frontiers in Energy Research

ORIGINAL RESEARCH
published: 16 August 2021
doi: 10.3389/fenrg.2021.725353

Selective Catalytic Reduction System Ammonia Injection Control Based on Deep Deterministic Policy Reinforcement Learning

Peiran Xie[1]*, *Guangming Zhang*[1], *Yuguang Niu*[1] and *Tianshu Sun*[2]

[1]State Key Laboratory of Alternate Electric Power System With Renewable Energy Sources, School of Control and Computer Engineering, North China Electric Power University, Beijing, China, [2]School of Control and Computer Engineering, North China Electric Power University (Baoding), Baoding, China

The control of flue gas emission in thermal power plants has been a topic of concern. Selective catalytic reduction technology has been widely used as an effective flue gas treatment technology. However, precisely controlling the amount of ammonia injected remains a challenge. Too much ammonia not only causes secondary pollution but also corrodes the reactor equipment, while too little ammonia does not effectively reduce the NOx content. In recent years, deep reinforcement learning has achieved better results than traditional methods in decision making and control, which provides new methods for better control of selective catalytic reduction systems. The purpose of this research is to design an intelligent controller using reinforcement learning technology, which can accurately control ammonia injection, and achieve higher denitrification effect and less secondary pollution. To train the deep reinforcement learning controller, a high-precision virtual denitration environment is first constructed. In order to make the virtual environment more realistic, this virtual environment was designed as a special structure with two decoders and a unique approach was used in fitting the virtual environment. A deep deterministic policy agent is used as an intelligent controller to control the amount of injected ammonia. To make the intelligent controller more stable, the actor-critic framework and the experience pool approach were adopted. The results show that the intelligent controller can control the emissions of nitrogen oxides and ammonia at the outlet of the reactor after training in virtual environment.

Keywords: selective catalytic reduction, deep reinforcement learning, deep deterministic policy, pollution control, nitrogen oxides

OPEN ACCESS

Edited by:
Siyi Luo,
Qingdao University of Technology, China

Reviewed by:
Safdar Hossain SK,
King Faisal University, Saudi Arabia
İsmail Altın,
Karadeniz Technical University, Turkey

***Correspondence:**
Peiran Xie
peiranxiencepu@163.com

Specialty section:
This article was submitted to Advanced Clean Fuel Technologies, a section of the journal Frontiers in Energy Research

Received: *15 June 2021*
Accepted: *06 August 2021*
Published: *16 August 2021*

Citation:
Xie P, Zhang G, Niu Y and Sun T (2021) Selective Catalytic Reduction System Ammonia Injection Control Based on Deep Deterministic Policy Reinforcement Learning. Front. Energy Res. 9:725353. doi: 10.3389/fenrg.2021.725353

INTRODUCTION AND BACKGROUND

In China, thermal power generation is still the main way of generating electricity (Tang et al., 2018). In recent years, with the awakening of people's awareness of environmental protection and increasingly strict environmental protection policies and regulations, pollutant emission control has become an urgent issue for thermal power plants. Among the many pollutants, nitrogen oxides (NOx) have attracted the attention of many scholars because they are highly associated with many serious environmental threats, such as acid rain and photochemical smog. Selective catalytic reduction technology (SCR) is widely used as an efficient denitrification method. The basic

principle of SCR technology is to reduce nitrogen oxides to nitrogen and oxygen by spraying a reductant such as ammonia in the flue gas denitrification reactor. The key to SCR technology is to spray a proper amount of reductant. Ammonia is used as a reducing agent in the SCR reaction, which has some negative effects while eliminating nitrogen oxides. These negative effects include ammonia escape, corrosion of equipment and cost increase. Excess reductant, on the one hand, will enter the atmosphere with the flue gas to cause new pollution, and on the other hand, it will produce corrosive $(NH_4)_2SO_4$ and NH_4HSO_4 in the exhaust gas and corrode the equipment (Strege et al., 2008; Du et al., 2017). However, it is difficult to strike a delicate balance between denitrification and side reactions. This is mainly due to fluctuations in NOx concentrations caused by frequent load adjustments in power plants, where the adjustment of the ammonia quantity is significantly slower than the actual demand due to the delayed characteristics of the measurement control system.

In recent years, artificial intelligence technology is considered to provide a new solution to the problems faced by the electricity industry (Khargonekar and Dahleh, 2018; Mishra et al., 2020). The promise of artificial intelligence technology in power systems stems from the great results it has already achieved in other areas. Deep Q-networks (DQN) has achieve results that exceeded human levels in an Atari game environment (Mnih et al., 2015). Deep reinforcement learning techniques have also performed well in autonomous driving tasks. Direct perception (Chen et al., 2015) and end-to-end (Mnih et al., 2016) control were showcased in the open racing car simulator (TORCS) car racing game using deep reinforcement learning. Further, by using the more realistic virtual environment World Rally Championship 6 (WRC6), the deep reinforcement agent was able to learn to drift through the corners, an advanced professional driving technique (Jaritz et al., 2018). Those researches have revealed the superior decision control capability of deep reinforcement learning. By using multi-agents collaboration, deep reinforcement learning is able to cope with more complex environments, where StarCraft is a strategy game in which players need to maintain a clever balance between competition and cooperation in order to achieve victory (Vinyals et al., 2019). Deep reinforcement learning techniques have not only achieved satisfactory results in computer games, but also perform well in real physical systems. Hwangbo et al. designed a deep reinforcement learning controller to control a real quadcopter (Hwangbo et al., 2017). The controller can control the quadcopter to fly stably along a set path, and can also keep the quadcopter stable in response to disturbances. In another research, a deep reinforcement learning controller can make a legged robot move quickly and save battery power. Further, the deep reinforcement learning controller was able to enable the legged robot to recover from a fall, which is usually difficult for traditional control methods (Hwangbo et al., 2019). Because of the excellent performance of deep reinforcement learning, this method is expected to be used in the control of flue gas denitrification process in thermal power plants. The superior performance of deep reinforcement learning in decision making is expected to determine the optimal amount of ammonia injection to balance the denitrification reaction and ammonia escape.

Reinforcement learning outperforms supervised learning in the control domain due to the "exploration-exploitation" style of learning that differs from supervised learning (Sutton and Barto, 1998). The virtual environment directly affects the performance of the intelligent controller, therefore building a compliant virtual environment is a very important part of this research. An inaccurate virtual environment can lead to bad results known as reality gap. A suitable virtual environment needs to meet both accuracy and responsiveness requirements. Some methods using numerical simulation perform well in terms of accuracy but are too time-consuming (Adamczyk et al., 2014; Stupar et al., 2015). in recent years, data-driven modeling method has been widely adopted by researchers. The main data-driven modeling approaches used in these studies include support vector machines and artificial neural networks (Zhou et al., 2012; Wei et al., 2013; Najafi et al., 2016; Lv et al., 2018). In this research, a model of the denitrification reaction needs to be constructed as an environment for the training of the agent. Since artificial neural networks have more advantages in nonlinear modeling, deep neural networks are used to construct NOx emission models. Details about the NOx model are elaborated in *Denitrification Environment Modeling* Reinforcement learning agent is used as intelligent controller, which generates actions according to the state of the environment. There are three main approaches to realize reinforcement learning agents: value-based, policy-based and actor-critic algorithms (Sutton and Barto, 1998). Value-based reinforcement learning is not suitable for continuous action space, and policy-based reinforcement learning is more suitable for continuous action space tasks. Generally, the policy can be gaussian distribution or SoftMax policy. Policy-based reinforcement learning tends to be less stable during training and has higher sampling variance. The actor-critic algorithm is to introduce a value function on policy-based reinforcement learning to improve the stability of convergence. Since it combines the advantages of both value-based and policy-based reinforcement learning, the actor-critic algorithm has been intensively studied and several variants have been born. Details about the reinforcement learning intelligent controller are described in *Reinforcement learning intelligent controller*.

In order to reduce the emissions of nitrogen oxides and avoid new pollution caused by excessive ammonia escape, this research attempts to use intelligent controller based on deep reinforcement learning to control the injection of reductant. In this research, many analyses and improvements were used to improve the accuracy of the virtual environment due to the importance of a suitable virtual environment. A reinforcement learning agent with the actor-critic framework was used as an intelligent controller for the denitrification reactor. To make the training process more stable, target networks and soft update methods of parameter updating are used. The whole outline of the paper is presented as follows. The purpose and method of constructing a virtual environment is shown in *Denitrification Environment Modeling Reinforcement learning intelligent controller* describes

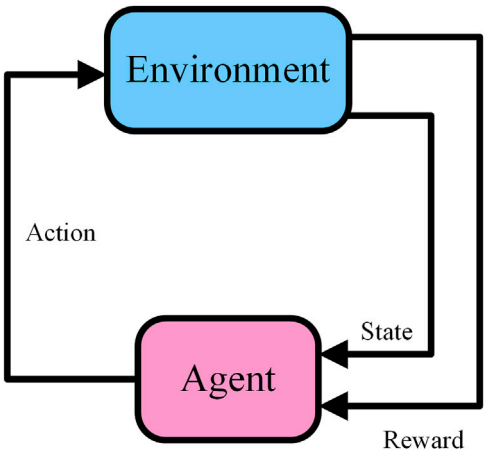

FIGURE 1 | Basic structure of reinforcement learning.

the theory and methods for constructing deep deterministic policy reinforcement learning agent. *Experiments and results* describes the details of training the virtual environment and reinforcement learning intelligent controller, respectively. Conclusions and necessary discussions are presented in *Discussion and Conclusion*.

DENITRIFICATION ENVIRONMENT MODELING

This section will focus on the necessity and challenge of building a virtual environment. According to the characteristics of selective catalytic reduction systems, a high precision virtual environment is designed.

The "exploration-exploitation" learning approach is a distinct marker that distinguishes reinforcement learning from other machine learning methods. Deep learning relies on a lot of labeled data, which are the correct results. There is no correctly labeled data in reinforcement learning. The information for reinforcement learning comes from the feedback of an agent's exploration of the environment. The basic principle of reinforcement learning is shown in **Figure 1**.

Because there are some dangerous results in the process of exploring the environment, and the low efficiency of acquiring experience, agents usually do not train in the real environment. The common research method is to train the agent to a satisfactory state in a virtual environment and deploy the agent in a real environment (Hwangbo et al., 2019). The advantage of using a virtual environment is not only to avoid some dangerous results, but also to improve learning efficiency. Although virtual environments have great advantages, the reality and responsiveness of virtual environments deserve special attention. Related researches indicate the importance of authenticity in virtual environments. The difference between virtual and real can lead to undesirable results known as reality gap (Zagal et al., 2004; Collins et al., 2019; Hwangbo et al., 2019), and a more realistic virtual environment helps agents learn more specialized skills (Chen et al., 2015; Jaritz et al., 2018). These researches indicate that the virtual environment should reflect the characteristics of the real system as much as possible, and that reducing the reality gap is an effective way to improve the control effectiveness of reinforcement learning controllers. On the other hand, since the agent interacts with the virtual environment several times during the learning process, it requires a fast responsiveness of the virtual environment. This requires that the virtual environment should be as simple as possible, using fewer parameters and neural network units. In general, a good virtual environment needs to have high model accuracy and low computational resource consumption.

There are many challenges to building a good virtual environment. These difficulties are mainly caused by the characteristics of the denitrification reaction. SCR systems are multivariate, nonlinear, large lag systems. Several researches have analyzed the effect of different variables on NOx emissions from various aspects such as secondary air, excess air coefficient (Díez et al., 2008; Ti et al., 2017; Stupar et al., 2019). The complex physical and chemical reactions that occur in SCR reactors also contain many nonlinear features. The complex physical and chemical reactions that occur in SCR systems make modeling the mechanism difficult and time-consuming. Some studies modeled SCR systems by numerical calculations and computational fluid dynamics methods (Díez et al., 2008; Adamczyk et al., 2014; Belošević et al., 2016; Wang et al., 2017; Mousavi et al., 2021). Although numerical calculations are highly accurate, their unsatisfactory responsiveness makes them unusable for building virtual environments. The data-driven modeling approach has received a lot of attention in recent years, and it usually has higher accuracy and better responsiveness.

Since the data stored in the distributed control system database of the power plant is time series data, Long Short-Term Memory (LSTM) neural networks, which are more suitable for processing time-series data, were chosen to construct the virtual environment (Tan et al., 2019; Yang et al., 2020). LSTM is a recurrent neural network. Different from other neural networks, recurrent neural network cells have connections with cells from previous time steps. Such a cell structure allows recurrent neural networks to have memory capabilities and to integrate information at different time steps. LSTM adds a gating control mechanism to the traditional recurrent neural network; the gating mechanism enables more efficient transfer of information from previous time steps. This improvement makes LSTM not suffer from "long-term dependency" problem. Specifically, the gate control mechanism consists of three gates, the forget gate input gate and the output gate. The forget gate is responsible for avoiding the over propagation of information from previous time steps. The input gate integrates the information from the current time step with the information from the previous time step and passes it to the output gate. The output gate finally combines the information from the previous and current time steps to produce a new message output. LSTM cell structure can be expressed in **Eq. 1**.

$$\begin{aligned} &\textit{forget gate: } f(t) = \sigma\left(W_{fx}x(t) + W_{fh}h(t-1) + b_f\right) \\ &\textit{input gate: } i(t) = \sigma\left(W_{ix}x(t) + W_{ih}h(t-1) + b_i\right) \\ &\hat{c}(t) = tanh\left(W_{cx}x(t) + W_{ch}h(t-1) + b_c\right) \\ &c(t) = f(t)c(t) + i(t)\hat{c}(t) \\ &\textit{output gate: } o(t) = \sigma\left(W_{ox}x(t) + W_{oh}h(t-1) + b_o\right) \\ &h(t) = o(t)tanh(c(t)) \end{aligned} \tag{1}$$

Where, x_t is the input at t moment. c_t is the cell state at t moment. h_t is the hidden state at t moment. f_t, i_t and o_t are forget gate, input gate and output gate, respectively. $\hat{c}_t$ is new candidate cell state that could be added to the cell state. w and b are the corresponding weights and biases.

Similar to other neural networks, LSTM can improve nonlinear fitting ability by stacking multiple layers. Some scholars have used multilayer LSTM neural networks to study NOx emission (Tan et al., 2019; Yang et al., 2020), but another network structure with better performance is end-to-end network structure. End-to-end model (Cho et al., 2014; Sutskever et al., 2014) was developed by two Google teams. Although the details of the two are slightly different, the main encoder-decoder structure is the same. The encoder summarizes the information of the input sequence data to generate context information represented by a vector. The decoder generates an output sequence based on the context information. Such structures are widely used in areas such as natural language processing and video analysis, and offer better performance than traditional multilayer neural network structures. The denitrification model needs to provide the data of both NOx concentration and ammonia concentration at the SCR reactor outlet for the agent. In order to calculate the two data more accurately to avoid interfering with each other, two decoders are set up for NOx concentration and ammonia concentration respectively, based on using one decoder to extract the information.

Further considering the effect of multiple variables on the denitrification reaction, attention mechanism was introduced to improve the accuracy of the virtual environment model. Li et al. designed an ammonia injection control method for SCR systems based on leading factor analysis (Gang et al., 2016). In particular, the calculations of the dominant factor analysis indicated that the influence coefficients of different factors with NOx concentrations at different time intervals were dynamically varying. The calculation method of influence coefficient is be expressed in **Eq. 2**.

$$E_m = \frac{1}{N}\sum_{t=1}^{N}\left\{\frac{|u_t - u_0|}{u_0} - \frac{|p_{t+m} - p_0|}{p_0}\right\}^2 \tag{2}$$

Where, N is the number of historical samples; u_t is the parameter affecting the reactor inlet NOx concentration at time t; p_{t+m} is the reactor inlet NOx concentration at time $t+m$; u_0 and p_0 are the average values of the parameters affecting the reactor inlet NOx concentration and the average values of the denitrification reactor inlet NOx concentration, respectively.

To be able to better cope with this dynamic change, the attention mechanism was introduced. Attention mechanisms appeared earlier in the field of computer vision, but have been more widely used in the field of natural language processing. There are two common attention mechanisms that have been widely used in sequence data, Bahdanau attention and Luong attention (Bahdanau et al., 2015; Luong et al., 2015). The core idea of both is the same, calculating additional weights for data at different time steps to highlight useful information. The two attention mechanisms differ only in the method of calculating attention weights. Similar to the reason for setting two decoders for the SCR reactor outlet NOx and ammonia concentrations separately, it is necessary to calculate the respective attention weights for NOx and ammonia concentrations. Basic structure of denitrification environment is shown in **Figure 2**.

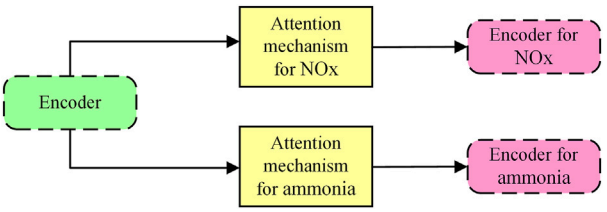

FIGURE 2 | Basic structure of denitrification environment.

In this section, the purpose and importance of the virtual environment is presented. By analyzing the characteristics of the SCR system, the requirements of the virtual environment are clarified. For accuracy and responsiveness, a single encoder dual decoder virtual environment with attention mechanism is proposed.

REINFORCEMENT LEARNING INTELLIGENT CONTROLLER

The objective of this research is to attempt to use artificial intelligence controllers to control ammonia injection in selective catalytic reduction systems to achieve a reduction in NOx emissions while reducing ammonia escape. In this research, the controller is the reinforcement learning agent. There are usually three ways to realize the agent, which are value-based, policy-based and actor-critic framework (Csáji and Monostori, 2008; Liu et al., 2020; Yu and Sun, 2020). As shown in **Figure 1**, the agent acts on the environment through actions and receives the immediate rewards (r) corresponding to the actions from the environment. The agent decides the action according to the state, and the mapping relationship between the state and the action is called a policy denoted by π.

$$\pi: S \rightarrow A \tag{3}$$

The goal of the agent is to find an optimal policy that allows the agent to obtain as many returns (R) from the environment as possible. Rewards and returns have the following relationship as shown in **Eq. 4**.

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \tag{4}$$

Where, R_t is the return at time step t. r_{t+1} is the immediate reward at time step $t+1$.γ is the discount factor.

Value-Based Intelligent Controller

To evaluate a policy, a value function needs to be defined. Such an evaluation is used to reflect how well the strategy is controlled. There are two kinds of value functions which are state value function and state-action value function. The state value function represents the expected return starting from state and then following policy, as shown in **Eq. 5**.

$$V^{\pi}(s) = E_{\pi}[R_t|\mathrm{s}_t = s] = E_{\pi}\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|\mathrm{s}_t = s\right] \tag{5}$$

The state-action value function represents the expected return starting from state, taking action and then following policy, as shown in **Eq. 6**.

$$Q^{\pi}(s,\mathrm{a}) = E_{\pi}[R_t|\mathrm{s}_t = s, a_t = a] = E_{\pi}\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|\mathrm{s}_t = s, a_t = a\right] \tag{6}$$

The state value function and the state-action value function have the following relationship, as shown in **Eq. 7**.

$$\begin{aligned} Q^{\pi}(s,a) &= E\left[r_{t+1} + \gamma V^{\pi}(\mathrm{s}_{t+1})\right] \\ V^{\pi}(s) &= E_{a\sim\pi(a|s)}[Q^{\pi}(s,a)] \end{aligned} \tag{7}$$

Solving the optimal policy is equivalent to solving the optimal value function. The method for solving the optimal value function is shown in **Eq. 8** and **Eq. 9**.

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r + Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)] \tag{8}$$

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[r + \max_a Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right] \tag{9}$$

The method of solving the optimal Q value using **Eq. 8** is called SARSA (Chen et al., 2007), and the method of solving the optimal Q value using **Eq. 9** is called Q-learning (Gomes and Kowalczyk, 2009). After the optimal Q value is obtained, this has two common policies for selecting actions based on the Q value, the greedy policy and the ε-greedy policy (Wang et al., 2019). The greedy policy is a deterministic strategy that always picks the largest value function. It is the use of known knowledge by the agent. The ε-greedy policy will probabilistically choose the non-maximal value function to represent the exploration of the unknown environment by the agent. As mentioned above, the reinforcement learning method that selects actions according to the value function is called value-based method.

The value-based reinforcement learning method requires that the state-action space is discrete and not too large. Usually, the value function is represented in the form of a table, so it is also called tabular reinforcement learning (Sutton and Barto, 1998). If the state-action space is too large, the table is difficult to converge. There are two main reasons. On the one hand, too large state-action space leads to too many elements in the table, so it is difficult to visit each element enough times to ensure convergence. On the other hand, from a practical point of view, it is very time-consuming to find an element in a very large table. The method of value function approximation can cope with such a shortcoming (Korda et al., 2016; Wang et al., 2019). The state-action value function is fitted using a function containing the parameter θ. The state-action value function is made to approximate the optimal Q value function by updating the parameter θ, as shown in **Eq. 10**.

$$Q^{'}(s,a,\theta) \approx Q^{\pi}(s,a) \tag{10}$$

Using a deep neural network as an agent to solve a reinforcement learning problem is deep reinforcement learning. DQN is representative of this approach, which uses two layers of convolutional neural networks and two layers of fully connected layer neural networks (Mnih et al., 2015). DQN generates 18 discrete actions based on the input high-dimensional data. DQN can calculate Q value more accurately because of the good fitting ability of deep neural networks to nonlinear functions. Meanwhile, because neural network has good generalization ability, for unexplored states, neural network can also give reasonable q values according to similar states. The process of fitting the state-action value function by DQN is supervised learning. Label data is indispensable for supervised learning, and the method of making label data is shown in **Eq. 11**.

$$y = r + \max_a Q(s,a,\theta) \tag{11}$$

Neural network can fit the optimal value function as long as the loss function is minimized. Although function approximation methods using neural networks greatly alleviate the limitations of value-based reinforcement learning methods in high-dimensional state-action space, they are still difficult to solve for continuous action space problems. In this research, the amount of ammonia injection controlled by the intelligent controller is a continuous variable and the value-based method is not suitable for continuous variables. However, the critic in the actor-critic method is usually composed of value-based methods, so the value-based intelligent controller is introduced.

Policy-Based Intelligent Controller

The policy-based reinforcement learning methods can better solve the problem of continuous action space (Lillicrap et al., 2016). Policy-based reinforcement learning and value-based reinforcement learning have the same input, but the output is the probability distribution of actions being selected in the action space. The policy can be represented by the following **Eq. 12**, where, θ is the parameter to be trained (Sutton et al., 2000).

$$\pi_{\theta}(s,a) = P(a|s,\theta) \approx \pi(s,a) \tag{12}$$

The method of parameter updating is given by the policy gradient theorem (Peters and Schaal, 2008), as shown in **Eq. 13**.

$$\begin{aligned} \nabla R_{\theta} &= \sum_{\tau} R(\tau)\pi_{\theta}(\tau)\nabla\log\pi_{\theta}(\tau) \\ &= \mathrm{E}_{\tau\sim\pi_{\theta}(\tau)}\left[R(\tau)\nabla\log\pi_{\theta}(\tau)\right] \end{aligned} \tag{13}$$

Where, τ is the trajectory of states and actions acquired by the agent as it explores the environment,$\tau = \{\mathrm{s}_1, \mathrm{a}_1, \mathrm{s}_2, \mathrm{a}_2 \ldots \mathrm{s}_t, \mathrm{a}_t\}$.

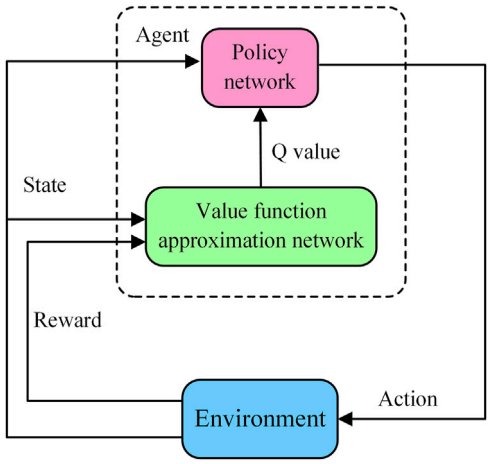

FIGURE 3 | Actor-critic framework schematic diagram.

TABLE 1 | The results of MIC analysis.

Variables	MIC
Reactor inlet NOx concentration	0.579951
Reactor inlet oxygen concentration	0.543274
Ammonia injection volume	0.469567
Temperature near ammonia sprayer	0.469331
Temperature near reactor entrance	0.532343
Temperature near boiler outlet	0.556201
Total air volume	0.612232
Active Power of Generator	0.595234
Oxygen Content in Boiler Flue Gas	0.584735
Total fuel consumption	0.549589

Actor-Critic Intelligent Controller

The policy-based approach also has its own drawbacks, such as the tendency to converge to a local optimum rather than a global optimum. The actor-critic approach, a combination of value-based and policy-based reinforcement learning methods integrates the advantages of both. The Actor-Critic method is obtained by replacing the return R in **Eq. 13** with the Q value. The new policy gradient is shown in **Eq. 14**.

$$\nabla R_\theta = \sum_{n=1}^{N} \sum_{t=1}^{T} Q^{\pi_\theta}\left(a_t^n, s_t^n\right) \nabla \log \pi_\theta\left(a_t^n | s_t^n\right) \tag{14}$$

Since Q value is an expectation, using Q value instead of the return R reduces the variance of the experience gained by the agent when exploring the environment, avoiding falling into a local optimum. When the Q value is larger, the gradient of the trainable parameter update is larger, which accelerates the convergence speed of the policy to the optimal direction. In the actor-critic method, the policy network that generates the actions is called the actor and the value function approximation network used to generate the Q value is called the critic. The framework of the actor-critic approach is shown in **Figure 3**.

This section focuses on the theoretical approach to constructing reinforcement learning intelligent controllers. Other details such as the structure of the intelligent controller will be elaborated in the next section.

EXPERIMENTS AND RESULTS

This section will elaborate on the experimental details of this study. The experiment consists of two main parts: building a virtual environment and training a reinforcement learning intelligent controller. As mentioned earlier, a high precision virtual environment is necessary and critical in order to train reinforcement learning intelligent controllers. In order to obtain a more accurate virtual environment, some effective measures are employed. These measures include data correlation analysis, hyperparameter optimization and unique step-by-step training. In the experiments to train the intelligent controllers, they were designed with different structures and a soft update approach in order to avoid coupling between the actor network and the critic network. Further, important details of the activation function and the reward function are elaborated.

Training Denitrification Environment

The learning process of reinforcement learning agent completely depends on the exploration of virtual environment, but there is inevitably a slight error between virtual environment and real environment. The error between the virtual environment and the real environment can cause a deterioration in the control of the agent, this phenomenon known as reality gap (Zagal et al., 2004; Collins et al., 2019; Hwangbo et al., 2019). It is important to build a high accuracy virtual environment for training intelligent controllers. In order to have a high accuracy of denitrification environment and reduce the reality gap, various measures including variable screening and validation of neural network models with various structures are taken.

Data-driven modeling approaches are usually data-sensitive, so it is necessary to filter the data to avoid irrelevant variables that reduce model accuracy. In this research, the maximum information coefficient (MIC), a statistical analysis tool, is used to analyze the correlation between variables and to select reasonable input variables to reduce the error caused by irrelevant variables (Reshef et al., 2011). The data used for modeling comes from the real historical data in the distributed control system of power plant, and is analyzed by MIC numerical value. Some variables with strong correlation were selected and the results of their MIC analysis are shown in **Table 1**.

Another factor that affects the accuracy of the neural network model is the structure and hyperparameters of the neural network. Although the properties of different neural networks are helpful to design models for the denitrification environment, designing the structure of the neural network and determining the hyperparameters need to be validated several times depending on the task. At first, in order to save computing resources and use more computing resources to train agents, a simple multilayer LSTM neural network was adopted. Although such a model consumes less computational resources, the accuracy is not satisfactory. Then, the end-to-end structure is used to replace

TABLE 2 | The raining errors for different models.

Network Structure	5 time steps (%)	10 time steps (%)	20 time steps (%)	
Multi-layer LSTM network	2 layers	7.9420	7.2094	7.0238
	4 layers	6.8226	6.5965	6.3765
	6 layers	6.2329	6.0883	6.1947
Single decoder with attention mechanism	2 layers	33384	3.0631	3.8497
	4 layers	2.9482	2.8395	2.6543
	6 layers	2.7389	2.3173	2.2879
Dual decoders with attention mechanism	2 layers	2.8274	2.4098	2.1062
	4 layers	1.9489	1.8159	1.7975
	6 layers	1.8037	1.7963	1.7498

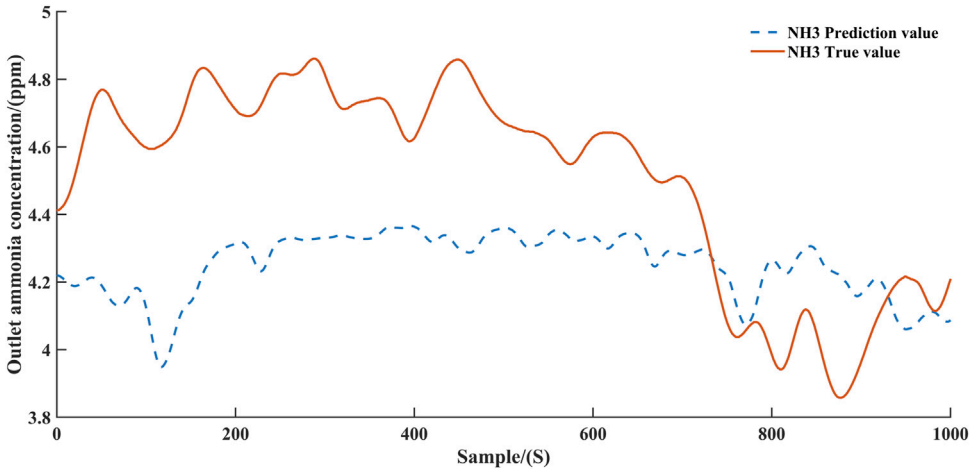

FIGURE 4 | Ammonia concentration prediction curve.

the simple multilayer neural network structure, and attention mechanism is introduced to improve the accuracy of the denitrification environment. The end-to-end structure with an attention mechanism improves the accuracy of the denitrification environment, but considering the phenomenon of reality gap, higher accuracy is still needed. Further, the new structure uses two decoders with attention mechanisms to decode NOx and ammonia separately, which avoids the coupling of ammonia and NOx and improves the model accuracy. Finally, the output of the decoder is improved by multi-layer fully connected layer neural network to improve the nonlinear fitting ability. Suitable hyperparameters can improve the accuracy of neural networks. The common hyperparameters include the number of layers and time steps of neural networks. There is no clear theoretical method to determine the most suitable hyperparameters, which needs to be set according to researchers' experience and confirmed by multiple verifications. In this research, multiple validations were implemented to determine the structure and hyperparameters of a high precision denitrification environmental model. The training errors for different structural and hyperparametric models are shown in the **table 2**.

According to the verification of denitrification environment model structure and hyperparameters, considering the accuracy of each model and the consumption of computing resources, the dual decoder structure with attention mechanism was selected. The time step is set to 20 and the number of neural network layers is set to 4. In order to make the virtual environment more realistic, a large amount of data was collected from the historical database of the power plant for training the virtual environment. These data include data under different load and operating conditions in order to give a more comprehensive description of the overall situation of the power plant.

To further improve the accuracy of the denitrification environment, a special approach is also taken in the training process of the model. In particular, the training of dual decoder model needs to be carried out step by step. Firstly, the whole model is trained to a relatively low error level. Secondly, the parameters of the encoder and the parameters of NOx concentration decoders are frozen and only the parameters of ammonia concentration decoder are trained. Finally, only the NOx concentration decoder parameters that were frozen in the second step are trained. The results of the denitrification environment model are shown in **Figure 4**,**Figure 5**, **Figure 6** and **Figure 7**. In training the virtual environment, the first step is the most time-consuming and usually lasts for several days, while

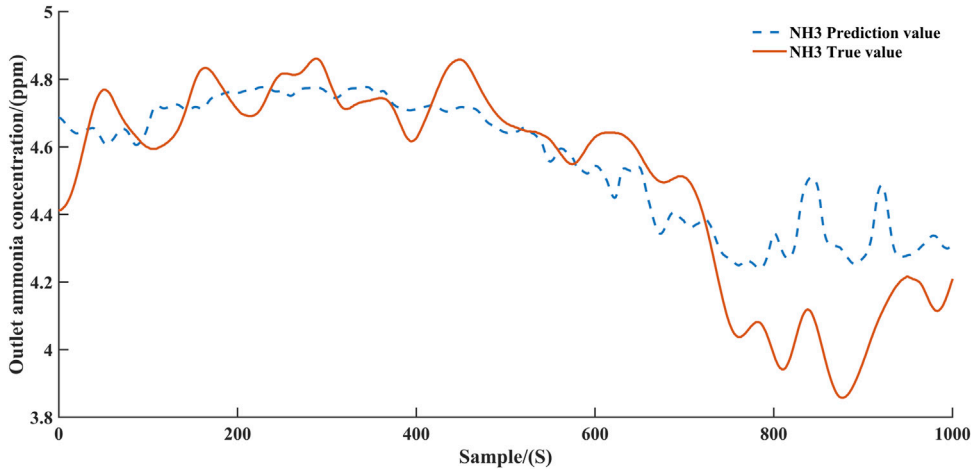

FIGURE 5 | Ammonia concentration prediction curve after parameter freezing training.

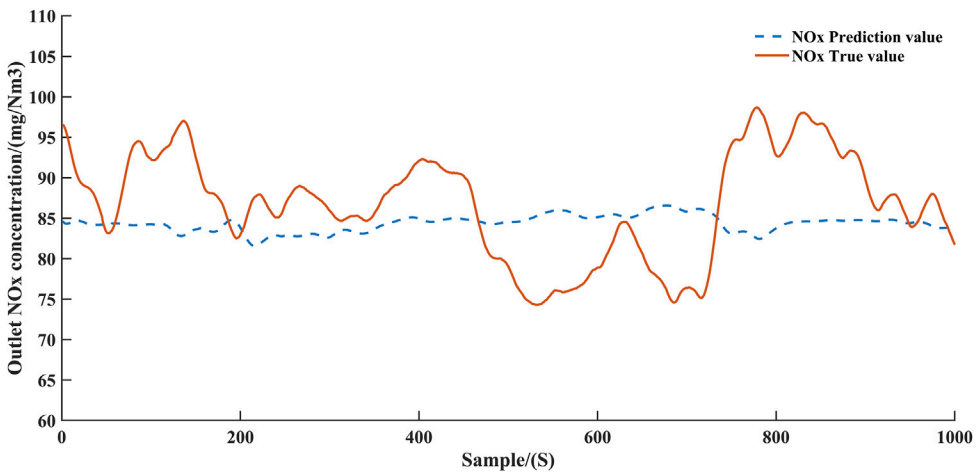

FIGURE 6 | NOx concentration prediction curve.

training the two decoders separately takes relatively less time. The reasons for this phenomenon will be discussed later.

Figure 4 and **Figure 6** demonstrate the predicted values of ammonia concentration and NOx concentration, respectively, after the first step of training. **Figure 5** and **Figure 7** demonstrate the predicted values of ammonia concentration and NOx concentration after parameter freezing training respectively. According to **Figure 4** and **Figure 5 Figure 6** and **Figure 7**, it can be concluded that the accuracy of the model can be improved by training the model step by step. This can be explained as follows. Since the two decoders share a common encoder, the parameters of the encoder are optimized by the gradients returned by both decoders in the first training step. In this case, the information extracted by the encoder will bring more errors to the two decoders. When the encoder and one decoder are frozen, optimizing the other decoder can prevent the error from propagating to the final predicted value. Obviously, separate models could be constructed for NOx and ammonia to improve accuracy, but this would increase the consumption of computational resources. The structure of single encoder and double decoder can balance the consumption of computing resources and the accuracy of the model, and devote more computational resources to training the agent.

Training Reinforcement Learning Intelligent Controller

In this research, the reinforcement learning agent is the controller that controls the spraying of ammonia into the SCR reactor. Since the amount of ammonia sprayed is a continuous action space, the deep deterministic policy gradient (DDPG) method is used as the reinforcement learning agent instead of other value-based

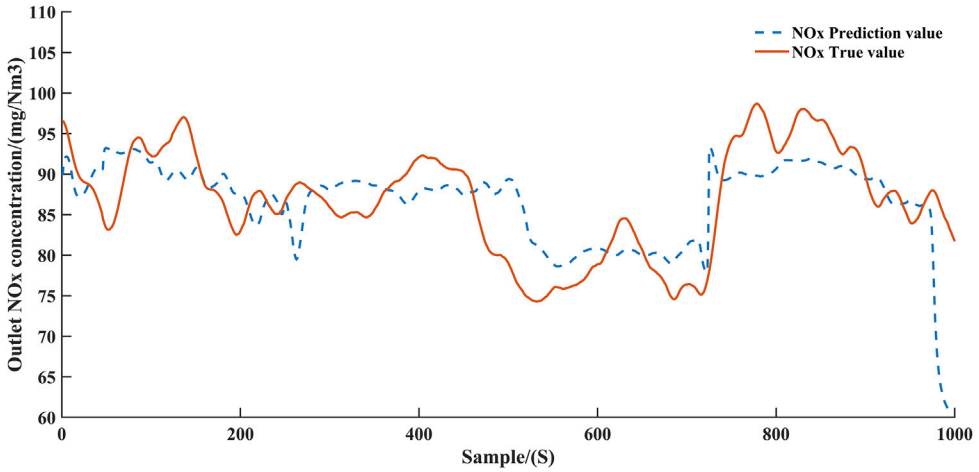

FIGURE 7 | NOx concentration prediction curve after parameter freezing training.

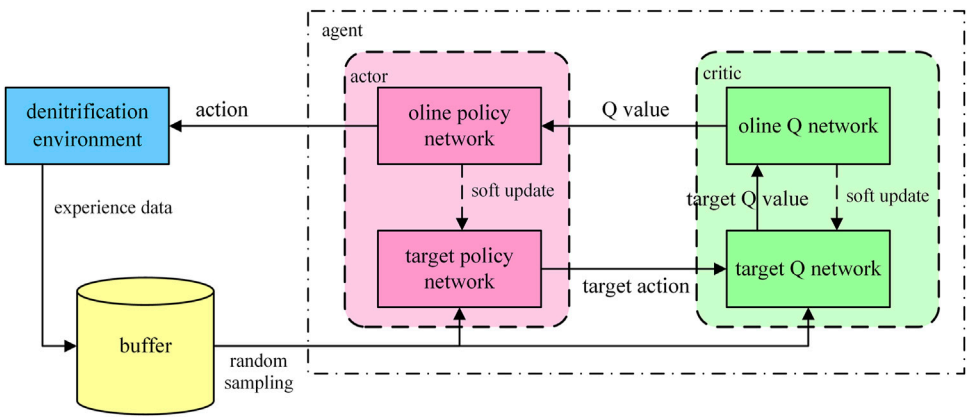

FIGURE 8 | DDPG algorithm flow chart.

reinforcement learning methods. The DDPG algorithm is based on the actor-critic framework, which contains a policy network as the actor and a Q network as the critic. The policy network generates action outputs and the Q network optimizes the parameters of the policy network by evaluating the actions generated by the policy network through Q values.

In order to improve the stability of the agent training, soft updating and buffering methods are adopted. DDPG creates two copies of the policy network and the Q network, called the target policy network and the target Q network, respectively. The target network parameters are updated using the soft update method as shown in **Eq. 15**.

$$\begin{aligned} \theta^{Q'} &\leftarrow \alpha\theta^{Q} + (1-\alpha)\theta^{Q'} \\ \theta^{\mu'} &\leftarrow \alpha\theta^{\mu} + (1-\alpha)\theta^{\mu'} \end{aligned} \tag{15}$$

Where, θ is the network parameter. Q' is the target Q network. μ' is the target policy network. $\alpha$is update step. Different from other policy gradient methods that use random policies, deterministic policies will only produce one action in one state, which is more suitable for industrial control requirement. Since a deterministic policy is used, the gradient of the policy network is shown in **Eq. 16** (Silver et al., 2014).

$$\nabla R_{\theta^{\mu}} = \frac{1}{N}\sum_{n=1}^{N} \nabla_a Q\left(a, s|\theta^{Q}\right)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^{\mu}} \mu\left(s|\theta^{\mu}\right)|_{s=s_i} \tag{16}$$

The experience produced by agents in exploring the environment has sequence correlation. In order to avoid the agent falling into local optimum caused by sequence correlation, the delay buffering method is adopted in the research. The experience data obtained by the agent exploring the environment are not trained directly by the agent, but are stored in a buffer. The data of training agent is generated by random sampling in buffer. The structure of DDPG is shown in the **Figure 8**.

As shown in **Figure 8**, the agent and the environment contain five neural networks, which require a lot of computational

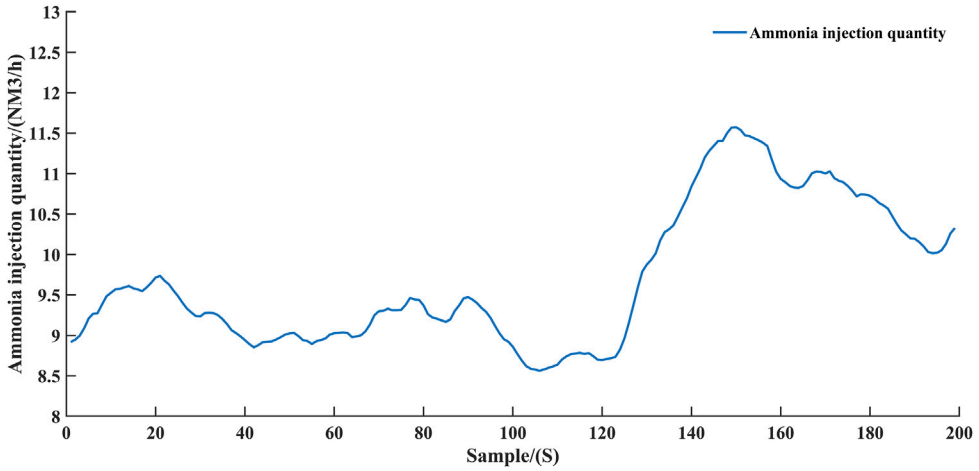

FIGURE 9 | Ammonia injection amount given by agent controller.

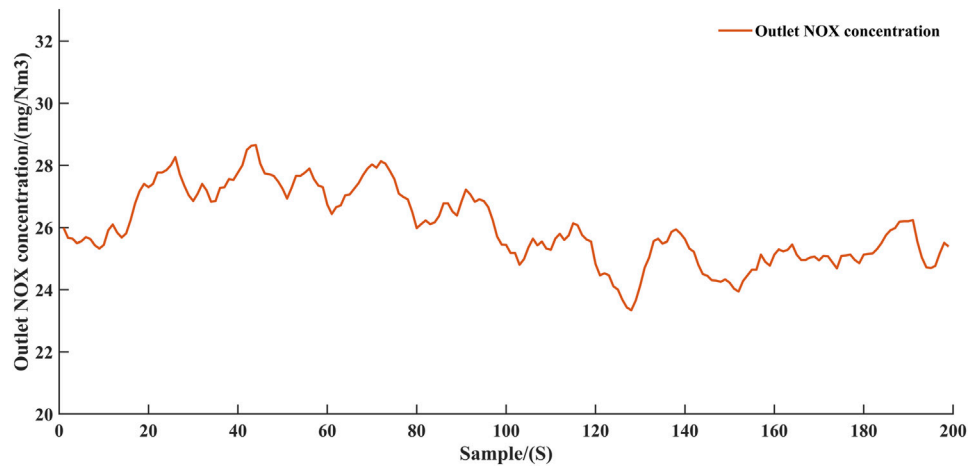

FIGURE 10 | NOx concentration at the outlet of reactor controlled by agent controller.

resources to support. This is the reason that the structure of the denitrification environment model is designed as two decoders sharing one encoder. Compared with separately designing models for NOx concentration and ammonia concentration at the outlet of denitrification reactor, this structure can save more computing resources. The computational resources need to be conserved for training the agent during the training agent phase.

After completing the overall architecture design of the reinforcement learning agent, there are still many details to be refined in the design of the actor network and the critic network inside the agent. The first concern is the network structure of the policy network and the value network. LSTM neural network is used as the policy network because the data processed are time series data. In particular, the policy network is designed as a three-layer LSTM neural network. To avoid the coupling caused by the same as the strategy network, the value network is designed as a three-layer one-dimensional convolutional neural network. Convolution is a classical digital signal processing method. Usually, two-dimensional convolution neural network is used to process image data, and one-dimensional convolution is used to process time series data (Abdeljaber et al., 2017; Antoshchuk et al., 2020).

Another detail worth noting is the activation function of the neural network. Depending on the range of the activation function, the activation function can be divided into saturated and unsaturated activation functions (He et al., 2015; Krizhevsky et al., 2017). The action of the actor network output is the flow of ammonia injected into the denitrification reactor, which reaches a maximum value when the valve is fully open and reaches zero when the valve is fully closed. Such an action range is more suitable using the sigmoid activation function, which can avoid the actor network to produce some unreasonable actions, such as negative or too large flow values. The sigmoid function is a common saturation activation function, whose upper limit is one

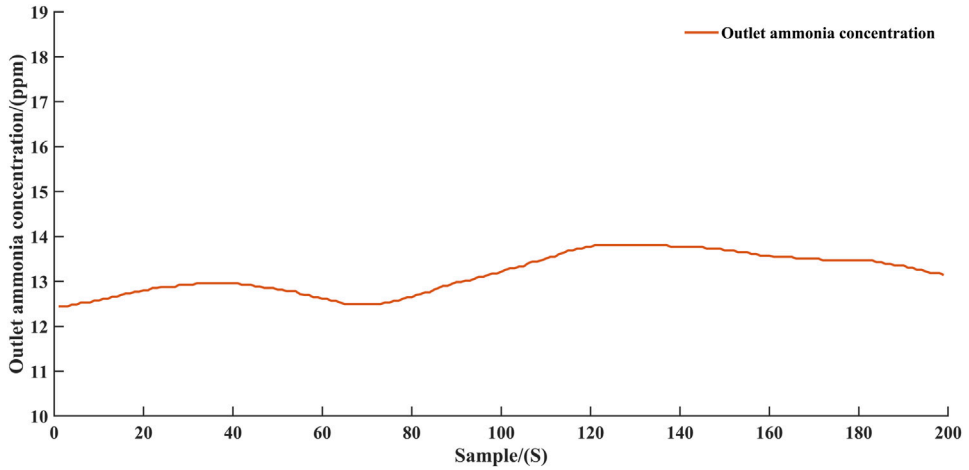

FIGURE 11 | Ammonia concentration at the outlet of reactor controlled by agent controller.

and lower limit is zero. However, the sigmoid activation function also has the drawback of causing the vanishing gradient. Therefore, the sigmoid function is only used as the activation function in the output layer of the actor network. The critic network outputs Q values and does not have upper and lower limits, thus using a non-saturating activation function.

In reinforcement learning, the goal of an agent is formally characterized as a special signal, called reward, which is usually a function of state. At each time step, the reward is a single scalar value. The role of reward is to guide the agent's policy toward the desired outcome. The goal of the controller in this research was to reduce both the NOx and ammonia concentrations at the outlet of the denitrification reactor. Therefore, the reward function is designed as in the **Eq. 17**.

$$r = -\alpha C_{NOx} - (1-\alpha)C_{NH3} \tag{17}$$

Where α is a number between 0 and 1 to tune the style of the reinforcement learning agent's policy. With the increase ofα, the agent tends to inject more ammonia to reduce the nitrogen oxide content at the outlet of the reactor. After interactive training with virtual environment, the agent has learned to control ammonia injection to reduce the concentration of nitrogen oxides at the outlet of denitrification reactor. The ammonia injection quality of the agent controller and the concentration of nitrogen oxides and ammonia at the outlet of the reactor are shown in the **Figure 9**, **Figure 10** and **Figure 11**, respectively. In this experiment, the parameterαin **Eq. 17** is set to 0.5, which means that the agent regards the control of nitrogen oxide concentration and ammonia concentration as equally important. The results show a noticeable increase in ammonia injection after 120 steps, which corresponds to a slight overall decrease in NOx concentration at the outlet of the SCR system. The ammonia escape from the system also tends to increase slightly after 120 steps. Such results are consistent with the empirical common sense that increasing ammonia injection would contribute to the reduction of NOx emissions but would increase the extent of ammonia escape. On the other hand, such results indicate that the virtual environment accurately reflects the dynamic characteristics of the system, and that the model structure of the virtual environment and the training methods to improve accuracy are successful and effective.

DISCUSSION AND CONCLUSION

The experiments of this study can be divided into two parts: training the virtual environment and training the intelligent controller. The method of freezing some parameters and setting up a double decoder during the training of the virtual environment significantly improves the accuracy of the virtual environment. In training the virtual environment, the first step is the most time-consuming and usually lasts for several days, while training the two decoders separately takes relatively less time. This is mainly due to the large number of model parameters that need to be optimized in the first training step, which contain the parameters of one encoder and two decoders, respectively. While in the stage of partial parameter freezing, the parameters of one encoder and one decoder are frozen and only the parameters of one decoder need to be optimized.

According to the experimental results, the intelligent controller was able to control the SCR system ammonia injection to reduce the nitrogen oxidation emissions while avoiding excessive ammonia escape. The experimental results validate the feasibility of reinforcement learning in the field of process control. In particular, as shown in **Figure 9**, the intelligent controller increases the ammonia injection after 120 time-steps. In response to this change, the NOx concentration at the outlet decreased slightly from around 28 mg/Nm3 to around 26 mg/Nm3 as shown in **Figure 10**. Such a change also verifies that the intelligent controller is interacting correctly with the virtual environment.

The main contribution of this research consists of two aspects, which are the virtual environment and the intelligent

controller. As mentioned before, the accuracy of the virtual environment has a critical impact on the control effectiveness. The selective catalytic reduction system as the object of research has characteristics such as large latency and multiple inputs and outputs. The model structure designed in this research is well adapted to these characteristics, especially the parameter freezing and step-by-step training methods improve the accuracy of the virtual environment. Since other systems in thermal power plants have similar characteristics to SCR systems, the model structure and training methods in this study can be extended to other systems in thermal power plants. The methods used to construct and train the virtual environment in this research can support more in-depth studies. Another contribution of this research is its validation of the feasibility and effectiveness of using deep reinforcement learning intelligent controllers to control thermal power plant systems. The potential of artificial intelligence techniques in power systems has been noticed by many scholars, but little research has been reported in this area. This research takes a hot artificial intelligence technique, deep reinforcement learning, as an intelligent controller in the field of pollutant emission control, which is currently of wide interest. The experimental results demonstrate that the intelligent controller is able to keep both NOx emissions and ammonia escaping at low levels. On the one hand, such results validate the effectiveness of the reinforcement learning intelligent controller for selective catalytic reduction systems, and on the other hand, this study reveals the feasibility of applying deep reinforcement learning techniques to other systems in power plants.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

PX: Conceptualization, Methodology, Software Data curation, Writing- Original draft. YN: Writing- Reviewing and Editing. GZ: Writing- Reviewing and Editing. TS: Writing- Reviewing and Editing.

FUNDING

This project was funded by the project supported by the National Key R&D Program of China (2016YFB0600205).

REFERENCES

Abdeljaber, O., Avci, O., Kiranyaz, S., Gabbouj, M., and Inman, D. J. (2017). Real-time Vibration-Based Structural Damage Detection Using One-Dimensional Convolutional Neural Networks. *J. Sound Vibration* 388, 154–170. doi:10.1016/j.jsv.2016.10.043

Adamczyk, W. P., Werle, S., and Ryfa, A. (2014). Application of the Computational Method for Predicting NO X Reduction within Large Scale Coal-Fired Boiler. *Appl. Therm. Eng.* 73, 343–350. doi:10.1016/j.applthermaleng.2014.07.045

Antoshchuk, S., Babilunha, O., Kim, T. T., Nikolenko, A., and Thi Khanh, T. N. (2020). Non-Stationary Time Series Prediction Using One-Dimensional Convolutional Neural Network Models. *Herald Adv. Inf. Techn.* 3, 362–372. doi:10.15276/hait01.2020.3

Bahdanau, D., Cho, K. H., and Bengio, Y. (2015). "Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473.

Belošević, S., Tomanović, I., Crnomarković, N., Milićević, A., and Tucaković, D. (2016). Numerical Study of Pulverized Coal-Fired Utility Boiler over a Wide Range of Operating Conditions for In-Furnace SO2/NOx Reduction. *Appl. Therm. Eng.* 94, 657–669. doi:10.1016/j.applthermaleng.2015.10.162

Chen, C., Seff, A., Kornhauser, A., and Xiao, J. (2015). "DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving," in Proceedings of the IEEE International Conference on Computer Vision, 2722–2730. doi:10.1109/ICCV.2015.312

Chen, Y., Mabu, S., Hirasawa, K., and Hu, J. (2007). "Genetic Network Programming with Sarsa Learning and its Application to Creating Stock Trading Rules," in 2007 IEEE Congress on Evolutionary Computation, 220–227. CEC 2007. doi:10.1109/CEC.2007.4424475

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation," in EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. doi:10.3115/v1/d14-1179

Collins, J., Howard, D., and Leitner, J. (2019). "Quantifying the Reality gap in Robotic Manipulation Tasks," in Proceedings - IEEE International Conference on Robotics and Automation (ICRA), 6706–6712. doi:10.1109/ICRA.2019.8793591

Csáji, B. C., and Monostori, L. (2008). Value Function Based Reinforcement Learning in Changing Markovian Environments. *J. Machine Learn. Res.* 9 (8).

Díez, L. I., Cortés, C., and Pallarés, J. (2008). Numerical Investigation of NOx Emissions from a Tangentially-Fired Utility Boiler under Conventional and Overfire Air Operation. *Fuel* 87, 1259–1269. doi:10.1016/j.fuel.2007.07.025

Du, X., Yang, G., Chen, Y., Ran, J., and Zhang, L. (2017). The Different Poisoning Behaviors of Various Alkali Metal Containing Compounds on SCR Catalyst. *Appl. Surf. Sci.* 392, 162–168. doi:10.1016/j.apsusc.2016.09.036

Gang, L. I., Jia, X., Wu, B., Niu, G., and Xue, D. (2016). Spraying Ammonia Flow Control System of SCR Denitration System Based on Leading Factor Analysis. *Therm. Power Generationeneration* 45, 99–102.

Gomes, E. R., and Kowalczyk, R. (2009). "Modelling the Dynamics of Multiagent Q-Learning with ε-greedy Exploration," in Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, 2, 1181–1182. (AAMAS).

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification," in Proceedings of the IEEE International Conference on Computer Vision, 1026–1034. doi:10.1109/ICCV.2015.123

Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V., et al. (2019). Learning Agile and Dynamic Motor Skills for Legged Robots. *Sci. Robot.* 4, eaau5872. doi:10.1126/scirobotics.aau5872

Hwangbo, J., Sa, I., Siegwart, R., and Hutter, M. (2017). Control of a Quadrotor with Reinforcement Learning. *IEEE Robot. Autom. Lett.* 2, 2096–2103. doi:10.1109/LRA.2017.2720851

Jaritz, M., De Charette, R., Toromanoff, M., Perot, E., and Nashashibi, F. (2018). "End-to-End Race Driving with Deep Reinforcement Learning," in Proceedings - IEEE International Conference on Robotics and Automation, 2070–2075. doi:10.1109/ICRA.2018.8460934

Khargonekar, P. P., and Dahleh, M. A. (2018). Advancing Systems and Control Research in the Era of ML and AI. *Annu. Rev. Control.* 45, 1–4. doi:10.1016/j.arcontrol.2018.04.001

Korda, M., Henrion, D., and Jones, C. N. (2016). Controller Design and Value Function Approximation for Nonlinear Dynamical Systems. *Automatica* 67, 54–66. doi:10.1016/j.automatica.2016.01.022

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* 60, 84–90. doi:10.1145/3065386

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., et al. (2016). "Continuous Control with Deep Reinforcement Learning," in 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings, arXiv:1509.02971.

Liu, C.-L., Chang, C.-C., and Tseng, C.-J. (2020). Actor-critic Deep Reinforcement Learning for Solving Job Shop Scheduling Problems. *IEEE Access* 8, 71752–71762. doi:10.1109/ACCESS.2020.2987820

Luong, T., Pham, H., and Manning, C. D. (2015). "Effective Approaches to Attention-Based Neural Machine Translation," in Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, arXiv:1508.04025. doi:10.18653/v1/d15-1166

Lv, Y., Romero, C. E., Yang, T., Fang, F., and Liu, J. (2018). Typical Condition Library Construction for the Development of Data-Driven Models in Power Plants. *Appl. Therm. Eng.* 143, 160–171. doi:10.1016/j.applthermaleng.2018.07.083

Mishra, M., Nayak, J., Naik, B., and Abraham, A. (2020). Deep Learning in Electrical Utility Industry: A Comprehensive Review of a Decade of Research. *Eng. Appl. Artif. Intelligence* 96, 104000. doi:10.1016/j.engappai.2020.104000

Mnih, V., Badia, A. P., Mirza, L., Graves, A., Harley, T., Lillicrap, T. P., et al. (2016). "Asynchronous Methods for Deep Reinforcement Learning," in 33rd International Conference on Machine Learning, 1928–1937. (ICML 2016).

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level Control through Deep Reinforcement Learning. *Nature* 518, 529–533. doi:10.1038/nature14236

Mousavi, S. M., Fatehi, H., and Bai, X.-S. (2021). Numerical Study of the Combustion and Application of SNCR for NO Reduction in a Lab-Scale Biomass Boiler. *Fuel* 293, 120154. doi:10.1016/j.fuel.2021.120154

Najafi, G., Ghobadian, B., Moosavian, A., Yusaf, T., Mamat, R., Kettner, M., et al. (2016). SVM and ANFIS for Prediction of Performance and Exhaust Emissions of a SI Engine with Gasoline-Ethanol Blended Fuels. *Appl. Therm. Eng.* 95, 186–203. doi:10.1016/j.applthermaleng.2015.11.009

Peters, J., and Schaal, S. (2008). Reinforcement Learning of Motor Skills with Policy Gradients. *Neural Networks* 21, 682–697. doi:10.1016/j.neunet.2008.02.003

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., et al. (2011). Detecting Novel Associations in Large Data Sets. *Science* 334, 1518–1524. doi:10.1126/science.1205438

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). "Deterministic Policy Gradient Algorithms," in 31st International Conference on Machine Learning, 387–395. (ICML 2014).

Strege, J. R., Zygarlicke, C. J., Folkedahl, B. C., and McCollor, D. P. (2008). SCR Deactivation in a Full-Scale Cofired Utility Boiler. *Fuel* 87, 1341–1347. doi:10.1016/j.fuel.2007.06.017

Stupar, G., Tucaković, D., Živanović, T., and Belošević, S. (2015). Assessing the Impact of Primary Measures for NOx Reduction on the thermal Power Plant Steam Boiler. *Appl. Therm. Eng.* 78, 397–409. doi:10.1016/j.applthermaleng.2014.12.074

Stupar, G., Tucaković, D., Živanović, T., Stevanović, Ž., and Belošević, S. (2019). Predicting Effects of Air Staging Application on Existing Coal-Fired Power Steam Boiler. *Appl. Therm. Eng.* 149, 665–677. doi:10.1016/j.applthermaleng.2018.12.070

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). "Sequence to Sequence Learning with Neural Networks," in Advances in Neural Information Processing Systems, 3104–3112.

Sutton, R. S., and Barto, A. G. (1998). Reinforcement Learning: An Introduction. *IEEE Trans. Neural Netw.* 9, 1054. doi:10.1109/tnn.1998.712192

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (2000). "Policy Gradient Methods for Reinforcement Learning with Function Approximation," in Advances in Neural Information Processing Systems, 1057–1063.

Tan, P., He, B., Zhang, C., Rao, D., Li, S., Fang, Q., et al. (2019). Dynamic Modeling of NOX Emission in a 660 MW Coal-Fired Boiler with Long Short-Term Memory. *Energy* 176, 429–436. doi:10.1016/j.energy.2019.04.020

Tang, N., Zhang, Y., Niu, Y., and Du, X. (2018). Solar Energy Curtailment in China: Status Quo, Reasons and Solutions. *Renew. Sustain. Energ. Rev.* 97, 509–528. doi:10.1016/j.rser.2018.07.021

Ti, S., Chen, Z., Li, Z., Min, K., Zhu, Q., Chen, L., et al. (2017). Effect of Outer Secondary Air Vane Angles on Combustion Characteristics and NO Emissions for Centrally Fuel Rich Swirl Burner in a 600-MWe wall-fired Pulverized-Coal Utility Boiler. *Appl. Therm. Eng.* 125, 951–962. doi:10.1016/j.applthermaleng.2017.05.180

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., et al. (2019). Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning. *Nature* 575, 350–354. doi:10.1038/s41586-019-1724-z

Wang, J., Zheng, K., Singh, R., Lou, H., Hao, J., Wang, B., et al. (2017). Numerical Simulation and Cold Experimental Research of a Low-NOx Combustion Technology for Pulverized Low-Volatile Coal. *Appl. Therm. Eng.* 114, 498–510. doi:10.1016/j.applthermaleng.2016.11.204

Wang, Y., Zibaeenejad, A., Jing, Y., and Chen, J. (2019). "On the Optimality of the Greedy Policy for Battery Limited Energy Harvesting Communication," in IEEE Workshop on Signal Processing Advances in Wireless Communications 2019 (SPAWC). doi:10.1109/SPAWC.2019.8815586

Wei, Z., Li, X., Xu, L., and Cheng, Y. (2013). Comparative Study of Computational Intelligence Approaches for NOx Reduction of Coal-Fired Boiler. *Energy* 55, 683–692. doi:10.1016/j.energy.2013.04.007

Yang, G., Wang, Y., and Li, X. (2020). Prediction of the NO Emissions from thermal Power Plant Using Long-Short Term Memory Neural Network. *Energy* 192, 116597. doi:10.1016/j.energy.2019.116597

Yu, M., and Sun, S. (2020). Policy-based Reinforcement Learning for Time Series Anomaly Detection. *Eng. Appl. Artif. Intelligence* 95, 103919. doi:10.1016/j.engappai.2020.103919

Zagal, J. C., Ruiz-Del-Solar, J., and Vallejos, P. (2004). Back to Reality: Crossing the Reality gap in Evolutionary Robotics. *IFAC Proc. Volumes* 37, 834–839. doi:10.1016/s1474-6670(17)32084-0

Zhou, H., Pei Zhao, J., Gang Zheng, L., Lin Wang, C., and Fa Cen, K. (2012). Modeling NOx Emissions from Coal-Fired Utility Boilers Using Support Vector Regression with Ant colony Optimization. *Eng. Appl. Artif. Intelligence* 25, 147–158. doi:10.1016/j.engappai.2011.08.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.