processes

Article

Intelligent Temperature Control of a Stretch Blow Molding Machine Using Deep Reinforcement Learning

Ping-Cheng Hsieh

Artificial Intelligence and Robotics Laboratory (AIR Lab), Department of Mechatronics Engineering, National Changhua University of Education, Changhua 50074, Taiwan; pchsieh@cc.ncue.edu.tw

Abstract: Stretch blow molding serves as the primary technique employed in the production of polyethylene terephthalate (PET) bottles. Typically, a stretch blow molding machine consists of various components, including a preform infeed system, transfer system, heating system, molding system, bottle discharge system, etc. Of particular significance is the temperature control within the heating system, which significantly influences the quality of PET bottles, especially when confronted with environmental temperature changes between morning and evening during certain seasons. The on-site operators of the stretch blow molding machine often need to adjust the infrared heating lamps in the heating system several times. The adjustment process heavily relies on the personnel's experience, causing a production challenge for bottle manufacturers. Therefore, this paper takes the heating system of the stretch blow molding machine as the object and uses the deep reinforcement learning method to develop an intelligent approach for adjusting temperature control parameters. The proposed approach aims to address issues such as the interference of environmental temperature changes and the aging variation of infrared heating lamps. Experimental results demonstrate that the proposed approach achieves automatic adjustment of temperature control parameters during the heating process, effectively mitigating the influence of environmental temperature changes and ensuring stable control of preform surface temperature within ±2 °C of the target temperature.

Keywords: stretch blow molding; reinforcement learning; deep learning; intelligent temperature control

check for updates

Citation: Hsieh, P.-C. Intelligent Temperature Control of a Stretch Blow Molding Machine Using Deep Reinforcement Learning. *Processes* **2023**, *11*, 1872. https://doi.org/10.3390/pr11071872

Academic Editor: Xiong Luo

Received: 7 May 2023
Revised: 5 June 2023
Accepted: 20 June 2023
Published: 22 June 2023

1. Introduction

In order to effectively enforce food hygiene and safety protocols, the incorporation of food packaging becomes imperative as an indispensable and critical element. Food packaging can be classified into four distinct categories; namely, paper packaging, plastic packaging, metal packaging, and glass packaging. In recent years, there has been a remarkable surge in the utilization of plastic packaging, particularly, the prevalence of polyethylene terephthalate (PET) bottles, within the many countries worldwide. This surge can be primarily attributed to the escalating consumer demand for bottled water, driven by convenience factors. Currently, PET packaging is widely used for carbonated and non-carbonated beverages due to its advantages over other packaging materials, such as aluminum cans or glass bottles, in terms of energy consumption during production. As a result, PET bottles have a smaller carbon footprint, which is increasingly important from an environmental perspective [1]. Therefore, the requirements for PET bottle production equipment with regards to speed, quality, and energy efficiency have become increasingly strict.

Stretch blow molding is the main technology used in the production of PET bottles. The initial step in the production process involves injecting PET resin into a tube-shaped mold, resulting in the formation of specific structurally preforms. Subsequently, the preforms undergo heating in an infrared heating box, surpassing the glass transition temperature. Concurrently, they are subjected to stretching with a stretch rod and inflation using high-pressure air, thereby achieving the desired bottle shapes. Cooling of the bottles occurs within the mold, followed by their ejection. The entire deformation process is completed within a few seconds. The performance characteristics of PET bottles produced through this method are influenced by three primary variables: the initial shape of the preform, the initial temperature of the preform, and the balance between stretching and blowing rates [2]. During the initial stage (injection molding stage) of the production process, two types of deformations occur, namely, volumetric shrinkage and warpage. Warpage in PET leads to an uneven material distribution across the surface of the preform wall, causing variations in wall thickness. When carbonated drinks are filled into the preforms, high pressure is applied, particularly at areas with minimal wall thickness, leading to the formation of high-stress concentrations. Consequently, under elevated pressure, preforms experience rupture at locations with maximum warpage (areas of stress concentration), resulting in loss of both the beverage and the preform. The findings of Ref. [3] demonstrate that ambient and melting temperatures are the most critical parameters contributing to warpage.

Taking a general stretch blow molding machine as an example, it includes the preform infeed system, transfer system, heating system, molding system, bottle discharge system, etc. Accurate temperature control plays a vital role in PET bottle forming technology. Failure to maintain the heating temperature of PET preforms within an optimal range can result in substandard product outcomes, as the heated PET preforms may not possess the necessary qualities for producing satisfactory end products. In certain seasons, specific regions encounter substantial variations in ambient temperature, where the disparity between morning and evening temperatures can exceed 20 °C. In such scenarios, the on-site operators of the blow molding equipment often need to adjust the infrared heating lamps in the heating system several times. The effectiveness of these adjustments heavily relies on the expertise and experience of the operators.

In addition, the heating system is composed of multiple heating boxes. Each heating box is equipped with several infrared heating lamps on one side, and a reflector made of aluminum alloy on the other side. The reflector of the heating box can efficiently reflect heat energy to the PET preform. It not only ensures the heating effect of the PET preform, but also reduces energy consumption. When the PET preform is heated by the infrared lamp, the heat energy will be transferred from the outside of the preform to the inside. In order to avoid excessive temperature difference between the inside and outside of the preform, a blower is usually installed in the heating box as a cooling element, so that the temperature in the heating box can be uniform and excess heat energy can be taken away. This will prevent the temperature on the outside of the preform from rising too quickly, thereby producing a product with a more uniform thickness and closer to the ideal. However, the heating box is a semi-open structure, which is easily affected by the interference of ambient temperature changes. It has become an automatic production challenge for bottle manufacturers. Furthermore, with the development trend of environmental protection and green manufacturing, preforms made of recycled materials are gradually used, and their requirements for temperature control are more precise and stricter.

In this paper, a novel approach utilizing deep Q-network (DQN) has been introduced as an intelligent technique for adjusting temperature control parameters. The aim is to address issues such as the interference caused by fluctuations in environmental temperature and the gradual degradation of infrared heating lamps over time. From a practical standpoint, the proposed approach offers the advantage of not necessitating any modifications to the existing hardware architecture of the heating system. As a result, it presents a cost-effective and easy-to-implement solution.

2. Literature Review

Proportional-integral-derivative (PID) controllers have gained widespread adoption in various industrial control applications [4,5]. Nonetheless, the most dynamical model of a control system has less parameter uncertainty, which can potentially degrade system performance and even render the closed-loop control system unstable. Adaptive control is a methodology that can address such problems. Fundamentally, an adaptive control system possesses the capability to mitigate the impact of parameter uncertainty within the system through automated adjustments of control parameters.

The parameter adjuster within an adaptive control system ensures the asymptotic stability of the system during steady-state conditions. Nevertheless, during the transient stage, the system may exhibit shaking behavior as the control parameters have not yet converged to their optimal values. Consequently, certain researchers have endeavored to integrate adaptive control theory with variable-structure design to enhance the transient response of the adaptive control system. A previous study had shown that [6,7] adaptive control theory coupled with variable-structure design can yield robustness and superior transient tracking performance, even in situations where the condition of persistent excitation is not satisfied. This approach represents a viable method for addressing the limitations inherent in adaptive control. In addition to parameter uncertainties, control systems are susceptible to the influence of external disturbances or non-parameter uncertainties, encompassing unmodeled dynamics and measurement noise. The above-mentioned issues cause the systematic control parameter to drift and the instability of the closed-loop control system.

With the increasing prevalence of cross-domain integration research, numerous studies have emerged to explore the combination of diverse control theories or machine learning methods with the aim of enhancing the overall performance of control systems. As an illustration, a significant breakthrough was achieved in 2014 through the fusion of deep neural networks and reinforcement learning [8]. This network, called the deep Q-network (DQN), can learn to play 49 Atari games with different rules, objectives, and structures through the same control network architecture. Afterwards, deep reinforcement learning techniques have gained widespread application across diverse domains [9], such as robotic pick-and-place control for random objects [10,11], multi-robot flocking control [12], medical data diagnostics [13], self-driving car control [14], building HVAC (heating, ventilation, and air conditioning) control [15] and building energy management [16], and so on. For example, as mentioned in the literature [16], classical model predictive control (MPC) has demonstrated effectiveness in building energy management. However, it involves the drawbacks of labor-intensive modeling and complex online control optimization. Therefore, the combination of classical MPC with reinforcement learning (RL)-based prediction approaches in model-based RL presents a favorable balance between reliability and practicality of implementation.

In the application of intelligent temperature control, RL has been gradually applied to HVAC control systems. In Ref. [17], an RL algorithm with dual safety policies is proposed for energy savings. This RL model incorporates safety measures into the optimization process by imposing penalties on actions that violate safety constraints. The integration of safety into the RL framework enhances the safety of the controlled HVAC systems while concurrently achieving energy savings. Experimental results show that this RL model reduces energy consumption of an HVAC system by more than 15.02% compared with PID control. Fang et al. [18] utilized the DQN method in a variable air volume (VAV) system to achieve energy savings while ensuring indoor thermal comfort. The effectiveness of the DQN method was evaluated in comparison to rule-based control (RBC) by controlling the setpoints of the air supply temperature and chiller water supply temperature. Their findings demonstrated that, in most cases, the DQN method outperformed RBC in terms of control efficiency. In Ref. [19], the authors propose a model-free actor-critic RL controller, which incorporates a variant of artificial recurrent neural networks known as long-short-term memory (LSTM) networks. The RL controller aims to optimize thermal comfort while minimizing energy consumption. The training and validation results demonstrate that this RL controller improves thermal comfort by an average of 15% and energy efficiency by an average of 2.5% when compared to other alternative strategies. As mentioned earlier, the manufacturing technique of PET bottles has gained widespread popularity. Nevertheless, there has been little research on intelligent temperature control methods within the manufacturing process.

3. Research Method

For stretch blow molding, the heating temperature of the preform stands out as one of the most important parameters in ensuring the quality standards of PET bottle production. Nevertheless, in conventional manual temperature control approaches, operators encounter challenges in promptly and effectively adjusting preform heating parameters in response to fluctuations in ambient temperature and aging variations of heating lamps. The machine operator only intervenes to adjust the heating parameters when a deterioration in the production quality of PET bottles is detected. Hence, the application of artificial intelligence techniques to substitute manual temperature control holds the potential to overcome issues such as the interference of environmental temperature changes and the aging variation of heating lamps, thereby offering avenues for improvement. Ultimately, it possesses the potential for further advancement into an intelligent temperature control technology with industrial 4.0 principles such as self-adjust for variation and self-optimize for disturbance [20].

In this paper, an intelligent temperature control technique based on DQN has been introduced for heating system of stretch blow molding machine. The concept of the proposed approach is as shown in Figure 1. To be more specific, this research focuses on utilizing real-time temperature sensing data obtained from various sources, including the heated preform surface, inside of heating box, and the machine's surrounding environment. By employing the DQN method, the proposed approach enables self-learning of the control strategy (i.e., to build a decision-making agent) for intelligently adjusting the power of the infrared heating lamp and the operating frequency of the blower, thereby ensuring optimal control of the temperature in the stretch blow molding process. In other words, the objective of the learned control strategy is to achieve the temperature variation in the heating box or on the preform surface within a predefined range.

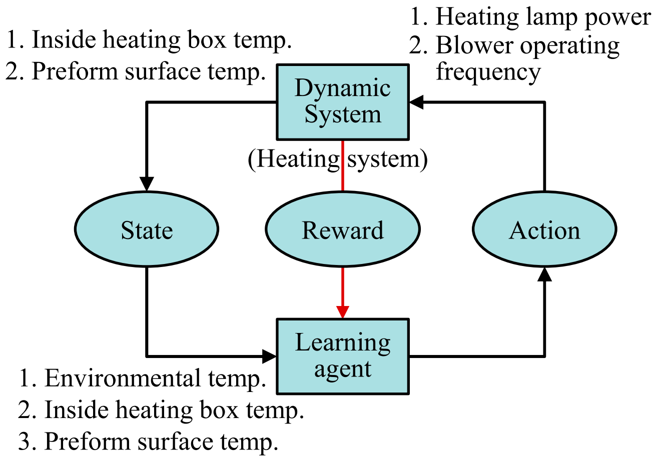

Figure 1. The concept of the proposed approach based on DQN.

Deep Q-Network is based on Q-Learning of an ordinal reinforcement learning (RL) and is an off-policy method. The architecture of DQN was described on Figure 2, where s_t denotes the state at time t and a_t indicates the action of choosing the next state s_{t+1}. Reward r_t depend on the state s_t and action taken at the previous time step $(t-1)$. The main components of DQN include: the environment, main Q-network, target Q-network, replay memory, and loss function, in which the two neural networks share the same structure [5]. During the training process, the parameter θ in main Q-network update in each time step, while the parameter θ^- in target Q-network update periodically. This updating strategy can improve the training stability. Moreover, in order to update parameter θ, the loss function can be defined as the difference between the target value (parameter θ^-) and evaluated value (parameter θ) which is expressed as

$$L(\theta) = E\Big[(z_t - Q(s_t,\ a_t;\theta))^2\Big] \tag{1}$$

where z_t is obtained as

$$z_t = r_t + \gamma \max_{a+1} Q\big(s_{t+1},\ a_{t+1};\theta^-\big) \tag{2}$$

where γ denotes the discount factor (a value between 0 and 1) which is the effect of future rewards to the current decision-making. Finally, the loss function can be optimized by using the stochastic gradient descent (SGD) algorithm.

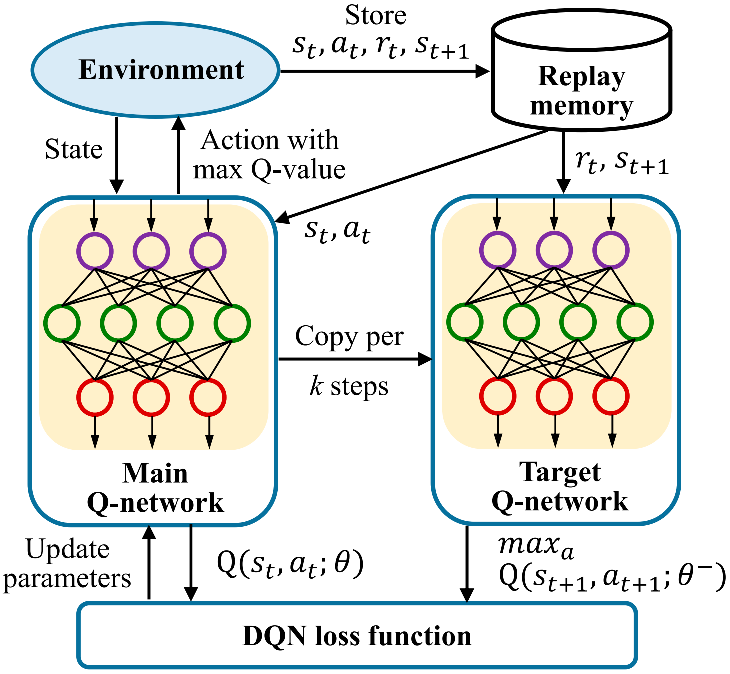

Figure 2. An architecture of DQN method.

As far as technical implementation is concerned, a communication module that can accomplish two-way information transmission with the machine controller is firstly established. As depicted in Figure 3, the communication module can acquire real-time temperature-sensing data from the controller and transmits the decision command learned by the decision-making agent back to the controller. Figure 4 displays the positions at which the temperature sensors were installed in the experiments. A non-contact temperature sensor (CT LT22) is positioned at the terminal of heating Section 2 to measure the temperature of the preform surface. The PT100 sensor is installed within the heating box of heating Section 2 to accurately sense and monitor the internal temperature of the heating box. The temperature and humidity sensor (PR-3002-WS-I20) is installed within the buffering section of the heating system to measure the ambient temperature of the surrounding environment. Based on the deep Q-network method, the neural network structure employed in this paper is depicted in Figure 5, including the input layer, hidden layer, output layer, and action selector.

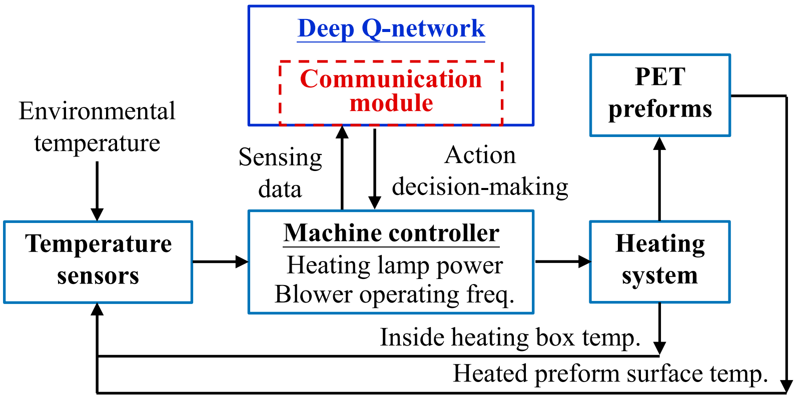

Figure 3. Technical implementation architecture diagram.

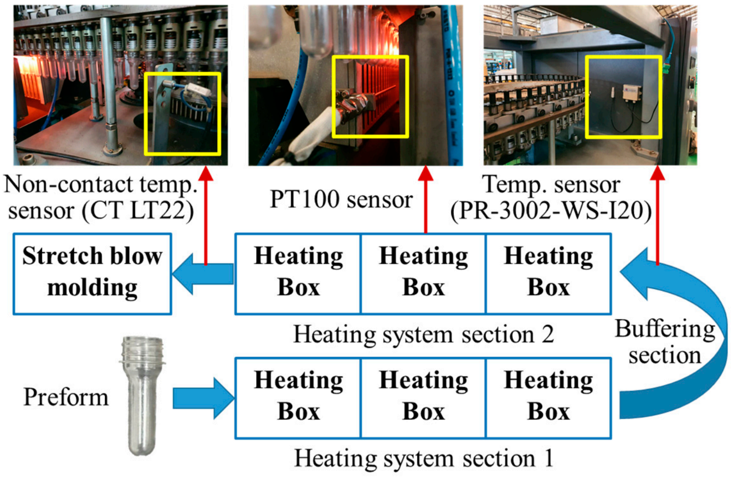

Figure 4. The installation positions of the temperature sensors.

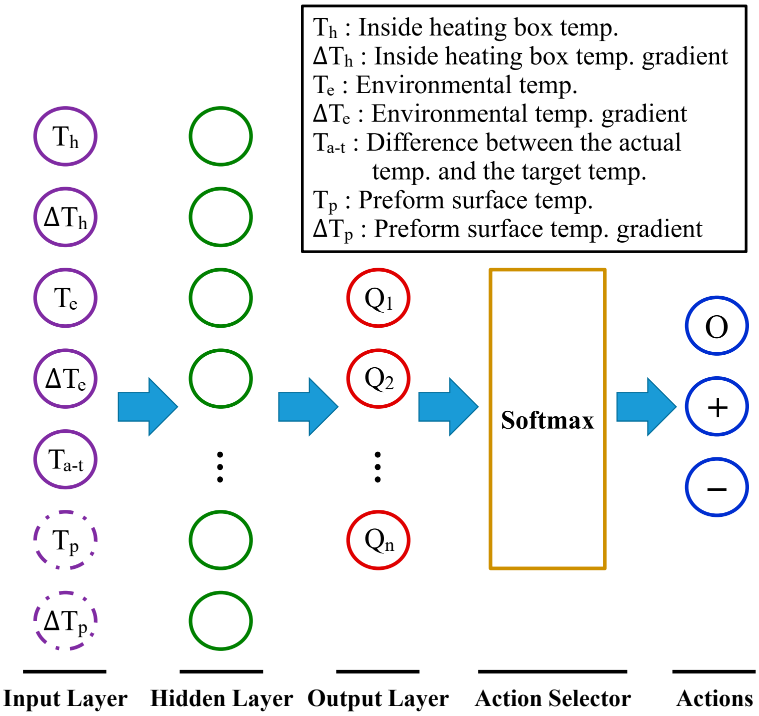
T_h : Inside heating box temp.
ΔT_h : Inside heating box temp. gradient
T_e : Environmental temp.
ΔT_e : Environmental temp. gradient
$T_{a\text{-}t}$: Difference between the actual temp. and the target temp.
T_p : Preform surface temp.
ΔT_p : Preform surface temp. gradient
T_h
ΔT_h
T_e
ΔT_e
$T_{a\text{-}t}$
T_p
ΔT_p
Q_1
Q_2
Q_n
Softmax
O
+
−
Input Layer
Hidden Layer
Output Layer
Action Selector
Actions

Figure 5. The used neural network architecture.

To mitigate the consumption of preforms during the training of the deep Q-network method, the following experimental design is used for the purpose of DQN training and validation. This design aims to effectively save material cost and to reduce training time. Given the significance of the temperature in the heating box on the quality of PET bottle formation, the preform surface temperature will theoretically achieve stability when the temperature in the heating box remains stable. Consequently, the experiments are divided into two distinct stages.

In the first stage, the machine is operated without preforms, with the inside heating box temperature serving as the control target of DQN. In the second stage, which is the same as the official production, the machine is operated with preforms and the control target of DQN is the preform surface temperature. Therefore, in the first stage of experiments, the input layer has five parameters, including the inside heating box temperature, the inside heating box temperature gradient, the environmental temperature, the environmental temperature gradient, and the difference between the actual temperature and the target temperature. In the second stage of the experiments, two additional parameters, namely, the preform surface temperature and the preform surface temperature gradient, were incorporated into the input layer in addition to the original five parameters.

The proposed approach was implemented by using Python 3.9.7 and PyTorch. The hidden layer of the neural network used 60 neurons in both experimental stages. The learning of neural network weights is performed using stochastic gradient descent. The Q-values of the output layer are finally obtained, after which Softmax is used as the action selector. There are three action modes that the decision-making agent can execute, including not adjusting the heating lamp power (indicated by the notation "O"), increasing the heating lamp power (indicated by the notation "+"), and reducing the heating lamp power (indicated by the notation "$-$"). The way to adjust the power every time is 1% of the maximum power. The hyperparameters used by the proposed approach are summarized in Table 1, including the total number of training steps is set to 10,800 steps each time, the batch size is 300, the discount factor (γ) is 0.9, and the learning rate is 0.001. Every 5000 training steps, the weights of the main Q-network are copied to the target Q-network.

Table 1. The hyperparameters setting.

Hyperparameter	Value
Training steps	10,800
Batch size	300
Discount factor	0.9
Learning rate	0.001

Before applying the proposed method, it is crucial to recognize the significance of the reward function's design in RL. This is because it significantly influences both the final performance of the RL model and its convergence speed. The reward function is an incentive mechanism that tells the decision-making agent what is correct and what is wrong using reward and punishment. The decision-making agent in RL always learns how to maximize the reward. Sometimes we need to sacrifice immediate rewards in order to maximize the total rewards. Thus, the reward function should be designed to ensure that the maximization of rewards is consistent with achieving the objectives. In this study, the control objective of the decision-making agent is to keep the temperature variation in the heating box or on the preform surface within ±2 °C of target temperature. The basic principle of reward function design is that when the measured temperature is closer to the target temperature, the reward is higher. Therefore, we propose a reward function based on three factors: (1) accuracy: if the controlled object's current state (temperature) equals the target setpoint (target temperature), the highest reward is given for the current state and action; (2) stability: if the current state is closer to the target setpoint than the previous state, a high reward is given for the current state and action. Otherwise, the reward decreases;

and (3) constraint: if the decision-making agent fails to keep the target temperature within the allowable range, negative rewards are given.

Based on the above description, the pseudo-code of proposed reward function is shown in Algorithm 1. First, the difference value between the current temperature and the target temperature is calculated, and the corresponding reward values are given according to the difference value. If the difference value is within the temperature allowable range, the reward value is set to 1. If the temperature difference is within the allowable range but more than the last time, the reward value is equal to the last time reward value multiplied by the absolute value of (1 − the difference value). If the difference value is outside the temperature allowable range but lower than the last time, the reward value is set to 0.1. If the difference value is outside the temperature allowable range, the reward value is equal to −0.1, multiplied by the difference value.

Algorithm 1 The pseudo-code of proposed reward function

```
while status update:
    difference = abs(current temperature − target temperature)
    max_difference = abs(temperature allowable range)
    if $0 \leq$ difference $\leq$ max_difference:
        reward = 1
        if difference $\geq$ last_difference:
            reward = last_reward $\times$ abs(1 − difference)
    elif difference $<$ last_difference:
        reward = 0.1
    else:
        reward = $-0.1\times$ difference
```

To expedite the algorithm’s training process, the blower in the heating box is employed as a disturbance source to simulate temperature variations resulting from the alternating day and night environment. As shown in Figure 6, the operating frequency of the blower is gradually increased from 10 Hz to 30 Hz, and subsequently gradually decreased from 30 Hz back to 10 Hz. This deliberate manipulation can induce a temperature differential exceeding 10 °C within the heating box. This approach effectively simulates scenarios with environmental temperature differentials surpassing 10 °C. Throughout the experiments, the blower’s operating frequency is adjusted in increments of 4 Hz every 6 min. In essence, this approach enables the simulation of an entire day’s (eight-hour workday) worth of environmental temperature changes within a one-hour timeframe, significantly reducing the required training time for the algorithm.

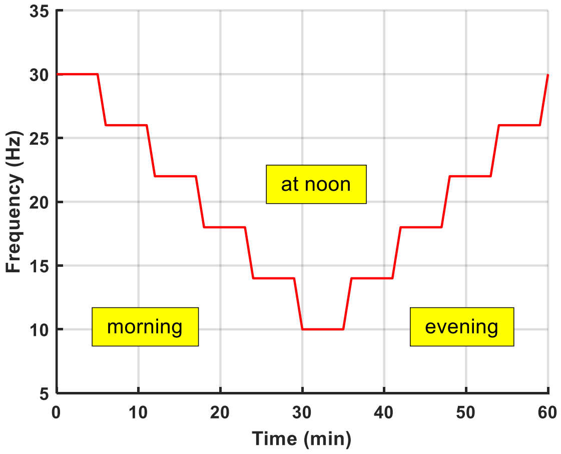

Figure 6. Temperature variation disturbance simulation by adjusting blower operating frequency.

4. Results and Discussion

The specifications of the PET preform used in the experiments are: 16 g in weight, 2.5 mm in thickness, 8.8 cm in total length, and 600 mL of capacity after molding, as shown in Figure 7. In the first stage of experiments, the target temperature of the DQN is set at 45 °C when the initial environmental temperature is 32.4 °C. During the training of the DQN, the temperature variation disturbance simulation shown in Figure 6 is repeatedly performed three times. The experimental results are shown in Figure 8. When comparing the training outcomes during the initial hour, it can be observed that the decision-making agent has progressively learned a better control strategy under the perturbation caused by the substantial temperature fluctuations in the heating box. The experimental results further indicate that during the third hour of training, the inside heating box temperature (depicted by the blue line in the Figure 7) has been able to stably control within ±1.5 °C of the target temperature (45 °C).

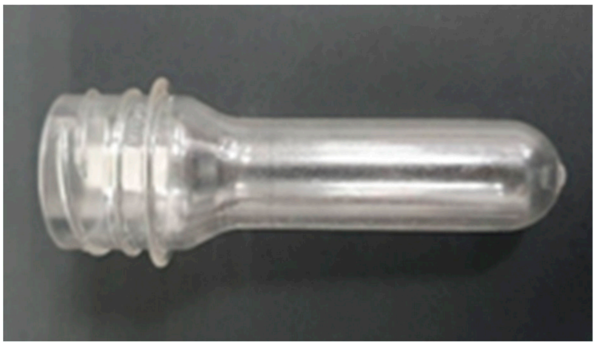

Figure 7. The PET preforms used in the experiments.

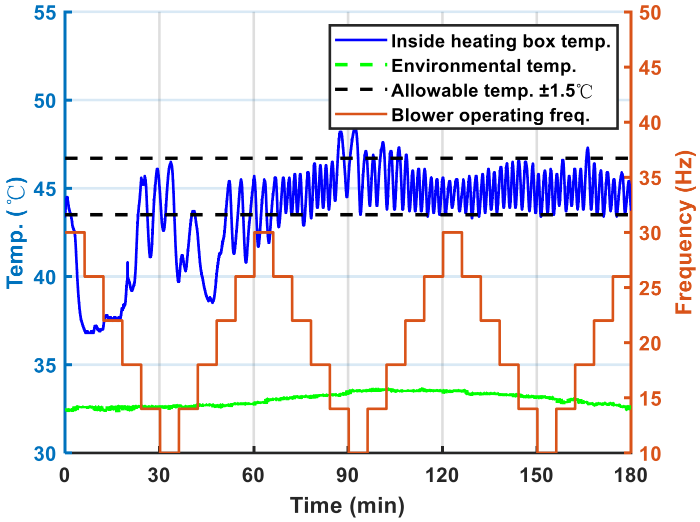

Figure 8. The decision-making agent training result in the first stage of experiment.

To further validate the capabilities of the trained decision-making agent model, identical experiments were conducted on different days. A similar result can be seen in Figure 9, where the trained decision-making agent could effectively control the temperature in the heating box within a range of ±1.5 °C from the target temperature (46.5 °C).

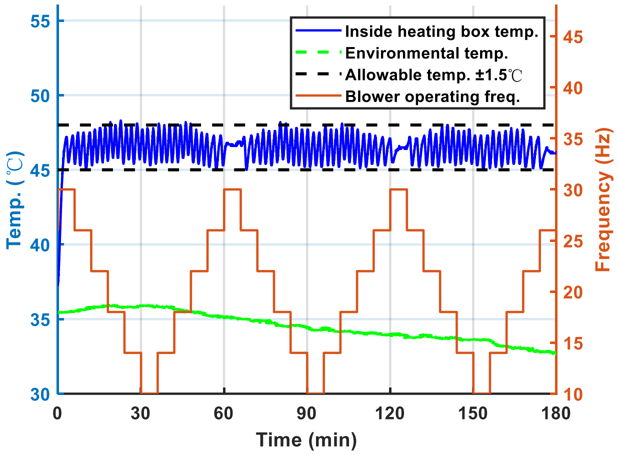

Figure 9. The trained decision-making agent test result in the first stage of experiment.

In the second stage of the experiments, the DQN's target temperature is set at 41.5 °C. Throughout the DQN training process, the temperature variation disturbance simulation illustrated in Figure 5 is repeatedly conducted three times. The experimental results are depicted in Figure 10. In contrast to the first stage of experiments, the input layer of DQN was increased by two new parameters: the preform surface temperature and the preform surface temperature gradient. With the increase in the number of input parameters in the DQN, the training complexity of the DQN model also escalates. Consequently, the experimental results reveal that the training of the DQN reaches a stable state after two hours. The temperature of the PET preform surface can be controlled within the allowable range (within ±2 °C of the target temperature). Furthermore, the variation trend of the preform surface temperature demonstrates a positive correlation with the temperature change trend in the heating box.

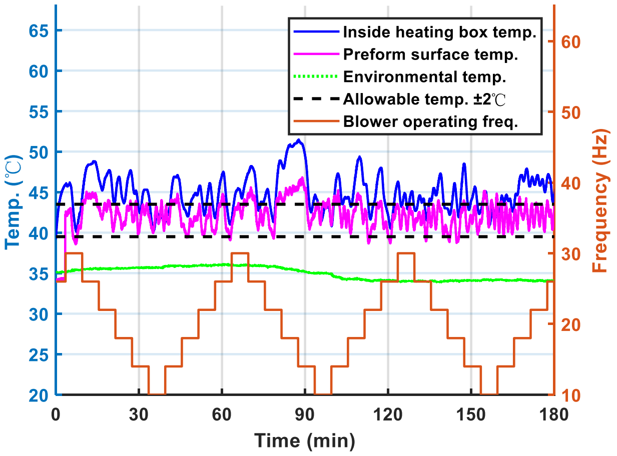

Figure 10. The decision-making agent training result in the second stage of experiment.

Since the PT100 sensor was installed in the heating Section 1 in the previous experiments, the difference between the preform surface temperature and the inside heating box temperature is obvious (as shown in Figure 10). In order to optimize the control performance, the installation position of the PT100 sensor was adjusted to the heating Section 2. To further validate the capabilities of the trained decision-making agent model, identical experiments were conducted on different days. A similar result can be seen from Figure 11, where the trained decision-making agent could effectively control the PET preform surface temperature within a range of ±2 °C from the target temperature. Furthermore, the difference between the preform surface temperature and the inside heating box temperature was significantly reduced as expected. Hence, the proposed approach based on DQN can autonomously regulate the temperature control parameters during the heating process, effectively mitigating the impact of environmental temperature fluctuations.

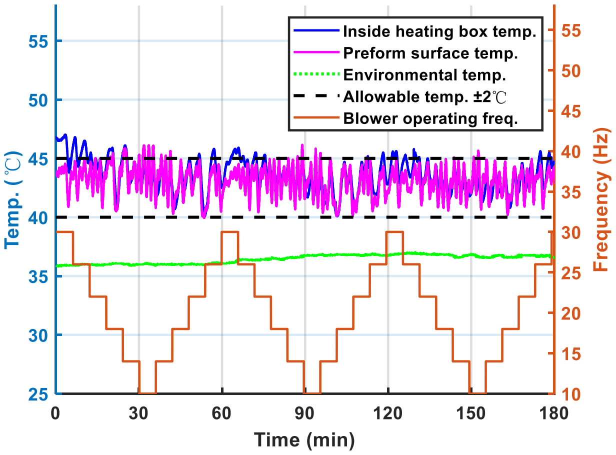

Figure 11. The trained decision-making agent test result in the second stage of experiment.

5. Conclusions

In this paper, we propose an intelligent temperature control approach for a stretch blow molding machine. This approach is based on the deep Q-network method and can solve problems such as the interference of environmental temperature changes and the aging variation of infrared heating lamps. The experimental results reveal that the proposed approach can automatically learn a control strategy to overcome the interference of significant changes in ambient temperature, so that the preform surface temperature can be stably controlled within the range of the target temperature ±2 °C (which meets the requirements of the cooperative manufacturer). Hence, compared with conventional method of manually adjusting the temperature control parameters, the proposed approach can effectively ensure production quality and reduce labor costs. Moreover, it is worth mentioning that this study proposes using the blower in the stretch blow molding machine's heating system as a disturbance source to simulate environmental temperature variations. From the experimental results, it can be observed that this approach indeed reduces the training time of the algorithm. Therefore, this method is believed to serve as a reference for future research aiming to develop automatic parameter adjustment algorithms for heating systems, air conditioning systems, and other control systems, in order to shorten the algorithm's training time.

Furthermore, this paper currently focuses on the temperature control of a specific point on the preform surface as the controlled object. In future investigations, the focus will be extended to include temperature control of multiple points on the preform surface as the controlled variables. Nevertheless, it should be noted that the heating lamps within the heating system have interactive effects on the temperature control of multiple points on the preform surface. Confronted with this situation, the proposed approach holds the potential for greater flexibility in handling multivariable control compared to traditional control architectures.

Funding: This work was supported by the National Science and Technology Council of Taiwan under grant no. MOST 111-2221-E-018-018 and MOST 111-2218-E-008-006.

Data Availability Statement: The data presented in this study are available in the manuscript.

Acknowledgments: The author gratefully acknowledges the support provided to this work by the National Science and Technology Council of Taiwan, under grant no. MOST 111-2221-E-018-018 and MOST 111-2218-E-008-006.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gomes, T.S.; Visconte, L.L.Y.; Pacheco, E.B.A.V. Life cycle assessment of polyethylene terephthalate packaging: An overview. *J. Polym. Environ.* **2019**, *27*, 533–548. [CrossRef]
2. Yang, Z.J.; Harkin-Jones, E.; Menary, G.H.; Armstrong, C.G. Coupled temperature-displacement modelling of injection stretch-blow moulding of PET bottles using Buckley model. *J. Mater. Process. Technol.* **2004**, *153–154*, 20–27. [CrossRef]
3. Chen, J.; Cui, Y.; Liu, Y.; Cui, J. Design and parametric optimization of the injection molding process using statistical analysis and numerical simulation. *Processes* **2023**, *11*, 414. [CrossRef]
4. Tsai, C.C.; Tsai, C.T. Digital command feedforward and PID temperature control for PET stretch blow molding machines. In Proceedings of the 11th Asian Control Conference (ASCC), Gold Coast, QLD, Australia, 17–20 December 2017; pp. 1128–1133.
5. Zhao, Z.; Zeng, J.; Zhou, W. Application of fuzzy control with PID in plastics temperature control of hollow molding machine. In Proceedings of the Fourth International Conference on Innovative Computing, Information and Control (ICICIC), Kaohsiung, Taiwan, 7–9 December 2009; pp. 1009–1012.
6. Hsu, L. Variable structure model reference adaptive control (VS-MRAC) using only input output measurements: The general case. *IEEE Trans. Autom. Control.* **1990**, *35*, 1238–1243. [CrossRef]
7. Chien, C.J.; Sun, K.C.; Wu, A.C.; Fu, L.C. A robust MRAC using variable structure design for multivariable plants. *Automatica* **1996**, *32*, 833–848. [CrossRef]
8. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
9. Li, Y. Reinforcement learning applications. *arXiv* **2019**, arXiv:1908.06973.
10. Gu, S.; Lillicrap, T.; Sutskever, I.; Levine, S. Continuous deep q-learning with model-based acceleration. In Proceedings of the International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016.
11. Kahn, G.; Villaflor, A.; Ding, B.; Abbeel, P.; Levine, S. Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 5129–5136.
12. Zhu, P.; Dai, W.; Yao, W.; Ma, J.; Zeng, Z.; Lu, H. Multi-robot flocking control based on deep reinforcement learning. *IEEE Access* **2020**, *8*, 150397–150406. [CrossRef]
13. Zhu, M.; Zhu, H. Learning a diagnostic strategy on medical data with deep reinforcement learning. *IEEE Access* **2021**, *9*, 84122–84133. [CrossRef]
14. Kiran, B.R.; Sobh, I.; Talpaert, V.; Mannion, P.; Sallab, A.A.A.; Yogamani, S. Deep reinforcement learning for autonomous driving: A survey. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 4909–4926. [CrossRef]
15. Wei, T.; Wang, Y.; Zhu, Q. Deep reinforcement learning for building HVAC control. In Proceedings of the 54th ACM/EDAC/IEEE Design Automation Conference (DAC), Austin, TX, USA, 18–22 June 2017.
16. Zhang, H.; Seal, S.; Wu, D.; Bouffard, F.; Boulet, B. Building energy management with reinforcement learning and model predictive control: A survey. *IEEE Access* **2022**, *10*, 27853–27862. [CrossRef]
17. Lin, X.; Yuan, D.; Li, X. Reinforcement learning with dual safety policies for energy savings in building energy systems. *Buildings* **2023**, *13*, 580. [CrossRef]
18. Fang, X.; Gong, G.; Li, G.; Chun, L.; Peng, P.; Li, W.; Shi, X.; Chen, X. Deep reinforcement learning optimal control strategy for temperature setpoint real-time reset in multi-zone building HVAC system. *Appl. Therm. Eng.* **2022**, *212*, 118552. [CrossRef]
19. Wang, Y.; Velswamy, K.; Huang, B. A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems. *Processes* **2017**, *5*, 46. [CrossRef]
20. Lee, J.; Bagheri, B.; Kao, H.A. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manuf. Lett.* **2015**, *3*, 18–23. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.