Multi-Agent Hierarchical Graph Attention Reinforcement Learning for Grid-Aware Energy Management

FENG Bingyi, FENG Mingxiao, WANG Minrui,
ZHOU Wengang, LI Houqiang
(University of Science and Technology of China, Hefei 230026, China)

DOI: 10.12142/ZTECOM.202303003
https://kns.cnki.net/kcms/detail/34.1294.TN.20230829.1650.002.html, published online August 30, 2023
Manuscript received: 2023-06-10

Abstract: The increasing adoption of renewable energy has posed challenges for voltage regulation in power distribution networks. Grid-aware energy management, which includes the control of smart inverters and energy management systems, is a trending way to mitigate this problem. However, existing multi-agent reinforcement learning methods for grid-aware energy management have not sufficiently considered the importance of agent cooperation and the unique characteristics of the grid, which leads to limited performance. In this study, we propose a new approach named multi-agent hierarchical graph attention reinforcement learning framework (MAHGA) to stabilize the voltage. Specifically, under the paradigm of centralized training and decentralized execution, we model the power distribution network as a novel hierarchical graph containing the agent-level topology and the bus-level topology. Then a hierarchical graph attention model is devised to capture the complex correlation between agents. Moreover, we incorporate graph contrastive learning as an auxiliary task in the reinforcement learning process to improve representation learning from graphs. Experiments on several real-world scenarios reveal that our approach achieves the best performance and can reduce the number of voltage violations remarkably.

Keywords: demand-side management; graph neural networks; multi-agent reinforcement learning; voltage regulation

Citation (Format 1): FENG B Y, FENG M X, WANG M R, et al. Multi-agent hierarchical graph attention reinforcement learning for grid-aware energy management [J]. ZTE Communications, 2023, 21(3): 11 - 21. DOI: 10.12142/ZTECOM.202303003
Citation (Format 2): B. Y. Feng, M. X. Feng, M. R. Wang, et al., "Multi-agent hierarchical graph attention reinforcement learning for grid-aware energy management," *ZTE Communications*, vol. 21, no. 3, pp. 11 - 21, Sept. 2023. doi: 10.12142/ZTECOM.202303003.

1 Introduction

The increasing shortage of fossil fuels and growing awareness of the need for environmental protection have made the adoption of solar photovoltaic (PV) power generation an important trend in the development of renewable energy. In recent years, more and more PV systems have been integrated into power distribution networks, owing to their low-carbon, clean, and economical benefits. However, the growing popularity of PV systems poses significant challenges to the stability of the power grid voltage. Thus, the need to make optimal use of the existing controllable resources in the power grid to ensure safe and reliable operation, reduce energy waste, and improve the acceptance of renewable energy has gained widespread attention. Prior research has suggested that using an inverter to control PV power conversion can alleviate this issue[1 - 2]. In addition, various energy storage and energy demand responses are also recommended as a means of voltage regulation[3 - 4]. Therefore, a comprehensive scheme is required to coordinate the control among these resources to ensure the stable operation of the entire power system with high PV penetration, which is referred to as grid-aware energy management[5].

Meanwhile, multi-agent reinforcement learning (MARL) has demonstrated impressive efficacy not only in games[6 - 8] but also in real-world applications[9 - 10]. Recently, MARL has also been employed to tackle issues in the power grid[11]. Under a data-driven and model-free setting, MARL does not need precise environment modeling and can be applied in situations with high PV penetration compared with traditional methods[11]. Moreover, using MARL in the power grid also potentially reduces costs and is regarded to have plug-and-play capability[12].

For grid-aware energy management, buildings established on a specific node in a power distribution network are considered as agents, which need to control the charge/discharge rate or electric energy conversion rate of multiple components, such as PV and battery. As the electric energy consumed or

This work is supported by National Key R&D Program of China under Grant No. 2022ZD0119802 and National Natural Science Foundation of China under Grant No. 61836011.
ZHOU Wengang and LI Houqiang are the corresponding authors.

generated will pass through the distribution network and cause voltage fluctuations, the goal of grid-aware energy management is to control these components inside buildings to keep the voltage within the safe range while satisfying building users' energy demands. And grid-aware energy management can be formulated as a cooperative task since all agents share one common objective which is to stabilize voltages at every node in the whole distribution network. Ref. [5] applied deep reinforcement learning to grid-aware energy management as they used independent proximal policy optimization and rule-based control to stabilize voltage.

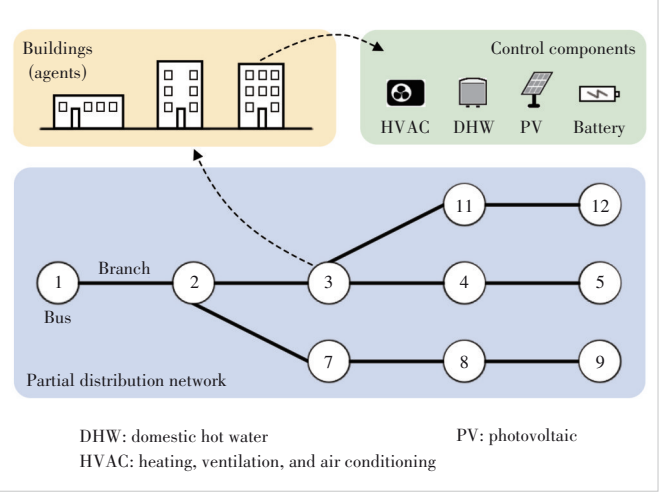
However, it is a non-trivial task to directly apply reinforcement learning algorithms to grid-aware energy management, because of the following challenges. 1) Cooperation among agents. The distribution network is a complex and nonlinear system, which results in a ripple effect to the voltage of all nodes within the distribution network if one agent takes action. Agents in previous work have been limited in their ability to learn cooperation by only utilizing their own observations during both the training and execution phases. This has resulted in difficulty in stabilizing voltage across all nodes. 2) Large state space from the large-scale agent system. There are hundreds of households on the power distribution network in reality. Directly learning a centralized agent system in a training process requires handling large state space and high-dimensional environments, which will cause serious scalability and efficiency problems[13]. 3) Topology of the distribution network. In the distribution network, each node is connected with some other nodes, forming a tree graph structure. The voltage of each node is affected by all other nodes, but the impact declines as the distance increases. Therefore, introducing the topology of the distribution network to algorithms can assist the agents in learning better correlations with each other. 4) The importance of different agents. Each building is regarded as an agent, but the building types are various and different types of buildings have different energy demands. For instance, typically, restaurants have more energy demand at noon for people to have meals, which indicates restaurants must pay more attention than offices when making decisions at noon.

To address the above challenges, we propose a multi-agent hierarchical graph attention reinforcement learning framework (MAHGA) to better stabilize voltage in a power distribution network. Our major contributions are summarized as follows: 1) We approach this task with the paradigm of centralized training and decentralized execution, enabling agents to learn better cooperation. 2) We model the whole distribution network as agent-level topology and bus-level topology. Based on these topologies, we construct an elaborate hierarchical graph attention architecture to extract correlations from agents and power grids. And it can facilitate the MARL-based methods deployed to the realistic power system. 3) Graph contrastive learning with two graph augmentations considered RL characteristics is designed as an auxiliary task in the reinforcement learning (RL) process to improve representation learning from graphs. 4) To the best of our knowledge, this is the first work to consider the topology characteristics to tackle voltage and energy tasks with large-scale agents. Experiments on several real-world datasets reveal that our approach achieves the best performance and can significantly mitigate voltage violations. The paper is organized as follows: In Section 2, we give the background of grid-aware energy management with the MARL formulations and introduce centralized training and decentralized execution. We describe the details of our method in Section 3. In Section 4, we demonstrate the results of the experiments, and in Section 5, we give a literature review of the related work. We conclude our work in Section 6.

2 Problem Formulation

2.1 Grid-Aware Energy Management

In the power system field, power distribution networks are modeled as a tree graph structure, where the node and edge represent a bus and a branch, respectively[14]. More specifically, a bus refers to a node in a power distribution network where power lines, buildings, and other electrical devices join together, and electrical power will be generated, distributed, or consumed here. The distribution network example is shown at the bottom of Fig. 1. For instance, the third bus in this figure is connected with the second bus, the fourth bus, and the eleventh bus. Hundreds of various buildings are distributed on these buses, and each building contains multiple controllable components: 1) HVAC: heating, ventilation, and air conditioning system, which consumes electricity primarily to control the temperature, humidity, and purity of the air inside a building affiliated with the storage to save cooling or thermal energy; 2) DHW: domestic hot water system, which can generate hot water by consuming electricity, affiliated with a tank to store hot water; 3) Battery: used for electricity storage or electricity supply to other equipment; 4) PV: photovoltaics, which is a micro-



▲Figure 1. An illustrative example of grid-aware energy management

generation device comprising solar cells.

The example of buildings and controllable components are shown at the top of Fig. 1. For instance, there are some buildings located on the third bus, and each building has the above four components to control to stabilize voltage after satisfying users' energy demands. Energy demand, including the use of HVAC and DHW, and other electric equipment/appliances (non-shiftable loads), as these components may constantly consume electricity from the power grid.

In terms of constructing the power models, grid-aware energy management environment GridLearn[5], grid models and AC power flows, etc., are modeled using Pandapower. The Pandapower library models the loads of the buildings with real and apparent power specifications; the PV arrays (and corresponding inverters) are modeled as PQ-controlled generators, which are defined to hold the active power P and reactive power Q constant while the voltage is allowed to vary over the limited range. It also calculates real and reactive power at each bus, load, and generator along with voltages at each bus. These values can be adapted to the state space or reward function. And they apply the preconfigured IEEE network model in it.

A large number of PV inside buildings will continuously inject power into the power grid and the power grid also needs to supply power frequently to meet the various users' energy demands, which may lead to frequent undervoltage or overvoltage problems in the power grid. Specifically, the voltage of each bus will be varied if it is injected with active power and reactive power. The exact numerical change of voltage is calculated with these two types of power through certain power flow formulas in the power flow model[5]. The formulas with physical quantities in the distribution network are complicated and non-linear in order to satisfy power system dynamics regulations[2].

The traditional control techniques for large-scale, complex, and non-linear systems are inadequate for real-time decision-making, particularly in systems with high penetration of renewable energy sources[11]. As a result, the employment of deep reinforcement learning algorithms has emerged as a potential and effective method in the literature to mitigate these difficulties.

2.2 MARL Formulations

The cooperative control process of grid-aware energy management can be modeled as a decentralized partially observable Markov decision process (DEC-POMDP)[15]. A DEC-POMDP is an extension of an MDP in decentralized multi-agent settings with partial observability. It can be defined by $\langle S, A, O, R, P, N, \gamma \rangle$, where S is the state space, A_i is the action space for agent i, $o_i = O(s; i)$ is the local observation for agent i at global state s, $P(s'|s, A)$ denotes the transition probability from S to S' given the joint action $A = (a_1, \cdots, a_n)$ for all N agents, $R(s, A)$ is the shared reward function and can also be called a global reward function, and $\gamma \in [0,1)$ is the discount factor. In a DEC-POMDP, each agent takes observation from the environment and executes an action generated by its policy to the environment. In turn, the environment provides one global feedback reward to all agents. During the interaction with the environment, the agents constantly adjust their policies to achieve the best decisions according to the rewards. Considering the grid-aware energy management problem, we describe specific elements in the DEC-POMDP in detail as follows, similar to Ref. [5].

Agent: As shown in Fig. 1, each building is regarded as an agent and will make control decisions on four components to maintain the voltage of all buses within a safe range.

Observation: The agent's observation incorporates 18 state spaces such as outdoor temperature, indoor temperature, voltage magnitude at the located bus, electricity generated by photovoltaic current, electricity consumed by base loads, current energy demand, time of day and the charging states of an HVAC storage device, a DHW storage device, and a battery.

Action: Each building controls four components, namely HVAC energy storage, DHW energy storage, battery storage, and inverters. The action made on each component is continuous and is all set in range $[-1,1]$. For the three energy storage components, the action denotes the increase (action>0) or decrease (action<0) of the energy's rate stored in the corresponding storage device. For the inverter, the action made on the inverter is used to scale the active power and reactive power supplied by PV and the battery.

Reward function: The reward function is mainly based on the voltage deviation from 1 p.u. for each bus. The term p.u. referred to "per unit" is used to express the voltage level in terms of a percentage of the nominal voltage. To alleviate the overvoltage and undervoltage problem across all buses, the reward function is calculated through the voltages on all buses. Specifically, let B denote the set of all buses in the distribution network, v_i denote the voltage on i's bus, and $\boldsymbol{\delta}_i$ a weighting factor to approximately normalize the reward function. The global reward function is calculated as follows:

$$R = -\sum_{i \in B} \left(\delta_i \left(v_i - 1 \right) \right)^2. \tag{1}$$

Note that this function limits the reward to 0 or negative and is devised to penalize the voltage rise deviation and the voltage drop deviation from 1 p.u. followed by Ref. [5]. Voltage deviations are typically measured from 1 p.u. (or 100% of the nominal voltage). For instance, if a 4% voltage deviation is allowed, the voltage safe range is from 0.96 p.u. to 1.04 p.u.

2.3 Centralized Training Decentralized Execution

Centralized training and decentralized execution (CTDE) is one of the paradigms in MARL which assumes that global in-

formation is available during training and that each agent can only use local information during execution to achieve decentralized execution[6 – 7,16]. In this paper, grid-aware energy management is formulated as a cooperative task because all agents share one common objective, which is to stabilize voltages at every bus in the whole distribution network. If each agent only observes local information on its located bus in the training phase, it is usually difficult to learn to control voltage within the safety range and guarantee service quality[2, 11]. One reason is that the environment is non-stationary if only considering the local observation where one agent's action can actually affect the whole distribution network[17].

As a result, we approach grid-aware energy management with the paradigm of CTDE. In CTDE, agents' information is shared in the training phase. In the execution phase and the time we evaluate the algorithm performance, agents are only allowed to make decisions based on their local observation. Specifically, in this paper, we improve and introduce our algorithms all based on the actor-critic class. After combining the structure of CTDE with actor-critic RL algorithms, the critic mainly assists the actor in learning during training, and the input of the critic is global information; while the input of the actor is local information, and the actor needs to make decisions independently in the execution phase. The advantages of CTDE for grid-aware energy management are twofold. On one hand, the centralized training process can motivate multiple agents to learn cooperation by perceiving a more comprehensive landscape. On the other hand, the execution process is fully decentralized without requiring complete information in the training phase, which guarantees efficiency and flexibility in online management. By applying CTDE, the learned strategies can be deployed to the power grid and achieve cooperative control without any communication device. Note that the paradigm of centralized training and centralized execution does not apply to this task due to commercial settings and users' privacy provision[18].

3 Method

In order to address the aforementioned challenges, we approach this grid-aware energy management task with the CTDE paradigm and propose a novel MAHGA approach. In the following, we first introduce the construction of the graph topology. After that, we discuss our hierarchical graph attention architecture for the critic to better extract agents' correlations. Finally, graph contrastive learning is devised as an auxiliary task in the training process to improve representation learning from graphs. The overview of MAHGA is shown in Fig. 2, where the agent takes action depending on its own observation by using the policy. In the training phase, the critic predicts global value based on all agents' observations and is updated by RL loss and graph contrastive loss. The policy is updated by corresponding RL loss with the predicted value from the critic. When in the execution phase, only the policy is used and it makes decisions by solely using agents' local observation.

3.1 Graph Topology Modeling

To capture the correlation between agents, we consider the unique characteristics of distribution networks and construct two graph structures, agent-level graph topology $G^1\left(V^1,D^1\right)$ and bus-level graph topology $G^2\left(V^2,D^2\right)$, respectively. Note that G represents graph topology, V represents the set of all nodes in the graph, and D represents the adjacency matrix which indicates how nodes are connected. For instance, if node i is connected to node j, D_{ij} equals 1; otherwise, D_{ij} equals 0. As for the agent-level graph topology, every agent is modeled as a node. The node set V_1 consists of all agents in the environment. We devise two types of operations to connect edges. The first is the operation of nodes on the same bus where all nodes on the same bus are connected with each other. Nodes on the same bus form a complete graph. The first operation for connecting edges is defined as follows:

$$D_{ij}^{1'}=\begin{cases}1, & b(i)=b(j)\\ 0, & \text{otherwise}\end{cases}, \tag{2}$$

where $b(i)$ denotes the bus, on which node i is located.

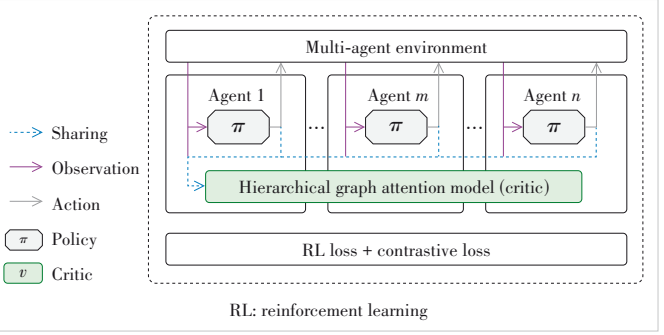
The second operation is to connect nodes from the adjacent buses. In detail, all nodes on bus i will be connected to the nodes on bus j, if bus i and bus j are connected in the distribution network. Different from the first operation, this operation makes all nodes on two adjacent buses form a complete bipartite graph. The second operation for connecting edges is defined as follows:

$$D_{ij}^{1''}=\begin{cases}1, & \text{if } b(i) \text{ and } b(j) \text{ are adjacent}\\ 0, & \text{otherwise}\end{cases}. \tag{3}$$

Then the adjacency matrix obtained from these two operations is taken as the union to form the final adjacency matrix for agent-level graph topology:

$$D_{ij}^{1}=D_{ij}^{1'}\cup D_{ij}^{1''}. \tag{4}$$

To sum up, the first operation is to model the relationship of all the buildings on the same bus, and the second is to model



▲ **Figure 2. Overview of multi-agent hierarchical graph attention (MAHGA), where each agent has one policy and shares the same critic**

the relationship of different buildings on adjacent buses.

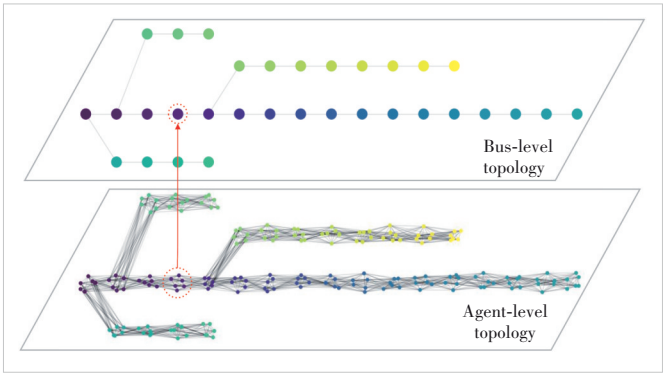
As for the bus-level graph topology, the agents from the same bus are treated as a cluster and thus every bus is modeled as a node. The node set V_2 consists of all the buses in a power distribution network. If two buses are connected in the distribution network, the corresponding node is set to be connected in G_2. The operations for connecting edges are defined as follows:

$$D_{ij}^2 = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

For better illustration, we visualize one example of graph topology in Fig. 3, where nodes enclosed in the red dotted circle are one of the buses and all the agents located on it.

3.2 Hierarchical Graph Attention Architecture

We now present the architecture that exploits the graph topology to handle various observations. The pipeline of the architecture is shown in Fig. 4. The architecture consists of four main components: 1) the agent-level attention module that extracts agent-level representations from the agents' observations based on the agent-level graph topology $G^1\left(V^1,D^1\right)$; 2) the aggregation layer that clusters the agent-level nodes together and aggregates the representations to the embedding from the buses' point of view; 3) the bus-level attention module that extracts bus-level representations with the bus-level graph topology $G^2\left(V^2,D^2\right)$; 4) the readout layer and concatenation that scales down the size of representations and aggregates the representations from the above two attention layers to distill the final representations. The hierarchical characteristics of our architecture are mainly reflected in the different graph attention modules and readouts with corresponding pooling operations.



▲ **Figure 3. Visualized example of agent-level topology and bus-level topology in one case**

3.2.1 Agent-Level Attention Module

We first extract representations from agents' observations through an agent-level attention module using the agent-level graph topology mentioned above. Similar to Ref. [19], in graph attention networks, the importance of node j's feature to node i is calculated as:

$$e_{ij}^k = \vec{c}^T\left(W^k\vec{o}_i \| W^k\vec{o}_j\right), \tag{6}$$

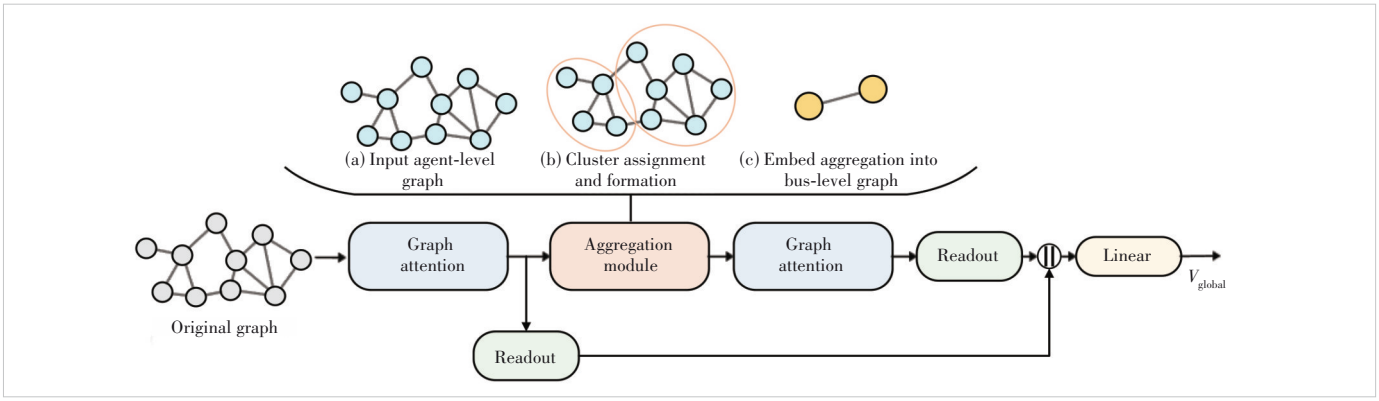
where $\vec{c}^T$ and W^k are learnable parameters, k is the k-th head among K multi-attention heads, $\cdot^T$ represents transposition and $\|$ is the concatenation operation. Then, the coefficients computed by the attention mechanism is defined as:

$$\alpha_{ij}^k = \frac{\exp\left(\text{LeakyReLU}\left(e_{ij}^k\right)\right)}{\sum_{v\in\mathcal{N}_i^1}\exp\left(\text{LeakyReLU}\left(e_{iv}^k\right)\right)}, \tag{7}$$

where $\mathcal{N}_i^1$ represents the set of node i's one-hop neighbor nodes in the graph topology G^1 and the LeakyReLU nonlinearity is applied.

Note that the mask graph attention is adopted and only the neighbor node is allowed to participate in the node i's attention coefficient calculations. The final output of node i in the attention network is formulated as:

$$\vec{h}_i^1 = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_j^1}\alpha_{ij}^k W^k\vec{o}_j\right), \tag{8}$$



▲ **Figure 4. An overview of hierarchical graph attention architecture**

where σ represents the softmax nonlinearity. By applying the above steps to every node in G^1, we can get the agent-level representations h^1 for all nodes.

3.2.2 Aggregation Module and Bus-Level Attention Module

The implementation of a graph attention network is intrinsically flat, as it only propagates information across the edges of a graph. The purpose of this architecture is to define a strategy in a power distribution network that allows one to use two or more graph attention networks hierarchically to extract representations from the graph structure. Formally, given the input embedding, which is the output of the upper network, we seek to define a strategy to output a new coarsened graph embedding.

The new graph embedding contains fewer nodes and node connectivity and can then be used as input to another graph attention module.

As a result, the aggregation module is designed primarily to cluster the agent-level nodes into the classes of buses which means that the agent-level embedding $\vec{h}^1$ will be transformed to the bus-level embedding $\vec{h}^{1'}$ via this module. The aggregation process is demonstrated in the upper part of Fig. 4. Specifically, bus j' s embedding $\vec{h}_j^{1'}$ is the calculation of the embedding of all the agents situated on bus j.

$$\vec{h}_j^{1'} = \phi\left(\left\|_{v,b(v)=j}^{V^1}\left(\vec{\mathrm{h}}_v^1\right)\right.\right), \tag{9}$$

where ϕ denotes the projection function in the aggregation module and $b(v)$ indicates the bus where node v is located.

Then, the bus embedding $\vec{h}^{1'}$ is transformed into the bus representation $\vec{h}^2$ through the bus-level attention module with the graph topology G_2. The structure of the bus-level graph attention module is similar to that of the agent-level attention module which utilizes Eqs. (6) – (8) to calculate the representation but the topology inserted is G_2.

3.2.3 Readout Layer and Concatenation

Inspired by JK-net architecture[20], which has proposed a readout layer that aggregates node features to make a fixed-size representation, we apply a permutation-invariant readout layer to an extracted and integrated representation of agents. The summarized output feature of the readout layer after the agent-level attention module is as follows:

$$\vec{f}^1 = \frac{1}{\left|V^1\right|}\sum_{i=1}^{|V^1|}\vec{h}_i^1 \Big\| \max_{i=1}^{|V^1|}\vec{h}_i^1, \tag{10}$$

where $\|$ is the concatenation operation. The pooling operation in the readout layer is mainly to distill essential information of the state into latent representation while dropping redundant information.

Similarly, we can obtain the integrated representation f^2 after the bus-level graph attention module. As shown in Fig. 4, we apply a readout layer after each attention module. Then the concatenation of each readout layer is applied to aggregate the features.

$$\vec{h}^3 = \left[\vec{f}^1 \middle\| \vec{f}^2\right]. \tag{11}$$

The final representation $\vec{h}^3$ is then fed into the linear function to predict the global state value.

3.3 Graph Contrastive Learning

According to the large-scale energy management environment where hundreds of agents lead to a high dimension of model input, it can be difficult to learn representations through RL objectives as it only depends on reward from the environment. Graph contrastive learning, which has proven its effectiveness on graph prediction tasks[21], has not yet been explored in reinforcement learning, mainly due to the different nature of the problem. Inspired by graph contrastive learning already used in graph prediction tasks and image contrastive learning used in a pixel-based environments[22], we devise a graph contrast learning objective as an auxiliary task in our reinforcement learning task. The objective is devised mainly to stimulate the MAHGA to learn better representation from high-dimensional and various observation inputs.

To apply graph contrastive learning to MARL, we first introduce augmentations that should be made to the graph. The graph augmentation methods include: 1) Observation masking. We randomly select agents and mask certain ratios of agents' observations. Observation masking drives models to recover masked agent observation using their unmasked information. The underlying assumption is that missing partial node attributes does not influence the model performance much. 2) Edge dropping. It is devised to remove the connectivity in G_1 by randomly dropping a certain ratio of edges. It indicates that the semantic meaning of G_1 has certain robustness to the edge connectivity pattern variances. We also follow an independent and identically distributed (i.i.d.) uniform distribution to drop each edge.

Specifically, given the graph data Z_q composed of graph topology G_1 and observations of all nodes from the training batch of the size N, Z_q will undergo graph data augmentations mentioned above to obtain two correlated graphs Z_q^i as a positive pair. The other N–1 graphs in the batch are also augmented to generate N–1 augmented graphs. Then, we utilize the normalized temperature-scaled cross-entropy loss (NT-Xent) for graph Z_q as:

$$l_n = -\log\frac{\exp\left(Z_i^q W_s Z_j^q/\tau\right)}{\sum_{k=1,k\neq q}^{N}\exp\left(Z_i^q W_s Z_j^k/\tau\right)}, \tag{12}$$

where we employ a bilinear product to evaluate the similarity

of pairwise instances. In the formula, τ denotes the temperature parameter, N denotes the size of the training batch and W_s are learnable parameters. Objective l_n will be taken as an auxiliary task to be jointly optimized with the RL objective. Here we select multi-agent proximal policy optimization (MAPPO)[8] as the base algorithm to describe the RL objective. Specifically, following the settings in PPO's clipped surrogate objective[23], we let $r_t(\theta)$ be the probability ratio calculated by the agent's policy and $\hat{A}_t$ be an estimator of the advantage function at timestep t calculated by the global state value. ε is a hyperparameter that implicitly restricts Kullback-Leibler (KL) divergence[23]. The RL objective l_r is defined as:

$$l_r = \min\left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t\right). \tag{13}$$

4 Experiments

In this section, we first introduce the experiment setup. Then we demonstrate and analyze experiment results about overall performance and ablation study. All the experiments have been conducted based on the GridLearn open-source platform[5] with the IEEE 33-bus system[24].

4.1 Experiment Setup

4.1.1 Data Description

We conduct experiments on four real-world scenarios with different climate zones respectively. Each scenario includes 192 buildings distributed on buses, and the corresponding data for the whole year in the specific climate zone[12](climate zone 2A: hot-humid; climate zone 3A: warm-humid; climate zone 4A: mixed-humid; climate zone 5A: cold-humid).

For each scenario, we select four months from four different seasons for training, as different seasons have quite different temperatures, humidity, solar radiation, and users' energy demands, which probably leads to different control strategies. The four training months are mixed and used to train algorithms until convergence. The rest eight months are used for testing. Each month containing 2 880 timesteps is regarded as an episode. The training phase lasts sixteen episodes and each experiment is conducted using 5 random seeds. After the training phase, we evaluate the learned strategy on the test dataset.

4.1.2 Comparison Algorithms

The methods that we evaluated include rule-based control (RBC), independent advantage actor-critic (IA2C), independent proximal policy optimization (IPPO), multi-agent advantage actor-critic (MAA2C), and MAPPO. Specifically, RBC devised by the used environment[5, 12] makes decisions mainly based on the time of day. For example, at 6 a.m., the battery charges and the charge value is 0.138 3. Most of the devices will choose to discharge in the daytime and early evening, and charge at night. The IA2C and IPPO are actor-critic algorithms that directly apply single-agent reinforcement learning algorithms A2C[25] and PPO[23] to MARL. All agents are completely independent. The critic network approximates the expected return only depending on agent-specific observation. MAA2C and MAPPO[8], as an extension of A2C and PPO, are actor-critic algorithms but are in the CTDE paradigm. As extensions of independent algorithms, their critic learns a joint state value function where this centralized critic conditions on all agents' observations rather than the individual observation. And their actor can only use local observation to generate actions same as IA2C and IPPO. In contrast to MAA2C, MAPPO's main advantage is its combination of on-policy optimization with its surrogate objective function.

4.1.3 Evaluation Metrics

Following the evaluation settings in GridLearn[5], we use four metrics to evaluate the performance of algorithms. For better demonstration, we name these four metrics as follows.

1) The number of soft voltage violations (NSVV). It calculates the number of all buses' voltage that is not under control within the soft safe range. Note that the soft safe range of voltage is between 0.96 p.u. and 1.04 p.u.

2) Soft reduction rate (SRR). It calculates the proportion of the algorithm to reduce the number of voltages compared with the rule-based control strategy with the soft safe range. Specifically, for the learned algorithm C, SRR is defined as:

$$\text{SRR}_\text{C} = \frac{\text{NSVV}_\text{RBC} - \text{NSVV}_\text{C}}{\text{NSVV}_\text{RBC}}, \tag{13}$$

where NSVV_RBC and NSVV_C represent the number of soft voltage violations using the rule-based control and the learned algorithm C.

3) The number of hard voltage violations (NHVV). It calculates the number of all buses' voltage that is not under control within the hard safe range. Note that the hard safe range of voltage is between 0.97 p.u. and 1.03 p.u.

4) Hard reduction rate (HRR). It calculates the proportion of the algorithm to reduce the number of voltages compared with the rule-based strategy with the hard safe range. Similar to SRR, for the learned algorithm C, HRR is defined as:

$$\text{HRR}_\text{C} = \frac{\text{NHVV}_\text{RBC} - \text{NHVV}_\text{C}}{\text{NHVV}_\text{RBC}}, \tag{14}$$

where NHVV_RBC and NHVV_C represent the number of hard voltage violations using rule-based control and the learned algorithm C.

Note that NSVV and NHVV evaluate how the algorithm can do to prevent the voltage of all buses from getting out of the safe range, and the lower number represents the better. SRR and HRR describe how much performance the algorithm can enhance compared with the rule-based control method and the higher represents the better. There are two safe voltage ranges:

the soft one and the hard one. The hard range is a more challenging one to evaluate algorithms' performance. Exceeding the safe range frequently will cause lots of problems such as equipment damage and regional power outages.

4.2 Overall Performance

Table 1 reports the median NSVV, SRR, NHVV, and HRR of all algorithms. HMAA2C and HMAPPO refer to MAA2C and MAPPO applied with MAHGA. As shown in the table, our MAHGA framework improves MAA2C and MAPPO and is superior to all other baseline algorithms on four different scenarios concerning four metrics. Owing to the CTDE paradigm and the better learned representations correlative to the grid-aware energy management task, HMAPPO and HMAA2C achieve the best performance among all other algorithms. These two algorithms reduce the number of voltage violations significantly and increase the reduction rate considering both the soft safe range and the hard safe range.

CTDE algorithms, like MAPPO and MAA2C, all perform better compared with independent learning algorithms like PPO and A2C. This proves that centralized critic integrating all agents' observations to have a global perspective can assist agents to implicitly learn better cooperation. Furthermore, HMAPPO and HMAA2C consistently perform better than MAPPO and MAA2C, which validates that MAHGA can make further improvements and motivate the agent to learn a better policy in multiple ways. More analysis of MAHGA will be discussed in an ablation study. As RBC is a well-crafted strategy, independent learning algorithms only show slightly better performance, especially in climate zone 4A.

4.3 Ablation Study

In this section, we conduct an ablation study on MAHGA to further verify the significance of each component. As MAPPO performs better in most scenarios than MAA2C, we choose MAPPO as a representative algorithm to conduct ablation experiments. And the experimental result that MAPPO outperforms PPO according to Table 1 shows that cooperation is implicitly learned and plays an important role in decision making. The following variants of HMAPPO are evaluated on all scenarios: 1) HMAPPOS removes the hierarchical graph attention architecture but uses a single graph attention network with the corresponding readout layer so as to only extract the agent-level representations from the graph attention network; 2) HMAPPOC removes the auxiliary task of graph contrastive learning. As can be seen in Fig. 5, removing any component will cause performance degradation. If we do not consider extracting representations from the bus-level topology, the performance will be significantly degraded. If the bus level and agent level are neither considered, where the algorithm is the original MAPPO, the algorithm will suffer from a large state space where all agents' observations are concatenated and are unaware of the two topologies, which finally leads to low performance. Moreover, introducing attention mechanisms into graphs can implicitly let agents learn how to make decisions with the surrounding agents of different types. These demonstrate that the application of a hierarchical graph attention framework in grid-aware energy management tasks is significantly effective. Besides, we can observe that removing auxiliary tasks of graph contrastive learning will also lead to performance degradation, which indicates that graph contrastive learning can assist the framework to learn representations better.

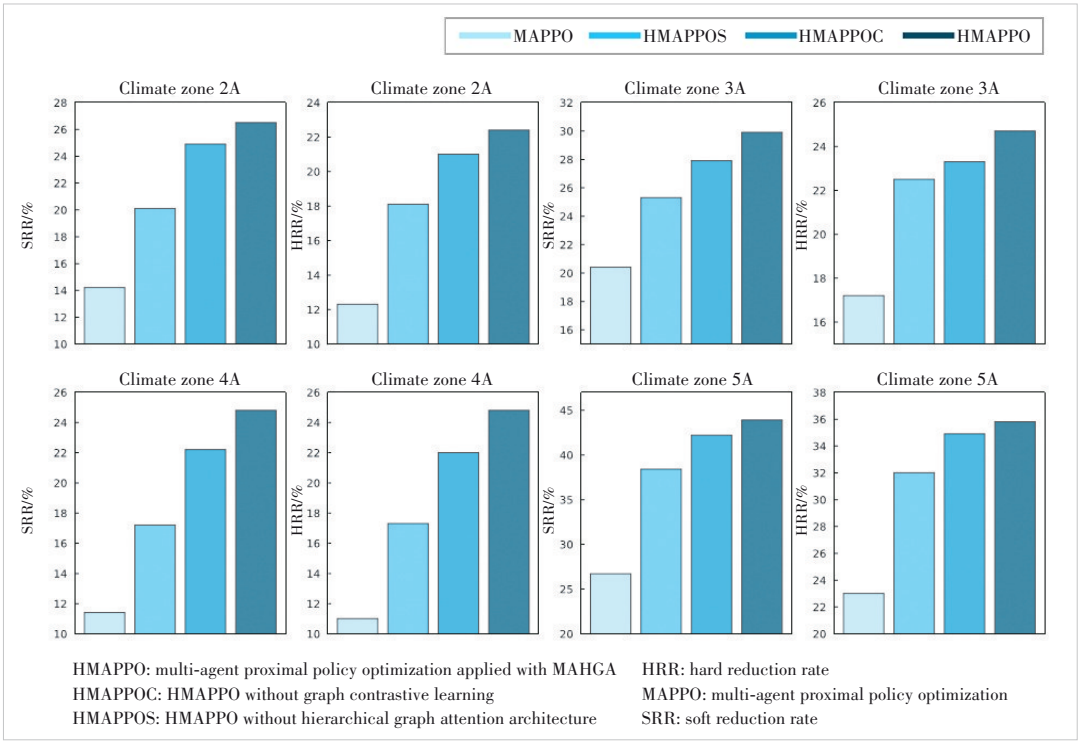
5 Related Work

1) Multi-agent reinforcement learning in power systems. Recently, efforts have been made to apply reinforcement learning to power systems for voltage regulation and energy management due to the progress of machine learning. Ref. [2, 26–27] introduce reinforcement learning in the active voltage control tasks. These works have considered managing a small number of agents and optimizing only reactive power components. In Refs. [2, 26], the load is inflexible and only the PV

▼**Table 1. Overall performance on four scenarios, where HMAA2C and HMAPPO refers to MAA2C and MAPPO applied with multi-agent hierarchical graph attention (MAHGA) (↓ denotes the lower the better, and ↑ denotes the higher the better)**

	Climate Zone 2A				Climate Zone 3A				Climate Zone 4A				Climate Zone 5A			
	NSVV ↓	SRR ↑	NHVV ↓	HRR ↑	NSVV ↓	SRR ↑	NHVV ↓	HRR ↑	NSVV ↓	SRR ↑	NHVV ↓	HRR ↑	NSVV ↓	SRR ↑	NHVV ↓	HRR ↑
RBC	86 181	0.0%	158 736	0.0%	110 902	0.0%	193 751	0.0%	83 648	0.0%	162 076	0.0%	106 823	0.0%	195 277	0.0%
A2C	79 905	7.3%	154 662	2.6%	101 102	8.8%	185 201	4.4%	81 648	2.4%	158 902	2.0%	93 365	12.6%	174 671	10.6%
PPO	79 601	7.6%	153 849	3.1%	100 954	9.0%	184 365	4.8%	81 224	2.9%	155 645	4.0%	92 920	13.0%	173 997	10.9%
MAA2C	73 264	15.0%	139 654	12.0%	89 423	19.4%	162 249	16.3%	74 569	10.9%	144 274	11.0%	79 369	25.7%	154 786	20.7%
MAPPO	73 919	14.2%	139 210	12.3%	88 236	20.4%	160 345	17.2%	74 126	11.4%	144 316	11.0%	78 314	26.7%	150 322	23.0%
HMAA2C	**64 516**	**25.1%**	**125 497**	**20.9%**	**78 158**	**29.5%**	**146 392**	**24.4%**	**63 105**	**24.6%**	**122 568**	**24.4%**	**60 766**	**43.1%**	**127 494**	**34.7%**
HMAPPO	**63 320**	**26.5%**	**123 116**	**22.4%**	**77 724**	**29.9%**	**145 946**	**24.7%**	**62 865**	**24.8%**	**121 829**	**24.8%**	**59 887**	**43.9%**	**125 386**	**35.8%**

A2C: advantage actor critic
HMAA2C: multi-agent advantage actor critic applied with MAHGA
HMAPPO: multi-agent proximal policy optimization applied with MAHGA
HRR: hard reduction rate
MAA2C: multi-agent advantage actor critic
MAPPO: multi-agent proximal policy optimization
NHVV: number of hard voltage violations
NSVV: number of soft voltage violations
PPO: proximal policy optimization
RBC: rule-based control
SRR: soft reduction rate



▲**Figure 5. Ablation studies on four scenarios**

inverter can be controlled. In Ref. [27], demand is regarded to be constant. CityLearn[12] is a platform that satisfies residential energy demands by controlling various shiftable components inside buildings. This environment has been widely used mainly to experiment with various reinforcement learning algorithms and their improvements, in comparison with reinforcement learning baselines and rule-based control. Besides, GridLearn[5], as an extension of CityLearn, considers both the grid-level and building-side objectives and introduces a more realistic environment where hundreds of agents are involved, and agents should control multiple components with corresponding resources to stabilize voltage in a power distribution network after satisfying the residential energy demand. IA2C[25] and IPPO[23] are used, which shows some effectiveness compared with the rule-based control method. In our paper, all experiments are conducted based on GridLearn. Apart from the power system field, in game-like environments, works have been proposed to better motivate the cooperation of agents, such as MAA2C and MAPPO[8], and they have shown good performance in some multi-agent game-like environments. Another approach to achieving agents' cooperation is to learn communication among multiple agents[28 – 30]. However, such approaches always lead to high communication overhead because of the large amount of information transfer.

2) Graph neural networks (GNNs) have been applied successfully to solve prediction and classification tasks in many real-world applications, including recommender systems, chemistry and bioinformatics[21, 31 – 32]. However, GNN has not yet been fully explored in MARL, mostly due to the different nature of the problem. Former works in Refs. [33 – 35] attempt to apply GNN to extract better representations from agents in game-like environments and Ref. [35] considers the unique nature of competitive games to extract information hierarchically. Their methods show encouraging performance with a few agents in games.

However, it is not yet clear whether MARL with GNN can still achieve competitive performance if applied to fully cooperative tasks with large-scale agents in real-world applications such as power systems. In the smart grid field, there are few works related to GNN. In Refs. [36 – 37], mask mechanisms and graph convolutional networks are applied to regulate voltage in single-agent reinforcement learning tasks.

6 Conclusions and Future Work

In this paper, we propose MAHGA, a novel multi-agent reinforcement learning framework for grid-aware energy management. Specifically, we first resolve the problem with the CTDE paradigm aiming to stimulate agents to learn cooperation strategy. Then, depending on modeling the distribution network to two different kinds of topology, we propose a hierarchical attention architecture to better extract agents' correlations from high-dimensional environment and capture the characteristics of grid. In addition, graph contrastive learning is designed to learn a more effective representation in the reinforcement learning training phase. Extensive experiments on four large-scale real-world scenarios have demonstrated the effectiveness of MAHGA where voltage violations can be significantly reduced compared with other baselines.

In our future work, we will explore the generalization over different climates and grid topologies, as well as the possibility of adding more energy control components that buildings can control, like electric vehicles.

References

[1] LIU H T, WU W C. Online multi-agent reinforcement learning for decentralized

inverter-based volt-VAR control [J]. IEEE transactions on smart grid, 2021, 12 (4): 2980 – 2990. DOI: 10.1109/TSG.2021.3060027
[2] WANG J H, XU W K, GU Y J, et al. Multi-agent reinforcement learning for active voltage control on power distribution networks [EB/OL]. (2021-10-27) [2023-06-10]. https://arxiv.org/abs/2110.14300
[3] WANG D X, MENG K, GAO X D, et al. Coordinated dispatch of virtual energy storage systems in LV grids for voltage regulation [J]. IEEE transactions on industrial informatics, 2018, 14(6): 2452 – 2462. DOI: 10.1109/TII.2017.2769452
[4] YI J, WANG P, TAYLOR P C, et al. Distribution network voltage control using energy storage and demand side response [C]//The 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe). IEEE, 2013: 1 – 8. DOI: 10.1109/ISGTEurope.2012.6465666
[5] PIGOTT A, CROZIER C, BAKER K, et al. GridLearn: multiagent reinforcement learning for grid-aware building energy management [J]. Electric power systems research, 2022, 213: 108521. DOI: 10.1016/j.epsr.2022.108521
[6] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments [C]//The 31st International Conference on Neural Information Processing Systems. ACM, 2017: 6382 – 6393. DOI: 10.5555/3295222.3295385
[7] RASHID T, DE WITT C S, FARQUHAR G, et al. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning [C]//35th International Conference on Machine Learning. ICML, 2018: 6846 – 6859. DOI: 10.48550/arXiv.1803.11485
[8] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of PPO in cooperative multi-agent games [J]. Advances in neural information processing systems, 2022, 35: 24611 – 24624. DOI: 10.48550/arXiv.2103.01955
[9] MA Y, HAO X, HAO J, et al. A hierarchical reinforcement learning based optimization framework for large-scale dynamic pickup and delivery problems [J]. Advances in neural information processing systems, 2021, 34: 23609 – 23620
[10] ZHANG H C, FENG S Y, LIU C, et al. CityFlow: a multi-agent reinforcement learning environment for large scale city traffic scenario [C]//The World Wide Web Conference. ACM, 2019: 3620 – 3624. DOI: 10.1145/3308558.3314139
[11] CHEN X, QU G N, TANG Y J, et al. Reinforcement learning for selective key applications in power systems: recent advances and future challenges [J]. IEEE transactions on smart grid, 2022, 13(4): 2935 – 2958. DOI: 10.1109/TSG.2022.3154718
[12] VAZQUEZ-CANTELI J R, DEY S, HENZE G, et al. Citylearn: standardizing research in multi-agent reinforcement learning for demand response and urban energy management [EB/OL]. (2020-12-18) [2023-06-10]. https://arXiv. org/abs/2012.10504
[13] PAPOUDAKIS G, CHRISTIANOS F, RAHMAN A, et al. Dealing with non-stationarity in multi-agent deep reinforcement learning [EB/OL]. (2019-6-11) [2023-06-10]. https://arXiv.org/abs/1906.04737
[14] BAKER K. Learning warm-start points for Ac optimal power flow [C]//The 29th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2019: 1 – 6. DOI: 10.1109/MLSP.2019.8918690
[15] OLIEHOEK F A, AMATO C. Finite-horizon dec-POMDPs [M]. A concise introduction to decentralized POMDPS. Cham: Springer, 2016: 33 – 40. DOI: 10.1007/978-3-319-28929-8_3
[16] FOERSTER J, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients [C]//The AAAI Conference on Artificial Intelligence. ACM, 2018: 2974 – 2982. DOI: 10.1609/aaai.v32i1.11794
[17] CHEN T Y, BU S R, LIU X, et al. Peer-to-peer energy trading and energy conversion in interconnected multi-energy microgrids using multi-agent deep reinforcement learning [J]. IEEE transactions on smart grid, 2022, 13(1): 715 – 727. DOI: 10.1109/TSG.2021.3124465
[18] GLATT R, DA SILVA F L, SOPER B, et al. Collaborative energy demand response with decentralized actor and centralized critic [C]//The 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. ACM, 2021: 333 – 337. DOI: 10.1145/3486611.3488732
[19] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks [EB/OL]. (2017-10-30) [2023-06-10]. https://arxiv.org/abs/1710.10903
[20] XU K, LI C, TIAN Y, et al. Representation learning on graphs with jumping knowledge networks [C]//International conference on machine learning. PMLR, 2018: 5453 – 5462. DOI: 10.48550/arXiv.1806.03536
[21] YOU Y N, CHEN T L, SUI Y D, et al. Graph contrastive learning with augmentations [C]//The 34th International Conference on Neural Information Processing Systems. ACM, 2020: 5812 – 5823. DOI: 10.5555/3495724.3496212
[22] SRINIVAS A, LASKIN M, ABBEEL P. CURL: contrastive unsupervised representations for reinforcement learning [EB/OL]. (2020-04-08) [2023-06-10]. https://arxiv.org/abs/2004.04136
[23] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [EB/OL]. (2017-08-28) [2023-06-10]. https://arXiv. org/abs/1707.06347
[24] BARAN M E, WU F F. Network reconfiguration in distribution systems for loss reduction and load balancing [J]. IEEE power engineering review, 1989, 9 (4): 101 – 102. DOI: 10.1109/MPER.1989.4310642
[25] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning [C]//The 33rd International Conference on International Conference on Machine Learning. ACM, 2016: 1928 – 1937. DOI: 10.48550/arXiv.1602.01783
[26] CAO D, HU W H, ZHAO J B, et al. A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters [J]. IEEE transactions on power systems, 2020, 35(5): 4120 – 4123. DOI: 10.1109/TPWRS.2020.3000652
[27] ZHANG Y, WANG X N, WANG J H, et al. Deep reinforcement learning based volt-VAR optimization in smart distribution systems [J]. IEEE transactions on smart grid, 2021, 12(1): 361 – 371. DOI: 10.1109/TSG.2020.3010130
[28] SUKHBAATAR S, SZLAM A, FERGUS R. Learning multiagent communication with backpropagation [C]//The 30th International Conference on Neural Information Processing Systems. ACM, 2016: 2252 – 2260. DOI: 10.5555/3157096.3157348
[29] FOERSTER J N, ASSAEL Y M, DE FREITAS N, et al. Learning to communicate with Deep multi-agent reinforcement learning [C]//The 30th International Conference on Neural Information Processing Systems. ACM, 2016: 2145 – 2153. DOI: 10.5555/3157096.3157336
[30] JIANG J C, LU Z Q. Learning attentional communication for multi-agent cooperation [C]//The 32nd International Conference on Neural Information Processing Systems. ACM, 2018: 7265 – 7275. DOI: 10.5555/3327757.3327828
[31] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural message passing for Quantum chemistry [C]//The 34th International Conference on Machine Learning. ACM, 2017: 1263 – 1272. DOI: 10.5555/3305381.3305512
[32] LEE J, LEE I, KANG J. Self-attention graph pooling [C]//International conference on machine learning. PMLR, 2019: 3734-3743. DOI: 10.48550/arXiv.1904.08082
[33] IQBAL S, SHA F. Actor-attention-critic for multi-agent reinforcement learning [C]//International Conference on Machine Learning. PMLR, 2019: 2961 – 2970. DOI: 10.48550/arXiv.1810.02912
[34] JIANG J C, DUN C, HUANG T J, et al. Graph convolutional reinforcement learning [EB/OL]. (2018-10-22) [2023-06-10]. https://arxiv. org/abs/1810.09202
[35] RYU H, SHIN H, PARK J. Multi-agent actor-critic with hierarchical graph attention network [C]//The AAAI Conference on Artificial Intelligence. AAAI, 2020: 7236 – 7243. DOI: 10.1609/aaai.v34i05.6214
[36] YOON D, HONG S, LEE B J, et al. Winning the L2RPN challenge: power grid management via semi-markov afterstate actor-critic [EB/OL]. (2021-01-13) [2023-06-10]. https://openreview.net/pdf?id=LmUJqB1Cz8
[37] HOSSAIN R R, HUANG Q H, HUANG R K. Graph convolutional network-based topology embedded deep reinforcement learning for voltage stability control [J]. IEEE transactions on power systems, 2021, 36(5): 4848 – 4851. DOI: 10.1109/TPWRS.2021.3084469

Biographies

FENG Bingyi received his BE degree in computer science from Anhui University, China in 2021. He is working towards his MS degree at University of Science and Technology of China. His research interest focuses on deep reinforcement learning, multi-agent reinforcement learning, and machine learning systems.

FENG Mingxiao received his BE degree in computer science from University of Science and Technology of China in 2017. Now he is working towards his PhD degree with the School of Information Science and Technology, University of Science and Technology of China. His research interests mainly include deep reinforcement learning, multi-agent reinforcement learning, and large language model.

WANG Minrui received his BE degree in computer science from Anhui University, China in 2020, and his MS degree from the University of Science and Technology of China, in 2023. His research interests mainly include deep reinforcement learning, multi-agent reinforcement learning, and machine learning for recommendation systems.

ZHOU Wengang received his BE degree in electronic information engineering from Wuhan University, China in 2006, and PhD degree in electronic engineering and information science from the University of Science and Technology of China (USTC) in 2011. From September 2011 to September 2013, he worked as a postdoc researcher in Computer Science Department at the University of Texas, USA. He is currently a professor at the EEIS Department, USTC. His research interests include reinforcement learning, multimedia information retrieval, and computer vision. In those fields, he has published over 100 papers in IEEE/ACM Transactions and CCF Tier-A International Conferences.

LI Houqiang (lihq@ustc.edu.cn) received his BS, ME, and PhD degrees in electronic engineering from University of Science and Technology of China (USTC) in 1992, 1997, and 2000, respectively. He is a professor and the Vice Dean of the School of Information Science and Technology, USTC, and the Director of MOE-Microsoft Key Laboratory of Multimedia Computing and Communication. He is a fellow of IEEE. His research interests include deep learning, reinforcement learning, image/video coding, image/video analysis, and computer vision, etc. He has authored and co-authored over 300 papers in journals and conferences, and holds over 60 granted patents.