable for real-time applications in partially-observable spectrum. Although these show good performance compared to traditional RL techniques such as Q-learning, most of the proposed approaches in the above mentioned references can not be applied easily in a noncontiguous wideband spectrum and suffer from high computational complexity.

The purpose of this paper is to enable the radio to find spectrum opportunities over noncontiguous wideband spectrum and allow spectrum agility to avoid interference and other harmful signals in real-time. The wideband spectrum is divided into a number of sub-bands where each sub-band consists of a number of channels. The radio's objective is to select the best sub-band for communications and then select the best channel inside the sub-band that maximizes the signal-to-interference-plus-noise ratio (SINR). The problem is formulated as a multi-task DRL problem where each sub-band represents a single task. The radio uses a double deep Q-network (DDQN) with CNN to learn an efficient policy for each task. Transfer learning is used between tasks to reduce the training time for any new task.

The rest of this paper is organized as follows: Section II presents the system model. The details of the proposed spectrum-agile approach for interference avoidance using DDQN is discussed in Section III. We provide simulation results in Section IV and conclude this work in Section V.

II. System Model

Let us consider two radios that form a communications link and try to preserve the connectivity between each other. The wideband spectrum consists of N non-overlapping non-contiguous channels with equal bandwidth. The bandwidth of each channel can be set equal to the bandwidth of the signals of interest. The receiver of the link of interest is assumed to be a wideband autonomous cognitive radio (WACR) having the ability to sense spectrum, learn and take decisions [4], [16]. The spectrum might be accessed by other systems that may cause interference making the task of finding the spectrum opportunities challenging. Since these systems access the spectrum of interest at particular frequencies and at different times, the interference signals can be distributed widely in both frequency and time domains.

Fig. 1 shows a typical heterogeneous wideband spectrum where the solid lines represent the signals of our link of interest and the dashed ones represent signals from other sources that might cause interference. The objective of the WACR is to choose a frequency channel with highest SINR for communications at every time instant. The frequency synchronization between the receiver and the transmitter is done through a secured common control channel. Let us assume that the transmitter sends the signal of interest with a given power P_s. The channel power gain from the transmitter to the WACR is given by h_s. The interference source i sends its signals with a given power $P_{I,i}$, while the channel power gains to the WACR is $h_{I,i}$. The received SINR (in dB) of the WACR at channel n and time t can be expressed as

$$\mu_n(t) = 10\log_{10}\frac{h_s P_s}{\sigma^2 + \sum_i h_{I,i} P_{I,i}} \ , \qquad (1)$$

where σ^2 is the receiver noise power, assuming additive white noise.

In this work, we assume a wideband system that may cover several hundred MHz that is much wider than other wideband systems in the literature [6], [7], [16]. Thus, based on the channel bandwidth, the number of channels could be several tens or hundreds. At the same time, due to hardware and signal processing limitations, the WACR might not be able to sense all frequency bands of interest simultaneously. One of the solutions is to do spectrum segmentation by dividing the wideband spectrum into a number of sub-bands [4], [17]. The sub-band width should not exceed the maximum spectrum width that the WACR can sense and process in real time. To the best of our knowledge, the maximum instantaneous bandwidth available in the market is about 200 MHz (e.g. supported by the USRP-N320 from Ettus Research [18]).

Thus, we divide the spectrum into K non-overlapping sub-bands. These sub-bands might not have the same bandwidth, i.e, each may contain a different number of channels. The kth sub-band is assumed to contain N_k channels, $\forall k \in \{1, \cdots, K\}$, such that the total number of channels $\sum_{k=1}^{K} N_k$ is N. To keep the notation simpler, however, we may assume equal bandwidth sub-bands each with N/K channels. Sub-banding the spectrum of interest would divide the problem into two parts: First, the WACR needs to decide whether to continue on the same sub-band or to leave. Moreover, in case of leaving the current sub-band, which sub-band to choose that would have the best spectrum opportunities.

Second, once the sub-band is decided, the WACR needs to select a communications channel that would enable it to avoid interference and maximize SINR.

Let $b(t)$ denotes the sub-band index chosen by the WACR at time t for transmission. The WACR stays on the same sub-band at time t if $b(t) = b(t-1)$ and it moves to a new one otherwise. Let $c_{b(t)}(t)$ represents the index of the chosen channel for communications inside the sub-band $b(t)$. Then, at each time t, the WACR chooses an action, denoted by $a(t) = c_{b(t)}(t) \in \mathcal{A}$, where $\mathcal{A}$ is the action set whose size is $|\mathcal{A}| = N/K$. On other hand, let μ_{th} denote the required SINR threshold for successful communications over any channel (it is straightforward to generalize to different thresholds). Then, an indication function for communications over channel n at time t can be defined as follows:

$$f(\mu_n(t)) = \begin{cases} \lambda\mu_n(t), & \text{if } \mu_n(t) > \mu_{th} \text{ (success)} \\ -\lambda, & \text{if } \mu_n(t) \leq \mu_{th} \text{ (failure)} \end{cases} \quad (2)$$

where $\lambda > 0$ is a design parameter selected to provide sufficient contrast between the two cases for efficient learning.

Most commercial software defined radio (SDR) platforms include two or more different RF ports where each one has a separate RX chain [18]. This might allow the radio to do two parallel communications operations at different frequencies. In our proposed WACR, one RF port is used to receive the signal of interest. An additional RF port is used to perform dedicated sensing so that the WACR can explore the surrounding RF spectrum and collect information without interrupting the ongoing communications on the other port. This information can be used afterwards to select the best communications channel and its corresponding sub-band for avoiding interference and achieving high SINR. At any time t, the WACR can estimate the power spectral density (PSD) in a sensed sub-band. Then, using spectral activity detection, the WACR can identify the availability of channels in the corresponding sub-band by comparing the PSD with an appropriate threshold that is designed based on noise floor estimation [4].

Let the vector $\varphi_k(t) = [\phi_1^k(t), \phi_2^k(t), \cdots, \phi_{N/K}^k(t)]$ indicate the availability of channels inside the sensed sub-band k, where the function $\phi_n^k(t) = \eta$ for channel n being available and equal to $-\eta$ otherwise, where $\eta > 0$ is weighting factor. In this work, the sensing is assumed to follow a predefined strategy in which it alternates between sensing the current operational sub-band and sensing other sub-bands. By sensing the current sub-band, the WACR can keep track of the availability of the channels inside this sub-band. When the current channel cannot meet the Quality-of-Service (QoS) requirement due to interference, the WACR switches the communications to another channel on the same sub-band. On the other hand, sensing other sub-bands will help the WACR to select the best sub-band for communications if sub-band switching is required.

In order to select other sub-bands for sensing, a learning-aided sub-band selection algorithm could be applied [17], [19]. However, for simplicity, in this work sub-bands are chosen either sequentially or randomly. Using $\varphi_k(t)$, the WACR can identify the percentage of available channels in the kth sub-band, denoted by $0 \leq g(\varphi_k(t)) \leq 1$, where $g(\varphi_k(t)) = 1$ if the whole sub-band is available. A list $\mathbf{B}$ can be created that includes the percentage of the available channels and the best communication channel in each sub-band as shown in (3). Every time the kth sub-band is sensed, it's corresponding $g(\varphi_k)$ and c_k^* in $\mathbf{B}$ are updated. Hence, the list $\mathbf{B}$ will always keep the most updated values for each sub-band as follows:

$$\mathbf{B} = \begin{bmatrix} g(\varphi_1) & c_1^* \\ g(\varphi_2) & c_2^* \\ \vdots & \\ g(\varphi_K) & c_K^* \end{bmatrix}, \quad (3)$$

where c_k^* denotes the index of the best channel that can be used for communications in the kth sub-band.

III. Proposed Spectrum-agile Cognitive Communications Protocol using A Multi-task Double Deep Q-network (DDQN)

As mentioned earlier, there are two levels of actions that WACR needs to take at each time slot. First, stay on the current sub-band or choose a new one. Second, choose the best channel for communications in the corresponding sub-band. The first action is triggered by the indication function (2). In particular, if $f(\mu_n(t))$ is evaluated to be $-\lambda$ for J successive time slots in the current sub-band k, for any $n \in \{1, \cdots, N_k\}$, then the WACR switches to a new sub-band. This means that if the WACR failed to receive the signal of interest with acceptable SINR on the chosen channel/s in the current sub-band for J successive time slots it will move to a new sub-band. For sub-band switching, the WACR chooses the sub-band with the largest $g(\varphi_k)$ from the list $\mathbf{B}$, $\forall k \in \{1, \cdots, K\}$. Once the sub-band is decided, the problem becomes which channel to choose for communications inside this sub-band.

The interference avoidance problem inside each sub-band can be modeled as a Markov decision process (MDP) in which the WACR selects action $a(t)$ at time t. Therefore, the overall problem inside the wideband spectrum of interest can be considered as a multi-task DRL, where each sub-band represents a single task. In this work we propose using DDQN with CNN for a WACR to select an interference-free channel in each sub-band. The state at time t, denoted by $\mathbf{S}(t)$, consists of the indicies of the current sub-band and the channel, the corresponding indication function and the most updated sensing results of that sub-band, i.e., $\mathbf{S}(t) = [b(t), a(t), f(\mu_{a(t)}), \varphi_{b(t)}(t)]$. On the other hand, the reward function of choosing channel $a(t)$ for communications while in state $\mathbf{S}(t)$ is defined as follows:

$$r(\mathbf{S}(t), a(t)) = \begin{cases} \beta\mu_{a(t)}, & \text{if communications over channel } a(t) \text{ was successful} \\ -\beta e^{-\frac{\mu_{a(t)}}{\beta}}, & \text{otherwise} \end{cases} \quad (4)$$

where $\beta > 0$ is a weighting factor. The function $r(\mathbf{S}(t), a(t))$ is designed to obtain a certain reward value proportional to the received SINR if the communications over channel $a(t)$ was successful while a penalty is received in case of communications failure.

Algorithm 1 summarizes the proposed DDQN-based interference avoidance approach inside a single sub-band. The WACR is currently assumed to operate over the kth sub-band, i.e, $b(i) = k, \forall i \in \mathbb{Z}$. Furthermore, it selected the

Algorithm 1 DDQN-aided proposed spectrum-agile algorithm for a single sub-band

Initialize:
  Parameters λ, μ, γ
  The weights θ of the Q-network
  The weights θ^- of the target Q-network
for each time slot t **do**
  Determine $\mu_n(t)$
  Estimate $f(\mu_n(t))$
  Obtain $\varphi_k(t)$ from sensing operation
  Create state $\mathbf{S}(t)$
  With probability $\epsilon(k)$:
    Choose $a(t) = c_k(t) \in \mathcal{A}$ at random
  Otherwise:
  Obtain $Q(\mathbf{S}(t), a')$ from the proposed CNN $\forall a' \in \mathcal{A}$
    Select $a(t) = \arg\max_{a'} Q(S(t), a'; \theta(t))$
  Use channel $a(t)$ inside sub-band k to transmit signal
  Store new experience $\mathbf{e}(t-1) = \{\mathbf{S}(t-1), a(t-1), r(\mathbf{S}(t-1), a(t-1)), \mathbf{S}(t))\}$ in data set $\mathcal{D}$
  for i=1,$\cdots$,I **do**
    Select $\mathbf{e}(i) = \{\mathbf{S}(i), a(i), r(\mathbf{S}(i), a(i)), \mathbf{S}(i+1)\} \sim U(\mathcal{D})$
    Compute target Y via (7)
    Compute the gradient of the loss function (6)
    Update $\theta(t)$
  end for
  Reset $\theta^-(t) = \theta(t)$ for every L iterations.
end for

nth channel to transmit at time slot $t-1$ ($\mathbf{a}(t-1) = n$). For each time slot t, the WACR estimates the SINR of the received signal $\mu_n(t)$ to determine whether the transmission was succssful or not using $f(\mu_n(t))$ given in (2). On the other hand, through sensing, the WACR obtains vector $\varphi_k(t)$ to determine the availability of other channels inside the kth sub-band. Both information from communications and sensing are used to create current state $\mathbf{S}(t)$. The state is reshaped into a $(\lceil N_k/3 \rceil + 1) \times 3$ matrix and taken as the input to the CNN.

The CNN estimates the Q-values for all possible actions. The optimal action $a(t)$ is chosen with probability $\epsilon(k) - 1$, and a random action $a(t) \sim U(\mathcal{A})$ is selected uniformly with probability $\epsilon(k)$ to avoid staying in a local optima. The exploration rate $\epsilon(k)$ is defined by using $\nu(k)$ which is the number of visits to sub-band k, as follows:

$$\epsilon(k) = \frac{1}{\log_2(\frac{\nu(k)}{N_k} + 2)}. \quad (5)$$

For DDQN training, experience replay is used in which WACR's experiences $\mathbf{e}(t) = \{\mathbf{S}(i), a(i), r(\mathbf{S}(i), a(i)), \mathbf{S}(i+1)\}$ at each time instant t are stored in an experience replay buffer $\mathcal{D}(t) = [\mathbf{e}(1), \cdots, \mathbf{e}(t)]$. During learning at time t, we draw uniformly at random an experience $\mathbf{e}(i) \sim U(\mathcal{D}(t))$, for $1 \leq i \leq t$, from the set of the stored experiences. The weights $\theta(t)$ of the Q-network (CNN) at time t are then updated according to the stochastic gradient descent (SGD) using the loss function below:

$$L(\theta(t)) = \mathbb{E}_{\mathbf{e}(i) \sim U(\mathcal{D}(t))}[(Y - Q(\mathbf{S}(t), a(t); \theta(t))^2], \quad (6)$$

Table I
SIMULATION PARAMETERS

Spectrum	400 to 600 MHz, 800 to 1200 MHz, 1860 to 2060 MHz, 2300 to 2700 MHz
N	60 channels
K	6 sub-bands
μ_{th}	10 dB
γ	0.8
J	10
I, L	5
λ, η, β	10

where Y is the target Q-value. This process is repeated for I times at each time instant t in which $\theta(t)$ is updated according to I randomly selected experiences. The DDQN maintains two Q-networks: the Q-network (with weights θ) and the target Q-network (with weights θ^-). The learning network $Q(\mathbf{S}, a; \theta)$ keeps the current parameters which may get updated several times at each time-step, while target Y is computed by the target network with old parameters i.e., $Q(\mathbf{S}, a; \theta^-)$ as follows:

$$Y = r(\mathbf{S}(t), a(t)) + \gamma\, Q(\mathbf{S}(t+1), a^*; \theta(t)^-), \quad (7)$$

where

$$a^* = \arg\max_{a'} Q(\mathbf{S}(t+1), a'; \theta(t)).$$

The parameter $0 < \gamma < 1$ denotes the discount factor. The old parameters θ^- are set equal to parameters θ every L iterations.

As mentioned earlier, the problem is modeled as a multi-task DDQN where each sub-band represents a single task. One solution is to use different Q-networks for each task (i.e. each sub-band), at the expense of increased computational complexity and memory requirements especially when there is a large number of sub-bands. In this work we propose using transfer learning (TL) between tasks [20]. TL allows a network that is proven to do well on some task to be adapted to a separate but potentially related task by reusing what is learned. In our case, we use a network that has performed very well at estimating Q-values for one sub-band and adapt it, with a little more training, to estimate Q-values for another sub-band.

Hence, since the objective of each task is identical and the number of channels in each sub-band is the same, a single Q-network can be used for all tasks instead of using separate networks for each task. If the channel-availability dynamics of the new sub-band is close to that of the previous sub-band, TL may speed up learning of the new task. However, if they were to be significantly different (which is to be expected), this may not be the case. To address this problem, we propose using a single Q-network, but with different output layers and separate replay memory buffers dedicated for each task. Note that, having separate replay memory buffers assigned for each task is essential so that each task is trained according to it's own experience. The sub-band index is used to switch between the different output layers and the replay memory buffers.

IV. SIMULATION RESULTS

In the following we used the data collected from 30 MHz to 3 GHz at a single location in Vienna, VA for 87 hours [21]. Table I shows the simulation parameters. We considered

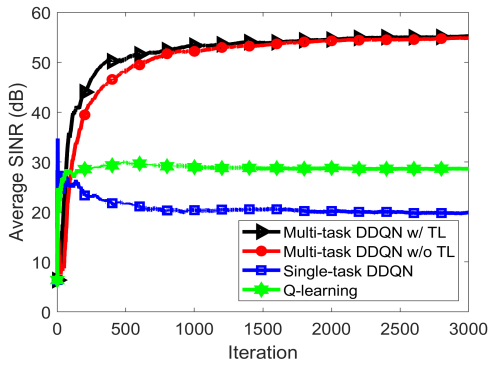

Figure 2. Comparison of the average SINR values.

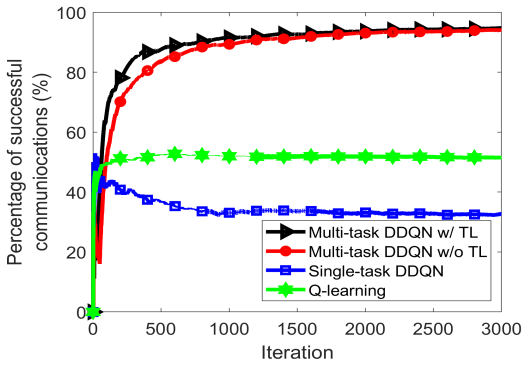

Figure 3. The percentage of successful communications.

noncontiguous wideband spectrum that is divided into 6 sub-bands each with 200 MHz bandwidth. Each sub-band consist of 10 channels with 20 MHz each. Thus, the total number of channels is 60. Based on the observations of the spectral activities in [21], the noise floor in each channel is set equal to -100 dBm. To incorporate fading effects, we consider different received signal strength for each channel. In particular, the received signal strength at any channel can take any value between -30 dBm to -60 dBm.

From the description above, the state $\mathbf{S}(t)$ at any time t is reshaped into a 5×3 matrix, where any empty locations are filled with zeros. Then, it is used as the input for the CNN. The proposed CNN architecture for the DDQN algorithm consists of one convolutional layer that includes 10 filters with size 2×2 and stride 1. This is followed by an average pooling layer of size 2×1 and a rectified linear unit (ReLU). Finally, a fully connected layer with 10 rectified linear units is used.

In the following, the proposed algorithm is used with two configurations: with and without transfer learning, named multi-task DDQN w/ TL learning and multi-task DDQN w/o TL, respectively. In the first configuration, only one CNN is used for all tasks, albeit with different output layers and experience replay buffers for each task. In the second configuration, six different CNNs are used, one per each task. For comparison, we also implemented two other techniques: First is the single-task DDQN based on the technique introduced in [16]. Using the same parameters described above with a memory depth equal to 10, the state of [16] is represented by a 11×60 matrix. Second is the Q-learning algorithm, where the state only consists of 2 values: state that includes only the index of the communications channel and the indicator for successful communications. Note that, in both single-task DDQN and Q-learning there is only one level of action, where the WACR directly selects one channel for communications among N channels without any spectrum sub-banding.

Fig. 2 shows the average estimated SINR values for the four techniques. From Fig. 2, we observe that both configurations of the proposed multi-task DDQN achieves the highest SINR compared to the single-task DDQN and the Q-learning. It is noticed also that the Q-learning outperforms the single-task DDQN. This is due to the state definition in [16] in which the majority of the values are zeros as the number of channels becomes large. This could negatively affect the feature extraction process in the CNN. On the other hand, Fig. 3 shows the percentage of time of achieving successful communications, i.e. SINR>10 dB. It can be seen from the figure that the proposed technique outperforms both the single-task DDQN and the Q-learning.

After 3000 iterations, the proposed multi-task DDQN approaches w/ and w/o TL were able to preserve the connectivity over the link of interest and perform successful communications for 95% and 94.5% of the time, respectively. Although both configurations achieve similar performance, using TL is clearly preferable from the computational complexity perspective. This is because TL allows using a single Q-network for all tasks, approximately reducing the computational complexity by a factor of 6 (in this example) compared to that w/o TL. In general, when there are N_b sub-bands, if the computational complexity of the multi-task DDQN w/o TL is $\mathcal{O}(M)$, then by using multi-task DDQN w/ TL it will approximately be reduced to $\mathcal{O}(M/N_b)$, ignoring the costs associated with implementing separate output layers.

V. Conclusion

In this paper we have proposed a multi-task DRL, based on DDQN, for WACR to achieve spectrum agility over non-contiguous wideband heterogeneous spectrum. The proposed approach divided the wideband spectrum of interest into a number of non-overlapping sub-bands and the problem is solved at the sub-band level where each sub-band represents a single task. In particular, the WACR uses DDQN to learn a policy to select the best communications channel inside each sub-band that will maximize SINR. In order to reduce the computational complexity and speed-up the learning process, transfer learning is allowed among tasks. A single deep Q-network is considered for all tasks with different output layers and experience replay buffers for each task. With simulation results, it is shown that the proposed approach is able to achieve better reliability performance compared to the single-task DDQN and Q-learning. After 3000 iterations, the proposed multi-task DDQN was able to avoid interference signals for 95% of the time.

Acknowledgment

This research was sponsored in part by the Army Research Laboratory and was accomplished under Grant Number W911NF-17-1-0035. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to

reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] M. McHenry, "NSF Spectrum Occupancy Measurements Project Summary," Shared Spectrum Co., tech. rep., Aug. 2005.

[2] V. Valenta, R. Maršálek, G. Baudoin, M. Villegas, M. Suarez, and F. Robert, "Survey on Spectrum Utilization in Europe: Measurements, Analyses and Observations," 2010 Proceedings of the Fifth International Conference on Cognitive Radio Oriented Wireless Networks and Communications, Cannes, France, Jun. 2010.

[3] J. Mitola, III and G. Maguire, Jr., "Cognitive radio: making software radios more personal," IEEE Personal Communications, vol. 6, no. 4, Aug. 1999.

[4] S. K. Jayaweera, "Signal Processing for Cognitive Radios," John Wiley & Sons, Hoboken, NJ, USA. ISBN: 978-1-118-82493-1, 2014.

[5] B. Farhang-Boroujeny, "Filter Bank Spectrum Sensing for Cognitive Radios," IEEE Transactions on Signal Processing, vol. 56, no. 5, pp. 1801–1811, May 2008.

[6] Z. Tian and G. Giannakis, "Compressive Sensing for Wideband Cognitive Radios," 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07), Honolulu, HI, USA, Apr. 2007.

[7] H. Sun, A. Nallanathan,C. X. Wang and Y. Chen, "Wideband spectrum sensing for cognitive radio networks: a survey," IEEE Wireless Communications, vol. 20, no. 2, pp. 74–81, Apr. 2013.

[8] B. Wang, Y. Wu, K. Liu, and T. Clancy, "An anti-jamming stochastic game for cognitive radio networks," IEEE Journal on Selected Areas in Communications, vol. 29, no. 4, pp. 877–889, Apr. 2011.

[9] S. Dastangoo, C. E. Fossa, Y. L. Gwon and H. Kung, "Competing Cognitive Resilient Networks," IEEE Transactions on Cognitive Communications and Networking, vol. 2, no. 1, pp. 95–109, Mar. 2016.

[10] M. A. Aref, S. K. Jayaweera and S. Machuzak, "Multi-agent Reinforcement Learning Based Cognitive Anti-jamming," IEEE Wireless Communications and Networking Conference (WCNC'17), San Francisco, CA, USA, Mar. 2017.

[11] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," MIT Press, 1998.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski, "Human-level control through deep reinforcement learning," Nature, vol. 518, no. 7540, Jan. 2015.

[13] G. Han, L. Xiao, and H. V. Poor, "Two-dimensional anti-jamming communication based on deep reinforcement learning," IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, Mar. 2017.

[14] L. Xiao, D. Jiang, X. Wan, W. Su, and Y. Tang, "Anti-jamming underwater transmission with mobility and learning," IEEE Communications Letters, vol. 22, no. 3, pp. 542–545, Mar. 2018.

[15] X. Liu, Y. Xu, L. Jia, Q. Wu, and A. Anpalagan, "Anti-jamming communications using spectrum waterfall: A deep reinforcement learning approach," IEEE Communications Letters, vol. 22, no. 5, pp. 998–1001, May 2018.

[16] M. A. Aref and S. K. Jayaweera, "Spectrum-agile Cognitive Interference Avoidance through Deep Reinforcement Learning," 14th EAI International Conference on Cognitive Radio Oriented Wireless Networks (CROWNCOM'19), Poznan, Poland, Jun. 2019.

[17] Y. Li, S. K. Jayaweera, M. Bkassiny and C. Ghosh, "Learning-aided Sub-band Selection Algorithms for Spectrum Sensing in Wide-band Cognitive Radios," IEEE Transactions in Wireless Communications, vol. 13, no. 4, pp. 2012–2024, Apr. 2014.

[18] https://www.ettus.com/all-products/usrp-n320/

[19] M. A. Aref, S. Machuzak, S. K. Jayaweera and S. Lane, "Replicated Q-learning based sub-band selection for wideband spectrum sensing in cognitive radios," IEEE/CIC International Conference on Communications in China (ICCC), Chengdu China, Jul. 2016.

[20] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[21] General RF Band Survey, "General Survey of Radio Frequency Bands (30 MHz to 3 GHz): Vienna, Virginia, September 1–5, 2009," Shared Spectrum Company, Vienna, VA, USA, Sept. 2010. http://www.sharedspectrum.com/papers/spectrum-reports/