Medical Equipment Supply Chain Optimization and Stability Study using Deep Reinforcement Learning

Zhuoxun Chen *

Newtown, Pennysulvania, USA

* Corresponding Author Email: chenzhuoxun0324@gmail.com

Abstract. Medical equipment is a critical resource during the COVID-19 pandemic. An efficient and stable supply chain of medical equipment (masks, goggles, protective coveralls, etc.) enables medical workers and first responders to effectively and safely fight against this highly contagious disease. In my research, I design and investigate two agents based on the traditional (s,Q) policy and the Deep Reinforcement Learning (DRL) algorithm and apply them respectively to optimize a two-echelon medical equipment supply chain where one distribution center and multiple retailers are involved. To my knowledge, this is the first implementation of DRL algorithm for medical supply chain optimization. I implement the DRL algorithm in Python using Ray and RLlib packages and conduct experiments using Google Colab with GPU support. To maximize the DRL algorithm's potential, I optimize the reward function and the hyperparameters of this algorithm. By testing the agents in different environment initializations, I find that the DRL algorithm outperforms the static (s, Q) agent, which is one of the most commonly used methods in many inventory optimization systems, by returning a 17.33% greater cumulative reward on average. Additionally, the relative standard deviation of baseline (s,Q) Policy is 1.97% and that of DRL is 2.49% based on ten repeated trials. Thus, the DRL approach is not only stable but can also significantly improve the retailer's profit. My DRL model can be further applied to more complicated multi-echelon supply chain systems and lays a solid foundation for optimizing large-scale medical supply chains [TF18].

Keywords: (s, Q) policy; Deep Reinforcement Learning; medical supply chain.

1. Introduction

The Coronavirus disease (COVID-2019), "an infectious disease caused by a newly discovered coronavirus" affecting more than 213 countries and territories globally, has created global health and economic disruption. Global virus containment responses include closing non-essential businesses, social distancing, and requiring populations to shelter in place [1]. It has also affected the existing supply chain and the way it was being managed (Khan). Due to the pandemic, public's general recognition of protection and sanitation in response to virus transmission has raised significantly. As a result, the demand for medical equipment such as facial masks and ventilator has experienced an unprecedented increase, and hospitals and many other healthcare carriers have been hit particularly hard. As a result, the pandemic left the health care system in crisis: "hospitals on the verge of collapse with their capacity overflowed, critical item supply chains interrupted, and federal and state agencies struggling to take palliative and preventative measures." [2]

The coronavirus is still being efficiently treated and eradicated, and supply chains are still working to enable the delivery of patient care in the event of a recurrence or future pandemic. Despite ongoing efforts to successfully treat and eradicate the coronavirus, the supply chain and distribution system for various medical supplies has faced several difficulties.

Due to the pandemic, the production capacity continued to fail to meet the demand, resulted in shortage in the supply chain system. As a result, the number of daily infections in major cities such as New York and Philadelphia were increasing significantly, further dramatically increased the number of patients in the hospital, resulting in a shortage of ventilators and other medical equipment.

As Odenkirk as pointed out, accurate forecasting data is essential of solving potential shortages, in tandem with rightsizing and moving inventory based on projections of product demand. Going forward, forecasting will need to be more accurate considering the constant process of projecting demand and tweaking forecast scenarios while factoring in the cost of inventory, time and

transportation. The data is even more significant as it determines the industry to also drill down on cost based on geography.

Supply chain management that is efficient adds value and positions businesses strategically. In recent years, there has been an increase in interest in supply chain management's scientific components. The deep learning method, the long- short-term-memory network, is among one of the most popular methods. Research has shown that the deep learning method has consistently "showed promising results on timeseries forecasting problems". The deep learning method has been used as a foundation to estimate future multi-echelon supply chain behavior because it is capable of addressing all three aspects of supply chain forecasting (items, time, and channel) in. It handles the linear and nonlinear time-based variations in complex time-series with higher efficiency and accuracy. Beyond generality, the deep learning algorithm has been implemented to forecast demand onto real life. They have obtained data from SOK Market in Turkey with 6700 stores, 1500 products, and 23 distribution centers and discovered that the deep learning model provided the most accurate prediction among all tested model. Additional research has also shown that the deep learning approach offers the best efficiency for stock optimization, cost reduction, and growth in sales, profit, and customer loyalty. The framework provides the short-term up to long-term demand forecasts for strategic, tactical, and operational planning levels in a supply chain.

The synchronization of inventory policies implemented by various roles, including suppliers, manufacturers, and distributors, is one of the key components of inventory management. The reinforcement learning algorithm is developed to manage the inventory decisions at all stages in an integrated fashion. A near-optimal policy under an average reward algorithm is found using the RL algorithm, a simulation-based stochastic approach. Complex Markov Decision Processes that represent real-world systems have been successfully solved using the artificial intelligence approach of reinforcement learning. It is effective at teaching agents the best control strategy through reward iteration and simulation. While implementing RL agent onto a multi supply chain environment which makes the acceptance decision and a deterministic scheduling component, they have found that the RL agent outperforms a simple learning heuristic in all training states. While comparing the RL agent with the traditional (s, Q) model and SARSA algorithm onto a Markov Decision Process that determines how many products should be produced in a factory and how many products should be shipped to different warehouses, the result has indicated that the RL agent results in the most optimal strategy regardless of the complexity of situation.

Existing work has implemented both deep learning and reinforcement learning methods onto the supply chain system, and combining with real-world cases has shown significant improvement on both policies making and distribution algorithm. While when combining the two methods together, there are much less experimented cases. The deep reinforcement learning methods combines both reinforcements learning algorithm, allowing computational agent to learn to make decisions both on trials and errors, and the deep learning into its optimal solution, which further enables the agent to make decisions without any manual engineering.

Besides, our work is also the first deeply analyzing the supply chain correlating to the medical field and constructing a near optimal distributing system using the reinforcement learning. Under the frame of unprecedented pandemics such as the Covid-19, constructing an algorithm for better distributing limited facial mask supply becomes extremely critical. It not only maximizes suppliers' profit of producing facial masks but more importantly dramatically decrease the frequency of virus transmission. Previous research has implemented deep learning algorithm.

As the result of our experiment, the deep reinforcement learning agent not only accurately predicts future covid cases number based on current data but also shows the highest efficiency when distributing facial masks and outperforms any other agents when maximizing the retailer's profit.

The organization of the paper is as following:

In II, we introduce data statistics.

In III, we discuss the research method.

In IV, we analyze the experimental result.

In V, we conclude the result.
a) Literacy of covid 19 affection
b) Elaborate on traditional methods
c) Citation

2. Datasets / Data Sources

The 5 CVS retail stores located in the state of Pennysulvania are shown on the map below and the store information is shown Table I:

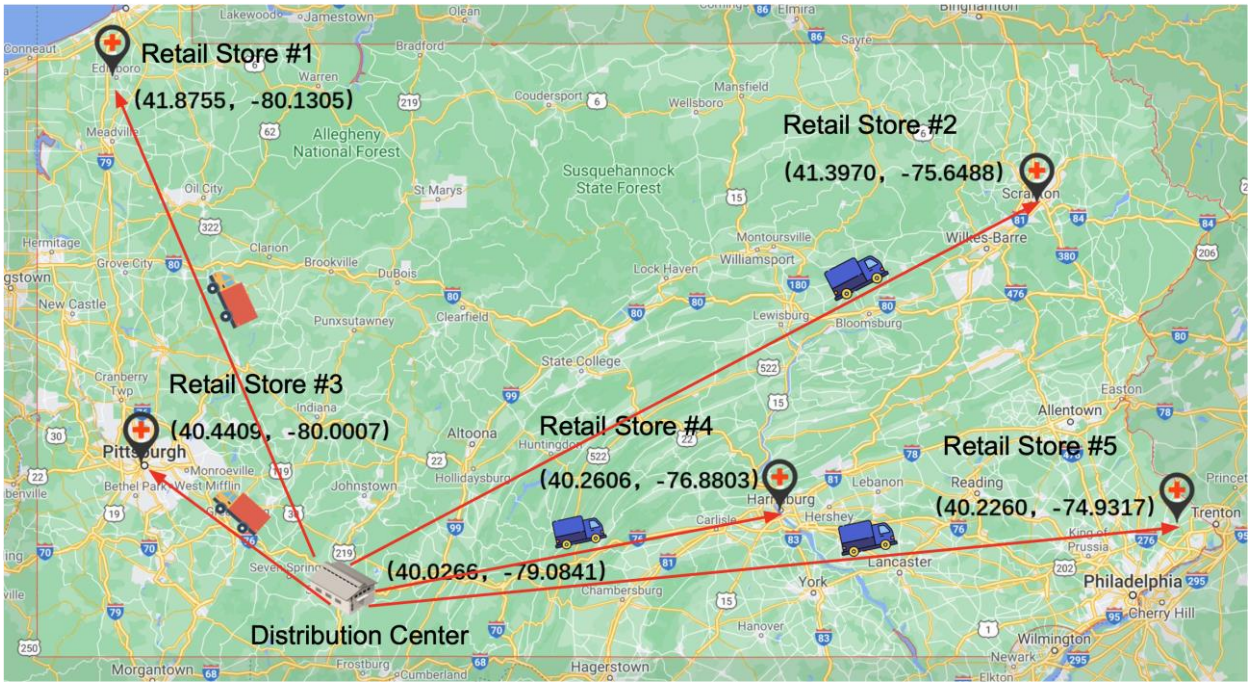


Figure 1. The map of 5 CVS retail stores.

The store list is listed as following:

Table 1. Store Information.

Retail Store Number	Address
1	124 Erie St, Edinboro, PA 16412
2	110 Green Ridge St, Scranton, PA 18509
3	4610 Centre Ave, Pittsburgh, PA 15213
4	3865 Derry St, Harrisburg, PA 17111
5	200 S Lincoln Ave, Newtown, PA 18940

All information regards the covid case number are being download and process through New York Times. Price of facial masks is acquired from CVS.com. Population accounting data of each city is acquired from each city's official website. Retail Store Capacity are being generalized by since cvs store are generally 11,000-15,000 square feet. Number of stores, unit price, cost, storage capacity, storage and transportation cost, penalty cost are all included in the model.

3. Method

3.1. Background

The decision variables s (reorder point) and q together create the baseline (s, q) policy (order quantity). It is frequently used to protect the manufacturing plan against unforeseen circumstances

(e.g. random demands, machine breakdowns, late deliveries). The size of the reorder point s determines the time point at which replenishment orders are placed under the (s, q) policy, although the order quantity q remains constant throughout time. The (s, q) policy's optimum implementation would involve constant monitoring of the inventory situation. The total amount of merchandise on hand as well as that which is ordered, less any unfulfilled backorders, is the inventory position. According to the following decision rule, the inventory management system will initiate a replenishment order of size q if, at the time of a review instant, the inventory position has reached the reorder point s (from above). [3]

By interacting with the environment and learning the best course of action through trial and error within the framework of a reward scheme, an agent uses reinforcement learning (RL). It has a lengthy history and mixes deep learning and reinforcement learning citekonstantinos2020forecasting. Finance has benefited greatly from recent outstanding deep learning advances in sectors such as massive data, powerful computing, novel algorithmic approaches, mature software packages, and architectures [2].

The deep learning network has been extensively used in robotics, films, gaming, and other applications. AlphaGo and the development of neural architecture are recent examples. Between the input and output layers in deep learning, there may be one or more hidden layers. The total of the weighted units from the preceding layer is used to determine each unit's input for each input layer. The input from the preceding layer is typically represented by applying a nonlinear transformation. A backward computation of the error derivative might be done after computations move forward from input to output. The weighted might be modified to optimize a loss function using a back-propagate gradient approach. [4]

The reinforcement learning network starts with an agent interacting with an environment over time. At each time step$t$, the agent receives a state s_t in a state space $\mathcal{S}$ and chooses an action a_t from an action space$\mathcal{A}$, following a policy$\pi(a_t|s_t)$, which is the agent's behavior. After each action is being accomplished, the agent will receive a reward and transition to the next states_{t+1}. The process continues until the agent reaches a terminal state and then it restarts. The return $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$ is the discounted, accumulated reward with the discount factor$\gamma \in (0, 1]$. The agent aims to maximize the expectation of such long-term return from each state. [5]

When we approximate any of the following reinforcement learning components using deep neural networks, we produce deep reinforcement learning (deep RL) methods: value function, $\hat{v}(s; \theta)$ or$\hat{q}(s, a; \theta)$, policy$\pi(a|s; \theta)$, and model (state transition function and reward function). A RL agent uses the value function, policy, and model as its main building blocks to carry out a series of actions while observing states and rewards. A RL problem can be expressed as a prediction, control, or planning problem, and there are model-free and model-based solution approaches that can include value functions and/or policies.

3.2. Define the Environment

First, we define the following as initialization variables:
a) Maximizing Capacity of the warehouse and each retail store
b) Demand and Stock Level at each retail store each day
c) Unit Production, Storage, and Transportation Cost
Continuing, we define the state space which involves:
a) Stock Level of retail stores and distribution center
b) Demand History
Then, we define the action space, which includes variables:
a) Production Level
b) Number of masks shipping to each retail store
An algorithm of deep reinforcement learning is illustrated as below:

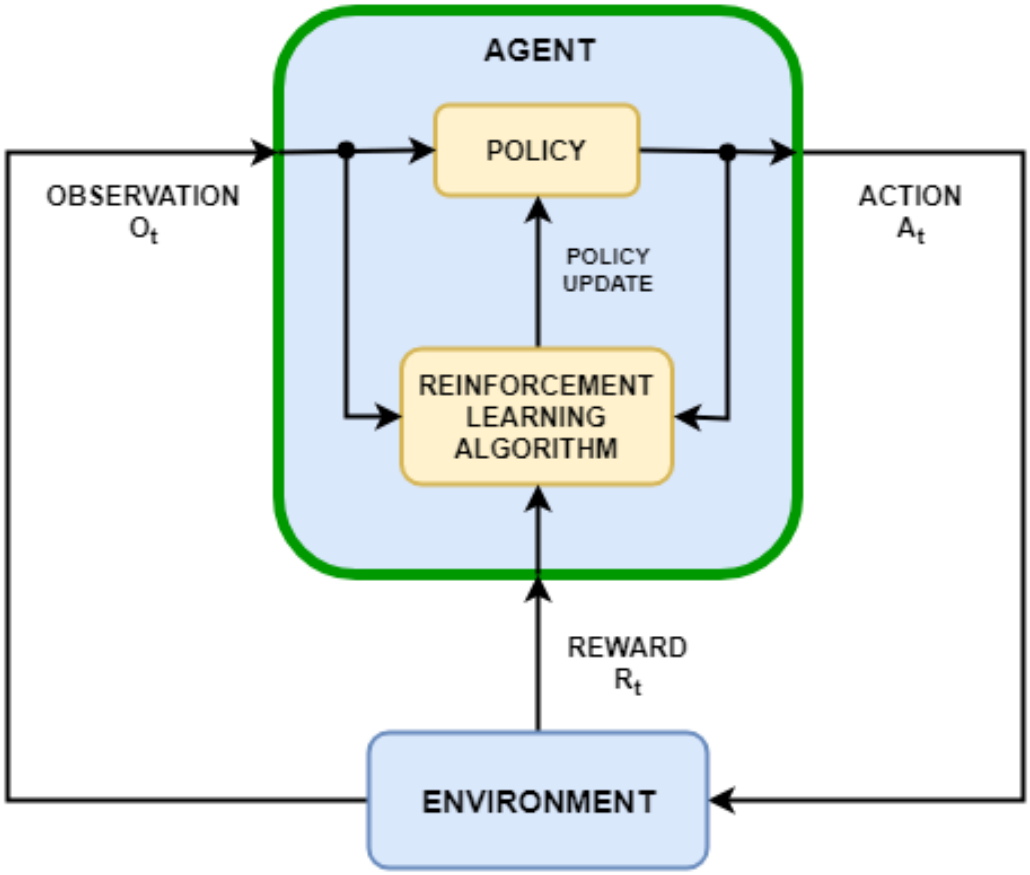


Figure 2. Environment of Reinforcement Deep Learning.

All notations are summarized in Table 2.

Table 2. Store Information.

Notations	Description
r	The reward function
p	Unit retail price of facial mask
d_k	demand of facial mask
v	Unit production cost of facial mask
p_k	Production level of facial mask
$z_{s,k}$	storage cost of retailer store k
$z_{tr,k}$	transportation cost of retailer store k
q_k	stock level of retailer store k
$z_{p,k}$	penalty cost of retailer store k

3.3. Data Collection and Demand Simulation

To simulate the demand in different area, we collect local data including:

a) Unit Production Cost and Retail Price

b) Unit Storage and Transportation Cost

c) Storage Capacity

d) City population

e) Covid Case Number

f) Number of CVS Retail Stores in each location

For unit production cost and retail price, we collect data from local retailer and CVS stores. We borrow data from Internet to make assumption of unit storage and transportation cost. We use satellite

map software to calculate the capacity of each store. Lastly, we borrow data from each state's official population tracker as well as real-time Covid case number from New York Times.

Given information needed, we can then formulate the approximated demand:

$$Demand = \frac{Population\ Number * Covid\ Case\ Number}{25000 * Number\ of\ Retail\ Stores} \quad (1)$$

The demand of each location is graphed below:

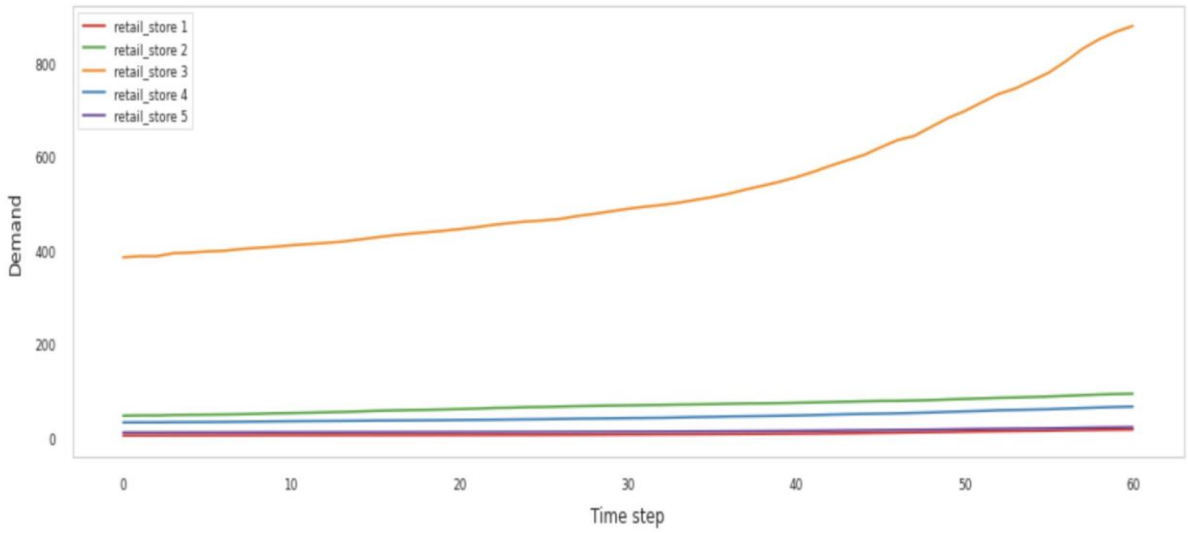


Figure 3. Demand of Facial Masks of each location.

3.4. Reward Function

Since our goal is to maximize profit, we set the reward as the profit. Referring to the environment, state space and action space described in previous subsection. We formulate the reward function as:

$$\text{Reward Function} = \text{Revenue} - \text{Storage Cost} - \text{Transportation Cost} + \text{Penalty Cost} \quad (2)$$

$$r = p\sum_{j=1}^{W} d_j - \sum_{j=0}^{W} z_j^S \max q_j, 0 - \sum_{j=1}^{W} z_j^T a_j + \sum_{j=1}^{W} z_j^P \min q_j, 0 \quad (3)$$

To calculate revenue, we multiply the unit price of the product by the quantity of sales. To calculate the storage cost, we multiply the unit storage cost by the quantity of product needed to be storage. To calculate the transportation cost, we multiply the quantity of product being transported each time by the unit cost of transportation. Lastly, the penalty cost is calculated by multiplying the unit penalty cost by the quantity of unfilled order (or 0 if there is no unfilled order). Notice that since the penalty cost is either 0 or negative, we put a plus sign in the equation [6].

4. Results and Discussions

Below is the result for both (s,q) policy and deep reinforcement learning:

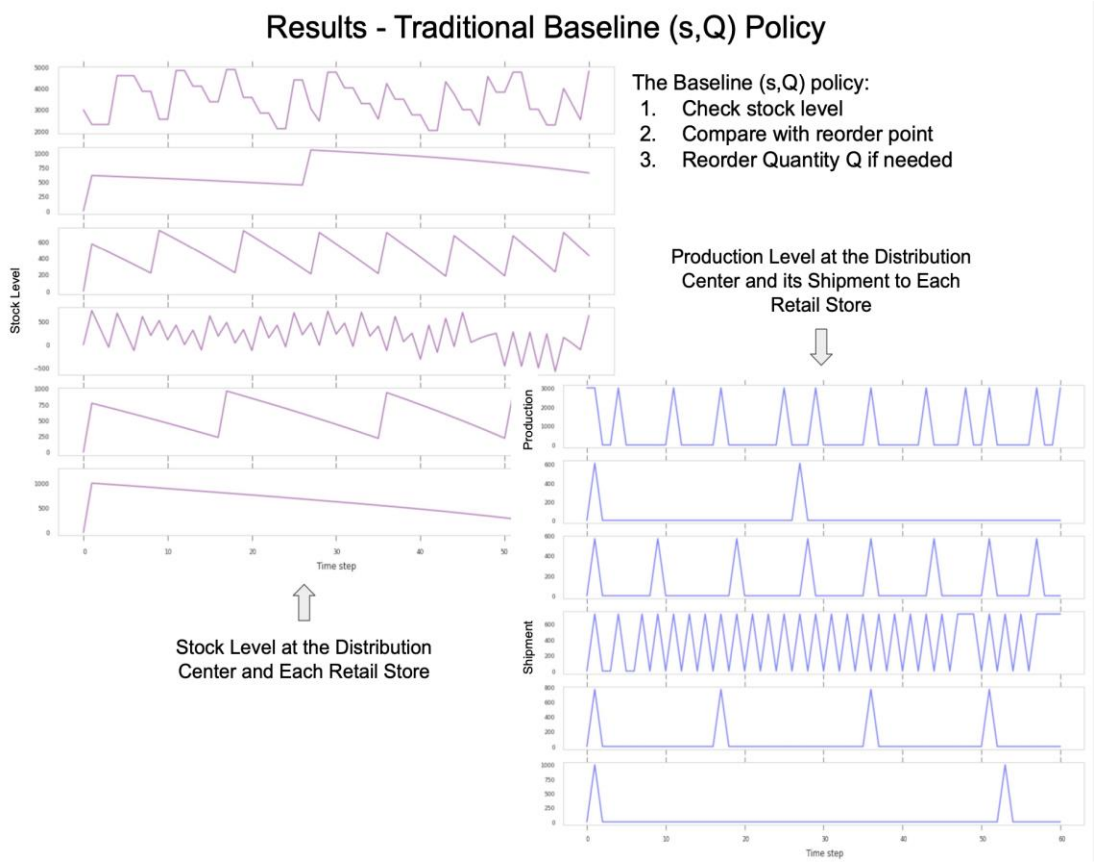


Figure 4. Each store's stock level under (s,Q) agent is on the top left and production level under (s,Q) agent in on the bottom right.

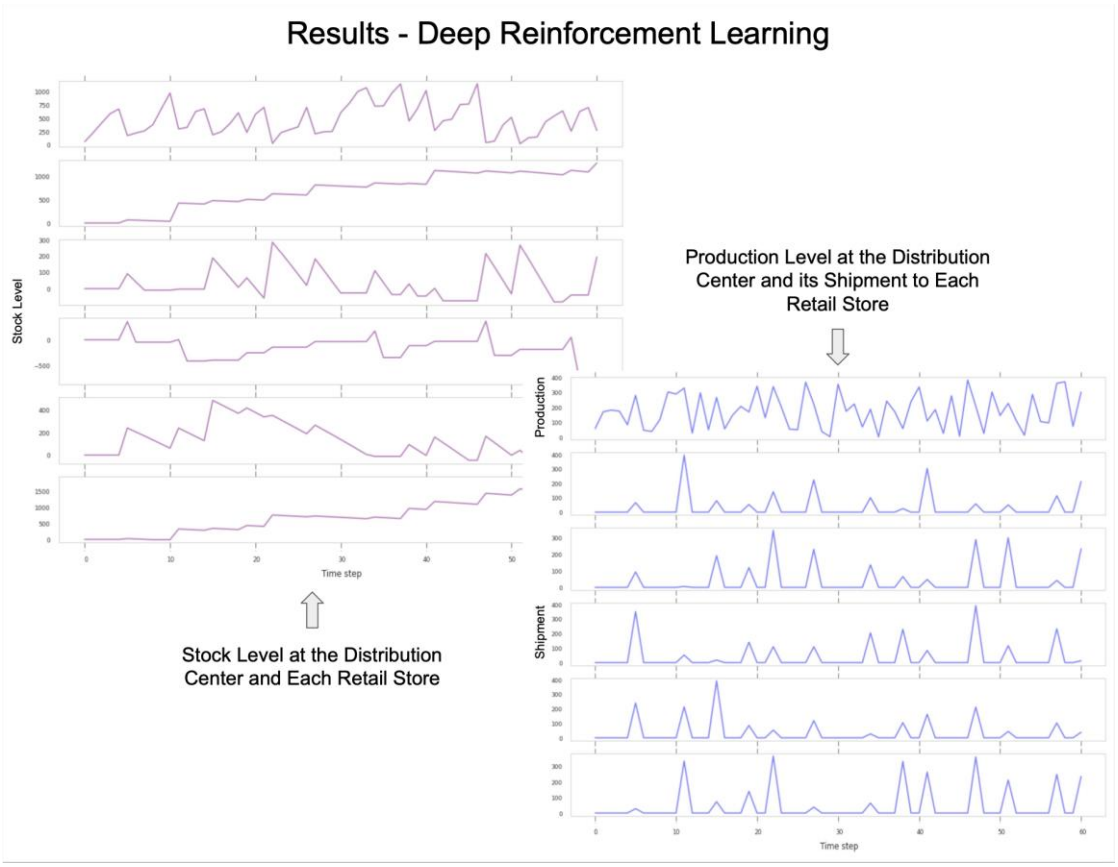


Figure 5. Each store's stock level under DRL agent is on the top left and production level under DRL agent in on the bottom right.

Shown on the graph above, for the (s, Q) agent each retail store's stock level decreases to a certain point and then gets pushed back to a higher point and repeats the pattern for several times. While there is not an obvious pattern or default restock rule for the DRL agent. As comparison, the (s, Q) agent restocks more frequently especially in location with large population. The result is as expected since cities with large population always need more facial masks more Covid patients and avoid passing of the pandemics.

Below is the reward simulation result, which monitors the total amount of profit for each agent.



Figure 6. (s,Q) agent reward result is at the top and DRL agent reward result is at the bottom.

The reward from the s,Q policy is about 0.9 million and that from DRL is about 1.05 million. And the relative standard deviations are very close. However, the cumulative profit of DRL is much greater and improved by 17.3% then s,Q policy, which means it greatly outperforms the traditional method.

5. Conclusion

As conclusion, this paper is dedicated to help medical agents and retailers better distribute their equipment under large-scale pandemics. In this paper we monitor 5 different CVS stores in the state of Pennsylvania to train the DRL agent over the period of 2 months and compare the result with traditional (s,Q) agent. The result as shown that the DRL agent performs better at restocking goods over time and maximizing profits.

The paper continues from previous literature and further applies to much more complicated and unpredicted environments. It also exploits new aspects of DRL and lays the foundation for future research. The model can be applied to much more complicated supply chain system which may involve a greater number of retailers and distribution centers. And there can be a more unpredicted environment and demand existing. The model can also be applied to other medical devices such as goggles, protective coveralls.

References

[1] Pierpaolo Pontrandolfo Ilaria Giannoccaro. Inventory management in supply chains: a reinforcement learning approach. International Journal of Production Economics, 78:153–161, 200Mariana Krakovsky. Reinforcement renaissance.

[2] Andreas Schafers Christos Tsinopoulos Chrysovalantis Vasilakis ¨ Konstantinos Nikolopoulos, Sushil Punia. Forecasting and planning during a pandemic: Covid-19 growth rates, supply chain disruptions, and governmental decisions. European Journal of Operational Research, 290:99–115, 2020.

[3] Christopher Krauss Thomas Fischer. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 239:N/A, 2018.

[4] ”masks on cvs.com. https://www.cvs.com/search?searchTerm=masks. Accessed: 2022-04-07.

[5] Mariana Krakovsky. Reinforcement renaissance.

[6] Coronavirus (covid-19) data in the United States. https://github.com/nytimes/covid-19-data. Accessed: 2022- 04-07.

[7] Aikaterini Dedeilia, Marinos G Sotiropoulos, John Gerrard Hanrahan, Deepa Janga, Panagiotis Dedeilias, and Michail Sideris. Medical and surgical education challenges and innovations in the covid-19 era: a systematic review. In vivo, 34(3 suppl):1603– 1611, 2020.

[8] Uma Gaur, Md Anwarul Azim Majumder, Bidyadhar Sa, Sankalan Sarkar, Arlene Williams, and Keerti Singh. Challenges and opportunities of preclinical medical education: Covid-19 crisis and beyond. SN comprehensive clinical medicine, 2(11):1992– 1997, 2020.

[9] Tom M Mitchell and Tom M Mitchell. Machine learning, volume 1. McGraw-hill New York, 1997.

[10] Raniah Zaheer and Humera Shaziya. A study of the optimization algorithms in deep learning. In 2019 third international conference on inventive systems and control (ICISC), pages 536–539. IEEE, 2019