

A Deep Reinforcement Learning-Based Caching Strategy for IoT Networks with Transient Data

Hongda Wu, *Student Member, IEEE*, Ali Nasehzadeh, Ping Wang, *Fellow, IEEE*

Abstract—The Internet of Things (IoT) has been continuously rising in the past few years, and its potentials are now more apparent. However, transient data generation and limited energy resources are the major bottlenecks of these networks. Besides, minimum delay and other conventional quality of service measurements are still valid requirements to meet. An efficient caching policy can help meet the standard quality of service requirements while bypassing IoT networks' specific limitations. Adopting deep reinforcement learning (DRL) algorithms enables us to develop an effective caching scheme without the need for any prior knowledge or contextual information. In this work, we propose a DRL-based caching scheme that improves the cache hit rate and reduces energy consumption of the IoT networks, in the meanwhile, taking data freshness and limited lifetime of IoT data into account. To better capture the regional-different popularity distribution, we adopt a hierarchical architecture to deploy edge caching nodes in IoT networks. The results of comprehensive experiments show that our proposed method outperforms the well-known conventional caching policies and an existing DRL-based solution in terms of cache hit rate and energy consumption of the IoT networks by considerable margins.

Index Terms—Deep Reinforcement Learning, Edge Caching, Energy Efficiency, Internet of Things

I. INTRODUCTION

THE world is flooding with data and data transmissions; simultaneously, the demand for more reliable and faster connections is souring. This concurrency results in either terribly congested networks or the need for more resources and more intelligent networking schemes. With the emergence of the Internet of Things (IoT), there is an exponential growth in the number of connected devices in the world, which again emphasizes the importance of better networking schemes [1] [2]. Caching is a networking technique that enables networking nodes to store frequently requested files, mitigating network traffic and improving the response time. In cases where the downlink is more heavily used, such as in content delivery networks (CDNs), extremely large files need to be transmitted to users. Caching can prevent multiple transmissions from the source by responding to users from the edge nodes where the files have been downloaded earlier. *Edge caching* is a new technology that enables edge nodes, e.g., base stations or user's devices, to be a part of the caching schemes and store files. Being close to the end-users means that the request does not necessarily go all the way up to the source to fetch

the response because one of the edge nodes might already have the desired file in its caching memory and is ready to fulfill the request, which in turn will significantly shorten the response time and simultaneously ease the load on the backhaul link [3]. Numerous caching methods have been proposed previously; some well-known methods include Least Frequently Used (LFU) and Least Recently Used (LRU) methods [4]. The basic idea for both caching algorithms is to rank the files in the nodes' caching memories. The lowest-ranked file will be replaced with the new one if the node has to store a new file that was not stored before. The difference between these two methods lies in how they approach the ranking phase. Authors in [5] proposed a dynamic probabilistic caching design to solve the tasks with time-varying content popularity, which is difficult for LRU to apply since LRU does not exploit any content distribution. However, the probabilistic caching design requires the user request preference as the prior information. Zhang *et al.* deployed a greedy algorithm for cooperative edge caching problem to minimize the average transmission delay, where the dynamic update of contents is not considered [6]. Both LFU, LRU, and the mentioned greedy algorithm are not sensitive to popularity distribution, making them unsuitable for IoT caching problems due to the fact that rates of request in the edge nodes of IoT are time-variable [7], [8]. Moreover, traditional methods are not applicable to delay-sensitive IoT applications, e.g., quick-responding biometric applications in e-Health [9]. Thus, the limited data lifetime needs to be considered when designing caching policy for IoT networks. With the promising advancements in artificial intelligence and machine learning in recent years, Reinforcement Learning (RL), a machine learning paradigm, has recently gained significant attention [10]. The goal of RL algorithms is to achieve an optimal policy that can be utilized for control/optimization purposes, i.e., a sequence of mapping from states to actions that leads to a maximized benefit (a.k.a. cumulative reward). Without providing prior knowledge of networks (e.g., popularity distribution of files) and explicitly defining essential features of learning problems, RL is inherently suitable to solve caching problems with dynamic characteristics. Further, Deep Reinforcement Learning (DRL) unfolds countless possibilities to address problems with huge searching space [11], which can be adopted to solve caching problems with a massive of devices in the network [12].

Many existing works have applied DRL to caching problems for different scenarios, including Content Delivery Network (CDNs) [7], [13], mobile networks [14], [15], and IoT networks [16]–[24]. Nevertheless, most of them did not take the limited data lifetime and constrained device energy into

Preliminary version of this paper was accepted to IEEE/CIC International Conference on Communications in China (ICCC), August 2020 [1].

Hongda Wu, A. Nasehzadeh, and Ping Wang are with the Department of Electrical Engineering and Computer Science, York University, Toronto, ON, M3J 1P3 Canada e-mail: hwu1226@yorku.ca, arian92@eecs.yorku.ca, pingwu@yorku.ca.

account. To make IoT caching problem more practical, we focus on the freshness feature of IoT data and take energy consumption into consideration. In addition, we adopt a hierarchical architecture for caching rather than using the conventional single-layer architecture to better capture the heterogeneous popularity distribution in different regions. To the best of our knowledge, this paper is the first one that brings hierarchical caching architecture to IoT networks. Our main contributions in this paper are as follows.

- We formulate the IoT caching problem as a Markov decision process (MDP) while considering real restrictions in this scenario, i.e., limited lifetime of caching data and constrained energy of IoT devices. The newly defined reward function incorporates the data freshness, so as to better imitate the real IoT networks when designing the DRL solution. We adopt the proximal policy optimization solver to get the optimal policy.
- We apply a more practical architecture for deploying the caching node in IoT networks. The adopted two-layer hierarchical architecture (i.e., parent-leaf node architecture) takes regional-different popularity distribution into consideration, aside from being more realistic, it also leads to better system performance.
- We conduct extensive experiments, gather comprehensive results and compare the proposed caching strategy with different caching methods and architectures. Experimental results indicate that the proposed caching strategy can improve the cache hit rate and reduce energy consumption compared to the conventional caching schemes and existing DRL solution [24]. Compared with existing benchmarks, the proposed algorithm demonstrates superior performance in handling the limited lifetime of data.

II. RELATED WORKS

Existing works have leveraged DRL algorithms to find a good caching policy for CDNs with huge files [7], [13] and mobile networks [14], [15]. In particular, the authors of [13] considered a hierarchical system model consisting of parent and leaf nodes. They developed a cooperative caching scheme where each node's caching decision is affected by the decisions of others. The proposed Hyper Deep Q-Network (HDQN) requires stacks of deep Q-networks (DQN) to produce a caching decision. The output of DQNs forms a cost vector, which leads to the DRL agent returning a caching decision. In [7], the authors considered the scenario that multiple edge nodes are cooperative and competitive. The proposed Wolpertinger is based on actor-critic architecture with a shrunken action space by the K-nearest neighbor algorithm. The smaller action space allows the algorithm to have a faster run time without significant performance degradation. The authors of [14] proposed a DRL-based caching strategy for vehicular networks leveraging the multi-view videos from street cameras to proactively cache the required contents and deliver a higher video quality. The term "proactively" refers to the caching policy in which a networking node can ask for a file and store the file in its caching memory just because it is anticipating to receive the file request in the near future

instead of actually receiving the request. The authors of [15] proposed a joint computation offloading, caching, and resource allocation optimization by taking advantage of mobile social networks' data. The main feature of their work is the definition of a trust index based on the interaction history among network users. This index indicates which user can share its resources, e.g., memory, to help with caching.

While the majority of related works concentrate on the caching problem in the context of mobile cloud networks or CDNs, the exponential growth of IoT devices has brought the corresponding caching problem into the spotlight [16]–[25]. Some existing works considered one node, e.g., router or base station, to function the caching behavior [16], [24], while another series of works involve multiple edge nodes to formulate the IoT caching problem [17]–[23]. In addition, based on the way to solve IoT caching problems, the latter works can be categorized into two streams, i.e., non-cooperative [17]–[20] and cooperative [21]–[23] edge caching, depending on whether there is a cooperation between multiple nodes with caching capability.

Nath *et.al.* [16] and the authors in [17] considered a joint optimization for resource allocation, computation offloading and caching for IoT networks using a DRL-based solution. The former considers the scenario of one caching node and the latter scheme involves multiple caching nodes. The authors of [18] brought a new point of view to the caching problem; they used a DRL algorithm to assign proper caching capacity and transmission rate to the nodes in a content-centric IoT network. In [19] the authors proposed a DRL-based caching mechanism for a content-centric IoT network, in the meanwhile, focusing on the quality of experience.

Different from single caching node and multiple nodes without cooperation, authors in [21], [22] considered the multiple edge nodes for caching and assumed that multiple edge nodes can exchange the information among those edge nodes [21] or with an additional central server [22]. In another recent work, Wang *et.al.* [23] deployed a federated deep reinforcement learning method in order to realize cooperative edge caching. Federated learning trains a global model across multiple decentralized edge devices holding local data samples and sharing information (via neural network model) in a privacy-preserved way. The proposed method outperforms the conventional methods such as LRU and LFU, and it achieves similar performance to the centralized DRL approach with a low-performance degradation.

However, the newly introduced features of IoT networks bring new challenges to the IoT caching problem, one of which is the periodic data generation by IoT devices. Considering the scenario where an IoT sensor might broadcast the room temperature every five minutes, meaning that the data has a limited lifetime in this scenario, which also indicates the ephemeral significance of those data. Thus, data freshness must be taken into account for caching in IoT networks, which is not satisfactorily reflected by the above mentioned works [16]–[19], [21]–[23]. Furthermore, energy consumption is also critical in IoT networks because IoT devices are generally battery-powered. Without a caching mechanism, all the user requests will go directly to the IoT device side, which triggers

the device to leave its energy-saving mode, sense and generate the data, transmit the data, and eventually go back to the energy-saving state [26].

In this work, we focus on the data transiency in IoT networks and take the limited lifetime and data freshness into consideration. Ma *et al.* proposed a framework that utilizes information from the queues of user requests for different dynamic contents that need to be served in the near future [20], where the indicator adopted is the age of information (a similar concept to data freshness). Differently, we argue that files are valid for a short period of time and their values can be constantly changing in dynamic networks; this proactive caching does not offer enough value for IoT networks. Another similar work [24] mentioned the data transiency in IoT networks. To the best of our knowledge, [24] is the only work that simultaneously applies DRL and considers the limited lifetime and data freshness to find a solution to IoT caching problem. On the other hand, there is no mention of the energy consumption rate or how it is affected by the caching policy. Moreover, the system model considered in [24] comprises only one networking node (a single router). Even though it is an innovative work, the scalability of such a model is of question. In this work, aside from focusing on the freshness of IoT data, we also adopt a more practical architecture for deploying the caching nodes in IoT networks, which is a hierarchical (i.e., parent-leaf) caching system. Assume a file that is not the most popular one in any location, but it has moderate popularity across all of the regions; in this case, none of the leaf nodes see any reason to cache this file, but the parent node who can observe this general popularity will cache this file and improve the cache hit rate significantly.

III. SYSTEM MODEL

A. Network Architecture

In real-world IoT network architectures, there may exist multiple layers in a network. In this work, we consider a model consisting of two layers with a parent node at the upper layer and multiple edge/leaf nodes at a lower layer, all capable of caching¹. Figure 1 depicts the general architecture based on which we have implemented our work. There are different entities in this IoT network, namely IoT devices, parent node, leaf (edge) nodes, and users. In what follows, we briefly introduce each of them.

- **IoT Devices:** These devices are deployed in the physical world, and they are supposed to sense their environment or control a part of it. Light sensors and smart locks are two examples of such devices. As an indispensable part of their functionality, IoT sensors produce some type of

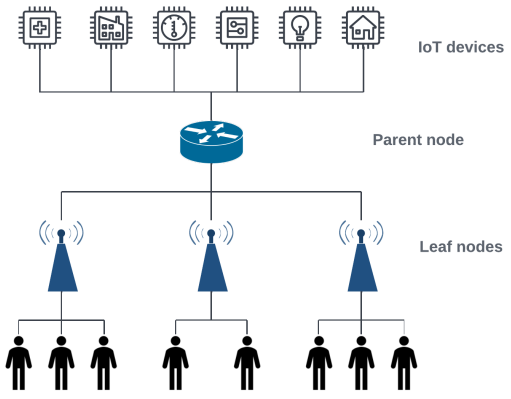


Fig. 1. System Model

data, which users ask for. One specific attribute to IoT systems is that the data produced by respective devices have a *limited lifetime*, meaning that the data is only valid for a limited time duration after their generation. This is a challenging characteristic to handle for data caching. For instance, a popular file might have an extremely short life span, making it not worth caching.

- **Parent Node:** This type of node shapes the inner layers of a network, and it acts as a gateway that connects all the IoT devices at one side and all the edge nodes at the other side. The parent node receives the data requests from the edge nodes, fetches the data from the IoT devices, and then forwards the data to the edge nodes.
- **Leaf (edge) Nodes:** These are the networking nodes placed close to the end users. Their coverage is limited to one region with some users, and they are connected to the parent node.
- **Users:** Users could be the applications on some devices such as personal computers, smartphones, or even air conditioning machines. These applications request the data generated by the IoT devices and use them for analysis or control objectives, e.g., a user can monitor the room temperature with a smartphone application using data generated by an IoT sensor. Owing to the short request response time, it is reasonable to assume limited movement for the users and thus a stable connection between users' devices and leaf nodes. This assumption means that requests will be responded and users will not leave the edge leaf's coverage before responses arrive.

In the absence of a caching scheme in a network, users submit their request for a specific type of data (we refer to as a file hereinafter) to the local edge node. The submitted request will then traverse its way up to the parent node and then the primary source to evoke a response from the corresponding IoT device. In contrast, when a caching scheme is deployed, after receiving a request, the edge node will look up its cache memory for the requested file. If the file is available, the edge node will respond to the user itself and prevent further communications. Otherwise, the edge node will pass the request to the parent node, and now, the parent will go through the same process of checking its cache memory. In this

¹The proposed hierarchical caching architecture can be viewed as a variant of the general one-layer architecture with multiple edge nodes. Since the popularity distribution (see detailed illustration in Section III-B) across different regions is different, one advantage behind the proposed architecture is that the parent node can capture the file popularity (with request) in the sense of global view, which may not be reflected by all regional/local edge nodes in one layer. However, assumptions on parent/edge nodes, i.e., parent node has sufficient storage space to cache all content files (e.g., [25]) and multiple edge nodes can exchange the information (e.g., [21], [22]), are not needed. The empirical results in Section VI show the effectiveness of the proposed architecture when an equal sum of memory space is dispatched.

way, the data source needs to generate and transmit a response only when there is no valid response available in the edge or the parent nodes. It is readily evident that the response time would be significantly shorter by avoiding multiple hops of communication. Besides, when data sources are less frequently forced to wake, generate and transmit a response, they will consume less energy.

Without loss of generality, we consider N IoT devices that have the role of data sources and each of them can only produce one type of data. Given a file i , denoted as f_i , a ternary tuple $\{id_i, t_{gen}^i, t_{life}^i\}$ is introduced to represent the unique ID, the corresponding generation timestamp, and lifetime of that file. Lifetime is the duration of time that the content is still valid after being generated. Each type of data has its own lifetime, e.g., smoke detector's data might be valid for just twenty seconds, because it is crucial to act upon a sudden change of smoke, while humidity data might be valid for hours. Freshness, i.e., $freshness_i$, is another metric that can measure the performance of a caching scheme and is defined as $freshness_i = (t_{current} - t_{gen}^i)/t_{life}^i$, where $t_{current}$ is the current time. When a file reaches its expiration time, i.e., $freshness_i \geq 1$, the file is no longer valid and will be deleted from the cache memory.

B. Popularity Distribution

Different from other works [21], [22] that include multiple edge nodes (with one-layer architecture) for cooperative edge caching by exchanging cache information, we have considered multiple caching nodes to form a multi-layered networking architecture. The motivation behind the proposed hierarchical caching architecture is the regional-different popularity distribution². The popularity of files follows different distributions from region to region. It is possible for a file $f_i^{\mathcal{X}}$ to be moderately popular in some regions \mathcal{X} but not popular enough to be cache-worthy. None of the edge nodes caches this file in such a scenario and thus sends every request for f_i to the parent node to be fetched from its source. However, from the parent node's point of view, things are different; the parent receives the accumulated requests for f_i and might find out that f_i is a cache-worthy file. In other words, the popularity distribution in the parent node is unique to itself, and thus we can also benefit from caching in the parent node.

Aside from being a more realistic system model, our empirical results (see Section VI-B for the detail) show that it is more beneficial to have a caching scheme in both edge and parent nodes rather than only at the edge nodes (even assuming an equal sum of memory space).

IV. PROBLEM FORMULATION

In this section, we first explain how we formulate the caching problem to a Markov Decision Process (MDP) and then introduce the ingredients of solving the MDP problem, including the design of state, action, and reward function.

²Regions are defined as the area that the edge nodes can cover and popularity refers to the frequency with which file has been requested.

A. Markov Decision Process Modeling

Conventionally, at each time step n , MDPs are represented by an information tuple of $\{s_n, a_n, p(s_{n+1}|s_n, a_n), r_n\}$ that respectively denotes the current state $s_n \in \mathcal{S}$, chosen action $a_n \in \mathcal{A}$, a distribution function $p(s_{n+1}|s_n, a_n) \in \mathcal{P}(\mathcal{S}, \mathcal{A})$ which gives the probability of a next state s_{n+1} given the current state and action, and finally a reward r_n that evaluates how good an action a_n is in the state s_n . Implementing a DRL algorithm means that we are seeking a policy, in our case a caching policy $\pi(a_n|s_n)$, which gives the probability of choosing an action a_n in the state s_n . It is intuitively clear that besides the immediate reward, caching decisions can also influence the rewards of the later steps, so it is only reasonable to consider long-term returns while making decisions. To do so, we define a cumulative reward G_n as below

$$G_n = \sum_{\tau=0}^T \gamma^t r_{n+\tau}, \quad (1)$$

where n refers to the time step, r_τ is the received reward in time step τ , T is the final time step. $\gamma \in (0, 1]$ is the discount factor that determines the effect of future rewards to current caching decisions. A smaller value of γ emphasizes the short-term rewards and the greater values emphasize the long-term returns. The ultimate goal here is to find the optimal policy π^* , under which the expected cumulative reward $\mathbb{E}[G_n|\pi]$ can be maximized

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \mathbb{E}[G_n|\pi]. \quad (2)$$

To find the optimal policy, we define the value function $V^\pi(s_n)$ and the action-value function $Q^\pi(s_n, a_n)$, respectively. $V^\pi(s_n)$ denotes the expected cumulative reward when starting from the state s_n and following the policy π after that. The action-value function $Q^\pi(s_n, a_n)$, also referred to as Q function, denotes the expected cumulative reward starting from the state s_n , taking action a_n , and after that, following the policy π . Using Bellman equations [10], we can write these two functions as in (3) and (4),

$$V^\pi(s_n) = \sum_{a_n \in \mathcal{A}} \pi(a_n|s_n) \sum_{s_{n+1} \in \mathcal{S}} p(s_{n+1}|s_n, a_n) [r_n + \gamma V^\pi(s_{n+1})], \quad (3)$$

$$Q^\pi(s_n, a_n) = \sum_{s_{n+1} \in \mathcal{S}} p(s_{n+1}|s_n, a_n) \cdot \left[r_n + \gamma \sum_{a_{n+1} \in \mathcal{A}} \pi(a_{n+1}|s_{n+1}) Q^\pi(s_{n+1}, a_{n+1}) \right]. \quad (4)$$

If any of these two functions is known, we can find the definite optimal policy π^* . The reason is that in each state s_n we would know what action will yield the greatest immediate reward r_n and the expected future value $V(s_{n+1})$, and we can simply define the optimal policy as choosing the most rewarding action in each time step. To obtain an optimal policy through value function or Q -function is known as the value-based reinforcement learning approach. However, traditional

RL is limited to the conditions that state space is fully-observable and with low dimensions. However, considering the freshness feature of IoT data makes the IoT system more complex and dynamic. Therefore the prerequisites such as handcrafted features or low-dimensional search spaces (e.g., \mathcal{S} and/or \mathcal{A}) is unavailable. Besides, the transition function $\mathcal{P}(\mathcal{S}, \mathcal{A})$ is also unknown, making it difficult to obtain or even approximate those value functions.

To conquer the above challenges, we leverage the deep reinforcement learning technique to approximate the value function or the Q function in Section V. One potent tool that helps us approximate those functions is the neural networks (NNs). NNs let us avoid manual feature extraction and train the agents on raw (and potentially high-dimensional) observations instead. Moreover, by deploying neural networks, it is possible to parametrize the policy $\pi_\theta(a_n|s_n)$ (a.k.a. π_θ), with θ being the set of parameters. Then we can directly tune the parameters of the policy function in search of the optimal one based on the gradient of some performance measure with respect to the policy parameters θ .

B. State, Action, and Reward

Through interactions with the environment, the RL agent receives multiple observations (i.e., samples of the state space), and after analysing the observation and remembering from previous experiences, the agent then takes its action, which gradually achieves its ultimate goal, maximizing the value function. Since the freshness of transient data is a key factor in IoT caching problem, the corresponding consideration in terms of the state space and reward function are shown. In what follows, we describe the ingredients of MDP, including the space of State \mathcal{S} and Action \mathcal{A} , as well as the design of reward function.

1) *State \mathcal{S}* : In order to form our observations, we first define the memory status of each caching node to be a part of state-space. Memory status is a three-row matrix representation for all cached files $\mathcal{M} = \{1, 2, \dots, M\}$ on this node. Each column corresponds to a file, e.g., f_i , and contains the file ID, the freshness value as calculated in (6), and the total cache hits k_i for that file in the cache memory.

$$\text{MemStatus} = \begin{bmatrix} id_1 & \dots & id_M \\ freshness_1 & \dots & freshness_M \\ k_1 & \dots & k_M \end{bmatrix}, \quad (5)$$

$$freshness = \frac{t_{current} - t_{gen}}{t_{life}}. \quad (6)$$

If there is no content stored in a memory at some time slot, the file's ID, freshness and cache hit value are all set to zero. The ID of requested file (e.g., \hat{f}_i) and its lifetime account for the other part of the observation. We put these two values together as the request variable

$$\text{ReqVariable} = [id, t_{life}]. \quad (7)$$

In this way, the RL agent can observe whether the requested file is available in the cache memories or not. The freshness of stored items is also available, so the agent might decide to replace a file that is going to expire in the near future with the requested file.

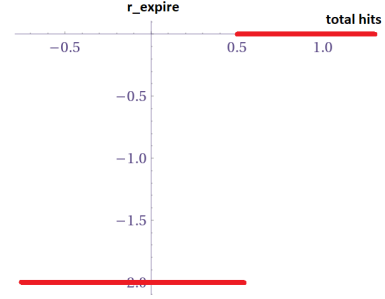


Fig. 2. r_{expire} function vs total number of hits ($C_1 = 1$)

2) *Action \mathcal{A}* : Assume a user has asked for a file f , and neither the corresponding edge node nor the parent has the file in its cache memory; thus, the response should be fetched from the source node. In this case, the networking nodes, i.e., parent and edge nodes, should make a caching decision. There are two possible actions and the action space is $\mathcal{A} = \{0, 1\}$. If $a = 1$, the node will cache the file at hand; otherwise, the file will not be stored. If a caching node decides to store a file when there is enough free space to do so, the process is a straightforward saving action, but if the cache memory is full, the new file has to replace an existing file. In the latter case, the proposed caching strategy will point out which file has to be replaced³.

3) *Reward*: As we mentioned earlier, data may have a limited lifetime in IoT networks, which raises concerns about data freshness⁴. As an example for the importance of freshness measure, a DRL agent might deduce that since a file is reaching its expiration, it can be a beneficial decision to replace it with the new and coming data. However, we should also note that freshness does not offer any intrinsic value; files are valid before expiration. To reflect the above points in our work, we *keep the freshness of cached files in the observation, but we do not include it in the reward function*. However, for each file in the cache memory that expires without being used at least once, we give a negative reward (punishment) to the agent for storing that file. We keep track of the number of times a cached file has been hit; thus, we can calculate the reward function defined in (8) for each file in the cache memory after it expires.

$$r_{expire}^i = [\text{sign}(k_i - 0.5) - 1] \times C_1, \quad (8)$$

where $\text{sign}(x)$ is the function that returns 1 if $x \geq 0$ and -1 otherwise. C_1 is the constant we choose as the punishment coefficient. Note that the total number of hits is an integer, thus by using (8), we are giving a negative reward only if a

³We clarify that the proposed caching scheme is not a proactive approach, i.e., no file will be cached prior to its corresponding request as in [14], [20]. In dynamic networks such as IoT, files are valid for a short period of times and their values can be constantly changing, proactive caching does not offer enough value and might add an extra burden to the backhaul links, which deviates from the purpose.

⁴Expired data is worthless to respond to users' requests and should be evacuated from cache memory. It is readily evident that repeatedly caching a file with a limited lifetime cannot be very beneficial, even though it might be fairly popular. Moreover, storing files with a long lifetime and inferior popularity rank is not efficient either since it is very possible that the file will be expired without receiving any request during its lifetime.

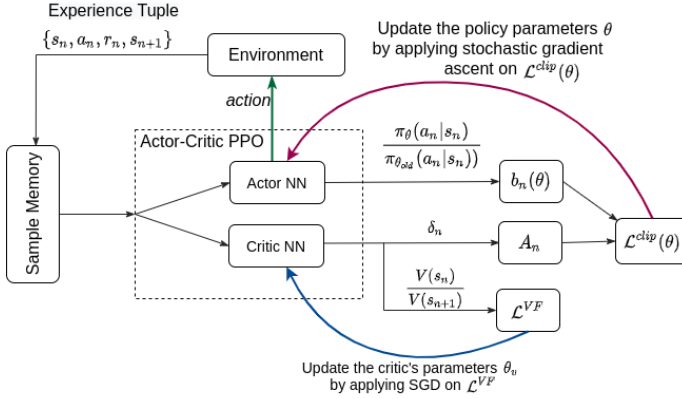


Fig. 3. PPO with Actor-Critic style

file has a total number of hits equal to zero. Refer to Figure 2 for a better illustration.

The ultimate objective of any good caching policy is to improve the cache hit rate, i.e., the rate at which users' requests are fulfilled locally by the caching nodes, and consequently, the communication cost is alleviated (less energy consumption). As such, increasing cache hit rate translates into shorter response time and higher energy efficiency. Delay (i.e., response time) is the time duration that users have to wait for their response. The longer distance a request should traverse to evoke a response, the longer the user should wait. Thus, it is most desirable if the first networking node has the requested file in its caching memory because the delay would be minimum. So, in order to encourage higher cache hit rates, we use the following reward function

$$r_{hit} = \frac{K \times C_2}{t_{current}}, \quad (9)$$

$$K = \sum_{t=0}^{t_{current}} \sum_{i=1}^M k_i, \quad (10)$$

where $t_{current}$ is the current time step, K is the sum of all the cache hits up until the current time step, and C_2 is a constant hyper-parameter that we choose arbitrarily. The final reward function is represented as follows

$$r = r_{hit} + \sum_{i=1}^M r_{expire}^i. \quad (11)$$

V. PROXIMAL POLICY OPTIMIZATION SOLVER

In this section, we have deployed one version of the Actor-Critic (AC) models named Proximal Policy Optimization (PPO) algorithm [27]. This algorithm is developed based on the Trust Region Policy Optimization (TRPO) algorithm [28] and aims to inherit the data efficiency and reliability of the TRPO while avoiding its second-order optimizations.

The PPO algorithm in this work is based AC method, by which we can simultaneously approximate the policy and value function using two separate neural networks named actor and critic, denoted as θ and θ_v , respectively. The actor decides which action should be taken and the critic evaluates the

Algorithm 1: PPO Algorithm, Actor-Critic Style

Input: initial NN parameters for actor and critic (θ and θ_v)

- 1: **for** iteration = 1, 2, ... **do**
- 2: Collect set of trajectories $\mathcal{T} = \{\tau_j\}$ by running the policy π_θ in the environment for T times
- 3: Estimate advantage values $\hat{A}_1, \dots, \hat{A}_T$
- 4: Receive the rewards and compute $\hat{G}_1, \dots, \hat{G}_T$
- 5: Optimize surrogate loss w.r.t. θ (e.g., stochastic gradient ascent with Adam)

$$\theta \leftarrow \operatorname{argmax}_{\theta} \mathcal{L}^{clip}(\theta)$$
- 6: Fit value function by regression on mean squared error using stochastic gradient descent

$$\theta_v \leftarrow \operatorname{argmin}_{\theta_v} \mathcal{L}^{VF}$$

where

$$\mathcal{L}^{VF} = \frac{1}{|\mathcal{T}|T} \sum_{\tau \in \mathcal{D}} \sum_{n=0}^T (V_{\theta_v}(s_n) - \hat{G}_n)^2$$

7: **end for**

action produced by the actor by computing the value function $V^{\pi_\theta}(s_n; \theta_v)$. A brief description of how AC works is as follow.

Step 1: Sample $\{s_n, a_n\}$ using policy π_θ from actor network; Execute a_n and observe s_{n+1}, r_n ;

Step 2: Evaluate the advantage function (a.k.a. TD error δ_n from critic network) by (12)

$$A_n = r_n + \gamma V^{\pi_\theta}(s_{n+1}; \theta_v) - V^{\pi_\theta}(s_n; \theta_v). \quad (12)$$

Step 3: Update the policy (i.e., θ) via gradient ascent method, based on gradient of advantage function in Step 2.

Different from traditional AC, PPO optimizes a clipped Surrogate objective function to update the policy θ , which is defined as

$$\mathcal{L}^{clip}(\theta) = \mathbb{E}_n[\min\{b_n(\theta)A_n, \text{clip}(b_n(\theta), 1 - \epsilon, 1 + \epsilon)A_n\}], \quad (13)$$

$$b_n(\theta) = \frac{\pi_\theta(a_n|s_n)}{\pi_{\theta_{old}}(a_n|s_n)}, \quad (14)$$

where \min gives the minimum argument given two inputs, $\text{clip}(\cdot)$ function takes in the probability ratio $b_n(\theta)$ defined in (14) but clips its value to be no more than $1 + \epsilon$ and no less than $1 - \epsilon$, with ϵ being a hyperparameter, θ_{old} in (14) is the vector of policy parameters before the update, and A_n is the advantage function defined in (12).

Finally, we use the following rule to update θ

$$\theta \leftarrow \operatorname{argmax}_{\theta} \mathcal{L}^{clip}(\theta). \quad (15)$$

where argmax is usually achieved by the stochastic gradient ascent [27].

Algorithm 1 gives a more comprehensive elaboration on the details of the PPO algorithm. In order to update the critic network, we first need the cumulative rewards G_n as defined in (1). Then as in step 6, we apply stochastic gradient descent on the mean squared error and update the critic network's parameters θ_v . Putting Algorithm 1 and Figure 3 together should give us a clear image on how the PPO performs.

VI. SIMULATIONS & RESULTS

In order to conduct our experiment, we have used Python version 3.7, NumPy [29], TensorFlow 1.14.0 [30], Keras, OpenAI's Gym [31], and Stable Baselines [32] libraries⁵. We ran the simulation on the workstation with an Intel Core i7-9700 CPU and 16G RAM. Below, we first elaborate our simulation setup, including the popularity distribution of data, parameter setting of the proposed DRL solver, and the benchmarks for comparison (i.e., LRU, LFU, and DRL in [24]) in Section VI-A. The superiority of the proposed IoT caching design on the cache hit rate and energy consumption is provided and discussed in Section VI-B. In addition, we also demonstrate how our proposed design outperforms benchmarks when different average freshness and popularity of cached files are present.

A. Simulation Setup

For ease of illustration, we consider a hierarchical IoT caching structure with one parent node, two edge nodes, and one hundred IoT devices producing the data. Each IoT device only produces one type of file with a unique ID and a lifetime, which is sampled randomly from a uniform distribution with a minimum of 2 and a maximum of 14 time-steps. The popularity distribution follows Zipf's law, which means that a file f with a popularity rank of x is going to be requested with a probability of

$$\bar{p}(x; \alpha, F) = \frac{x^{-\alpha}}{\sum_{x=1}^F x^{-\alpha}}, \quad (16)$$

where F is the total number of files in the network, and α is the characterizing parameter to the distribution, which we call the skewness factor in this work. As we assign greater values to the α , the request frequency gap between the most popular and least popular files will grow larger. We have done our experiments with 24 different settings, resulting from four different request rate values w and six different values of popularity skewness factor α . The request rate of users w varies from 0.5 to 2.0 requests per time step, and the popularity distributions have a varying α between 0.7 to 1.2. Please note that these distributions are unknown to the DRL agent; they are used only for generating requests in the simulations. We assume that each request can be fulfilled before its corresponding user leaves the network. There is no assumption for the users' arrival model since the DRL agent does not use any information except the ones received from the interactions with its environment to make caching decisions.

The architecture of the DRL algorithm is depicted in Figure 3. The learning rate is set to 0.001, the discount factor γ is equal to 0.99, and the agent will experience 16 steps before performing an update, and the number of training mini-batches per update is set to 4. The actor and critic NNs are feedforward neural networks, with two hidden layers and 64 neurons in each layer, and the activation function in use is tanh.

In this work, we compare our method with two of the best-known conventional caching strategies: the least recently used

(LRU) and least frequently used (LFU) algorithms. In addition, we have also implemented an existing DRL-based caching method [24] with a different reward function from ours to observe the effects of our proposed reward function.

Now, let us briefly review how each of these methods works:

- 1) *LRU*: In LRU, the stored files are always ranked based on how recently they have been used, and when a new file should replace one of the files in the memory, the least recently used file will be deleted from the memory.
- 2) *LFU*: Similar to LRU, cached items are ranked, but the ranking criterion is the frequency of their requests. If the memory is full, the new file (will be cached) will replace the file with the least frequency of request.
- 3) *DRL*: In our work, we keep track of the number of hits of each file and use this information to design a reward function to encourage a high hit rate. On the other hand, related works such as [24] directly consider the freshness of files in their reward function. To investigate the performance difference owing to the use of different reward functions, we have implemented the DRL approach in [24] for comparison. We refer to this approach as *DRL* in our results.

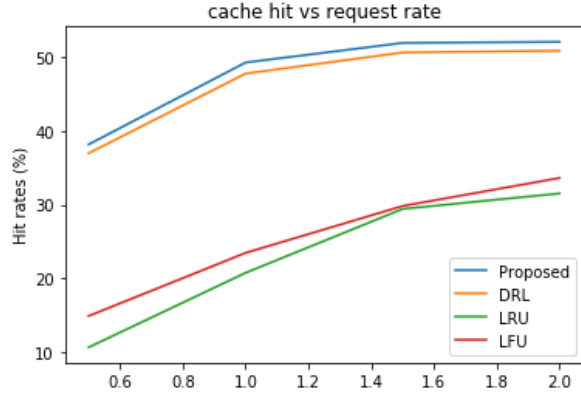
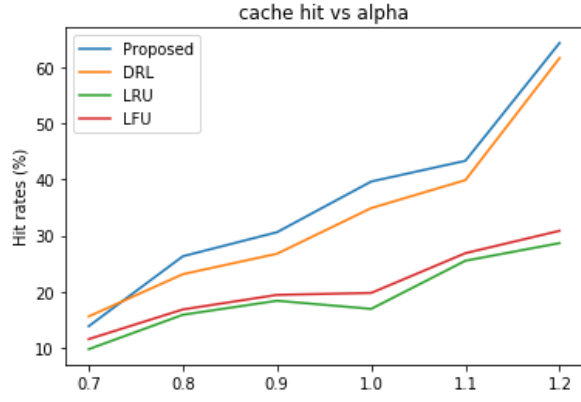
In the following section, we have evaluated our proposed method based on the cache hit rate and energy consumption. We have also monitored the average freshness and popularity of cached files.

B. Results and Discussions

Figures 4 and 5 show the cache hit rate versus varying request rate and popularity skewness factor, respectively. It is a clear trend that by increasing those two factors, the cache hit rate will increase as well. Higher request rate gives more chances to the cached files to be requested, and higher skewness factor makes the popularity gap between files even larger, which makes DRL agent easier to learn the pattern. We can see that our proposed method tremendously outperforms the well-known LRU and LFU methods, and it does a better job than the existing DRL method with a different reward function. The proposed method starts with about 38% cache hit, and it goes up to more than 52%. The other DRL approach follows the proposed one with about 1.3% gap, whereas the conventional method LFU, which outperforms LRU, shows a cache hit range of 16% to 34% for different request rates. The cache hit rate versus the skewness factor has a similar outcome for different approaches.

We have also simulated the energy consumption of the IoT devices for each caching scheme. Figure 6 and Figure 7 depict the simulation results with varying request rate and popularity skewness factor, respectively. We have normalized the energy consumption rates by our proposed method; values greater than 1 indicate that the energy consumption is higher than our proposed method for that specific configuration, and values less than one denote less energy consumption. We observe that our method uses the least energy for IoT devices among its counterparts. This result meets our expectation that the energy consumption is inversely proportional to the cache hit rate. The higher the cache hit rate, the lower the energy consumption.

⁵Stable Baselines is a library of reinforcement learning algorithm implementations based on OpenAI Baselines.


Fig. 4. Hit rate vs varying request rates w

Fig. 5. Hit rates vs popularity skewness α

In order to better understand what makes the proposed algorithm has a better cache hit rate, we took a look at the requests which have been responded to by a caching node. We categorize files based on their popularity and lifetime and then keep track of the requested files which have been hit in the cache memories. Figure 8 and Figure 9 show these results which the average over all requests and over 24 different settings of our experiments (the 24 settings vary in request rate and the popularity skewness factor as shown in Table VI-B). We can see that the DRL-based methods tend to store popular files, and also they favor files with a longer lifetimes, whereas conventional methods (i.e., LRU and LFU) might even favor files with shorter lifetimes. This shows that the DRL agent has learned that storing popular files with a longer lifetime is more beneficial, which is an entirely reasonable argument.

To evaluate the effects of the proposed reward function on the freshness of cached files, we have prepared Table VI-B. This table shows the average freshness for all the cache memory files during the whole simulation for each of the 24 different settings. Note that a freshness value of 1 means the file has expired, and a freshness value of 0 means the file has just been generated. The lower the freshness value, the fresher the file is. In the last row of the table, we can observe the mean value of all those settings. Since in our proposed method, the reward function does not include any punishment for the higher value of freshness, the cached files have slightly

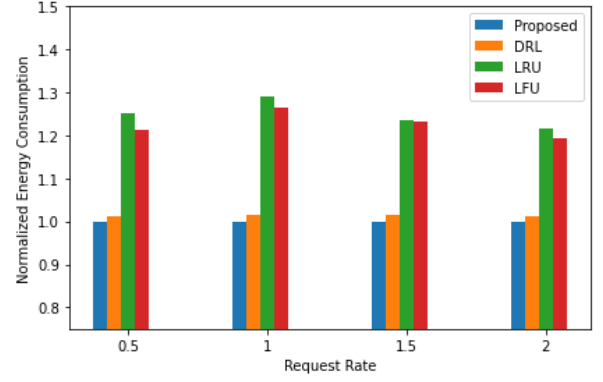
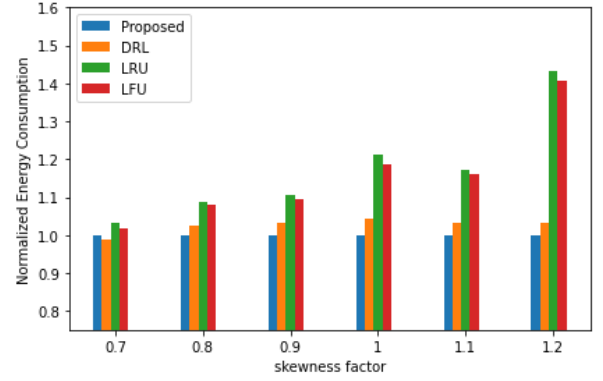
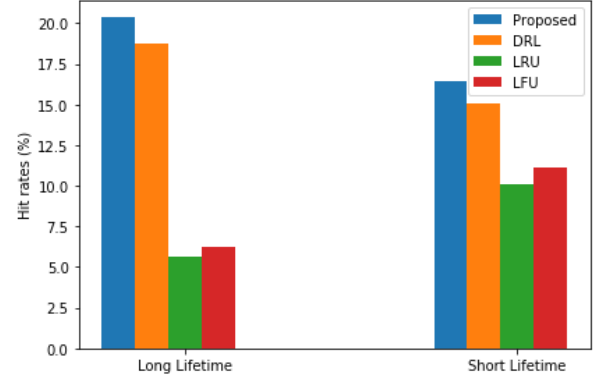

Fig. 6. Energy consumption vs request rates w

Fig. 7. Energy consumption vs popularity skewness α


Fig. 8. Hit rates vs lifetime

older files in the memory compared to the other DRL method, which explicitly considers freshness in its reward function. Nevertheless, the gap between these two methods is marginal.

Finally, in order to see the benefits of a hierarchical model architecture, we have implemented a single-layer caching scheme where there are two edge nodes capable of caching, and there is no parent node in the network. For the sake of fair comparison, we keep the sum of memory capacity of these two edge nodes on par with the cumulative memory capacity of the three nodes (one parent node and two edge nodes) in our hierarchical architecture. Figure 10 and Figure 11 show that the proposed hierarchical architecture can lead to higher cache

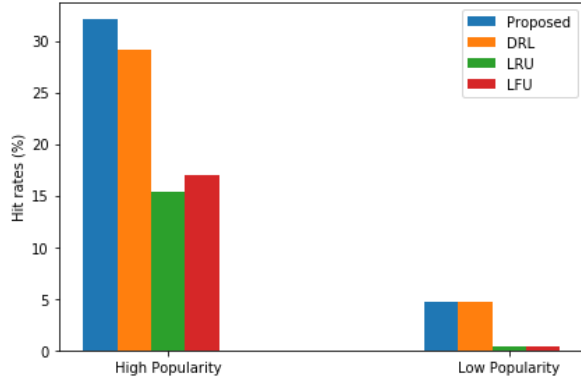


Fig. 9. Hit rate vs popularity

TABLE I
TABLE OF AVERAGE FRESHNESS

Setting	Proposed	DRL [24]
$w = 0.5, \alpha = 0.7$	0.497	0.17
$w = 0.5, \alpha = 0.8$	0.494	0.329
$w = 0.5, \alpha = 0.9$	0.536	0.53
$w = 0.5, \alpha = 1.0$	0.496	0.495
$w = 0.5, \alpha = 1.1$	0.532	0.532
$w = 0.5, \alpha = 1.2$	0.527	0.523
$w = 1, \alpha = 0.7$	0.469	0.468
$w = 1, \alpha = 0.8$	0.457	0.457
$w = 1, \alpha = 0.9$	0.422	0.422
$w = 1, \alpha = 1.0$	0.451	0.450
$w = 1, \alpha = 1.1$	0.411	0.409
$w = 1, \alpha = 1.2$	0.433	0.433
$w = 1.5, \alpha = 0.7$	0.247	0.355
$w = 1.5, \alpha = 0.8$	0.355	0.355
$w = 1.5, \alpha = 0.9$	0.296	0.296
$w = 1.5, \alpha = 1.0$	0.346	0.346
$w = 1.5, \alpha = 1.1$	0.333	0.293
$w = 1.5, \alpha = 1.2$	0.324	0.307
$w = 2, \alpha = 0.7$	0.182	0.27
$w = 2, \alpha = 0.8$	0.279	0.279
$w = 2, \alpha = 0.9$	0.225	0.228
$w = 2, \alpha = 1.0$	0.271	0.275
$w = 2, \alpha = 1.1$	0.249	0.251
$w = 2, \alpha = 1.2$	0.293	0.237
Average	0.380	0.362

hit rate and consequently lower energy consumption. The reason is that the parent node in the hierarchical architecture can effectively cache the files, which are commonly prevalent in multiple regions but might not be cache-worthy by edge nodes, thus increasing the overall cache hit rate.

VII. CONCLUSION

In this work, we have considered an IoT network with transient data files. Taking into account the limited lifetime of data and the energy consumption of IoT devices, we have formulated the caching problem into an MDP problem and then tackled the MDP problem with a deep reinforcement learning method using a proximal policy optimization algorithm. We have proposed a hierarchical architecture for implementing the DRL caching algorithms in two layers. The proposed method does not require any prior knowledge about the environment, users, or requests. The extensive experimental results show that our proposed DRL method and hierarchical architecture

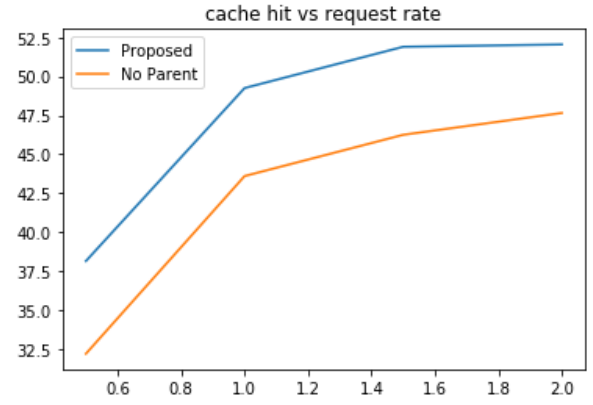
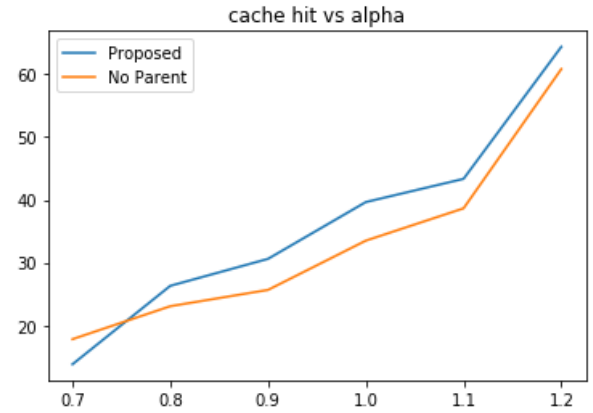
Fig. 10. Comparison of hierarchical architecture (proposed) and flat architecture (edge caching only). The figure shows the hit rate vs request rates ω 

Fig. 11. Comparison of hierarchical architecture (proposed) and flat architecture (edge caching only). The figure shows the hit rate vs the popularity skewness

can make smarter decisions in caching, significantly improve the cache hit rate, and reduce the energy consumption of the IoT devices in comparison to benchmark caching schemes including LRU and LFU, and an existing DRL solution.

ACKNOWLEDGEMENT

The authors would like to thank Aakash Agarwal for providing experimental results in the paper revision process.

REFERENCES

- [1] A. Nasehzadeh and P. Wang, "A deep reinforcement learning-based caching strategy for internet of things," in *Proc. IEEE/CIC International Conference on Communications in China (ICCC)*, 2020, pp. 969–974.
- [2] A. Nordrum, "Popular internet of things forecast of 50 billion devices by 2020 is outdated," *IEEE Spectrum*, Aug 2016.
- [3] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6524–6540, 2012.
- [4] D. Lee, J. Choi, J.-H. Kim, S. H. Noh, S. L. Min, Y. Cho, and C. S. Kim, "Lrfu: A spectrum of policies that subsumes the least recently used and least frequently used policies," *IEEE transactions on Computers*, vol. 50, no. 12, pp. 1352–1361, 2001.
- [5] J. Gao, S. Zhang, L. Zhao, and X. Shen, "The design of dynamic probabilistic caching with time-varying content popularity," *IEEE Transactions on Mobile Computing*, vol. 20, no. 4, pp. 1672–1684, 2020.

- [6] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao, and X. Shen, "Cooperative edge caching in user-centric clustered mobile networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 8, pp. 1791–1805, 2018.
- [7] C. Zhong, M. C. Gursoy, and S. Velipasalar, "A deep reinforcement learning-based framework for content caching," in *Proc. 52nd Annual Conference on Information Sciences and Systems (CISS)*, 2018, pp. 1–6.
- [8] M. Meddeb, A. Dhraief, A. Belghith, T. Monteil, and K. Drira, "How to cache in icn-based iot environments?" in *Proc. IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, 2017, pp. 1117–1124.
- [9] D. B. Santana, Y. A. Zocalo, and R. L. Armentano, "Integrated e-health approach based on vascular ultrasound and pulse wave analysis for asymptomatic atherosclerosis detection and cardiovascular risk stratification in the community," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 2, pp. 287–294, 2012.
- [10] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [11] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [12] H. Zhu, Y. Cao, W. Wang, T. Jiang, and S. Jin, "Deep reinforcement learning for mobile edge caching: Review, new features, and open issues," *IEEE Network*, vol. 32, no. 6, pp. 50–57, 2018.
- [13] A. Sadeghi, G. Wang, and G. B. Giannakis, "Deep reinforcement learning for adaptive caching in hierarchical content delivery networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 4, pp. 1024–1033, 2019.
- [14] Z. Zhang, Y. Yang, M. Hua, C. Li, Y. Huang, and L. Yang, "Proactive caching for vehicular multi-view 3d video streaming via deep reinforcement learning," *IEEE Transactions on Wireless Communications*, vol. 18, no. 5, pp. 2693–2706, 2019.
- [15] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 44–55, 2017.
- [16] S. Nath and J. Wu, "Deep reinforcement learning for dynamic computation offloading and resource allocation in cache-assisted mobile edge computing systems," *Intelligent and Converged Networks*, vol. 1, no. 2, pp. 181–198, 2020.
- [17] Y. Wei, F. R. Yu, M. Song, and Z. Han, "Joint optimization of caching, computing, and radio resources for fog-enabled iot using natural actor-critic deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2061–2073, 2018.
- [18] X. He, K. Wang, H. Huang, T. Miyazaki, Y. Wang, and S. Guo, "Green resource allocation based on deep reinforcement learning in content-centric iot," *IEEE Transactions on Emerging Topics in Computing*, vol. 8, no. 3, pp. 781–796, 2018.
- [19] X. He, K. Wang, and W. Xu, "Qoe-driven content-centric caching with deep reinforcement learning in edge-enabled iot," *IEEE Computational Intelligence Magazine*, vol. 14, no. 4, pp. 12–20, 2019.
- [20] M. Ma and V. W. Wong, "Age of information driven cache content update scheduling for dynamic contents in heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 8427–8441, 2020.
- [21] W. Jiang, G. Feng, S. Qin, and Y. Liu, "Multi-agent reinforcement learning based cooperative content caching for mobile edge networks," *IEEE Access*, vol. 7, pp. 61 856–61 867, 2019.
- [22] Y. Zhang, B. Feng, W. Quan, A. Tian, K. Sood, Y. Lin, and H. Zhang, "Cooperative edge caching: A multi-agent deep learning based approach," *IEEE Access*, vol. 8, pp. 133 212–133 224, 2020.
- [23] X. Wang, C. Wang, X. Li, V. C. Leung, and T. Taleb, "Federated deep reinforcement learning for internet of things with decentralized cooperative edge caching," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9441–9455, 2020.
- [24] H. Zhu, Y. Cao, X. Wei, W. Wang, T. Jiang, and S. Jin, "Caching transient data for internet of things: A deep reinforcement learning approach," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2074–2083, 2018.
- [25] C. Zhong, M. C. Gursoy, and S. Velipasalar, "Deep multi-agent reinforcement learning based cooperative edge caching in wireless networks," in *Proc. IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.
- [26] B. Rana, Y. Singh, and P. K. Singh, "A systematic survey on internet of things: Energy efficiency and interoperability perspective," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 8, p. e4166, 2021.
- [27] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [28] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. International conference on machine learning (ICML)*, 2015, pp. 1889–1897.
- [29] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.
- [30] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
- [31] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.
- [32] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dornmann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, 2021.



Hongda Wu is currently working toward the Ph.D. degree with the Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, Canada. His research interests include federated learning, reinforcement Learning, wireless networks, and the Internet of Things. He was the recipient of the Academic Excellence Fund from the Faculty of Graduate Studies and the Best Paper Award from IEEE Wireless Communications and Networking Conference (WCNC) from IEEE Communication Society in 2022.



Ali Nasehzadeh received his M.A.Sc degree in Computer Engineering from York University in 2021. He is currently active as a data scientist in the insurance industry, and his main research interests are natural language processing, deep learning, and reinforcement learning.



Ping Wang received her Bachelor and Master degree from Huazhong University of Science and Technology, in 1994 and 1997, respectively, and her Ph.D. degree from the University of Waterloo, Canada, in 2008, all in electrical and computer engineering. Currently, she is an Associate Professor at the Department of Electrical Engineering and Computer Science, York University, and a Tier 2 York Research Chair. Prior to that, she was with Nanyang Technological University, Singapore, from 2008 to 2018. Her research interests mainly include

wireless communication networks, cloud computing, and the Internet of Things. Her scholarly works have been widely disseminated through top-ranked IEEE journals/conferences. She was the recipient of the Best Paper Awards from IEEE Wireless Communications and Networking Conference (WCNC) in 2022, 2020 and 2012, from IEEE Communication Society: Green Communications & Computing Technical Committee in 2018, and from IEEE International Conference on Communications (ICC) in 2007. She is an IEEE Fellow and a Distinguished Lecturer of the IEEE Vehicular Technology Society.