A Deep Reinforcement Learning Method for Model-based Optimal Control of HVAC Systems

Zhiang Zhang[1,*], Chenlu Zhang[1] and Khee Poh Lam[1,2]

[1]Carnegie Mellon University, Pittsburgh, PA, USA
[2]National University of Singapore, SG

[*]*Corresponding email: zhiangz@andrew.cmu.edu*

ABSTRACT

Model-based optimal control (MOC) methods have strong potential to improve the energy efficiency of heating, ventilation and air conditioning (HVAC) system. However, most existing MOC methods require a low-order building model, which significantly limits the practicability of such methods. This study develops a novel model-based optimal control method for HVAC supervisory-level control based on the recently-proposed deep reinforcement learning (DRL) framework. The control method can directly use whole building energy model, a widely used flexible building modelling method, as the model and train an optimal control policy using DRL. By integrating deep learning models, the proposed control method can directly take the easily-measurable parameters, such as weather conditions and indoor environment conditions, as the input and controls the easily-controllable supervisory-level control points of HVAC systems. The proposed method is tested in an office building to control its radiant heating system. It is found that a dynamic optimal control policy can be successfully developed, and better heating energy efficiency can be achieved while maintaining the acceptable indoor thermal comfort. However, the “delayed reward problem” is found, which indicates the future work should firstly focus on the effective optimization of the deep reinforcement learning.

KEYWORDS

Deep Reinforcement Learning, HVAC Optimal Control, Energy Efficiency

INTRODUCTION

Heating, ventilation and air conditioning (HVAC) system is the major energy consumer in US office buildings. As a result, model-based optimal control (MOC) strategies become popular in the building industry in recent years due to its potential to save HVAC operation energy consumption. The most popular type of MOC is model predictive control (MPC) which, in real-time, uses a model to predict the future evolution of a process and calculates the optimal control decision for the current time step. However, MPC requires the model to be low-order, which is difficult for multi-zone buildings. This significantly limits its practicability. Therefore, some studies focus on using whole building energy model (BEM), a widely-used flexible physical-based building modelling method, in the predictive control of HVAC. However, the slow computational speed of BEM is the major limitation (Zhang and Lam, 2017).

Reinforcement learning (RL) is another alternative for the optimal control of HVAC systems. A standard RL control problem involves a learning agent (hereafter called ``RL agent") interacts with the environment in a number of discrete steps to learn how to maximize the returned reward from the environment, as shown in Figure 1 (Sutton and Barto, 2017). RL is a “model-free” method because the model is used as a simulator offline to let the RL agent learn the optimal control policy. The slow computational speed of BEM is no longer a problem because of the “model-free” feature. However, conventional RL methods need the expert knowledge to design the algorithm and cannot be easily scaled for complicated control problems.

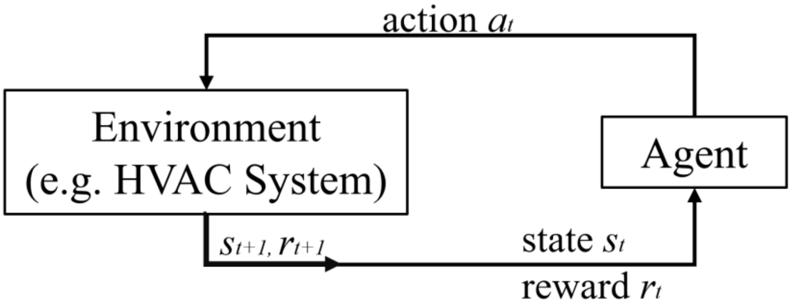


Figure 1. A Standard Reinforcement Learning Setting (Sutton and Barto, 2017)

Deep reinforcement learning (DRL), which uses deep learning models as the core of RL, becomes popular in recent years. DRL makes the "end-to-end" control possible, i.e. the DRL agent can take raw observations as the input and provide raw control actions as the output. Comparing with the conventional RL, DRL no longer requires the expert knowledge to design the algorithm and can be easily scaled for different control problems. However, HVAC optimal control using DRL is studied at an infant stage of research with very limited existing studies using only simple hypothetical building models as the case studies (Li et al., 2017; Wang et al., 2017; Wei et al., 2017).

This study proposes a framework (named “BEM-DRL”) that uses DRL to develop the optimal control policies of HVAC systems based on BEM. The processes, including building modelling, model calibration, DRL training, and control deployment, are explained by using a real-life office building as the case study. The energy efficiency and thermal comfort performance of the proposed control method are analysed using the building model. In addition, the optimization problems of the DRL training are discussed.

METHODS

The BEM-DRL framework is shown in Figure 2, which includes four steps:

1. BEM modelling: The building and its HVAC system are firstly modelled using BEM tools. In this study, EnergyPlus (Lawrence Berkeley National Laboratory, 2016) is used as the BEM tool.
2. HVAC model calibration: The BEM needs to be calibrated against real HVAC operation data. In this study, Bayesian calibration using the method proposed by Chong *et al.* (2017) is used.
3. DRL agent training: The DRL agent will be trained off-line using the calibrated BEM to learn the optimal control policy. The DRL state, reward and action design will be determined based on the building sensor data availability, control optimization objectives and HVAC system control capability. Asynchronous advantage actor critic (A3C) (Mnih et al., 2016) is the DRL training algorithm using this study. A3C surpasses other methods in its smaller computational memory usage and faster computational speed using only CPUs.
4. Control deployment: The trained DRL agent will be deployed to the actual HVAC system using the existing building automation system (BAS) infrastructure.

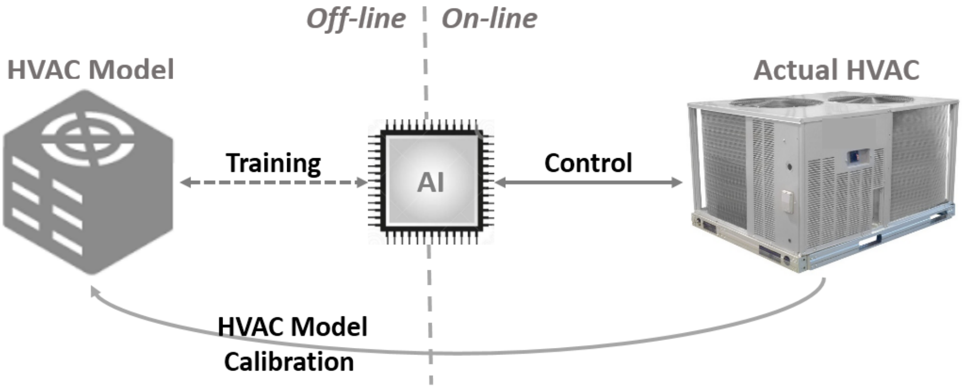


Figure 2. BEM-DRL Framework

RESULTS

The BEM-DRL framework is demonstrated using the Intelligent Workplace (IW) building located in Pittsburgh, PA, USA to control its heating system.

BEM Modelling and Calibration

The case study building, IW, is a one-level office building built on top of an existing building in 1997 with a complete building automation system installed. It has an area of approximately 600m^2, with about 20 regular occupants and a 30-person classroom.

IW uses a water-based radiant heating system that is integrated with the mullion of the window frames. The hot water for the radiant system is supplied by a steam-to-water heat exchanger in the campus. The hot water flow rate is constant while the hot water temperature is variable to respond to different heating demands. The hot water temperature setpoint is currently calculated by a PID controller based on the difference between the zones' average indoor air temperature and its setpoint.

Table 1. Modeling Errors after Calibration

Item	MBE	CVRMSE
Average Indoor Air Temperature	5-min: 0.52%	5-min: 4.82%
Heating Energy Consumption	Hourly: 0.43%	Hourly: 35.96%, Daily: 10.46%

MBE: Mean Bias Error, CVRMSE: Coefficient of Variation of the Mean Squared Error

The BEM of IW is calibrated for the heating energy consumption and indoor average air temperature. Three months measured data in the heating season of Pittsburgh, Jan 1st 2017 to Mar 31th 2017 with 5 minutes interval, is used for the calibration. The heating energy consumption is a calculated value based on the measured radiant system inlet/outlet water temperature and water mass flow rate. The modeling errors after the calibration is shown in Table 1. It is shown that the average indoor air temperature and aggregated heating energy consumption have been calibrated with the acceptable errors.

DRL Agent Training Setup

A DRL problem is defined by the design of the state, action and reward.

The state is a vector representing the environment observations, which includes the following items in this study:
[time, outdoor air temperature, outdoor air relative humidity, wind speed, wind direction, diffuse solar radiation, direct solar radiation, IW steam heat exchanger on/off status (HXOp),

IW average PPD (PPD)[1]*, IW Mullion system supply water temperature setpoint (MULLSSP), IW average indoor air temperature (IAT), IAT setpoint (IATSSP), IW occupancy mode flag (OCCU)*[2]*, IW heating demand (E_{hvac})]*
All items can be easily accessed through the BAS of IW.

The action space includes the discretized Mullion system supply water temperature setpoints from 20 ˚C to 65 ˚C with 5 ˚C interval, and an action to turn the IW steam heat exchanger off.

The reward function determines the control optimization objective. The objective of the study is to minimize the heating energy consumption and maximize the thermal comfort. The design of the reward functions may significantly affect the convergence of the deep reinforcement learning and the final control performance. An empirical reward function combining the heating energy consumption and the indoor thermal comfort is shown in Equation (1):

$$reward = -\begin{cases} \left[\tau * \left([PPD - 0.1]^{+} * \rho\right)^{2} + \beta * E_{hvac}\right]_{0}^{1} \Big|_{Occu=1} \\ \left[\tau * [Stpt_{low} - IAT]^{+} * \lambda + \beta * E_{hvac}\right]_{0}^{1} \Big|_{Occu=0} \end{cases} \tag{1}$$

where τ, β, ρ, λ, Stptlow are the tunable hyperparameters. τ and β control the relative weight between the HVAC energy efficiency and the indoor thermal comfort for the optimization; ρ is a scale factor to penalize large PPD value; λ is the penalty level for the indoor air temperature violation during the unoccupied hours, Stptlow is the indoor air temperature penalty threshold. Note all parameters are normalized.

The calibrated BEM of IW is used for DRL training and testing in this study. The DRL agent is trained in heating season (Jan to Mar) using the TMY3 weather data. The trained DRL agent is tested also for heating season but using the actual Jan-Mar weather data of 2017. The BEM simulation time step and the control time step are both 5 minutes. The Stptlow in Equation (1) is the indoor air temperature setpoint calculated by the IW BAS control logic.

Table 2. DRL Simulated Control Performance in Heating Season (selected experiments)

#	**Hyperparameters**		**Training Model**			**Testing Model**		
	Action Repeat[3]	τ, β, c[4]	Heating Energy (kWh)	PPD_{mean} (%)	PPD_{std} (%)	Heating Energy (kWh)	PPD_{mean} (%)	PPD_{std} (%)
Basecase	N/A	N/A	45302	10.48	4.48	43709	9.46	5.59
1	1	1.0, 1.5, 20	52806	8.72	4.31	47522	8.23	2.46
2	1	1.2, 2.5, 20	44549	11.58	5.35	39484	11.11	4.53
3	1	1.0, 2.5, 10	40101	16.09	9.46	37238	14.20	8.65
4	3	1.0, 1.5, 20	42255	11.46	4.26	38550	10.63	3.34
5	3	1.0, 1.5, 20	43532	10.63	4.23	39109	10.44	3.75
6	3	1.0, 2.5, 10	42104	11.49	4.24	37131	11.71	3.76

[1] PPD is short for Predicted Percentage of Dissatisfied, which is calculated by the BEM engine with the assumptions Clo = 1.0, Met = 1.2 and V_{air} = 0.137m/s.

[2] The occupancy mode flag is determined based on a fixed schedule (the occupancy mode flag is 1 between 7:00 AM and 7:00 PM of weekdays, and between 8:00 AM and 6:00 PM of weekends)

[3] Action repeat means the DRL agent repeats the same control action for multiple time steps.

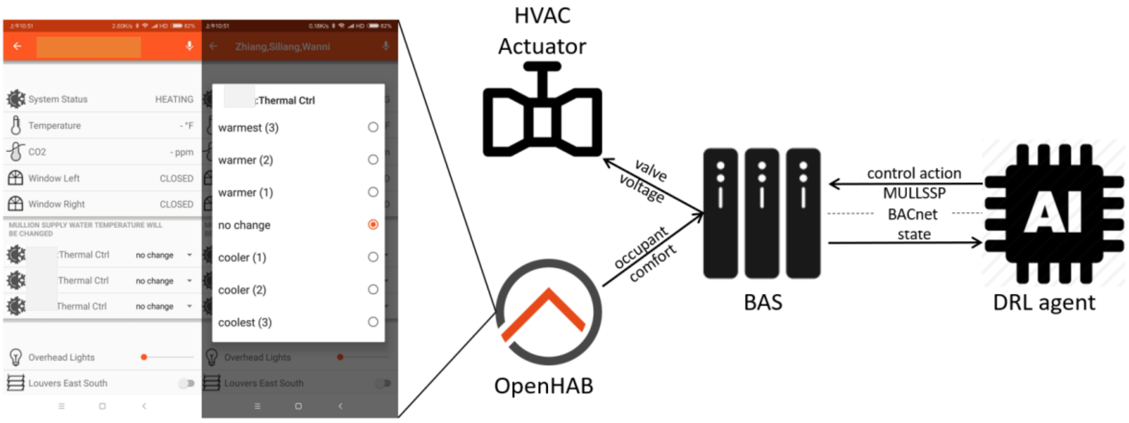
[4] ρ is determined by a function ρ = 1/(PPD_{thres}/100−0.1). ρ =20 and ρ =10 correspond to PPD_{thres}=15 and PPD_{thres}=20 respectively, meaning the reward function (Equation (1)) returns the minimum value if the PPD exceeds the PPD_{thres}.

Simulated Control Performance

The energy consumption and indoor thermal comfort (evaluated by the mean of PPD and standard deviation of PPD) of different DRL agents trained with different hyperparameters are shown in Table 2. The basecase in the table uses the current control logic in IW. Four key hyperparameters are tuned including action repeat, and τ, β, ρ in the reward function. Action repeat benefits the DRL training stability but lowers the flexibility of the trained optimal control policy. τ, β controls the relative weight on the energy efficiency and thermal comfort in the DRL training, and ρ determines the constraint for the maximum PPD during the DRL training. As shown in the table, the relationship between the hyperparameters and the control performance may not be intuitive. This is caused by the delayed reward problem of the DRL training, i.e. the control actions may not take effect soon because of the slow thermal response of the radiant heating system. Further study is needed for the delayed reward problem. Out of the six experiments in Table 6, case 6 performs comparably the best, which saves 15% of the heating energy with only slightly worse indoor thermal comfort quality in the testing model.

Control Deployment

The deployment is based on the existing building automation system (BAS) of IW using BACnet as the communication protocol. Figure 8 shows the deployment architecture. The DRL agent overwrites the Mullion supply water temperature setpoint every 15 minutes through BACnet, and the internal control logic of the BAS controls the three-way water valve to reach the defined water temperature setpoint. A phone app (OpenHAB) is used to collect the occupants' thermal sensation feedbacks. Occupants can select one of the 7 choices from "warmest" to "coolest" reflecting how they want the indoor thermal environment to be. The collected feedbacks are passed to the DRL agent for the control calculation.



System Status HEATING
Temperature - °F
CO2 - ppm
Window Left CLOSED
Window Right CLOSED
MULLION SUPPLY WATER TEMPERATURE WILL BE CHANGED
:Thermal Ctrl no change
:Thermal Ctrl no change
:Thermal Ctrl no change
Overhead Lights
Louvers East South

Zhiang,Siliang,Wanni

:Thermal Ctrl
warmest (3) ○
warmer (2) ○
warmer (1) ○
no change ●
cooler (1) ○
cooler (2) ○
coolest (3) ○

Figure 3. Deployment Architecture of the IW Case Study

DISCUSSIONS

The IW BEM has been calibrated with small error on most metrics, except the hourly CVRMSE for the heating energy consumption, as shown in Table 1. This may be because the hourly heating energy consumption of IW has large variance, which cannot be effectively captured by the BEM. However, aggregated energy consumption (e.g. daily) is more important than spot energy consumption (e.g. hourly) for energy efficiency research.

The slow thermal response of the IW radiant heating system causes the delayed reward problem in the DRL training. Therefore, the hyperparameters of the DRL must be tuned. It is found in Table 2 that some combinations of the hyperparameters can lead to better energy efficiency while maintaining the acceptable thermal comfort. However, no obvious trend has been found in the hyperparameter tuning. Structured method or guideline should be developed, such as better reward function that can re-reward previous actions based on the future evolution of the controlled process.

The DRL agent has been deployed to IW using the architecture proposed in Figure 3. However, by the time this paper is written, the actual control performance test is not yet finished. Future work will study the actual control performance of the DRL agent.

CONCLUSIONS

This study develops a method to use deep reinforcement learning for the BEM-based HVAC optimal control. The method is demonstrated in a case study building with a radiant heating system. An EnergyPlus model is created for the building and calibrated using the Bayesian method. A3C is used to train a DRL agent to develop the optimal control policy for the system supply water temperature setpoint. By simulation, it is found that the optimal control policy can save about 15% heating energy while maintaining the acceptable indoor thermal comfort. This agent is deployed in IW through BACnet protocol and a phone app is used to collect the thermal sensation feedbacks from the occupants. Future work includes analyzing the actual control performance of the IW case study and developing a structured method to solve the delayed reward problem of the DRL training.

ACKNOWLEDGEMENT

The first author would like to thank the China Scholarship Council (CSC) for the financial support for the author’s PhD study.

REFERENCES

Chong, A., Lam, K. P., Pozzi, M., & Yang, J. (2017). Bayesian calibration of building energy models with large datasets. *Energy & Buildings, 154,* 343–355. https://doi.org/10.1016/j.enbuild.2017.08.069

Lawrence Berkeley National Laboratory. (2016). EnergyPlus. Retrieved from https://energyplus.net/

Li, Y., Wen, Y., Guan, K., & Tao, D. (2017). Transforming Cooling Optimization for Green Data Center via Deep Reinforcement Learning. Retrieved from https://arxiv.org/abs/1709.05077

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., … Kavukcuoglu, K. (2016). Asynchronous Methods for Deep Reinforcement Learning. In *33rd International Conference on Machine Learning* (Vol. 48). New York, NY, USA. Retrieved from http://arxiv.org/abs/1602.01783

Sutton, R. S., & Barto, A. G. (2017). *Reinforcement Learning: An Introduction* (Second Edi). Cambridge, MA, USA: MIT Press.

Wang, Y., Velswamy, K., & Huang, B. (2017). A Long-Short Term Memory Recurrent Neural Network Based Reinforcement Learning Controller for Office Heating Ventilation and Air Conditioning Systems. *Processes, 5*(46). https://doi.org/10.3390/pr5030046

Wei, T., Wang, Y., & Zhu, Q. (2017). Deep Reinforcement Learning for Building HVAC Control. In *Proceedings of the 54th Annual Design Automation Conference 2017*. Austin, TX, USA.

Zhang, Z., & Lam, K. P. (2017). An Implementation Framework of Model Predictive Control for HVAC Systems: A Case Study of Energyplus Model-Based Predictive Control. In *ASHRAE 2017 Annual Conference*. Long Island, CA, USA.