Electric Power Systems Research 192 (2021) 106959

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

Multi-task deep reinforcement learning for intelligent multi-zone residential HVAC control

Yan Du[a], Fangxing Li[a,*], Jeffrey Munk[b], Kuldeep Kurte[b], Olivera Kotevska[b], Kadir Amasyali[b], Helia Zandi[b]

[a] Dept. of Electrical Engineering & Computer Science, The University of Tennessee, Knoxville, TN 37996, USA
[b] Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831, USA

ARTICLE INFO

Keywords:
Data-driven
Deep deterministic policy gradient (DDPG)
Multi-zone HVAC
Multi-task deep reinforcement learning

ABSTRACT

In this short communication, a data-driven deep reinforcement learning (deep RL) method is applied to minimize HVAC users' energy consumption costs while maintaining users' comfort. The applied deep RL method's efficiency is enhanced by conducting multi-task learning that can achieve an economic control strategy for a multi-zone residential HVAC system in both cooling and heating scenarios. The applied multi-task deep RL method is compared with a rule-based benchmark case and a single-task deep deterministic policy gradient algorithm to verify its effective and generalized application in optimizing HVAC operation.

1. Introduction

The latest development in machine learning such as deep learning and reinforcement learning techniques are being widely discussed in many critical areas that were once dominated by human intelligence, such as robotic control and autonomous driving [1], as well as in the field of power and energy [2]. In particular, the deep reinforcement learning (deep RL) method has been implemented for controlling heating, ventilation, and air conditioning (HVAC) systems to achieve both an economic benefit and improved customer comfort. In [3], a model-free deep Q network (DQN) is applied for joint data center and HVAC load control in mixed-use buildings to reduce energy consumption. In [4], the authors compare the value-based DQN method with the policy-based deep policy gradient (DPG) method in residential energy management, and demonstrate that the latter is more suitable to perform online scheduling of energy sources. Given that many control variables in HVAC thermal control are continuous, the deep deterministic policy gradient (DDPG) method is implemented in [5,6], to avoid the discretization of the control variables and to obtain better learning performance. In [7], the authors utilize imitation learning to pre-train the HVAC control agent on historical data to make it behave similarly to the existing controller. Following this, the RL agent continues to improve its policy during the online training using a policy gradient method proximal policy optimization (PPO). In [8], the authors further extend the deep RL algorithm to optimize multi-zone HVAC system control, where a set of actor network and critic network is designed for each thermal zone, and feature extraction from selected neighbor zones is collected to better capture the mutual thermal effects between different zones for improving the control policy.

While the effectiveness of the deep RL based HVAC control methods has been illustrated in the above existing researches, one deficiency is that the majority of the researches focus on learning a single HVAC control task by merely training the algorithm in either the cooling scenario or the heating scenario. Retraining of the algorithm is required when the scenario switches. It is widely known in the RL community that the training of the algorithm can be time-consuming and resource-consuming. Solving only one task at one time is less efficient and acceptable as more complex control problems emerge. Motivated by the above concern, in this short communication, we work on teaching the RL agent to master both the cooling and heating tasks simultaneously to

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the US Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

[*] Corresponding author.
E-mail address: fli6@utk.edu (F. Li).

https://doi.org/10.1016/j.epsr.2020.106959
Received 24 July 2020; Received in revised form 14 October 2020; Accepted 6 November 2020
Available online 19 November 2020

guarantee an optimal HVAC control regardless of the scenario. A multi-task DDPG algorithm is developed for this purpose and is further tested in a multi-zone residential HVAC system. Comparisons with a rule-based HVAC control strategy and a single-task DDPG algorithm demonstrate that the multi-task DDPG algorithm has higher generalization and enables lower energy consumption cost and less user comfort violation through intelligent scheduling.

2. Multi-task DDPG for multi-zone residential HVAC control

The changing of indoor temperature under the control of residential HVAC system can be formulated as a Markov Decision Process (MDP) [9], and the key parameters are defined as follows:

1) *State*: the outdoor temperature $T_{out}(t)$, the indoor temperature $T_{in,z}(t)$ for each zone z, and the retail price $\lambda_{retail}(t)$, where t is the index of time step; 2) *Action*: the setpoint $Setpt_z(t)$ for each zone z; 3) *Reward*: the total energy consumption cost plus the temperature violation penalty, as shown below:

$$r(t) = -\sum_{t^{'}=t-\Delta t}^{t} \lambda_{retail}(t^{'})E_{HVAC}(t^{'}) - \sum_{t^{'}=t-\Delta t}^{t} c_{penalty}(t^{'}) \tag{1}$$

In Eq. (1), Δt is the control interval; the first term is the energy cost of the HVAC system, where $\lambda_{retail}(t')$ is the retail price, and $E_{HAVC}(t')$ is the power consumption; the second term is the penalty for temperature violation, which is calculated as follows:

$$c_{penalty}(t^{'}) = \begin{cases} 1, \ for \ T_{in}(t^{'}) < T_{lower}(t^{'}) - T_{th} \ or \ T_{in}(t^{'}) > T_{upper}(t^{'}) + T_{th} \\ 0, \ else \ wise \end{cases} \tag{2}$$

In Eq. (2), T_{lower} and T_{upper} are the lower and upper bound of user comfort level, respectively; T_{th} is a threshold with a small value, which corresponds with the deadband setting of the HVAC system.

The state transition probability is not defined in the above MDP because the thermal dynamic process of the HVAC system is affected by various uncertain factors such as resistance/capacitor values of the building mass and the weather conditions. This hidden information fails the model-based analytical methods. We consider leveraging the DDPG to overcome the above model unobservability. The DDPG algorithm is well-known for handling complex control problems with continuous state space or action space [10]. In the HVAC control problem, both the temperature and the setpoint are continuous variables, which makes the DDPG a natural fit for the problem. The architecture of the applied multi-task DDPG is shown in Fig. 1, and the algorithm details are shown in **Algorithm 1:**

Fig. 1 is explained as follows: the algorithm first receives the state information for the external environment, such as the temperature and the retail price value. Apart from the state information, the algorithm also receives a task ID from the external environment, which is a 0–1 binary variable indicating whether it is cooling scenario or heating scenario. The task ID is an important indicator of the task that the actor is currently solving. The next step is to normalize the state parameters. Normalization is an essential step because the state parameters from the two tasks can be widely different. For example, the outdoor temperature in the cooling scenario is much higher than that in the heating scenario. Unnormalized data can result in algorithm divergence. The normalized state parameters are then concatenated with the task ID and sent to the deep neural networks. There are two types of the neural networks in the DDPG algorithm, the actor network and the critic network. The actor network is used to generate HVAC control action, and the critic network is used to calculate the Q value as an evaluation of the selected action. Also, for both actor network and the critic network, there is a behavior network and a target network. The behavior network produces the control action, and the target network produces a target value for the behavior network to learn, which resembles the labeled data in the supervised learning. The target network helps stabilize the training process. In total, there are four neural networks in the DDPG algorithm. The structure of each neural network is also revealed in Fig. 1. As can be

Algorithm 1
Multi-task DDPG method for multi-zone HVAC control.

```
Initialization: neural network parameters $\theta^{Q}$, $\theta^{\pi}$, $\theta^{Q'}$ and $\theta^{\pi'}$
    for episode = 1 to arbitrary number do
      Randomly select an HVAC control task (cooling or heating)
      Initialize the building environment based on the selected task
      fort $t = 1$ to $N_T$ do
        Select the control action Setpt$_z(t)$ based on state $s(t)$ and the task ID $\tau$
        Execute Setpt$_z(t)$, get the reward $r(t)$ and the next state $s(t+\Delta t)$
        Store the transition ($s(t)$, Setpt$_z(t)$, $r(t)$, $s(t + \Delta t)$, $\tau$) in the replay buffer
        Collect a mini-batch of transitions ($s^{(i)}(t)$, Setpt($i$) z ($t$), $r^{(i)}(t)$, $s^{(i)}(t + \Delta t)$,
    $\tau^{(i)}$) from the replay buffer
        Update $\theta^{\pi}$ and $\theta^{Q}$
        Soft update $\theta^{\pi'}$ and $\theta^{Q'}$
      end for
    end for
```

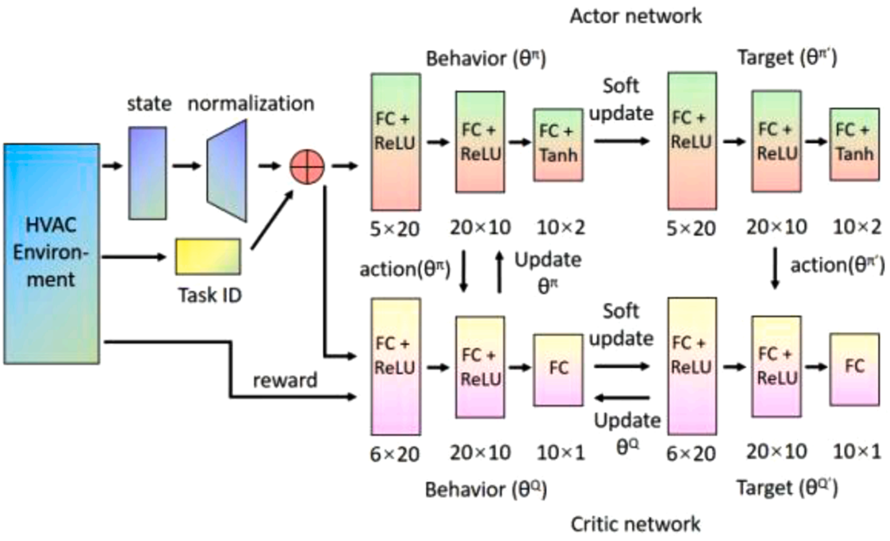


Fig. 1. Multi-task DDPG for multi-zone HVAC control.

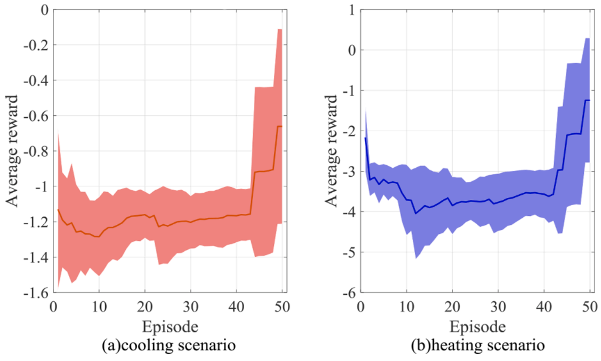
observed, each neural network has two hidden layers and one output layer. The numbers below each hidden layer indicate the number of neurons in that layer. For all hidden layers, ReLU is used as the activation function due to its quasi-linear feature. For the output layer, the two actor networks utilize Tanh as the activation function to confine the control action within a certain range; the two critic networks do not have an activation function for the output layer, since it is difficult to estimate the range of the Q value.

The above DDPG algorithm is trained in a mingled environment, where the algorithm needs to intelligently identify whether heating or cooling action is needed based on the state information and the task ID. Nevertheless, the two tasks share some similar properties such as the same reward function and the same state definition. It is believed that the joint learning of multiple tasks with common structures can boost feature extraction and action exploitation, and results in better learning performance. This deduction will be verified by the simulation results in the next section.

3. Simulation results

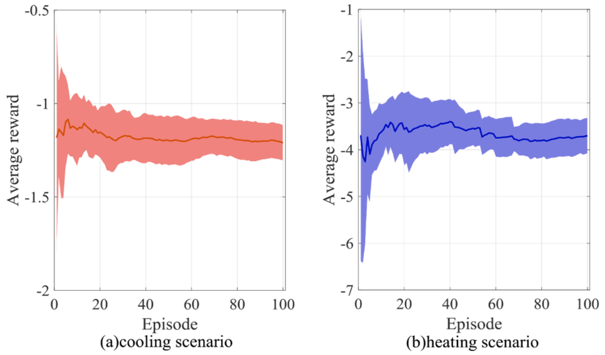
The above multi-task DDPG algorithm is tested in a two-zone residential HVAC building model [11]. The weather data and Georgia Power price data from [12-13] are used for algorithm training and testing. The Georgia Power price contains only two price values, a peak price value at 0.2$/kWh, and an off-peak price at 0.05$/kWh. For the cooling scenario, the algorithm is trained with data from Jul. 1st, 2019 to Jul. 31st, 2019 and tested with data from Aug. 1st, 2019 to Aug. 10th, 2019, and the user comfort level is set to 21 °C-24 °C; for the heating scenario, the algorithm is trained with data from Dec. 1st, 2019 to Dec. 31st, 2019, and tested with data from Jan. 1st, 2020 to Jan. 10th, 2020, and the user comfort level is set to 19 °C-22 °C. The structures of the designed neural networks have been shown in Fig. 1. Note that the action from the actor network is a normalized variable between −1 and 1 as a result of the Tanh activation function. The actual setpoint is calculated as $T_{lower}+(T_{upper}-T_{lower}) \times (action+1)/2$.

1) Learning efficiency of multi-task DDPG: Fig. 2 and Fig. 3 compare the learning efficiency of the multi-task DDPG and single-task DDPG. In the single-task DDPG case, a training process is conducted for the cooling scenario and heating scenario separately. Both algorithms are trained for 100 episodes, with each episode lasting for one day. Note that the x axis in Fig. 2 only shows 50 episodes. This is because in the multi-task DDPG algorithm, one episode learns either the cooling task or the heating task. On average each task accounts for 50 episodes. The average reward in the y axis is normalized so it is a unitless value. The shaded regions in the figures are one standard deviation over 5 runs. As observed in the figures, the multi-task DDPG can achieve a higher average reward with shorter training episodes than the single task learning method. This is because by jointly learning different tasks, a more generalized and robust feature representation that is common to each task can be captured by the hidden layers of the neural network, which in turn will lead to more adaptive action strategies than learning each task separately.



(a)cooling scenario
(b)heating scenario

Fig. 2. Average reward per episode based on multi-task DDPG.

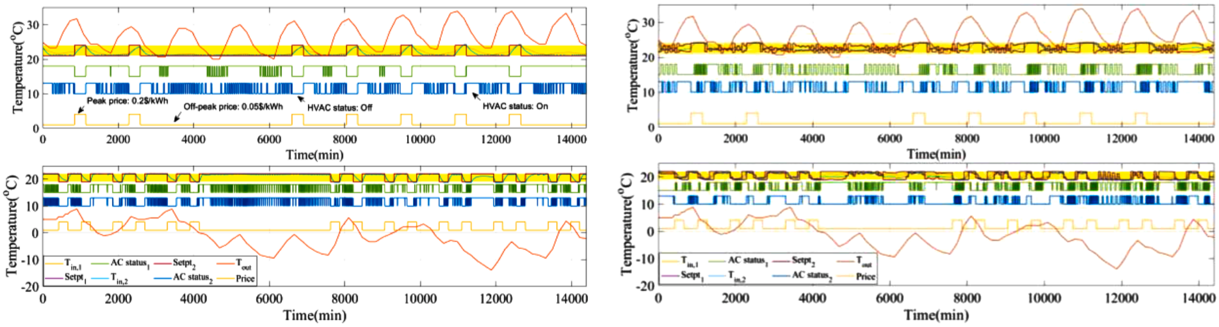


(a)cooling scenario
(b)heating scenario

Fig. 3. Average reward per episode based on single-task DDPG.

2) Comparison with benchmark cases: A rule-based control strategy is applied as the benchmark case to verify the effectiveness of the multi-task deep RL-based HVAC control strategy. The rule-based control strategy follows the simple logic that in the heating scenario, the setpoint is set at the highest value at the off-peak price hours, and the lowest value at the peak price hours, to realize the preheating effect to reduce energy cost, and vice versa for the cooling scenario. Fig. 4 shows the test results from both the rule-based control strategy and from the DDPG RL agent based on a 10-test-day simulation run, and the total energy consumption cost and temperature violation are summarized in Table I. The numbers in the brackets in the column of DDPG algorithm are the percentage of cost savings compared with the rule-based control strategy. As shown in the table, the RL-based setpoint control strategy has a lower energy consumption cost than the rule-based control strategy. The reason is that during the long off-peak price hours, the rule-based control strategy keeps the setpoints at the lowest/highest value, which leaves the HVAC system in “on” status for an unnecessarily long time and brings excessive energy consumption. By comparison, the RL agent coordinates its control strategy with both the price signal and the temperature factor and therefore obtains a more economic control strategy. For example, in the upper figure of Fig. 4(b), between the first price peak and the second price peak, there is a long period of price valley. Instead of always setting the setpoint at the lowest value as in the rule-based case, the RL agent adjusts its setpoint continuously based on the indoor and outdoor temperature, which results in a shorter “on” time for the HVAC system and consequently a lower energy consumption cost.

To verify the generalization of the well-trained multi-task DDPG algorithm, it is further tested with the PJM market price with no additional training. The PJM market price is far different from the above peak/off-peak price and is highly time-variant. For example, the highest PJM market price during the 10 test days in the cooling scenario is 0.31$/kWh, and the lowest price is 0.040$/kWh. The comparison of the DDPG RL agent with the rule-based case is shown in Fig. 5, and the optimization results are also shown in Table I. For the PJM market price, because it contains more than two price values, the average price is used as a threshold to identify the peak/off-peak price. From the figure, it can be observed that the RL agent is still able to follow the price signal to set the HVAC control action even with the dramatic change in the price pattern. For example, in the upper figure of Fig. 5(b), which is the cooling scenario, there is an obvious price peak within the time interval 4000–6000 min, and the RL agent sets the setpoint to the highest value accordingly to save energy consumption cost. In contrast, as shown in Fig. 5(a), in the rule-based case, the control agent sets the setpoint merely depending on whether the current price is lower or higher than

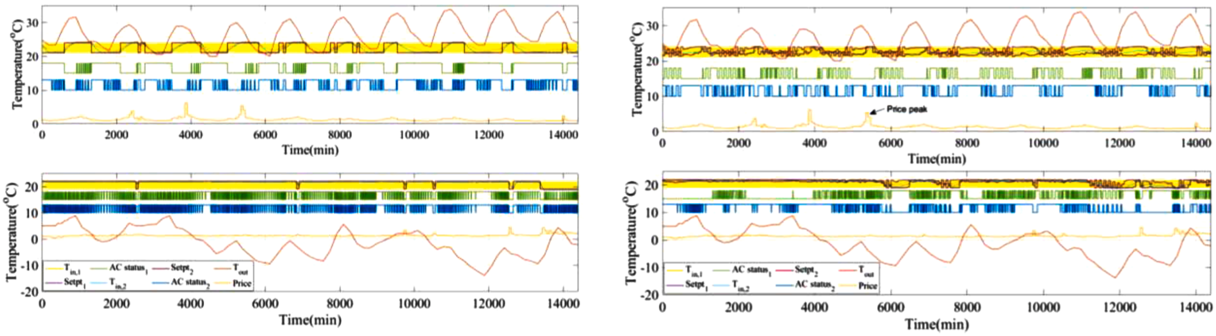


(a) Rule-based (upper: cooling; lower: heating) (b) Multi-task DDPG (upper: cooling; lower: heating)

Fig. 4. Comparison of control strategies under peak/off-peak price.

Table I
Comparison of HVAC control results.

Controlmethod	Peak/Off-peak price				PJM market price			
	DDPG		Rule-based		DDPG		Rule-based	
	Cooling	Heating	Cooling	Heating	Cooling	Heating	Cooling	Heating
Total cost ($)	8.85 **(−8.9%)**	38.70 **(−10.3%)**	9.71	43.15	8.69 **(−8.2%)**	45.16 **(−6.1%)**	9.47	48.11
$c_{penalty}$(minutes)	4	232	0	95	0	41	0	2



(a) Rule-based (upper: cooling; lower: heating) (b) Multi-task DDPG (upper: cooling; lower: heating)

Fig. 5. Comparison of control strategies under PJM market price.

the average price and without considering the other state information, which results in a higher energy consumption cost. The generalization and robustness of the applied multi-task DDPG algorithm is thus proven. Note that in Table I, the temperature violation time of the RL agent is occasionally longer than that of the rule-based case. However, the average temperature violation magnitude is around 0.6°, which is acceptable considering the resulting cost savings.

4. Conclusions

In this short communication, a multi-task DDPG method is applied to learn the setpoint control strategies of multi-zone residential HVAC systems in both cooling and heating scenarios. The multi-task learning process can lead to a more generalized feature extraction among different tasks that share some similarities and improves learning efficiency compared to single-task learning. Comparisons with rule-based control strategies demonstrate the economy and adaptability of the RL-based HVAC control strategy, which uncovers the potential of the multi-task RL algorithm in efficient parallel learning of diverse tasks.

Author statement

Yan Du: Concept development, algorithm development, algorithm implementation, and writing

Fangxing (Fran) Li: Concept development, algorithm development, technical supervision, and editing

Jeffery Munk: Concept development and algorithm development

Kuldeep Kurte: Concept development and algorithm development

Olivera Kotevska: Concept development and algorithm development

Kadir Amasyali: Concept development and algorithm development

Helia Zandi: Concept development, algorithm development, and technical supervision

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was funded in part by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Building Technology Office under contract number DE-AC05-00OR22725, in part by CURENT which is an Engineering Research Center (ERC) funded by the U.S. National Science Foundation (NSF) and DOE under the NSF award EEC-1041877, and in part by the U.S. NSF ECCS awards 1809458 and 2033910.

References

[1] R.S. Sutton, A.G. Barto, Reinforcement Learning: An introduction, MIT press, 2018.

[2] F. Li, Y. Du, From AlphaGo to power system AI, IEEE Power Energy Mag. 16 (2) (Mar. 2018) 76–84.

[3] T. Wei, S. Ren, Q. Zhu, Deep reinforcement learning for joint datacenter and HVAC load control in distributed mixed-use buildings, IEEE Trans. Sustain. Comput. (2019) early access.

[4] E. Mocanu, D.C. Mocanu, P.H. Nguyen, A. Liotta, M.E. Webber, M. Gibescu, J. G. Slootweg, On-line building energy optimization using deep reinforcement learning, IEEE Trans. Smart Grid 10 (4) (Jul. 2019) 3698–3708.

[5] L. Yu, W. Xie, D. Xie, et al., Deep reinforcement learning for smart home energy management, IEEE Internet Things J. 7 (4) (2019) 2751–2762.

[6] G. Gao, J. Li, Y. Wen, DeepComfort: energy-efficient thermal comfort control in buildings via reinforcement learning, IEEE Internet Things J. (2020) early access.

[7] B. Chen, Z. Cai, M. Bergés, Gnu-rl: a precocial reinforcement learning solution for building HVAC control using a differentiable mpc policy, in: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2019, pp. 316–325.

[8] L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, X. Guan, Multi-agent deep reinforcement learning for HVAC control in commercial Buildings, IEEE Trans. Smart Grid (2020) early access.

[9] N. Lu, An evaluation of the HVAC load potential for providing load balancing service, IEEE Trans. Smart Grid 3 (2012) 1263–1270.

[10] T.P. Lillicrap, J.J. Hunt, A. Pritzel, et al., Continuous control with deep reinforcement learning, in: Proc. 4th Int. Conf. Learn. Represent. (ICLR), San Juan, USA, May. 2016, pp. 1–14.

[11] B. Cui, J. Munk, R. Jackson, et al., Building thermal model development of typical house in US for virtual storage control of aggregated building loads based on limited available information, in: 30th International Conference on Efficiency, Cost, Optimisation, Simulation and Environmental Impact of Energy Systems, 2017.

[12] Clean Power Research. [online]: https://www.cleanpower.com/.

[13] Georgia Power. [online]: https://www.georgiapower.com/content/dam/georgia-power/pdfs/residential-pdfs/residential-rate-plans/2.20-tou-reo.pdf.