HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

THESIS

A Deep Reinforcement Learning based Online Charging Scheme for Target Coverage and Connectivity in WRSNs

BUI HONG NGOC
ngoc.bh164797@sis.hust.edu.vn

Major : Computer Science

Thesis advisor	:	Assoc. Prof. Do Phan Thuan
		Signature of advisor
Department	:	Department of Computer Science
Institute	:	School of Information and Communication Technology

Hanoi, 5-2021

© *2021 - Bui Hong Ngoc*
All rights reserved.
Re-distributed by Hanoi University of Science and Technology under license with the author.

This work is licensed under a Creative Commons "Attribution-NonCommercial-ShareAlike 3.0 Unported" license.

REQUIREMENTS OF THE THESIS

1. Student's information :
   Name : Bui Hong Ngoc.
   Phone : 0988 490 924 Email: ngoc.bh164797@sis.hust.edu.vn
   Class : CNTT 2.02 K61
   Affiliation : Hanoi University of Science and Technology.
   Duration : 11/02/2021 - 31/05/2021.

2. Thesis statement :
   This thesis aims to investigate the target coverage and connectivity problem in the settings of wireless rechargeable sensor network. A novel deep reinforcement learning-based online charging scheme is proposed and examined to overcome the drawbacks of offline and on-demand charging schemes.

3. Declarations/Disclosures :
   I – Bui Hong Ngoc – hereby warrants that the work and presentation in this thesis are performed by myself under the supervision of Prof. Do Phan Thuan. All results presented in this thesis are truthful and are not copied from any other works. All references in this thesis - including images, tables, figures, and quotes - are clearly and fully documented in the bibliography. I will take full responsibility for even one copy that violates school regulations

   Hanoi, date month year 2021
   Author

   Bui Hong Ngoc

4. Attestation of thesis advisor:
   . . . . . .
   . . . . . .

   Hanoi, date month year 2021
   Thesis Advisor

   Assoc. Prof. Do Phan Thuan

Acknowledgments

I would like to thank all the people who have inspired and supported me to accomplish this life milestone.

My special thanks to my thesis advisors, Prof. Do Phan Thuan and Dr. Nguyen Phi Le, whose initiation and suggestions were invaluable in formulating the research questions and methodology. Without their supervision and insightful advice, I could never have gotten to the point of writing this thesis.

I am also incredibly grateful to all supervisors I have had in my university time, Dr. Tran Viet Trung, Prof. Do Phan Thuan, and Ms. Nguyen Thi Tam. The experience and knowledge I have earned not only lay the groundwork for completing this thesis but also my future career.

I would like to acknowledge my friends, Tran Huy Hung and Do Duc Thai, for their support and review during the process of doing this thesis. I also want to thank all the friends who were there for me when I need a friend. The times when looking forward to playing games, football matches, or movies are precious and keep me going.

Finally, I dedicate this thesis to my family: my parent, who always support whatever I pursue; my girlfriend, who always with me when I was at my worst, listening and understanding my complaints about everything in the world; and my sister, who does nothing but bothering me, however, keeping my life livelier. All of those form the person who I am.

A Deep Reinforcement Learning based Online Charging Scheme for Target Coverage and Connectivity in WRSNs

Abstract

In recent decades, Internet of Things (IoT) has proliferated and become a fundamental component in the 4.0 industrial revolution. As one of the most critical technologies of IoT, wireless sensor networks have also attracted significant attention from researchers for their wide range of applications. However, prolonging network lifetime is one of the most crucial challenges due to the limited energy power of the battery deployed in the sensors. The recent breakthroughs of wireless energy transfer technology and rechargeable lithium battery make the idea of a wireless rechargeable sensor network more realistic. In WRSNs, a moving vehicle equipped with wireless charging devices, namely the mobile charger (MC), is employed to charge the sensors wirelessly.

In conventional WSNs, both coverage and connectivity can be considered as measurements of Quality of Service (QoS). However, it is not given due attention in WRSNs. In this thesis, we investigate the target coverage and connectivity problem in the configuration of WRSNs. A deep reinforcement learning-based mobile charging scheme, named DRL-TCC, is proposed to tackle the target coverage and connectivity problem. The model is a deep neural network with a pointing mechanism taking the network state as input and outputting the probability of each charging action. In particular, the requesting energy threshold in this charging paradigm is omitted that enables the mobile charger to decide the next charging destination among all sensors and the depot on its own, without requests from the sensor nodes.

To evaluate the performance of the proposed DRL algorithms, various experiments have been conducted and included in the thesis. The empirical results have demonstrated that the proposed algorithm outperforms two state-of-the-art algorithms, NJNP and INMA,

by a significant margin. Moreover, the results showed the adaptability and the generalization of the proposed method in various scenarios.

Một mô hình sạc dựa trên học tăng cường sâu cho bài toán bao phủ và kết nối mục tiêu trong mạng cảm biến sạc không dây

Tóm tắt đồ án

Trong những thập kỷ gần đây, Internet vạn vật (Internet of Things - IoT) phát triển nhanh chóng và trở thành nền tảng của cuộc cách mạng 4.0. Là một trong những công nghệ quan trọng nhất của IoT, mạng cảm biến không dây thu hút được rất nhiều sự chú ý từ các nhà nghiên cứu bởi tính ứng dụng cao của chúng. Tuy nhiên, việc kéo dài thời gian sống của mạng là một thách thức lớn trong bất kỳ ứng dụng nào do nguồn năng lượng giới hạn của pin được triển khai trong các cảm biến. Những đột phá gần đây của công nghệ truyền năng lượng không dây và pin lithium có thể sạc lại là tiền đề cho sự phát triển của mạng cảm biến sạc không dây (wireless rechargeable sensor network - WRSN) để giải quyết vấn đề năng lượng của mạng cảm biến truyền thống. Trong WRSN, một phương tiện di chuyển được trang bị thiết bị sạc không dây, được gọi là máy sạc di động (mobile charger - MC), được sử dụng để sạc từ xa cho các cảm biến. Từ đó, chúng ta có thể giải quyết ràng buộc về năng lượng trong mạng cảm biến truyền thống.

Trong mạng cảm biến truyền thống, tính bao phủ và kết nối thường được coi như là thước đo của chất lượng dịch vụ (Quality of Service - QoS). Tuy nhiên, chúng chưa được quan tâm đúng mực trong mạng cảm biến sạc không dây. Trong đồ án này, chúng tôi sẽ xem xét bài toán bao phủ và kết nối mục tiêu trong mạng cảm biến sạc không dây. Để giải quyết bài toán này, chúng tôi đề xuất một thuật toán sạc dựa trên phương pháp học tăng cường sâu (deep reinforcement learning - DRL). Trong thuật toán được đề xuất, mô hình học máy sẽ nhận trạng thái của mạng sẽ được là đầu vào và đưa ra quyết định sạc tiếp theo. Đặc biệt, khác với các mô hình sạc theo yêu cầu khác, chúng tôi sẽ loại bỏ ngưỡng yêu cầu sạc mà tại đó các cảm biến sẽ không gửi yêu cầu sạc cho máy sạc. Thay vào đó,

tất cả các cảm biến đều có thể được cân nhắc trong hành động sạc tiếp theo.

Để đánh giá hiệu quả của thuật toán đề xuất, chúng tôi đã xây dựng nhiều kịch bản thí nghiệm để đánh giá theo các khía cạnh được quan tâm. Các kết quả thực nghiệm cho thấy sự vượt trội của thuật toán đề xuất so với các thuật toán được so sánh. Không những vậy, các kết quả còn cho thấy khả năng thích ứng và tổng quát hoá của thuật toán đề xuất trong các kịch bản khác nhau.

Contents

List of Figures

List of Tables

List of Acronyms

BS base station. 1, 4, 6, 7, 23

DL deep learning. 6

DRL deep reinforcement learning. 6

DRL-TCC deep reinforcement learning approach for target coverage and connectivity problem. 37–44

IA intelligent agent. 6

INMA Invalid Node Minimized Algorithm. 38, 40, 41, 43, 44

MC mobile charger. 4–9, 12, 14, 23, 24, 26–30, 33, 37, 38, 41, 42

MDP Markov decision process. 27, 28

NJNP Nearest-Job-Next with Preemption. 5, 14, 38, 40, 41, 43, 44

QoS Quality of Service. 5

RL reinforcement learning. 6

SN sensor node. 1, 4, 5, 14, 24

WRSN wireless rechargeable sensor network. 4, 6, 7, 12, 25, 37

WSN wireless sensor network. 1–4, 6, 24, 25, 42

List of Notations

B_{MC} battery capacity of the MC.
B_s battery capacity of a sensor.
γ a discount factor (discount rate).
$\mathcal{A}$ a set of legal actions.
$\mathcal{M}$ Markov decision process.
$\mathcal{P}$ a set of deployed sensors.
$\mathcal{Q}$ a set of critical targets.
$\mathcal{R}$ a reward function.
$\mathcal{S}$ a state space.
$\mathcal{T}$ a transition model.
μ charging rate.
v velocity of the MC.
ω_{move} battery capacity of a sensor.
ω energy consumption rate.
π a policy.
τ charging trajectory.
a charging action.
e residual energy of a sensor.
m number of critical targets.
n number of deployed sensors.
p_0 base station.
p a sensor.
q a critical target.
r_c communication range.
r_s sensing range.
s^D state of the depot.
s^{MC} state of the mobile charger.
s^{SN} state of a sensor.
s a state.

Chapter 1

Introduction

1.1 Overview

A wireless sensor network (WSN) can be roughly defined as a self-configured and infrastructure-less wireless network comprised of spatially dispersed and dedicated sensors for monitoring and recording the physical conditions of the environment and organizing the collected data at a central location. The sensing data will be cooperatively transmitted through the network to a base station, also known as a *sink*, where the data can be observed and analyzed. A sink or base station (BS) here acts as an interface between users and the networks, as depicted in Fig. 1.1.1.

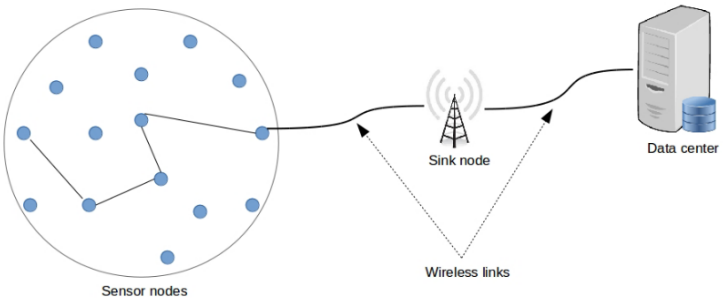


Figure 1.1.1: Wireless sensor network architecture (source: Bahri (2018)).

A wireless sensor network (WSN) can be built of a few to hundreds or thousands of low-cost, low-power sensor nodes (SNs), where each node connects to each other through the wireless network. Each SN is generally composed of four essential components, in-

cluding (1) a *sensor unit* (converting analog signal of the physical quantity into a digital signal), (2) a *preprocessing unit* (providing for computational and storage capability), (3) a *transceiver unit* (connecting the node to the network), (4) a *power unit* (usually in the form of an electrochemical battery) (Akyildiz et al., 2002). A simplified diagram of the architecture of a sensor node is shown in Fig. 1.1.2.

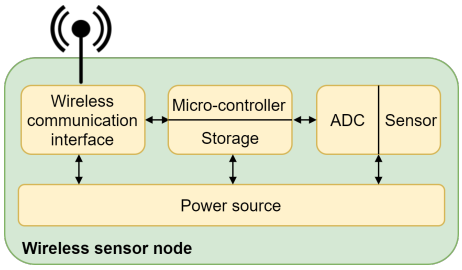


Figure 1.1.2: A block diagram of the architecture of the sensor node in the WSN.

The key advantage of WSNs is that the network can be deployed on the fly and can operate unattended, self-organizing without the need for any pre-existing infrastructure and with little maintenance. This allows random deployment in inaccessible terrains or disaster relief operations (Akyildiz et al., 2002). Combined with low deployment cost, those features give WSNs the flexibility to adapt to various applications.

Originally motivated by military applications, such as battlefield surveillance (Bokareva et al., 2006), WSNs are now widely applied in many civilian applications, including home automation (Akyildiz et al., 2002), environment monitoring (Oliveira and Rodrigues, 2011), smart agriculture (Li et al., 2011), health monitoring (Kim et al., 2007), transportation issues (Akan and Akyildiz, 2005). In the environment monitoring applications, the sensor nodes are deployed over a region where some phenomena or physical conditions are monitored. Depending on the application, sensor nodes can operate and send sensed data periodically to the sink or be triggered when an event being monitored (heat in forest fires detection (Yu et al., 2005), pressure in earthquake detection (Suzuki et al., 2007)) is detected. In the agriculture sector, one can use a wireless network to free the farmer from the maintenance of wiring in a harsh environment. Irrigation automation enables more efficient water use and reduces waste (Panchard et al., 2008). Beyond those originally designed goals, WSN-based systems now become an integral part of smart-city systems, where the factors affecting human life are closely monitored. For example, one can monitor the level of air pollution in critical points in the city and report the deterioration of the air quality (Khedo et al., 2010). Moreover, WSNs give birth to responsive traffic strategies by dynamically changing traffic lights based on real-time traffic congestion measurement (Kafi et al., 2013). Similarly, the traffic can be reduced by using systems that detect the

nearest parking space (Yoo et al., 2008). All of these help to reduce air pollution and traffic congestion, thus improving the life quality.

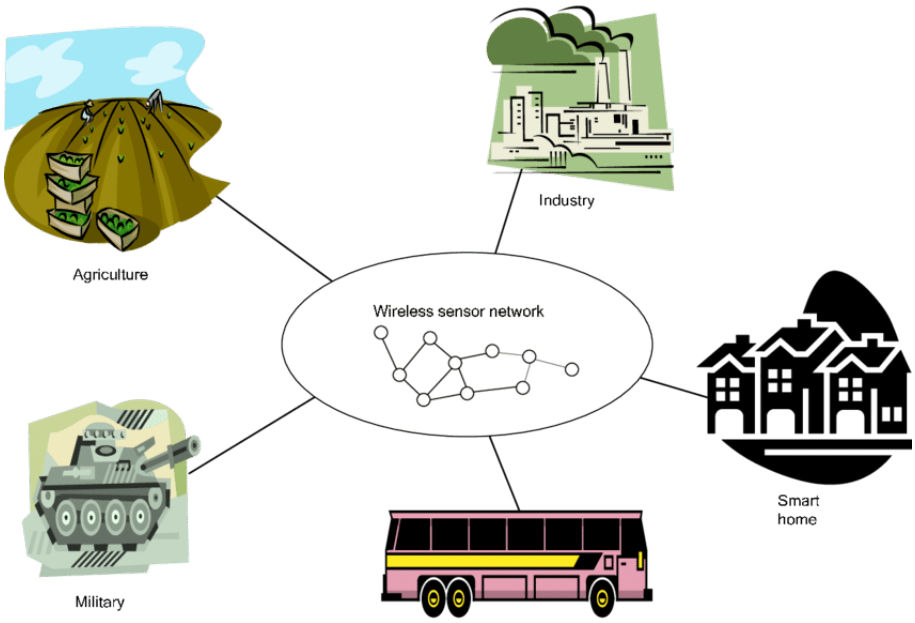


Figure 1.1.3: Applications of wireless sensor networks (source: Chen et al. (2010)).

In order to archive the flexibility to enable new applications and require no pre-existing infrastructure, WSNs have to trade-off several constraints. Sensor nodes are generally low-cost, resulting in a battery with a low capacity and usually non-renewable. When some sensors deplete their energy, the network will become fragmented; hence the data from some parts of the sensing field can no longer be extracted. In all cases for the design of any application, one of the main objectives is to keep the network alive and functional as long as possible. In the last two decades, a remarkable effort of researchers are put in designing network protocols to reduce the energy consumption in the network, which in turn prolongs the network lifetime. There are four main directions aiming to conserve energy usage in wireless sensor networks, including radio optimization (Masonta et al., 2012), data reduction (Goyal et al., 2019), sleep/wakeup schemes (Ba et al., 2013), and energy-efficient routing (Tam et al., 2019). However, none of the aforementioned solutions is universally applicable.

Furthermore, the energy conservation techniques can only extend the lifetime of the WSNs for a certain period of time. The battery will eventually be exhausted if there is no external source supplying the sensors. Another approach is to harvest energy from ambient sources (Adu-Manu et al., 2018). One can deploy an energy harvester or scavenger inside each sensor to convert energy from an external source (e.g., solar, thermal, wind, kinetic

energy; salinity gradients, radio frequency) into electrical energy recharging the battery. However, a drawback of this technique is the dependency on an ambient source which is usually unstable and uncontrollable.

The recent advances of wireless energy transfer technology (Kurs et al., 2007) and rechargeable lithium battery technology (Kang et al., 2006) have made the idea of a wireless rechargeable sensor network (WRSN) more realistic. Wireless rechargeable sensor networks (WRSNs) introduce one (or multi-) mobile charger (MC), which is equipped with a high-capacity battery and a transmitter coil. The primary task of the mobile charger (MC) is to go from one sensor to another and recharge the node's battery up to a certain level and comes back to the base to recharge its own battery when needed. With wireless charging technology, MCs (or drones) can wirelessly charge sensors in a WSN from a distance through magnetic resonance coupling (Beh et al., 2010). Using MCs (or drones) coupled with wireless charging techniques gives the potential to elongate network lifetime in inaccessible terrains (such as underground, underwater) or disaster relief operations. Ideally, the lifetime of the network would be extended indefinitely for perpetual operations. Unlike the traditional energy harvesting sensor networks (Adu-Manu et al., 2018), WRSNs offer agile, controllable, reliable, and predictable energy replenishment, thus enabling genuinely sustainable operations of sensor networks.

Those described features make WRSNs an emerging technology that has attracted much attention from researchers in recent years. One of the most critical challenges in WRSNs is constructing an efficient charging policy for the MC to meet the dynamic charging requirements of the sensors. Many algorithms are proposed to design an effective charging scheme. Those works are divided into two main categories as the *offline* and the *online* charging scheme. In the offline charging scheme (Jiang et al., 2017; Lyu et al., 2019; Ma et al., 2018; Xu et al., 2019), the energy consumption rate of each sensor is supposed to be constant and known in advance. Thus, an optimal charging trajectory is planned before the running phase. The MC then travels along the pre-optimized charging trajectory to recharge nodes in a *periodic* and *deterministic* manner. However, due to the close interaction with the surrounding environment, the nodes' energy consumption profiles show great diversity (He et al., 2013). Furthermore, this assumption is not applicable in the case of unpredictable node failures causing the network topology to be changed. That makes the pre-optimized charging trajectory no longer optimal.

On the other hand, an online charging scheme resolves that problem by making SNs periodically send their energy state to BS. Based on the long-range communication ability of MC and BS, MC can determine its charging actions based on real-time (or at least

near-real-time) information of SNs. In (He et al., 2013), the authors proposed a simple but efficient heuristic algorithm, namely Nearest-Job-Next with Preemption (NJNP), in an *on-demand* manner. The SN will send a charging request to MC if its energy is lower than a predefined threshold. The MC maintains a queue storing a list of requested SNs and then charges the spatially closest sensor in the queue. Contrary to its simplicity, NJNP showed the prospects for the on-demand charging scheme. Many later works are proposed based on NJNP results (Cao et al., 2021; Fu et al., 2015; Kaswan et al., 2018; La et al., 2020; Lin et al., 2017; 2019; Zhu et al., 2018). An example of on-demand charging is depicted in Fig. 1.1.4.

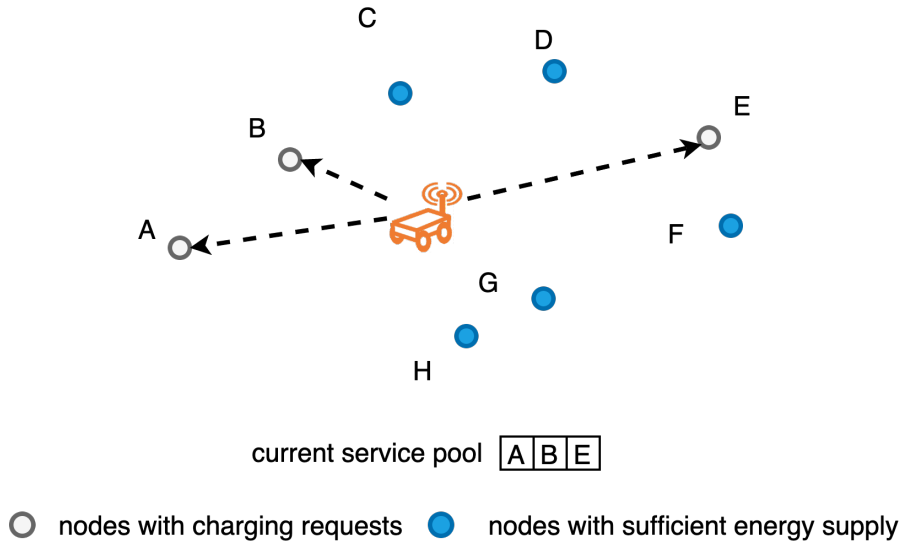


Figure 1.1.4: An example of on-demand charging scheme.

1.2 Motivation

Although many works have been proposed aiming to design an online strategy for the mobile charger, there are two significant problems with those current works. *First*, most of those works consider the role of the sensors in the network to be the same. In comparison, many applications require different roles for each sensor. For example, in target coverage and connectivity problem, there are a number of critical targets that require continuous surveillance. The sensor will be divided into two categories, *source sensor* and *relay node*. The former is the one covering and monitoring at least one target, while the latter is the sensor that only acts as a relay in the network to guarantee connectivity. In such cases, the coverage and connectivity are considered as a proportion of Quality of Service (QoS). Whereas the target coverage and connectivity is an essential problem in the conventional

WSNs (Zhao and Gurusamy, 2008), it is not given due attention in the WRSNs. *Second*, in the on-demand charging scenario, one needs to predetermine the energy requesting threshold where a sensor sends a charging request to BS if its residual energy is lower than that threshold. The performance of the on-demand charging scheme highly depends on the setting of the threshold. If the threshold is too low, the time left for MC to travel and charge the sensor is limited. On the contrary, a high threshold leads to the degradation of the efficiency of the charging scheme (Zhu et al., 2018). In the example shown in Fig. 1.2.1, after fully charging node F, the MC continuously chooses a requested node in the current service pool to charge (node A or B). If node E requests charging when the MC has just decided to go to A or B, the MC must move back and forth to serve the charging requests.

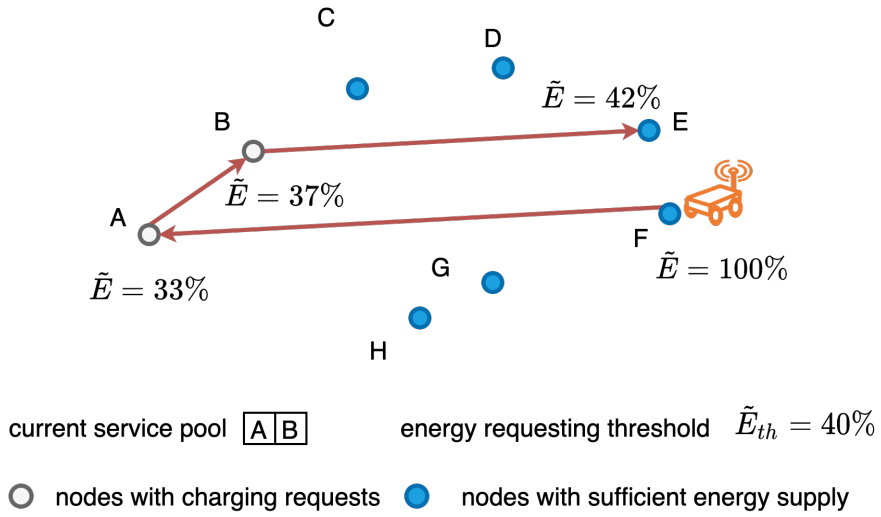


Figure 1.2.1: The drawback of on-demand charging scheme. Node E sends a charging request right after the MC decides to charge node A or B.

The recent surge of deep reinforcement learning (DRL), grounded on combining classical theoretical results in reinforcement learning (RL) with deep learning paradigm, provides a new promising opportunity to design an autonomous intelligent agent (IA) (e.g., mobile charger in this case). Different from traditional supervised or unsupervised approaches in deep learning (DL), RL's agent learns from interacting with an (uncertain) environment (instead of being instructed by a teacher/labeled dataset in supervised learning) to maximize reward function (instead of finding hidden patterns as in unsupervised learning). The incorporation of reinforcement learning and deep reinforcement learning gave birth to deep reinforcement learning (DRL) as a field of research and has created breakthroughs in the area of decision-making. Hence, the intelligent agent can learn highly complex strategies or even beat humans in the respective fields (Mnih et al., 2015; Silver

et al., 2016).

Inspired by those, this thesis aims to leverage deep reinforcement learning to tackle two aforementioned problems in an online charging scheme. In this thesis, we consider the connected target coverage problem, as one proposed by Zhao and Gurusamy (2008), in the setting of the WRSN paradigm. Specifically, a number of critical targets are required to be continuously monitored by a number of randomly scattered sensors. The objective is to maximize the network lifetime of the WRSN subject to the conditions: (1) each target is covered by at least one sensor, the sensor monitoring the target is called a source, (2) from each source to sink, there must exist at least one route traversing through only the active sensors. Furthermore, to overcome the limitation of the on-demand scheme, a truly online scheme, in which the requesting energy threshold is omitted, is proposed. The MC will choose the next charging destination among all sensors and the depot on its own, without requests from the sensor nodes. Finally, a deep reinforcement learning model deployed in the MC will be introduced to make charging decisions based on the network's status. This model will be trained through interactive experiences with the simulated environment so that the network's lifetime is maximized.

1.3 Problem statement

This thesis considers the deployment of a wireless sensor network as presented in the work of Zhao and Gurusamy (2008), which will be powered by a mobile charger (MC) as in the conventional wireless rechargeable sensor network (Zhu et al., 2018). A sensor network includes a base station (BS) denoted as p_0, a set of n randomly distributed sensor nodes $\mathcal{P} = \{p_1, p_2, ..., p_n\}$, and a set of m targets $\mathcal{Q} = \{q_1, q_2, ..., q_m\}$ which are required to be continuously monitored. Two sensors are said to be connected if their Euclidean distance is less than communication range r_c. Each sensor has a sensing area determined by its location $p_i = (x_i, y_i)$ and a sensing range r_s. Any target located in the sensing area of a sensor (say p_i) could be monitored. The sensor p_i then will be called a *source* sensor (covering at least one target). A source sensor will perform the monitoring task, periodically generate sensed data messages and transmit data to sink through multiple-hop communication. On the contrary, a sensor, which does not cover any target, may act as relay nodes to forward sensed data to the sink. When a sensor depleting its energy, it will deactivate itself and wait to be replenished. A network's state is considered *coverage* and *connectivity* if it satisfies the two following constraints: (1) each target is covered by at least one source sensor. (2) from each source to sink, there must exist at least one route

traversing active sensors only. Thus, the *network lifetime* is defined as a period of time the coverage and connectivity properties are held.

A mobile charger (MC), equipped with a high-capacity battery and a transmission coil, is employed to move around to charge sensors wirelessly or recharge its own energy when needed. An MC is represented by four factors that directly affect the efficiency of the charging model, including the battery capacity (B_{MC}), the traveling speed v, the charging rate to sensors (μ), and the energy consumption rate of the MC on one unit distance (ω_{move}). The ultimate objective is to design a charging strategy for MC to maximize the lifetime of the sensor network. In this work, the initial sensor network is assumed to satisfy both coverage and connectivity properties. An illustration of network structure is shown in Fig. 1.3.1.

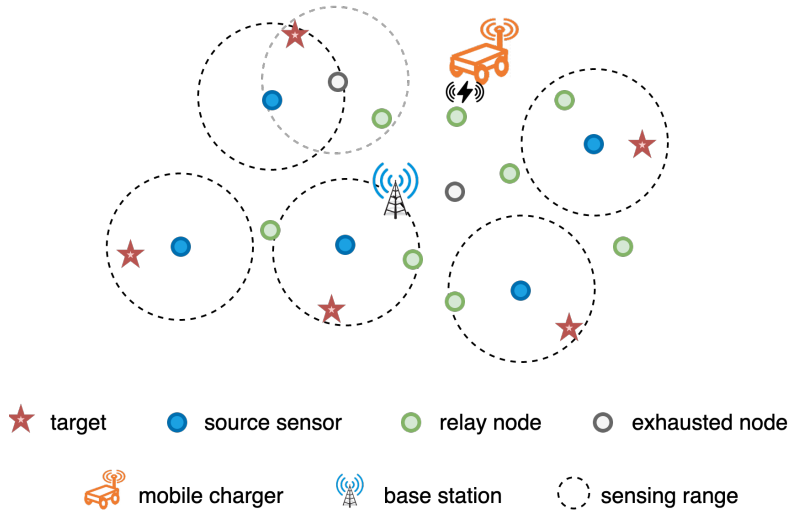


Figure 1.3.1: An illustration of the connected target coverage problem in wireless rechargeable sensor network.

It is worth noting that, different from the on-demand charging scheme, the requesting energy threshold is omitted in the charging paradigm. The MC will choose the next charging destination among all sensors and the depot on its own, without requests from the sensor nodes. Omitting the predefined parameter relaxes the potential charging destinations at a step to be any sensors, thereby expanding the potential trajectory space. However, it requires the charging algorithm to strike a balance between extending the network lifetime and the efficiency of the MC.

1.4 Thesis contributions and organization

The main contributions of this thesis can be summarized as follows:

- We investigate the target coverage and connectivity problem in the wireless rechargeable sensor network. In this paradigm, a novel online charging scheme is proposed by omitting the requesting energy threshold. The MC then makes charging decisions based on the latest state of the entire sensors in order to maximize the lifetime of the network.
- We propose a deep neural network model that takes the current status of the MC itself and all sensors, including their position, current energy, and energy consumption rate, to make charging decisions. We use the reinforcement learning method to train our model by trial and error.
- Finally, to demonstrate the efficiency of our proposed model, extensive experiments are conducted to compare the proposed approach to other state-of-the-art methods. The result shows the superiority of the proposed method over the others.

The remainder of the thesis is organized as follows: In *Chapter 2*, we briefly introduce some background of deep reinforcement learning and the attention mechanism used in our model. *Chapter 3* describes our simulation of the network environment. We present the proposed algorithms for the mobile charging scheduling problem in *Chapter 4*. The experiments and results are presented in *Chapter 5*. We also discuss some perspectives about our evaluation in this chapter. Finally, *Chapter 6* concludes the thesis and gives recommendations for future works.

Chapter 2

Literature review

Since this thesis aims to investigate the target coverage and connectivity problem in the configuration of WRSNs, the following section presents a literature review on the target coverage and connectivity problem. A survey of current approaches to charging schemes in WRSNs is shown in Section 2.2.

2.1 Target coverage and connectivity in WSNs

Prolonging network lifetime is one of the most crucial goals in any application in the wireless sensor network. There are many definitions of sensor network lifetime in the literature (Dietrich and Dressler, 2009), and n-of-n lifetime is one of the most used. In this definition, the network lifetime T_n^n ends as soon as the first node fails, thus:

$$T_n^n = \min_{v \in V} T_v. \tag{2.1}$$

where V is a set of sensor nodes, and T_v is the lifetime of node v. This is a favorable definition because of its straightness to compute, and the algorithms running in the network do not have to cope with topology changes. Nevertheless, it has some disadvantages. In most cases, the lifetime calculated by this metric will not be enough for the meaningful evaluation of sensor network applications. Because most networks whose node has several direct neighbors with the same sensing equipment are capable of dealing with the failure of one node in such case, but the metric cannot represent this kind of network redundancy. This metric can only be reasonably used if all nodes are of equal importance and critical to the network. Furthermore, the network can still provide helpful information for a long

time after the first node depletes its energy. This metric is also insufficient for evaluating scenarios that consider hardware failures because randomly distributed hardware failures might occur very early and thus distort the lifetime measure.

There are several variants of the T_n^n metric, which are stated in (Deng et al., 2005; Hellman and Colagrosso, 2006), to overcome some disadvantages of the T_n^n metric. However, these variants are very limited and insufficient. Defining network lifetime solely based on the number of alive nodes makes neither the ability to communicate measurements nor the ability to sense events in the region of interest are incorporated into these metrics.

Based on the specific characteristics of sensor networks, defining the network lifetime as the time that sensor nodes cover the region of interest is more practical. The coverage can be defined in different ways, depending on the composition of the region of interest and the achieved redundancy of the coverage. Area or volume coverage, target coverage, and barrier coverage are three typical coverage problems. However, defining network lifetime based on the achieved coverage is unsatisfactory for most practice scenarios since there is no guarantee that the monitored data can ever reach a sink node.

Another group of metrics considers the connectivity of the network. Blough and Santi (2002) define the lifetime as the minimum time when either the percentage of alive nodes or the size of the most significant connected component of the network drop below a specified threshold. Cărbunar et al. (2006) treat connectivity as the percentage of nodes with a path to the base station. The lifetime is defined as the number of successful data gathering trips in (Olariu and Stojmenovic, 2006) and (Giridhar and Kumar, 2005) add ”without any node running out of energy” constraint to the number of trips possible. Although integrating connectivity metrics is undoubtedly plausible, one needs to consider connectivity towards a base station, not just connections between arbitrary sensor nodes.

Several authors combine the coverage-based metrics with connectivity metrics. The network lifetime metric is defined by the time when either the coverage or the connectivity drop below a predefined threshold with coverage is measured in terms of α-coverage - requires that only a given percentage α of the region of interest is covered by at least one sensor, and connectivity is measured by the packet delivery ratio at the sink node. Cardei et al. (2005) define the lifetime to be the time until either coverage or connectivity is lost but the exact definition of coverage and connectivity is unmentioned.

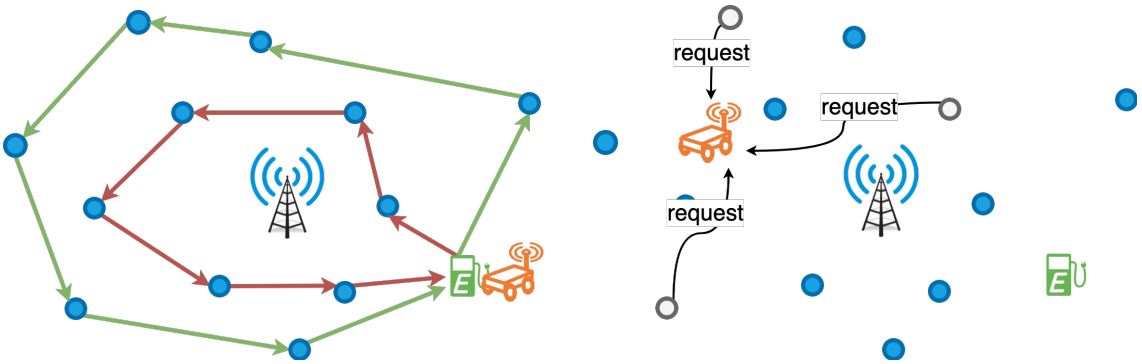
Practical applications require a high degree of reliability, so essential equipment or processes must be prioritized and uninterrupted. Since sensor nodes in WSNs are typically deployed at random, target coverage and connectivity among all sensors and the

sink node is a fundamental issue in target monitoring, which aims to cover the specified targets by a subset of the deployed sensor nodes with minimum resource consumption. Zhao and Gurusamy (2008) consider the connected target coverage (CTC) problem with the objective of maximizing the network lifetime by scheduling sensors into multiple sets, each of which can maintain both target coverage and connectivity. Li et al. (2007) form k-connected coverage of targets the minimal active nodes by addressing the k-connected augmentation problem and two heuristic algorithms to solve it. Qin et al. (2018) propose an optimized and lightweight energy-efficient connected coverage heuristic (OECCH) algorithm.

However, the target coverage and connectivity in the wireless rechargeable sensor network are not given due attention. To the best of our knowledge, there are only two related works. In (Zhou et al., 2017), the authors relax the strictness of perpetual operation by allowing some sensors to temporarily run out of energy while still maintaining target k-coverage in the network. The authors build a theoretical model and propose λ-GTSP Charging Algorithm to determine the optimal number of sensors to be charged in each cluster to maintain k-coverage in the network and derive the route for MC to charge them. La et al. (2020) introduce a novel on-demand charging scheme for WRSNs that optimize the charging time at each MC's charging location. They also leverage the Q-learning technique to maximize the number of monitored targets.

2.2 Energy conservation in WRSNs

One of the most critical challenges in WRSNs is constructing an efficient charging policy for MCs to meet dynamic charging requirements of the sensors. Many proposals are divided into two main categories as the *offline* and the *online* charging scheme. Fig. 2.2.1 illustrates an example of two charging schemes. Since the main topic of this thesis is about the online charging scheme, we only brief some of the most significant works of the offline scheme in the following subsection. The extensive survey on the online charging scheme is presented in Subsection 2.2.2.



(a) Offline charging scheme **(b)** Online charging scheme

Figure 2.2.1: A comparison of offline and online charging scheme.

2.2.1 Offline charging scheme

Although most existing mobile charging works belong to the offline scheme, those studies still have a critical drawback by making a solid and unrealistic assumption about constant energy consumption rate. Lyu et al. (2019) propose a periodic charging planning for mobile Wireless Charging Equipment with limited traveling energy. They propose a Hybrid Particle Swarm Optimization Genetic Algorithm (HPSOGA) because of the NP-hardness of the problem. In (Jiang et al., 2017), the authors jointly consider charging tour planning and MC depot positioning for large-scale WSNs. Their method consists of charging tour planning, candidate depot identification and reduction, depot deployment, and charging tour assignment. The charging scheme also considers the association between the MC charging cycle and the sensor nodes' lifetime. Ma et al. (2018) aim to minimize the sensor energy expiration time and the charging tour length of the mobile charger. They develop an approximation algorithm for the charging utility maximization problem if the energy consumption of the mobile charger on its charging tour is negligible and an efficient heuristic through a non-trivial reduction from a length-constrained utility maximization problem otherwise. Xu et al. (2019) deal with multiple mobile chargers for charging sensors. They formulate a novel longest delay minimization problem that is NP-hard and devise an approximation algorithm for the problem.

2.2.2 Online charging scheme

In (He et al., 2013), He et al. first lay the theoretical foundation for the on-demand mobile charging problem, where each sensor node requests charging from the mobile charger

when its energy is used up. The authors propose a simple but efficient heuristic algorithm NJNP in an *on-demand* manner. The SN will send a charging request to MC if its energy is lower than a predefined threshold. The MC maintains a queue storing a list of requested SNs and then charges the spatially closest sensor in the queue. Contrary to its simplicity, NJNP shows the prospects for the on-demand charging scheme. To overcome low charging request throughput or successful rate, which causes a major bottleneck in traditional scheduling strategies for on-demand WRSNs architecture because of undervaluing the unbalanced influences of spatial and temporal constraints posed by charging requests, Lin et al. (2019) introduce a Double Warning Thresholds with Double Preemption (DWDP) charging scheme, in which these thresholds are used when residual energy levels of sensor nodes fall below certain thresholds. Warning thresholds can be used to adapt charging priorities of different sensors, inform the following recharge deadlines, along with supporting preemptive scheduling.

Fu et al. (2015) propose a novel Energy Synchronized Mobile Charging (ESync) protocol, which simultaneously reduces the charger travel distance and the charging delay of sensor nodes. To overcome Traveling Salesman Problem (TSP)-based solutions' limitation that node's energy consumption are diverse, they construct a set of nested TSP tours based on their energy consumption, and only low-energy nodes are involved in each charging round. Moreover, the concept of energy synchronization is proposed to synchronize the charging requests sequence of nodes with their sequence on the TSP tours. In (Lin et al., 2017), the authors develop a temporal-spatial charging scheduling algorithm (TSCA) for the on-demand charging architecture. Their purpose is to minimize the number of fail nodes while maximizing energy efficiency to lengthen network lifetime. A feasible movement solution, an underlying path planning algorithm, optimizations, a node deletion algorithm, and a node insertion algorithm are introduced.

Kaswan et al. (2018) present a Linear Programming (LP) formulation for the MC scheduling problem in an on-demand wireless rechargeable sensor networks (WRSNs), then introduce an efficient solution based on a gravitational search algorithm (GSA) with a novel agent representation scheme and a potent fitness function. Zhu et al. (2018) consider the dynamic energy consumption rate of the sensor based on both its history statistics and real-time energy consumption. Two efficient online charging algorithms named PA and INMA are proposed. PA takes the next charging node according to the charging probability of the requesting nodes, whereas INMA always selects the nodes that make the least number of other requesting nodes endure energy deficiency as the charging candidates. For high charging efficiency, the node with the shortest time to finish the charging will be selected as the next charging node if the candidate set has more than one node.

In recent years, there are some reinforcement learning-based algorithms proposed for designing the on-demand charging scheme. Considering that sensors can be charged multiple times in one charging tour into account, Cao et al. (2021) present a new metric: charging reward, to measure the quality of sensor charging. The problem becomes scheduling the mobile charger to refill the sensors, such that the sum of charging rewards collected by the mobile charger on its charging tour is maximized. The sum is subject to the energy capacity constraint on the mobile charger and the charging time windows of all sensor nodes. A deep reinforcement learning technique is developed to attain the moving path for the mobile charger. In (La et al., 2020), the authors propose a novel on-demand charging scheme for the target coverage and connectivity problem in which the tabular Q-learning method is used to maximize the number of monitored targets.

Chapter 3

Preliminaries

3.1 Deep reinforcement learning

3.1.1 Reinforcement learning and key concepts

Reinforcement learning (RL) is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning. In RL, we have an agent in an environment, and this agent can obtain some rewards by interacting with the environment. The goal of RL is to help the agent learn a good strategy to maximize the rewards through trials and feedback. Thus, the agent can slowly adapt to changes in the environment to maximize future rewards. Fig. 3.1.1 sketches a basic paradigm in reinforcement learning.

There are some important key concepts in RL. First, the **environment** is the whole surrounding area that will react based on the agent's actions. Whenever the agent take

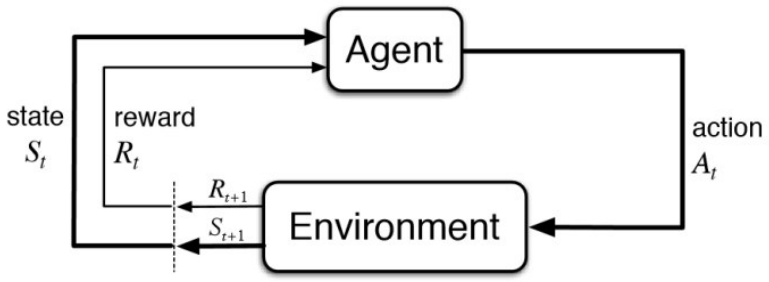


Figure 3.1.1: Reinforcement learning paradigm. (source: Sutton and Barto (2018))

an **action** $a \in A$, the environment will switch to a new **state** $s \in S$. One action can lead to one of many states, which is decided by transition probabilities between states P. The environment also returns a **reward** $r \in R$ as feedback for that action.

The reward function and transition probabilities of the environment are defined by a **model** which we may or may not be identified. This thesis focuses on the latter case when complete information about the environment has not been achieved.

In order to maximize the total rewards, the agent has to rely on two main factors: its **policy** $\pi(s)$, which helps to choose the optimal action in a certain state, and a **value function** $V(s)$, which evaluates a state by predicting the future rewards received in that state if following the policy.

The interaction between the agent and the environment involves a sequence of actions, which makes the environment switch state continuously. At each time step $t \in \{1, 2, ..., T\}$, the environment is in state S_t, the agent decides to take action A_t, which returns reward R_t. This sequence can be described as an **episode** that ends at a terminal state S_T:

$$S_1, A_1, R_2, S_2, A_2, ..., S_T.$$

Model. The model gives all necessary information about an environment: transition probability function P and reward function R. Both parts can be modeled as follows:

At state s, the agent chooses an action a, which leads to a new state s' and receives a reward r. This is called a **transition** step (s, a, s', r). The probability of this transition is formulated in function P:

$$P(s', r|s, a) = \mathbb{P}[S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a]. \tag{3.1}$$

The state transition function (transit from state s to s' after action a) can be defined as:

$$P^a_{ss'} = P(s|s', a) = \mathbb{P}[S_{t+1} = s'|S_t = s, A_t = a] = \sum_{r \in R} P(s', r|s, a). \tag{3.2}$$

The reward function R to predict reward for action a:

$$R(s, a) = \mathbb{R}[R_{t+1}|S_t = s, A_t = a] = \sum_{r \in R} r \sum_{s' \in S} P(s', r|s, a). \tag{3.3}$$

Policy. Policy π is the agent's behavior function that decides the action a to take in state s:

- Deterministic: $\pi(s) = a$.
- Stochastic: $\pi(a|s_{=}\mathbb{P}_{\pi}[A = a|S = s]$.

Value function. Value function predicts future reward that might be obtained by a state to evaluate its potential. The future reward of a step t is called **return** G_t and is represented by a total sum of discounted rewards after some consecutive states:

$$G_t = R_{t+1} + \gamma R_{t+2} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}. \tag{3.4}$$

The discounting factor $\gamma \in [0,1]$ reduces the future rewards depending on the particular problem: highly uncertain future, no immediate benefits, mathematical convenience, infinite loop avoidance, etc.

The expected return of state s at time t is the **state-value**:

$$V_{\pi}(s) = \mathbb{E}_{\pi}[G_t|S_t = s]. \tag{3.5}$$

Similarly, the quality of a state-action pair is the **action-value**:

$$Q_{\pi}(s,a) = \mathbb{E}_{\pi}[G_t|S_t = s, A_t = a]. \tag{3.6}$$

The above two functions can be connected by the probability distribution of the target policy π:

$$V_{\pi}(s) = \sum_{a \in A} Q_{\pi}(s,a)\pi(a|s). \tag{3.7}$$

Finally, the different returns between action and state make a profit for the action, also known as the action **advantage** value:

$$A_{\pi}(s,a) = Q_{\pi}(s,a) - V_{\pi}(s). \tag{3.8}$$

Optimal value and policy. The optimal value function produces the maximum return:

$$V_*(s) = \max_{\pi} V_{\pi}(s), Q_*(s,a) = \max_{\pi} Q_{\pi}(s,a). \tag{3.9}$$

The optimal policy achieves optimal value functions:

$$\pi_* = arg\max_{\pi} V_\pi(s), \pi_* = arg\max_{\pi} Q_\pi(s,a). \tag{3.10}$$

Thus, $V_{\pi_*}(s) = V_*(s)$ and $Q_{\pi_*}(s,a) = Q_*(s,a)$.

Bellman equations. Decomposing the value function into the immediate reward plus the discounted future values:

$$V(s) = \mathbb{E}[R_{t+1} + \gamma V(S_{t+1})|S_t = s], \tag{3.11}$$

$$Q(s,a) = \mathbb{E}[R_{t+1} + \gamma \mathbb{E}_{a\sim\pi} Q(S_{t+1}, a)|S_t = s, A_t = a]. \tag{3.12}$$

3.1.2 Markov Decision Process

RL problems can be represented as **Markov Decision Processes** (MDPs) since they are very similar. In both models, given the present state, the future and the past are **conditionally independent** since the current one contains all necessary statistics:

$$P[S_{t+1}|S_t] = P[S_{t+1}|S_1, ..., S_t]. \tag{3.13}$$

A MDP also comprises of five elements $M = \langle S, A, P, R, \gamma \rangle$, with the same roles as the key concepts of RL:

- S - set of states
- A - set of actions
- P - transition probability function
- R - reward function
- γ - discounting factor for future rewards

3.1.3 Policy Gradient method

There are many approaches for solving RL problems. Most of the common methods such as dynamic programming, Q-learning or SARSA methods focus on learning the state/action value function before selecting actions accordingly. In contrast, policy gradient methods learn the policy directly with a parameterized function with respect to θ, $\pi(a|s;\theta)$. We train the algorithm to maximize the reward function (which is the opposite of the loss function).

The reward function is defined as *the expected return* and formulated as follows. In discrete space with S_1 as the initial starting state:

$$\mathcal{J}(\theta) = V_{\pi_\theta}(S_1) = \mathbb{E}_{\pi_\theta}[V_1]. \tag{3.14}$$

Besides, in continuous space:

$$\mathcal{J}(\theta) = \sum_{s\in\mathcal{S}} d_{\pi_\theta}(s) V_{\pi_\theta}(s) = \sum_{s\in\mathcal{S}} \Big(d_{\pi_\theta}(s) \sum_{a\in\mathcal{A}} \pi(a|s,\theta) Q_\pi(s,a) \Big), \tag{3.15}$$

where $d_{\pi_\theta}(s)$ is stationary distribution of Markov chain for π_θ.

In this case, we use *gradient ascent* (as opposed to *gradient descent*) to find the optimal θ that gives the highest return. Naturally, these policy-based methods are more efficient in continuous space since the other value-based ones have to estimate an infinite number of actions and states.

In order to compute the gradient, θ can be perturbed by a small amount ε in the k-th dimension:

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta_k} \approx \frac{\mathcal{J}(\theta + \varepsilon u_k) - \mathcal{J}(\theta)}{\varepsilon}. \tag{3.16}$$

After some transformations with theories in Sutton and Barto (2018), we have the final equation:

$$\nabla \mathcal{J}(\theta) = \mathbb{E}_{\pi_\theta}[\nabla \ln \pi(a|s,\theta) Q_\pi(s,a)]. \tag{3.17}$$

This result is named *Policy Gradient Theorem* which becomes the foundation for various policy gradient algorithms.

3.1.4 Actor-critic method

In reinforcement learning, the value-based methods (such as Q-learning) evaluate the optimal cumulative reward and aim at finding an optimal policy π^* by obtaining an optimal value function in Eq. 3.9. Policy-based methods (such as REINFORCE) aim to estimate the optimal policy directly by optimizing a parametric function representing the policy. Actor-critic methods are hybrid approaches that combine the advantages of value-based and policy-based methods. In actor-critic methods, two networks are maintained. One is used for learning value function, namely the *critic*, and another learns the mapping between state and actions directly, known as the *actor*. The critic is used to criticize the actions made by the actor, and the actor adjusts its parameters in the direction suggested by the critic. The gradient of the actor is now given by:

$$\nabla \mathcal{J}(\theta) = \mathbb{E}_{\pi_\theta}[\nabla \ln \pi(a|s,\theta)(R_t + \gamma V_\psi(s) - V_\psi(s))], \tag{3.18}$$

where $V_\psi(s)$ is the estimated value of state s given by the critic.

3.2 Attention mechanism

In the last decade, *encoder-decoder* is one of the most prominent architectures in deep learning, which was originally introduced to solve the problem of mapping fixed-length input to output in sequence-to-sequence learning. In the vanilla encoder-decoder, a variable-input sequence is encoded to an internal, fixed-dimensional representation. The RNN-based decoder then uses this representation to produce a variable-length output gradually until a termination criterion is detected. A critical disadvantage of this fixed-length representation is the incapability of remembering long sentences. *Attention mechanism* emerged as a solution to this limitation of the traditional encoder-decoder architectures. Essentially, attention enables the decoder to use any of the encoder's hidden states instead of using the fixed-length representation produced by the encoder at the end of the input sequence. The idea is to create shortcuts that combine the entire input sequence to a context vector. These shortcuts are weighted to represent how much attention is devoted to each input. Mathematically, let us denote the encoder and decoder hidden states with

$(e_1, e_2, ..., e_n)$ and $(d_1, d_2, ..., d_n)$ a context vector at decoding time i is given by:

$$c_i = \sum_{j=1}^{n} a_j^i e_j, \tag{3.19}$$

where a^i is an *alignment* of the input vector, which is calculated by:

$$a_j^i = softmax(u_j^i), \quad j \in \{1, 2, ..., n\}, \tag{3.20}$$

where:

$$u_j^i = f(W_1 e_j + W_2 d_i), \quad j \in \{1, 2, ..., n\}. \tag{3.21}$$

This context vector c is later concatenated with decoder state d to make a prediction or compute a hidden vector for next steps of the recurrent model.

3.3 Pointing mechanism

Pointing mechanism is a technique first proposed in (Vinyals et al., 2015) with the aim of producing discrete outputs that correspond to positions in the input. For example, in the combinatorial problem - Travel Sailing Problem, the solution is a permutation of the input positions. In (Vinyals et al., 2015), the authors proposed a Pointer network which is an encoder-decoder LSTM model. The input, including a sequence of the node's position, is encoded by an LSTM encoder. In the decoder, instead of blending the encoder hidden states e_j into a context vector c at each decoder step, a reduction of attention mechanism is used to point to a member of the input sequence to be selected as the output:

$$u_j^i = f(W_1 e_j + W_2 d_i), \quad j \in \{1, 2, ..., n\}, \tag{3.22}$$

$$a_j^i = softmax(u^i), \quad j \in \{1, 2, ..., n\}, \tag{3.23}$$

where a_j^i is considered as the probability to select input j in the decoder step i. The node with the highest probability is chosen to be visited next. The procedure is iteratively repeated to obtain the final solution.

Chapter 4

System model

In this chapter, we describe the detail of the target coverage and connectivity problem in the configuration of a WRSN. The detail of the network structure, the energy model, the routing strategy, and the charging model are presented in the following sections.

4.1 Network structure

We consider the deployment of a wireless sensor network as in (Zhao and Gurusamy, 2008), which will be powered by a mobile charger (MC) as in the conventional wireless rechargeable sensor network Zhu et al. (2018). A sensor network includes a base station (BS) denoted as p_0, a set of n randomly distributed sensor nodes $\mathcal{P} = \{p_1, p_2, ..., p_n\}$, and a set of m targets $\mathcal{Q} = \{q_1, q_2, ..., q_m\}$ which are required to be continuously monitored. All nodes are deployed in a two-dimensional service area. The base station is a sink node assumed to be placed in the center of the sensing field with an unlimited power supply. Meanwhile, all sensors have a rechargeable lithium battery with the same capacity B_s and initial energy. Each sensor is also equipped with both the radio module and the sensing module to help the sensor act as a relay node for others and as an originator (source node) to generate sensing data.

Two sensors are said to be connected if their Euclidean distance is less than the communication range r_c. Each sensor has a sensing area determined by its location $p_i = (x_i, y_i)$ and a sensing range r_s. Any target located in the sensing area of a sensor could be monitored. A sensor is called a *source node* if it covers at least one target. A source sensor will perform the monitoring task, periodically generate sensed data messages. The sensing

data gathered by source sensors is transmitted to the sink by multi-hop communication over other active sensors. The sensors will also periodically send their residual energy to the sink, which in turn are forwarded to the mobile charger (MC) through long-range communication ability. When a sensor depleting its energy, it will deactivate itself and waiting to be replenished.

A network's state is considered *coverage* and *connectivity* if it satisfies the two following constraints: (1) each target is covered by at least one source sensor. (2) from each source to sink, there must exist at least one route traversing through only active sensors. Thus, the *network lifetime* is defined as a period of time that the coverage and connectivity properties are held.

4.2 Energy model

In spite of the fact that there are numerous models of energy dissipation in WSNs studied with different assumptions, the radio module is always the main component that causes battery depletion of sensor nodes (Rault et al., 2014). In this work, we use the same energy model as in (Gawade and Nalbalwar, 2016), which accounts for the dissipated energy at both the receiver and transmitter during transmission and omits the dissipated energy of the sensing and computing unit. The amount of consumption is dependent on the number of SNs sending data through the node, whereas the transmission power is proportionate to the distance of the transmit-receive pair. Specifically, the free space model (d^2 power loss) is used for proximal transmissions, and the multi-path fading model (d^4 power loss) is considered for large distance transmissions. The energy dissipated by the transmitter for transmitting an ℓ-bit packet to a distance d is given by:

$$\tilde{E}_t(d) = \begin{cases} \ell\varepsilon_{elec} + \ell\varepsilon_{fs}d^2, & \text{if } d \leq d_0 \leq r_c, \\ \ell\varepsilon_{elec} + \ell\varepsilon_{mp}d^4, & \text{if } d_0 < d \leq r_c, \\ \infty, & \text{if } r_c < d, \end{cases} \tag{4.1}$$

where $d_0 = \sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{mp}}}$ is the distance threshold for swapping amplification models and r_c indicates the range with which a node can communicate. In other words, there will be no connection established among the nodes that are out of this range.

The energy consumption of the receiver to receive an ℓ-bit packet is calculated as

Table 4.2.1: Network constants of the energy model.

Parameter	Value	Unit
ε_{elec}	50	nJ/bit
ε_{fs}	10	$pJ/bit/m^2$
ε_{mp}	0.0013	$pJ/bit/m^4$

follows:

$$\tilde{E}_r = \ell \varepsilon_{elec}. \tag{4.2}$$

The dissipated energy of a node receiving η packets and transmitting them to the parent node is calculated by the following formula:

$$\tilde{E}(\eta, \zeta, d) = \eta \tilde{E}_r + (\eta + \zeta)\tilde{E}_t(d), \tag{4.3}$$

where $\tilde{E}_t(d), \tilde{E}_r$ are calculated as in Equation 4.1 and 4.2, respectively. ζ is equal to 1 if the node is a source node, and 0 otherwise. d is the transmission distance. The network parameters shown in Table 4.2.1 are set as in (Wu and Liu, 2013). ε_{elec} is the energy dissipated per bit to run the transmitter or receiver circuit and is commonly set at 50 nJ/bit. ε_{fs} and ε_{mp} are the energy expenditures of transmitting one bit data at a short and a long distance respectively to achieve an acceptable bit error rate. Their common values are 10 $pJ/bit/m^2$ and 0.0013 $pJ/bit/m^4$, respectively.

4.3 Routing strategy

In conventional WSNs, the sensor nodes are typically randomly dispersed over a wide area and contain a limited power battery which can be depleted. It causes the network topology dynamically changed. Therefore, the routing protocol in the WSN must process the self-organizing capabilities. Moreover, the routing protocol highly affects network delay, throughput, and even the energy consumption of the sensors. Many studies have been proposed attempting to design a self-organizing protocol that provides a better quality of the network service and elongates the network lifetime (Liu et al., 2012; Singh et al., 2010; Youssef et al., 2002).

In WRSN, sensors are equipped with a rechargeable battery and can be wirelessly replenished by a mobile charger. Therefore, the most crucial task of designing a network protocol in WRSN is to guarantee self-organizing capabilities and the quality of service.

In this work, we adopt the centralized routing protocol as the one proposed by Tajeddine et al. (2012). The sensors periodically send their residual energy status towards the base station. Based on the latest status of the network, a Dijkstra's algorithm is performed to find the Shortest Path Tree (SPT) for every alive node towards itself. The SPT can be formed as an array where the value of the element i^{th} is the information of the parent node of node i in the SPT. This array is broadcasted to all nodes in the network. Each node will use the corresponding information to transmit or relay data to its parent in the SPT.

4.4 Charging model

An MC is equipped with a high-capacity battery and a transmission coil in our charging scheme, moving around to charge sensors wirelessly or going back to a depot to recharge its own energy when needed. There are four factors that directly affect the efficiency of the charging model, including the battery capacity (B_{MC}), the traveling speed (v), the charging rate to sensors (μ), and the energy consumption rate of the MC on one unit distance (ω_{move}). For the sake of simplicity, the velocity of the MC is assumed to be constant, and the path connecting two points in the sensor field is a straight line. A charging trajectory can be represented as a sequence of the MC's actions $\tau = \{a_{t_1}, a_{t_2}, .., a_{t_k}\}$ where a charging action of a sensor consists of two phases: (1) moving into its proximity and (2) fully charging the sensor. The time to fully charge a sensor is determined as follows:

$$t_{a_i} = \frac{B_s - e_i}{\mu - \omega_i}, \tag{4.4}$$

where B_s is the battery capacity; e_i and ω_i refers to the remaining energy and the energy consumption rate of sensor i, respectively.

Different from the on-demand charging scheme, we eliminate the requesting energy threshold. Initially, MC is located at the position of the depot. MC will choose the following action among all sensors or going back to the depot based on the latest status of the network and its own status. Each charging decision will be made when the previous charging action is completed.

Chapter 5

Deep reinforcement learning-based mobile charging scheme

Recent advances in deep reinforcement learning, grounded on combining classical theoretical results in reinforcement learning with the deep learning paradigm, set the stage for breakthroughs in decision-making. In this chapter, we propose a novel online mobile charging scheme based on deep reinforcement learning in which the MC is considered as an intelligent agent making charging decisions and moving around to replenish sensors or itself. The intelligent agent is modeled as a *policy* π that maps the network's states to charging decisions to be taken in those states. The MC is supposed to observe the current state of the network through long-range communication with the base station. Since the fully charging model is considered, the actions are simplified to be the next charging destinations. We model the learning agent by a deep neural network which takes the network's state as the input and produces the probabilities for each charging destination. This model is trained by interacting with the simulated environment, adjusting the policy by the return from the environment. The final goal is to prolong the lifetime of the given sensor network.

In the following sections, we first formally define the Markov decision process (MDP) for representing the interaction between the reinforcement learning agent and its environment. The detail of our deep neural network of the policy is presented in Section 5.2. Finally, we describe the Policy Gradient method for training the model in Section 5.3.

5.1 Learning model construction

We formally describe the Markov decision process (MDP) model for representing the interaction between the reinforcement learning agent and its environment as describe in Fig. 3.1.1. A mathematical model of MDP is typically defined as $\mathcal{M} = < \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma >$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ denotes the set of legal actions, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ represents the transition model, $\mathcal{R}$ refers to the reward function, and $\gamma \in [0, 1]$ is a discount factor for future rewards.

State. The state information represents the current status of the network and mobile charger which is divided into two groups, *static* and *dynamic* elements. The static elements contain information related to the properties of the mobile charger, the depot, or the sensors such as the position of sensors, battery capacity, the number of covering-targets. The dynamic elements include the residual energy of devices, the current position of the MC, and the estimated energy consumption rate of each sensor. Formally, we define a state $s \in \mathcal{S}$ as a tuple of $(s^{MC}, s^{D}, \bar{s}^{SN})$, where s^{MC} is a tuple of the static and dynamic information of the mobile charger, s_0^D contains the location of the depot in the sensor field, and $\bar{s}^{SN} = \{s_i^{SN}, i \in [1, n]\}$ is a sequence of tuples containing static and dynamic information of each sensor.

Action. In our charging scheme, the MC performs a charging action by two stages: (1) moving into the proximity of a sensor (or depot) and (2) fully charging (or recharging). Thus, we define $n + 1$ actions corresponding to $n + 1$ charging destinations (a depot and n sensors). An action a_t made at time t is an integer number $a_t \in \mathcal{A} = \{0, 1, 2, ..., n\}$, where we denote $a_t = i, i > 0$ with regard to the integral index of the sensors, and $a_t = 0$ corresponds to going back to the depot and recharging itself.

Reward. Since our objective is to prolong the lifetime of the given network, we define the reward $R(s_t, a_t)$ with respect to an action a_t at state s_t as the period of time doing charging action to hold the coverage and connectivity properties.

A charging trajectory can be represented as a sequence of the network state and the MC's actions $\tau = \{s_1, a_1, s_2, a_2, ...\}$, $a_t \in \mathcal{A}, s_t \in \mathcal{S}$. A stochastic policy $\pi(a|s)$ determines the probability of taking charging action a given network state s which models the MC's behavior at a given time. We are concerned with finding an optimal policy π^* that

maximizes the γ-discounted return:

$$G(\tau) = \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t). \tag{5.1}$$

5.2 Model architecture

In the proposed approach, we parameterize the stochastic policy $\pi_\theta(a_t|s_t)$ using a deep neural network aided by the attention mechanism. We leverage the neural network architecture proposed in (Hottung and Tierney, 2019) for Capacitated Vehicle Routing Problem (CVRP). The overall architecture is depicted in Fig. 5.2.1.

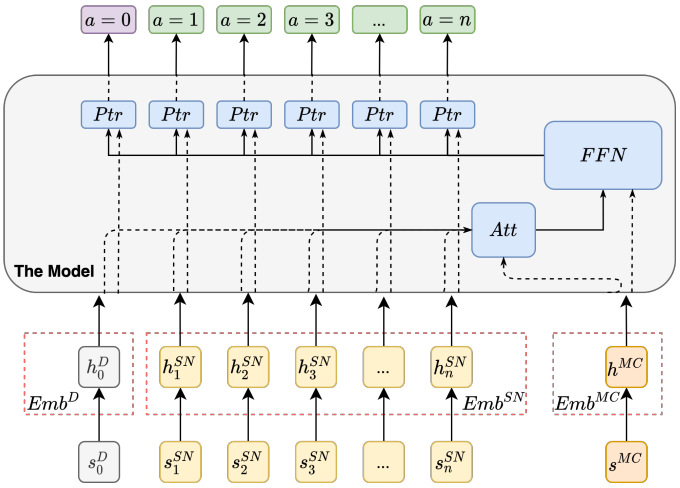


Figure 5.2.1: Model architecture.

The input to the model is the network state s_t at time t which is composed of the states at time t of the MC s^{MC}, and a sequence of charging destinations $(s^D, s_1^{SN}, ..., s_n^{SN})$. We embed the input by three transformations $(Emb^{MC}, Emb^D, Emb^{SN})$ used for the mobile charger, the depot, and the sensors, respectively. Note that we use the same transformation Emb^{SN} for all sensors, the embedding vector of each sensor is computed separately and identically. Precisely, let h^{MC}, h_0^D, h_i^{SN} be the embedded input corresponding to s^{MC}, s^D, s_i^{SN}. We denote $\bar{h}^C = \{h_0^D, h_1^{SN}, ...h_n^{SN}\}$ as a vector of embedded input of charging destinations. An attention layer is used to extract alignment vector $\bar{a}$, which specifies how much *attention* the MC might have for each charging destination given their current status. Precisely, the alignment vector is calculated by the following formula:

$$\bar{a} = softmax(u_0^H, u_1^H, ..., u_n^H), \tag{5.2}$$

where:

$$u_i^H = z^A \tanh\left(W^A[\bar{h}_i^C; h^{MC}]\right). \tag{5.3}$$

Here, $[;]$ denotes the concatenation of two vectors. The context vector c is provided by:

$$c = \sum_{i=0}^{n} \bar{a}_i \bar{h}_i^C. \tag{5.4}$$

The context vector is later concatenated with the embedded input of the MC to be the input of two layers feed-forward network (*FFN*) which, outputs a single, compatible vector q.

$$q = FFN_{W^B}([c; h^{MC}]). \tag{5.5}$$

The distribution of the policy over all actions given the state s_t is now given by:

$$\pi_\theta(a_t = i|s_t) = softmax(u_0, u_1, ..., u_n), \tag{5.6}$$

where:

$$u_i = z^C \tanh\left(\bar{h}_i^C + q\right), \tag{5.7}$$

and $\theta = \{z^A, W^A, W^B, z^C\}$ are trainable parameters.

Remark. Since the model is a deep learning model that adopts the pointing mechanism (Vinyals et al., 2015), which is literally designed to operate on the sequence-like inputs with different lengths, it is possible to generalize this model to different network settings without modifying the model's architecture. Furthermore, the replacement of the Gated Recurrent Unit (*GRU*) by a fully connected feed forward model (*FFN*) reduces the dependence of the MC's decision on the previous state. It enables the applications where the MC is deployed on the fly into an existing WSN.

5.3 Training method

We train the agent using a well-known policy gradient method in the reinforcement learning. Our objective is to maximize the expected total reward:

$$J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right], \tag{5.8}$$

where $p_\theta(\tau)$ is the distribution of Markov chain over all possible trajectories τ induced by the policy π_θ. Applying the REINFORCE approach proposed in (Williams, 1992), the gradient of Eq. 5.8 is given by:

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=0}^{\infty} \nabla_\theta \log \left(\pi_\theta(a_t|s_t) \right) \mathcal{A}_t \right], \tag{5.9}$$

where $\mathcal{A}_t = G_t - V_\theta(s_t)$ is the advantage function of taking action a_t given state s_t. This vanilla policy gradient update has two significant drawbacks. *First* it has no bias but high variance which leads to an unstable learning process. In order to deal with this problem, we use the Generalized Advantage Estimation (GAE) proposed by Schulman et al. (2015) to reduce the variance caused by the original advantage function at the cost of introducing bias. *Second*, due to the non-convex objective function, policy gradient usually suffers from local convergence. We hence use the entropy regularization as suggested in (Mnih et al., 2016) to encourage the exploration. The following formula computes the final policy gradient:

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=0}^{\infty} \nabla_\theta \log \left(\pi_\theta(a_t|s_t) \right) \hat{\mathcal{A}}_t^{GAE(\lambda)} + \beta \nabla_\theta \mathcal{H}(\pi_\theta(\cdot|s_t)) \right], \tag{5.10}$$

where $\mathcal{H}$ is entropy function, β is a hyperparameter controlling the strength of the regularization, and $\hat{\mathcal{A}}_t^{GAE(\lambda)}$ is the GAE function, which is estimated by:

$$\hat{\mathcal{A}}_t^{GAE(\lambda)} = \sum_{l=0}^{\infty} (\gamma\lambda)^l (R_{t+l} + \gamma V_\psi(s_{t+l+1}) - V_\psi(s_{t+l})), \tag{5.11}$$

where λ is a hyperparameter controlling the trade-off between variance and bias in advantage function. $V_\psi(s)$ is the estimated value function given state s.

During training, in the same manner as the other *actor-critic* approaches, we maintain two networks, one for the policy $\pi_\theta(a|s)$, and another for value function $V_\psi(s)$ with the trainable parameter θ and ψ, respectively. Similar to (Hottung and Tierney, 2019), we use a simple three-layer fully connected feed-forward network with ReLU activation in between for the value function's network. The policy is updated using gradients computed in Eq. 5.10. Meanwhile, we use Mean Square Error (MSE) to compute the loss function for the value function. Both networks are trained with *Adam* optimizer (Kingma and Ba, 2014). The pseudocode of our training process is described in Algorithm 1.

Algorithm 1 Actor-critic algorithm

initialize the actor network with random weight θ
initialize the critic network with random weight ψ
for *epoch* $= 1, 2, \ldots$ **do**
  reset gradient: $d\theta \leftarrow 0, d\psi \leftarrow 0$
  sample N network instances according to the distribution Φ
  for $n = 1, \ldots, N$ **do**
    initialize the environment on the network instance n^{th}.
    generate an episode following π_θ: $s_0, a_0, s_1, a_1, \ldots, s_{T-1}, a_{T-1}, s_T$
    $G \leftarrow 0, \mathcal{A}_t^{GAE} \leftarrow 0$
    $d\theta \leftarrow 0, d\psi \leftarrow 0$
    for $t = T-1, T-2, \ldots, 0$ **do**
      $G \leftarrow \gamma G + R_t$
      $\delta_t \leftarrow R_t + \gamma V_\psi(s_{t+1} - V_\psi(s_t)$
      $\mathcal{A}_t^{GAE} \leftarrow \gamma\lambda\mathcal{A}_t^{GAE} + \delta_t$
      $d\theta \leftarrow d\theta - \nabla_\theta \log(\pi_\theta(a_t|s_t)\mathcal{A}_t^{GAE} - \beta\nabla_\theta\mathcal{H}(\pi_\theta(\cdot|s_t))$
      $d\psi \leftarrow d\psi + \nabla_\psi \frac{1}{2} \left\| G - V_\psi(s_t) \right\|_2^2$
    end for
    $\theta \leftarrow ADAM(\theta, d\theta)$
    $\psi \leftarrow ADAM(\psi, d\psi)$
  end for
end for

Chapter 6

Experiments and results

6.1 Simulation settings

In the experimental studies, we assume that the sensor network is randomly deployed in a square area of interest $200m \times 200m$. The sink is supposed to be located in the center $(100, 100)$ of the field, and the depot is in the bottom-left corner $(0, 0)$ of the field. Initially, all sensors are equipped with a full and rechargeable battery with the capacity of $B_s = 10J$. Meanwhile, the MC is initially located at the depot has a high-capacity battery ($B_{MC} = 50J$) with an initial energy of $50J$. The reason for that setting is to encourage the exhausted state of the MC in some first episodes, which helps the MC learn when going back depot to recharge itself. That, in turn, accelerates the learning process.

For other settings, we adopt the parameters as in (He et al., 2013), where the system configuration is set as in Table 6.1.1, and the constants in the energy model are set as in Table 4.2.1. Specifically, the velocity of the MC is fixed to be $v = 5\ m/s$, and the charging rate between the MC and the sensor is $\mu = 0.04$. The energy consumed by traveling one unit distance is set to $\omega_{move} = 0.04$. The number of sensors in the experiments varies from 20 to 30 sensors, while the number of targets is from 10 to 20 targets. The energy consumption rate of each sensor ω is computed following the energy model described in 4.2. Regarding the configuration of the reinforcement learning model, the reward discount γ is set to 0.95. The hyperparameters of entropy regularization β and GAE λ are 0.02 and 0.9 respectively.

Table 6.1.1: Configuration.

Parameter	Value	Unit	Comment
n	$20 \sim 30$	$-$	number of deployed sensors
m	$10 \sim 20$	$-$	number of critical targets
B_{MC}	500	J	battery capacity of the MC
ω_{move}	0.04	J/m	battery capacity of a sensor
v	5	m/s	velocity of the MC
B_s	10	J	battery capacity of a sensor
r_s	40	m	sensing range
r_c	80	m	communication range
μ	0.04	J/s	charging rate

Concerning the implementation, the model is implemented in the PyTorch framework while we implement the environment simulation with the help of the Gym framework, which is well-known to be a basis for simulating reinforcement learning environments. Unlike the other simulators that imitate real-world behavior by generating and monitoring the transmissions of each packet in every second, we design a deterministic environment that can leap to the next state at which the network's topology is changed. To this end, several assumptions are made to omit the impact of the surrounding environment. Therefore, the simulation and training time can be significantly shortened. More importantly, we can increase the evaluations on different network topologies (typically, from hundreds to thousands), which highly affects the efficiency of the algorithms. Figure 6.1.1 shows an example of our simulation. Finally, all experiments are performed in a computer with Intel Intel(R) Core(TM) i7-6800K CPU @ 3.40GHz, 16 GB RAM, and GeForce GTX TITAN X GPU running on Ubuntu Linux 18.04.

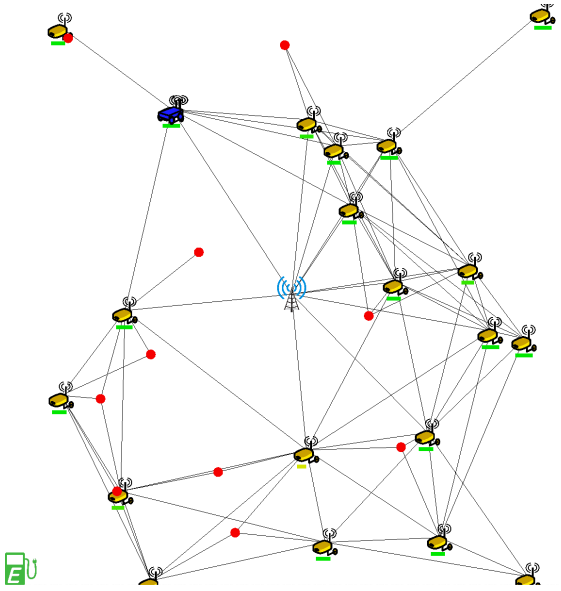


Figure 6.1.1: The graphical interface of a network topology.

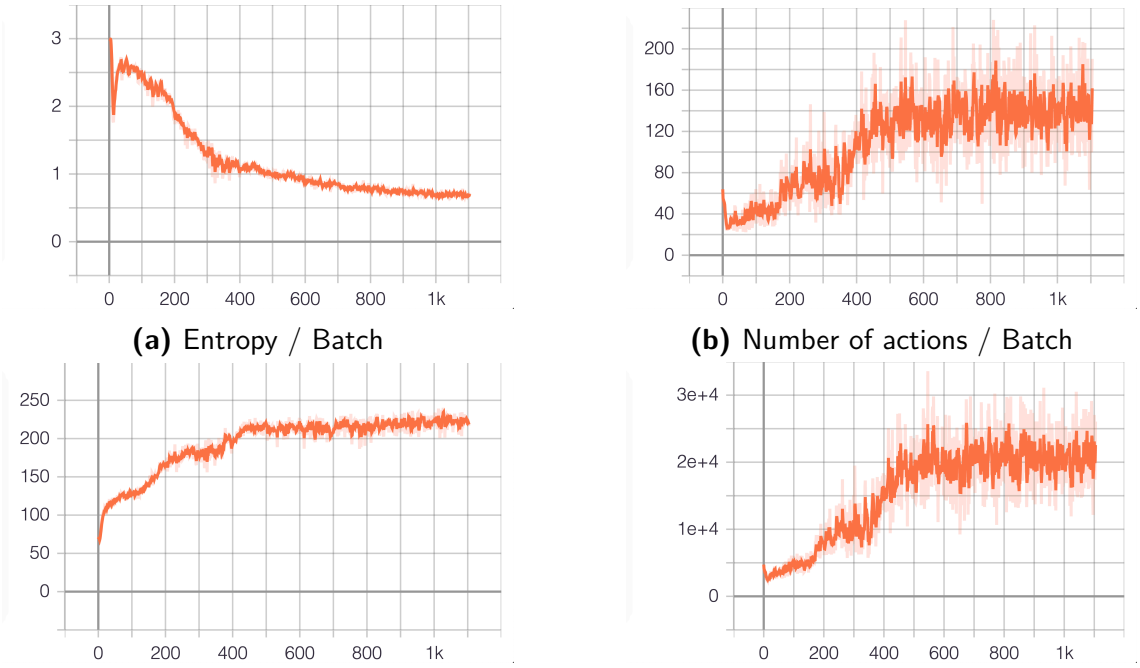
6.2 Datasets

Wireless sensor networks are generally deployed at random, which means the sensors are randomly dispersed in a given sensing field. In this thesis, we consider the problem that the positions of sensors are drawn in a square area $200m \times 200m$ according to the uniform distribution. In the training phase, we generate 10000 network instances with the default parameters for learning the agent. We use the same data distribution with different seeds to produce network instances for the testing phase. It should be noted that all instances are guaranteed to be covered and connected initially.

6.3 Learning process

We present the learning history of our model in Figure 6.3.1 in which four metrics are presented. Figure 6.3.1a shows the entropy history of the agent's decision. The entropy is high if the decision has low certainty. The decrease of the entropy metric represents the improvement of the agent where the agent is more confident of its decisions. Figure 6.3.1b depicts the number of actions the agent can perform on each episode, and Figure 6.3.1c shows the reward on average per action. Those increases point out that the agent gradually develops the ability to extend the network's lifetime by attending on low-power sensors (the increase of mean rewards per action). It leads to the improvement of the ability to

prolong the network lifetime showed in Fig. 6.3.1d. Finally, Fig. 6.3.2 summarizes the learning history of the agent on both training dataset and validation dataset.



(a) Entropy / Batch

(b) Number of actions / Batch

(c) Mean rewards per action / Batch

(d) Network lifetime / Batch

Figure 6.3.1: Learning history on each batch.

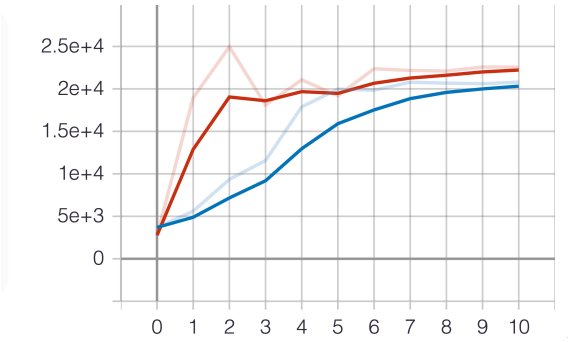


Figure 6.3.2: The network's lifetime improvement on each epoch where the blue line denotes learning process on the training instances, and red line is on the validation instances.

6.4 Baselines

We mainly compare our proposed with three baselines:

- Random: The agent chooses the next charging destination at random. We add a

simple estimation that helps the agent go back to the depot to recharge itself before becoming exhausted.

- NJNP: The algorithm is proposed in (He et al., 2013) with a simple but very efficient discipline that chooses the spatially closest requesting node as the next charging destination.

- IMNA: (Zhu et al., 2018) is a heuristic algorithm that selects nodes that make the least number of other requesting nodes enduring energy deficiency as the charging candidates. For high charging efficiency, the node with the shortest time to finish the charging will be selected as the next charging node if the candidate set has more than one node.

In what follows, we refer to *DRL-TCC* (a brief for deep reinforcement learning approach for target coverage and connectivity problem (DRL-TCC)) as our proposed method.

6.5 Results and discussions

In order to evaluate the proposed approach, we conduct extensive experiments to investigate the behavior of the algorithms on the various network topologies. We also vary different settings to investigate the impacts of the number of sensors, targets, or the packet generation probability to each algorithm. In each setting, numerous network instances are involved. Note that only one proposed model is trained with the default settings used to evaluate in all experiments.

In each experiment, three aspects of interest are discussed, including:

- *Network lifetime*: Due to some hardware constraints (velocity of the MC, the wireless charging rate between MC and sensor nodes), designing an effective strategy to prolong the network lifetime still is a primary objective in WRSN.

- *Sustainability*: For further understanding, there is a need for analyzing the ability to prolong for perpetual operation. Since it cannot be determined whether a mobile charger's strategy could extend the lifetime to infinitely long for a given network instance or not, we consider a system to be sustainable if the network is still active and guarantee coverage and connectivity after a large number M of charging actions of the MC. In this work, M is set to 2000.

- *Travel distance*: In the proposed online charging scheme, the MC is supposed to go to charge any sensor at any time. It is important to consider reducing travel distance as one of the concerns of this study.

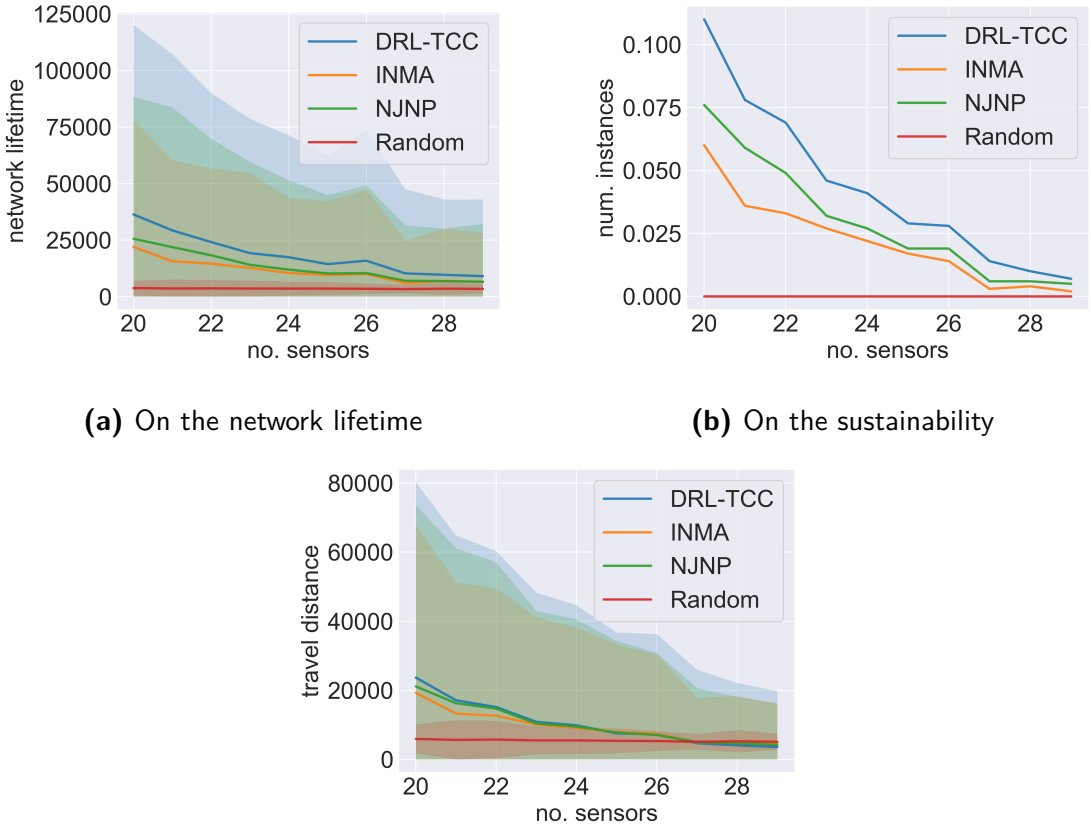
6.5.1 Impact of the number of sensors

In this experiment, we investigate the impact of the number of sensors on the efficiency of four algorithms. We generate 10 test set with the number of sensors varying from 20 to 29 sensors. Each test set has 1000 network instances drawn from the uniform distribution Φ. The results are shown in Fig. 6.5.1.

Fig. 6.5.1a shows the impact of the number of sensors on prolonging the lifetime of four algorithms. The proposed method outperforms three algorithms in terms of prolonging the network lifetime. When $n = 20$, the DRL-TCC extends the network lifetime to around $36375s$ on average 1000 network instances. This number of NJNP and Invalid Node Minimized Algorithm (INMA) are around $25600s$ and 22027. While the MC with random strategy only conserves the network to $3780s$ on average. The same order is maintained with other settings of sensors; however, the gap among those algorithms decreases when increasing the sensors. Another insight is the high variance observed in all settings. It demonstrates the high dependence on the network topologies of the charging algorithms. This is the main reason why numerous network instances are chosen to be evaluated in the paper.

In terms of sustainability, Fig. 6.5.1b presents the ratio of the number of the network instances so that the MC can elongate to a sustained state, over 1000 tested instances. The results show the superiority of the DRL-TCC compared to others. When $n = 20$, there are 110 network instances at which the MC with DRL-TCC's strategy elongates the operation over 2000 charging decisions. Ideally, those systems are expected to last for perpetual operation. Meanwhile, the corresponding numbers of NJNP and INMA are 76 and 60 instances. In comparison, no instances are considered to be sustainable with the Random strategy.

Fig. 6.5.1c compares the moving distance of four algorithms. The travel distance of the DRL-TCC is quite similar to those of NJNP and INMA, considering its improvement in terms of network lifetime.



(a) On the network lifetime

(b) On the sustainability

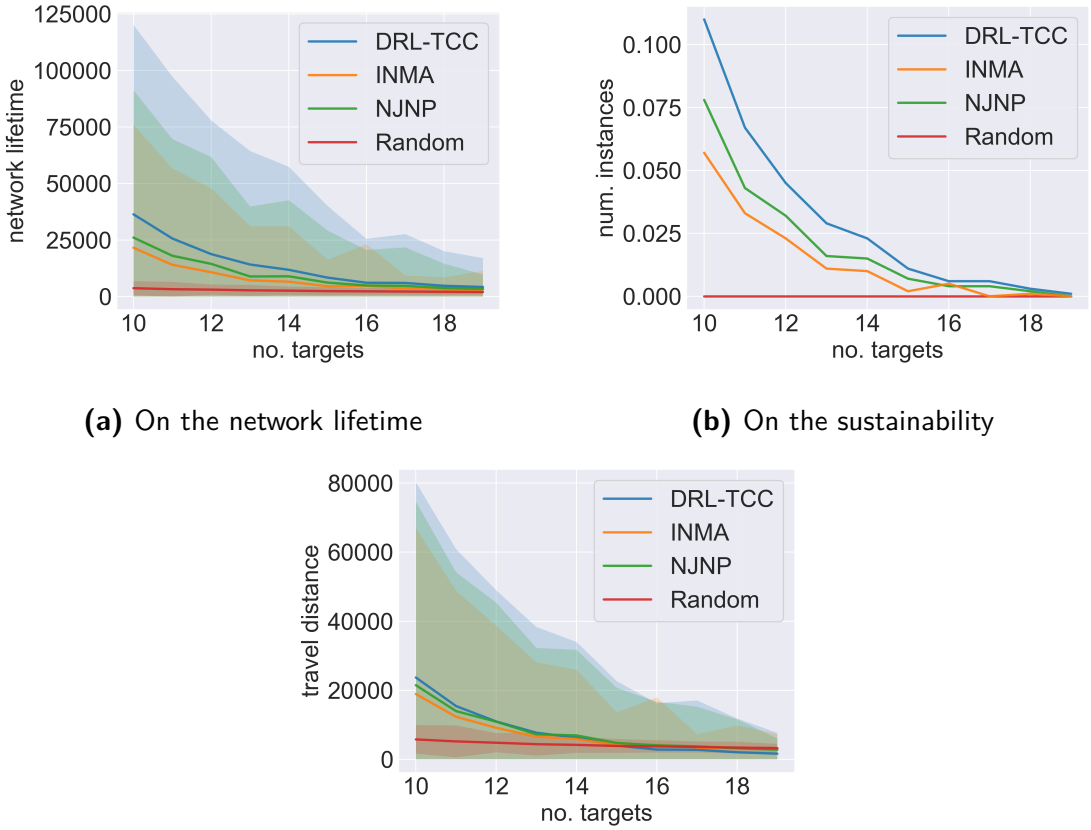
(c) On the travel distance

Figure 6.5.1: The impact of the number of sensor nodes.

6.5.2 Impact of the number of targets

This experiment aims to study the impact of the number of targets in the network on the charging algorithms. We consider the configurations in which the number of targets is from 10 to 19. The number of sensors is set to 20. The same as in the evaluation on the number of sensors, we generate 1000 network instances from uniform distribution Φ for each configuration. The results are shown in Fig 6.5.2.

The results are moderately similar to the results shown in evaluating the number of sensors since the number of source sensors is proportionate with the number of targets and the number of sensors in the network. We can observe the deteriorate trend when increasing the number of targets as well as the number of sensors. However, in all configurations, the DRL-TCC maintains superiority over other algorithms.



(a) On the network lifetime

(b) On the sustainability

(c) On the travel distance

Figure 6.5.2: The impact of the number of targets.

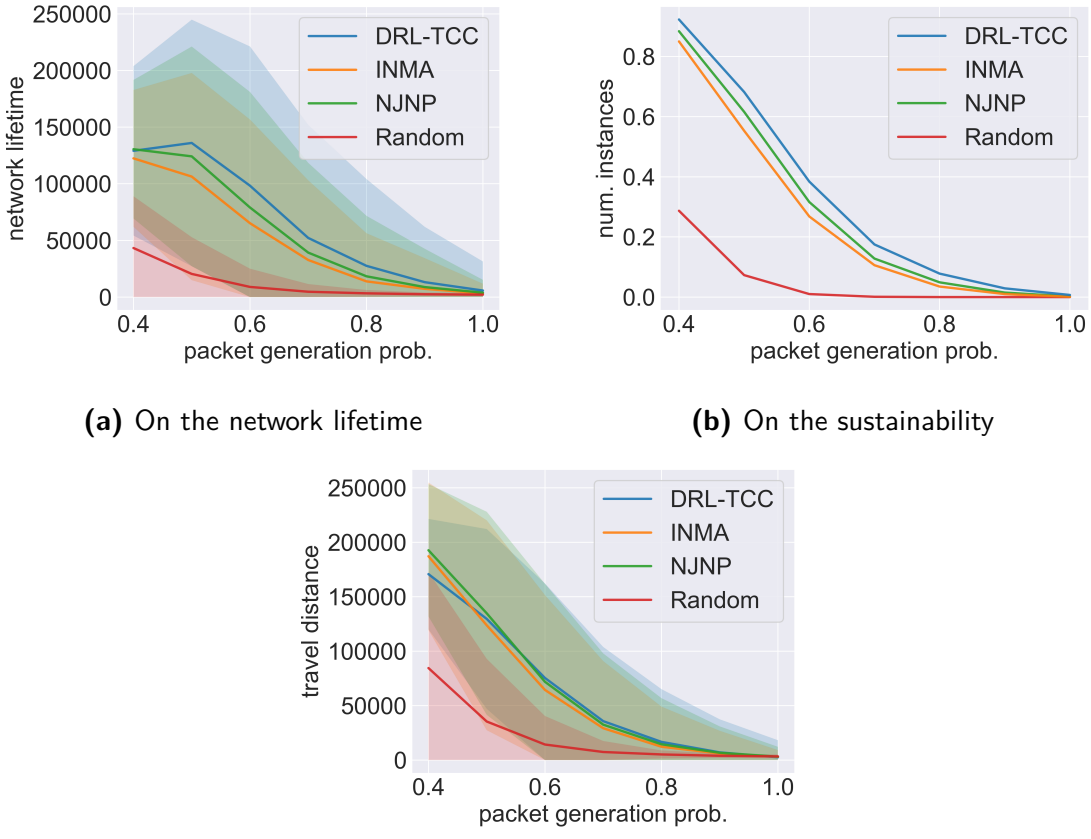
6.5.3 Impact of the packet generation probability

Since the energy consumption rate has a significant impact on the charging efficiency, we discuss the performance of these methods when varying the packet generation probability of each sensor. In this experiment, we consider the network's settings with 20 sensors and 10 targets. 1000 network instances are generated on this configuration. The probability of generating a packet varies from 0.4 to 1.0.

The results shown in Fig. 6.5.3 demonstrate the tremendous impact of the packet generation probability on the charging algorithms. The results on network lifetime (Fig. 6.5.3a) of the DRL-TCC deteriorate from 129119 to 5778 on average when varying the packet generation probability from 0.4 to 1.0. However, the result of DRL-TCC outperforms 30.57% on average compared to the results given by NJNP and 57.33% better than the result of INMA. Comparing the results of the NJNP and the INMA, NJNP outperforms 19.67% INMA. The Random strategy continues to show the worst results.

Regarding sustainability, we can observe the massive deterioration of all algorithms. When the packet generation probability is $p = 0.4$, DRL-TCC, NJNP, and INMA respectively maintain 92.3%, 88.4%, and 85% the number of instances, over 2000 charging actions. Even if the MC runs with the Random strategy, 28.7% of the number the instances survive. Hence, reducing the travel distance is more important than prolonging the network lifetime in those configurations. Fig. 6.5.3c points out that the DRL-TCC has the more effective trajectories in those instances. On the contrary, the results deteriorate rapidly in settings at which packet generation probability increases to 1.0. The works are too heavy for only one MC. However, the number of survived instances given by DRL-TCC is twice as that of NJNP and INMA (2.9% compared to 1.5% and 1%, respectively) in this configuration.

In general, the proposed algorithm outperforms other approaches in all evaluated aspects and configurations. The results of NJNP are better than the results given by INMA. The Random strategy performs worst in all results.



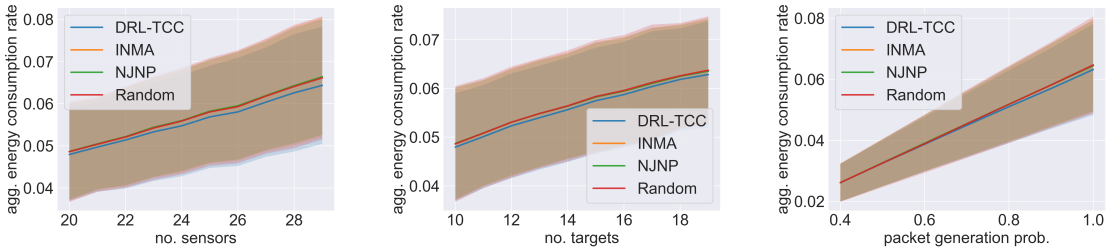
(a) On the network lifetime

(b) On the sustainability

(c) On the travel distance

Figure 6.5.3: The impact of the packet generation probability.

6.5.4 Discussion on the self-organizing capability

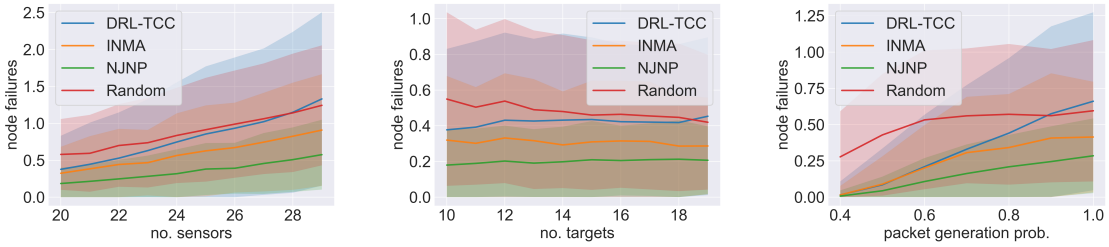


(a) Varying the number of sensors

(b) Varying the number of targets

(c) Varying packet generation probability

Figure 6.5.4: The comparison of the aggregated energy consumption rate.



(a) Varying the number of sensors

(b) Varying the number of targets

(c) Varying packet generation probability

Figure 6.5.5: The comparison of the number of node failures.

Self-organizing is considered an essential feature of WSNs that enables various applications of wireless sensor networks. This section discusses the MC's self-organizing capability for the target coverage and connectivity problem in wireless rechargeable sensor networks.

Fig. 6.5.4 shows the aggregated energy consumption rate, and Fig. 6.5.5 presents the average number of node failures in the three aforementioned experiments. We can observe that in the three experiments, the aggregated energy consumption rate of the sensor in the system running with the DRL-TCC strategy is slightly lower than that of the others. Note that there is no difference among the remaining algorithms. The cause is the higher number of node failures if the MC runs with the DRL-TCC strategy. However, in the target coverage and connectivity problem, only source sensors generate packets, and the network is considered to be terminated if the coverage and connectivity are no longer maintained. Considering the performance on prolonging network lifetime in those experiments, we can infer that those nodes failures are relay nodes that do not cause packet loss.

More importantly, it reduces the energy consumption rate in the network, which is crucial to prolong the network lifetime. In particular, Fig. 6.5.5c and 6.5.4c show that, when the packet generation probability is 0.4, the energy consumption rate and the number of node failures of DRL-TCC are quite similar to those of NJNP and INMA since the prolonging lifetime is not crucial in those configurations. Increasing the packet generation probability, the number of node failures of DRL-TCC also increases, and the aggregated energy consumption rate decreases. This means that when the dissipated energy becomes greater than the supply energy, the agent leaves some nodes considered as not crucial to exhausted to reduce the total energy consumption of the network. It demonstrates the adaptability to various scenarios of our proposed model, even though only one model was trained with the default settings. In contrast to the NJNP and INMA, the higher network lifetime corresponds to the lower node failures since both algorithms treat all sensors in the network equally.

Chapter 7

Conclusion and future works

7.1 Conclusion

In this thesis, we investigated the target coverage and target connectivity problem in wireless rechargeable sensor networks. We proposed a novel online charging scheme in which the requesting energy threshold is omitted. Unlike the traditional on-demand charging scheme, the MC will consider the next charging destination among all sensors and the depot based on the status of the network forwarded from sink through the long-range communication. A deep reinforcement learning-based mobile charging scheme ,namely DRL-TCC, is proposed to tackle the target coverage and connectivity problem in WRSNs. The model is a deep neural network with the pointing mechanism taking the network state as input and outputting the probability of each charging action. We trained our model by the actor-critic method in which the agent interacts with the network environment and adjusts its decisions by the feedback from the environment.

In the experiments, we conducted extensive experiments with different settings to compare the proposed method to two state-of-the-art methods in the on-demand charging scheme: NJNP and INMA. In each setting, numerous network instances are involved. The results demonstrate the superiority of the proposal to others in all evaluated settings. The network lifetime is significantly extended considering a reasonable travel distance trade-off. We also discussed the self-organizing capability of our method in those experiments. This demonstrates the adaptability and generalization to various scenarios of our proposed model, even though only one model was trained with the default settings.

7.2 Future works

With the limited amount of time, this work still has some limitations. *First*, our evaluation is only a small number of sensors, while in practice, the number of sensors in a sensing field is typically from hundreds to thousands. It is necessary to evaluate the efficiency of the deep reinforcement learning approaches on large and dense sensor networks. *Second*, the traveling efficiency problem has not been fully resolved in this work since the MC still has not an *idle* state. It requires the MC to go charging arbitrary sensors even not necessary. In the next stage of this work, we plan to introduce idle state as one of the MC's actions. It will reduce unnecessary charging action.

This work has shown the prospects of the truly online charging model in which the sensors do not need to send charging requests to the base station or the mobile charger. The mobile charger is now considered as an autonomous agent making charging decisions on a given sensor field instead of serving charging requests. With the recent advances in deep reinforcement learning, the agent can now learn highly complex strategies without hand-crafted features. Further investigation of other deep learning approaches for online mobile charging schemes is one of our works of interest.

References

Kofi Sarpong Adu-Manu, Nadir Adam, Cristiano Tapparello, Hoda Ayatollahi, and Wendi Heinzelman. Energy-harvesting wireless sensor networks (eh-wsns) a review. *ACM Transactions on Sensor Networks (TOSN)*, 14(2):1–50, 2018.

Ozgür B Akan and Ian F Akyildiz. Event-to-sink reliable transport in wireless sensor networks. *IEEE/ACM transactions on networking*, 13(5):1003–1016, 2005.

Ian F Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci. Wireless sensor networks: a survey. *Computer networks*, 38(4):393–422, 2002.

He Ba, Ilker Demirkol, and Wendi Heinzelman. Passive wake-up radios: From devices to applications. *Ad hoc networks*, 11(8):2605–2621, 2013.

Yosra Zguira Bahri. *Study and development of wireless sensor network architecture tolerant to delays*. PhD thesis, Université de Lyon; Université du Centre (Sousse, Tunisie), 2018.

Teck Chuan Beh, Takehiro Imura, Masaki Kato, and Yoichi Hori. Basic study of improving efficiency of wireless power transfer via magnetic resonance coupling based on impedance matching. In *2010 IEEE International Symposium on Industrial Electronics*, pages 2011–2016. IEEE, 2010.

Douglas M Blough and Paolo Santi. Investigating upper bounds on network lifetime extension for cell-based energy conservation techniques in stationary ad hoc networks. In *Proceedings of the 8th annual international conference on Mobile computing and networking*, pages 183–192, 2002.

Tatiana Bokareva, Wen Hu, Salil Kanhere, Branko Ristic, Neil Gordon, Travis Bessell, Mark Rutten, and Sanjay Jha. Wireless sensor networks for battlefield surveillance. In *Proceedings of the land warfare conference*, pages 1–8. Citeseer, 2006.

Xianbo Cao, Wenzheng Xu, Xuxun Liu, Jian Peng, and Tang Liu. A deep reinforcement learning-based on-demand charging algorithm for wireless rechargeable sensor networks. *Ad Hoc Networks*, 110:102278, 2021.

Bogdan Cărbunar, Ananth Grama, Jan Vitek, and Octavian Cărbunar. Redundancy and coverage detection in sensor networks. *ACM Transactions on Sensor Networks (TOSN)*, 2(1):94–128, 2006.

Mihaela Cardei, My T Thai, Yingshu Li, and Weili Wu. Energy-efficient target coverage in wireless sensor networks. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 3, pages 1976–1984. IEEE, 2005.

Chin-Ling Chen, I-Hsien Lin, et al. Location-aware dynamic session-key management for grid-based wireless sensor networks. *Sensors*, 10(8):7347–7370, 2010.

Jing Deng, Yunghsiang S Han, Wendi B Heinzelman, and Pramod K Varshney. Scheduling sleeping nodes in high density cluster-based sensor networks. *Mobile Networks and Applications*, 10(6):825–835, 2005.

Isabel Dietrich and Falko Dressler. On the lifetime of wireless sensor networks. *ACM Transactions on Sensor Networks (TOSN)*, 5(1):1–39, 2009.

Lingkun Fu, Liang He, Peng Cheng, Yu Gu, Jianping Pan, and Jiming Chen. Esync: Energy synchronized mobile charging in rechargeable wireless sensor networks. *IEEE Transactions on vehicular technology*, 65(9):7415–7431, 2015.

Rohit D. Gawade and S. L. Nalbalwar. A centralized energy efficient distance based routing protocol for wireless sensor networks. *Journal of Sensors*, 2016, 2016. ISSN 16877268. doi: 10.1155/2016/8313986.

Arvind Giridhar and PR Kumar. Maximizing the functional lifetime of sensor networks. In *IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005.*, pages 5–12. IEEE, 2005.

Nitin Goyal, Mayank Dave, and Anil K Verma. Data aggregation in underwater wireless sensor network: Recent approaches and issues. *Journal of King Saud University-Computer and Information Sciences*, 31(3):275–286, 2019.

Liang He, Yu Gu, Jianping Pan, and Ting Zhu. On-demand charging in wireless sensor networks: Theories and applications. In *2013 IEEE 10th international conference on mobile ad-hoc and sensor systems*, pages 28–36. IEEE, 2013.

Keith Hellman and Michael Colagrosso. Investigating a wireless sensor network optimal lifetime solution for linear topologies. *Journal of Interconnection Networks*, 7(01): 91–99, 2006.

André Hottung and Kevin Tierney. Neural large neighborhood search for the capacitated vehicle routing problem. *arXiv preprint arXiv:1911.09539*, 2019.

Guiyuan Jiang, Siew-Kei Lam, Yidan Sun, Lijia Tu, and Jigang Wu. Joint charging tour planning and depot positioning for wireless sensor networks using mobile chargers. *IEEE/ACM Transactions on Networking*, 25(4):2250–2266, 2017.

Mohamed Amine Kafi, Yacine Challal, Djamel Djenouri, Messaoud Doudou, Abdelmadjid Bouabdallah, and Nadjib Badache. A study of wireless sensor networks for urban traffic monitoring: applications and architectures. *Procedia computer science*, 19:617–626, 2013.

Kisuk Kang, Ying Shirley Meng, Julien Breger, Clare P Grey, and Gerbrand Ceder. Electrodes with high power and high capacity for rechargeable lithium batteries. *Science*, 311(5763):977–980, 2006.

Amar Kaswan, Abhinav Tomar, and Prasanta K Jana. An efficient scheduling scheme for mobile charger in on-demand wireless rechargeable sensor networks. *Journal of Network and Computer Applications*, 114:123–134, 2018.

Kavi K Khedo, Rajiv Perseedoss, Avinash Mungur, et al. A wireless sensor network air pollution monitoring system. *arXiv preprint arXiv:1005.1737*, 2010.

Sukun Kim, Shamim Pakzad, David Culler, James Demmel, Gregory Fenves, Steven Glaser, and Martin Turon. Health monitoring of civil infrastructures using wireless sensor networks. In *Proceedings of the 6th international conference on Information processing in sensor networks*, pages 254–263, 2007.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Andre Kurs, Aristeidis Karalis, Robert Moffatt, John D Joannopoulos, Peter Fisher, and Marin Soljačić. Wireless power transfer via strongly coupled magnetic resonances. *science*, 317(5834):83–86, 2007.

Van Quan La, Phi Le Nguyen, Thanh-Hung Nguyen, Kien Nguyen, et al. Q-learning-based, optimized on-demand charging algorithm in wrsn. In *2020 IEEE 19th International Symposium on Network Computing and Applications (NCA)*, pages 1–8. IEEE, 2020.

Deying Li, Jiannong Cao, Ming Liu, and Yuan Zheng. K-connected target coverage problem in wireless sensor networks. In *International Conference on Combinatorial Optimization and Applications*, pages 20–31. Springer, 2007.

Shining Li, Jin Cui, and Zhigang Li. Wireless sensor network for precise agriculture monitoring. In *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, volume 1, pages 307–310. IEEE, 2011.

Chi Lin, Jingzhe Zhou, Chunyang Guo, Houbing Song, Guowei Wu, and Mohammad S Obaidat. Tsca: A temporal-spatial real-time charging scheduling algorithm for on-demand architecture in wireless rechargeable sensor networks. *IEEE Transactions on Mobile Computing*, 17(1):211–224, 2017.

Chi Lin, Yu Sun, Kai Wang, Zhunyue Chen, Bo Xu, and Guowei Wu. Double warning thresholds for preemptive charging scheduling in wireless rechargeable sensor networks. *Computer Networks*, 148:72–87, 2019.

Anfeng Liu, Ju Ren, Xu Li, Zhigang Chen, and Xuemin Sherman Shen. Design principles and improvement of cost function based energy aware routing algorithms for wireless sensor networks. *Computer Networks*, 56(7):1951–1967, 2012.

Zengwei Lyu, Zhenchun Wei, Jie Pan, Hua Chen, Chengkai Xia, Jianghong Han, and Lei Shi. Periodic charging planning for a mobile wce in wireless rechargeable sensor networks based on hybrid pso and ga algorithm. *Applied Soft Computing*, 75:388–403, 2019.

Yu Ma, Weifa Liang, and Wenzheng Xu. Charging utility maximization in wireless rechargeable sensor networks by charging multiple sensors simultaneously. *IEEE/ACM Transactions on Networking*, 26(4):1591–1604, 2018.

Moshe Masonta, Yoram Haddad, Luca De Nardis, Adrian Kliks, and Oliver Holland. Energy efficiency in future wireless networks: Cognitive radio standardization requirements. In *2012 IEEE 17th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pages 31–35. IEEE, 2012.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for

deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

Stephan Olariu and Ivan Stojmenovic. Design guidelines for maximizing lifetime and avoiding energy holes in sensor networks with uniform distribution and uniform reporting. In *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, pages 1–12. Citeseer, 2006.

Luís ML Oliveira and Joel JPC Rodrigues. Wireless sensor networks: A survey on environmental monitoring. *JCM*, 6(2):143–151, 2011.

Jacques Panchard, Seshagiri Rao, Madavalam S Sheshshayee, Panagiotis Papadimitratos, Sumanth Kumar, and Jean-Pierre Hubaux. Wireless sensor networking for rain-fed farming decision support. In *Proceedings of the second ACM SIGCOMM workshop on Networked systems for developing regions*, pages 31–36, 2008.

Danyang Qin, Jingya Ma, Yan Zhang, Pan Feng, Ping Ji, and Teklu Merhawit Berhane. Study on connected target coverage algorithm for wireless sensor network. *IEEE Access*, 6:69415–69425, 2018.

Tifenn Rault, Abdelmadjid Bouabdallah, and Yacine Challal. Energy efficiency in wireless sensor networks: A top-down survey. *Computer Networks*, 67:104–122, 2014.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Shio Kumar Singh, MP Singh, Dharmendra K Singh, et al. Routing protocols in wireless sensor networks–a survey. *International Journal of Computer Science & Engineering Survey (IJCSES)*, 1(2):63–83, 2010.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Makoto Suzuki, Shunsuke Saruwatari, Narito Kurata, and Hiroyuki Morikawa. A high-density earthquake monitoring system using wireless sensor networks. In *Proceedings of the 5th international conference on Embedded networked sensor systems*, pages 373–374, 2007.

Ayman Tajeddine, Ayman Kayssi, and Ali Chehab. Center: a centralized trust-based efficient routing protocol for wireless sensor networks. In *2012 Tenth Annual International Conference on Privacy, Security and Trust*, pages 195–202. IEEE, 2012.

Nguyen Thi Tam, Huynh Thi Thanh Binh, Dinh Anh Dung, Phan Ngoc Lan, Bo Yuan, Xin Yao, et al. A hybrid clustering and evolutionary approach for wireless underground sensor network lifetime maximization. *Information Sciences*, 504:372–393, 2019.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. *arXiv preprint arXiv:1506.03134*, 2015.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Yin Wu and Wenbo Liu. Routing protocol based on genetic algorithm for energy harvesting-wireless sensor networks. *IET*, 3(2):112–118, 2013. ISSN 2043-6386. doi: 10.1049/iet-wss.2012.0117. URL `www.ietdl.org`.

Wenzheng Xu, Weifa Liang, Haibin Kan, Yinlong Xu, and Xinming Zhang. Minimizing the longest charge delay of multiple mobile chargers for wireless rechargeable sensor networks by charging multiple sensors simultaneously. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 881–890. IEEE, 2019.

Seong-eun Yoo, Poh Kit Chong, Taehong Kim, Jonggu Kang, Daeyoung Kim, Changsub Shin, Kyungbok Sung, and Byungtae Jang. Pgs: Parking guidance system based on wireless sensor network. In *2008 3rd International Symposium on Wireless Pervasive Computing*, pages 218–222. IEEE, 2008.

Moustafa A Youssef, Mohamed F Younis, and Khaled A Arisha. A constrained shortest-path energy-aware routing algorithm for wireless sensor networks. In *2002 IEEE Wireless Communications and Networking Conference Record. WCNC 2002 (Cat. No. 02TH8609)*, volume 2, pages 794–799. IEEE, 2002.

Liyang Yu, Neng Wang, and Xiaoqiao Meng. Real-time forest fire detection with wireless sensor networks. In *Proceedings. 2005 International Conference on Wireless Communications, Networking and Mobile Computing, 2005.*, volume 2, pages 1214–1217. Ieee, 2005.

Qun Zhao and Mohan Gurusamy. Lifetime maximization for connected target coverage in wireless sensor networks. *IEEE/ACM transactions on networking*, 16(6):1378–1391, 2008.

Pengzhan Zhou, Cong Wang, and Yuanyuan Yang. Leveraging target k-coverage in wireless rechargeable sensor networks. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pages 1291–1300. IEEE, 2017.

Jinqi Zhu, Yong Feng, Ming Liu, Guihai Chen, and Yongxin Huang. Adaptive online mobile charging for node failure avoidance in wireless rechargeable sensor networks. *Computer Communications*, 126:28–37, 2018.