Role of Reinforcement Learning in Financial Management Strategy

Soumya Shrivastava

Madhav Institute of Technology and Science, Gwalior (M.P.), 474009 India
er.soumyashrivastava[at]gmail.com

Abstract: ***Algorithm trading is a hot topic in machine learning. Reinforcement learning is widely used in financial trading because it can directly teach behavioural rules using rational rewards. A typical financial trading strategy application is flexible and expresses the state of multiple stocks. Limited trading activities for financial products were possible due to the difficulty of designing a flexible action space. Due to the inherent transformation of the market base, it is difficult to extract effective characteristics during trading battles from price fluctuations in financial markets, including perhaps noise, and models learned from past price data. It does not work well with unknown price data. Thus, we achieved effective feature extraction from raw price data by DNN, online reinforcement learning adaptation to unknown price factors, and flexible asset management across multiple stocks in this study. The Casual dilated convolution layer is a DNN layer architecture that can capture long-term dependencies. The proposed method did not outperform the conventional method that applied online deep reinforcement learning to the portfolio management method of financial products, but 1x5 Convolution is the optimum filter size for this method's convolution layer.***

Keywords: Machine Learning, Reinforcement Learning, DNN, Conventional Method

1. Introduction

Automate financial transactions is a hot topic in machine learning. It is possible to directly learn behavioural rules by using rational rewards in reinforcement learning, which is a type of machine learning. Financial price time series data is very unstable with lots of noise and jumps. Moving averages and statistics are known for reducing data noise and instability while summarising financial conditions. Also, mathematical finance has extensively studied the search for ideal financial indicators for technical analysis [5]. Technical analysis has never found any effective and generalised indicators. Moving averages, for example, can catch the trend but suffer significant losses when the market reverses.

2. Problem Formulation

To maximise long-term tribute behaviour and adapt to unknown markets, many studies have used online reinforcement learning [6] [3] [4]. Finance these methods are flexible because they can handle only discrete behavioural spaces because action selections increase dramatically to enable various buying and selling behaviours, learning is not stable. A practical financial trading strategy also handles multiple stocks simultaneously. Risk management is commonplace by diversifying assets.

So, in this study, we dealt with the difficulty of extracting financial characteristics. Deep neural networks (DNNs), which incorporate handcrafted feature engineering, have recently been used in image and voice recognition tasks. Online reinforcement learning is also used to address market volatility, trading behaviour restrictions, and risk management, which were issues when applied to conventional financial transaction strategies. Use deterministic policy gradient, a reinforcement learning method that handles large action spaces well.

3. Deep Learning

3.1 Convolutional Neural Network

A deep neural network is a multi-layer neural network with more intermediate layers. It is true that increasing the number of layers in a neural network increases the number of nodes, and thus the expressiveness of a nonlinear approximater, but backpropagation propagates the output error from the output layer to the input layer. Learning is difficult when the error gradually vanishes. The vanishing gradient problem. There are methods and architectures for avoiding the vanishing gradient problem in supervised deep neural networks.

3.1.1 Convolution Layer

The convolution layer extracts features from images by applying a ReLU filter. The filter used for this feature extraction is called the kernel, and the multiple features that are convoluted by the kernel but derived from the notification are called feature maps. Figure 1 illustrates the convolution operation.

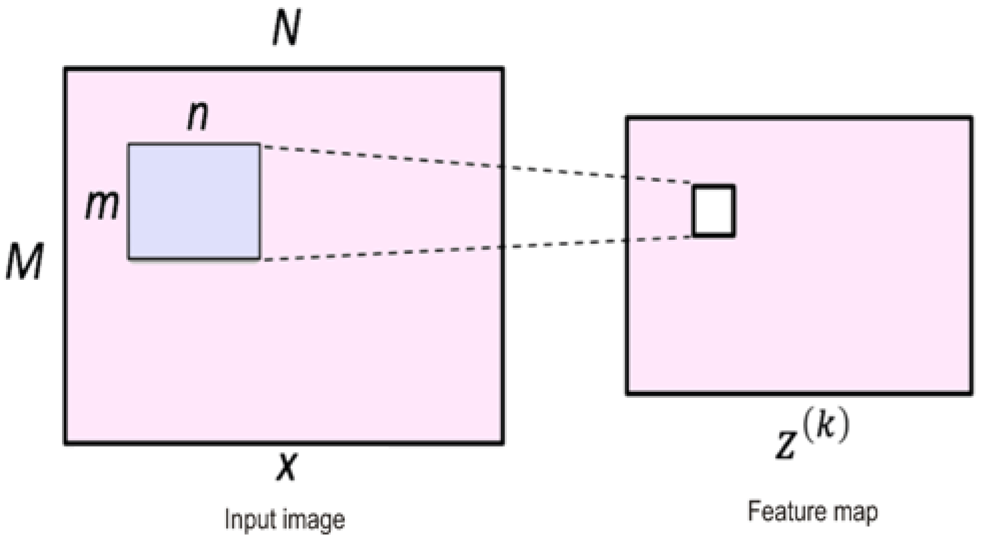


Figure 1: Schematic diagram of the convolution operation of convolution

Assuming that the image size is M × N and the kernel size is mxn, the convolution corresponding to the kth kernel can be expressed by equation (1).

$$\boldsymbol{z_{ij}^{(k)}} = \sum_{s=0}^{m-1} \sum_{t=0}^{n-1} \boldsymbol{w_{st}^{(k)}} \boldsymbol{x_{(i+s)(j+t)}} \qquad (1)$$

Here, w is the weight by the kernel, that is, the parameter of the model. When the convolution is finished, all the convolution values are activated by the activation function.

3.1.2 Pooling Layer

The most commonly used max pooling will be demonstrated here. When the activation open number output is disabled. Equation (max pooling) (2).

$$\boldsymbol{y_{ij}^{(k)}} = (\boldsymbol{a_{(l_1 i+s)\,(l_2 j+t)}^{(k)}}) \qquad (2)$$

4. Deep Reinforcement Learning

4.1 Overview

Deep reinforcement learning uses a deep neutral network to approximate value and policy functions. In traditional reinforcement learning, linear functions were used to represent value and policy functions. Because using a nonlinear function as a function approximator does not guarantee learning convergence. Also, traditional reinforcement learning necessitates feature design. But Deep Q-Network (DQN), a reinforcement learning method proposed by Mmih et al. [11][12] in 2013, is an Atari game. Employees with feature design using linear functions outperformed the reinforcement learning method of employees with feature design using linear functions in the task. He also outperformed the human average in over half of the Atari tasks. Because the game screen can be directly input and the game score is normalised and used as a reward, reinforcement learning is versatile enough to work for different Atari game tasks with the same design.

DDPG algorithm

```
Randomly initialize critic network $Q(s|\ \theta^Q)$ and actor $Q(s|\ \theta^\mu)$. with weights $\theta^Q$. and $\theta^\mu$
Initialize target network $Q$'and $\mu$'with weights $\theta^{Q}$ ' ← $\theta^{Q'}$, $\theta^{\mu'}$← $\theta^\mu$.
Initialize replay buffer $R$
for episode = 1, M do

Initialize a random process $\mathcal{N}$for action exploration
Receive initial observation state $s_1$
for t = 1, T do

Select action $a_t =\mu\ (s|\ \theta^Q) + \mathcal{N}_t$ according to the current policy and exploration noise
Execute action $a_t$and observe reward $r_t$and observe new state $S_{t+1}$
Store transaction ($S_t$, $a_t$, $r_t$, $S_{t+1}$) in $R$
Sample a random minibatch of transactions ($S_t$, $a_t$, $r_t$, $S_{t+1}$) from $R$
    $Set\ y_t = r(s_t, a_t) + \gamma Q'\left(s_{t+1}, \mu\left(s \mid \theta^{\mu'}\right) \mid \theta^{Q'}\right)$

Update critic by minimizing the loss:
    $L(\theta^Q) = \frac{1}{N} E[(Q(s,a \mid \theta^Q) - y_t)^2]$
Update the actor policy using the sampled policy gradient:
    $\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s,a \mid \theta^Q)\Big|_{s=s_i,a=\mu(s_i)} \nabla_{\theta^\mu}\mu(s \mid \theta^\mu)\Big|_{s=s_i}$
        Update the target networks:

    $\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'} \theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$
        end
    end
```

5. Related research

This section summarises deep reinforcement learning for financial transaction optimization. 5.1 uses the value iterative method. Section 5.2 explains the deep policy gradient method, and Section 5.3 summarises the open studies.

5.1 Value Iterative Method

5.1.1 Method by Welfare-Type Deep Reinforcement Learning [14]

Matsui et al. [14] proposed compound interest reinforcement learning, which extended Q-learning to deep reinforcement learning by Deep Q-Network. In Chapter 3, the sum of discount rewards expressed by the formula (3) was used as revenue.

$$G_t = \sum_{\tau=0}^{\infty} \gamma^{\tau} R_{t+1+\tau} \qquad 0 < \gamma < 1 \qquad (3)$$

Depending on the task, the maximum profit is not always the sum of discount rewards. The rate of return is important in financial transactions. To maximise the compound interest effect of the rate of return, deep strengthening learning uses the compound interest profit rate as the profit. Using the rate of return R, we get the compound interest rate of return.

$$G_t = \prod_{k=1}^{t} (1 + R_k) \ (28)$$

Buy (long position) and sell (short position) Japanese government bonds are the investment targets (short position). In both cases, all assets are invested. The composition is a fully connected neural network with 20 units in the middle layer and 1 unit in the output layer, with 3 layers in total.

5.2 Deep Policy Gradient Method

5.2.1 Deep Deterministic Policy Gradient Method [15]

Zhengyao et al. [15] use the Deep deterministic policy gradient [16]. So we proposed an optimal automated portfolio management method for multiple cryptocurrency stocks.

INN's input data refers to the state in reinforcement learning. As shown in Figure 2, the relative price data for Bitcoin for each cryptocurrency stock in the portfolio are obtained at 30-minute intervals.

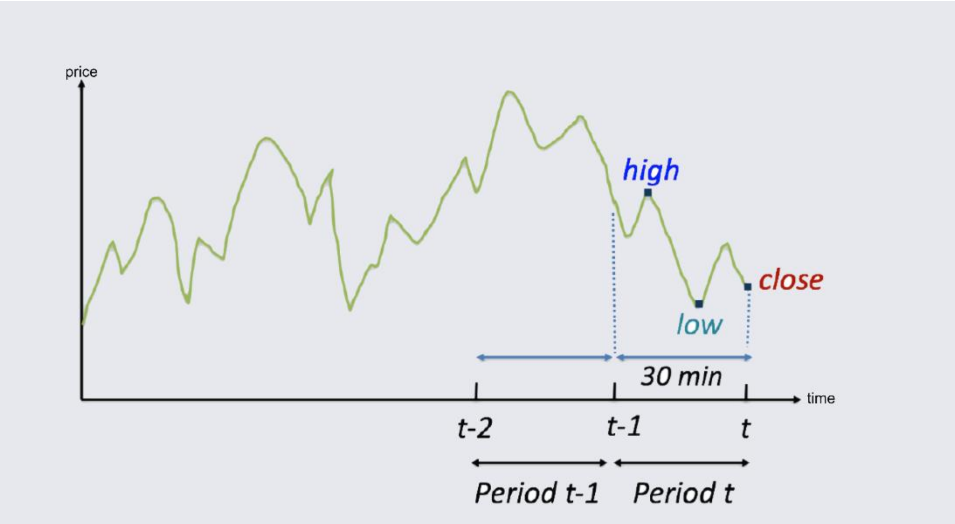
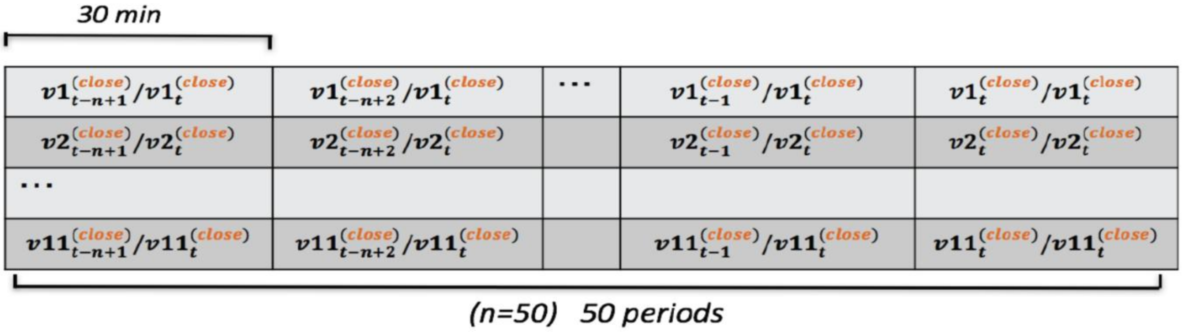


Figure 2: How to get price data for stocks using the method

Next, for each of the 11 types of virtual currency stocks p1 to v11, for example, the closing price in the period of the stock is expressed. Based on the latest period at this time, the latest 50 period is represented as a two-dimensional table as shown in 2.

Table 1: Two-dimensional table of closing prices



30 min

$v1_{t-n+1}^{(close)}/v1_t^{(close)}$	$v1_{t-n+2}^{(close)}/v1_t^{(close)}$	...	$v1_{t-1}^{(close)}/v1_t^{(close)}$	$v1_t^{(close)}/v1_t^{(close)}$
$v2_{t-n+1}^{(close)}/v2_t^{(close)}$	$v2_{t-n+2}^{(close)}/v2_t^{(close)}$		$v2_{t-1}^{(close)}/v2_t^{(close)}$	$v2_t^{(close)}/v2_t^{(close)}$
...				
$v11_{t-n+1}^{(close)}/v11_t^{(close)}$	$v11_{t-n+2}^{(close)}/v11_t^{(close)}$		$v11_{t-1}^{(close)}/v11_t^{(close)}$	$v11_t^{(close)}/v11_t^{(close)}$

(n=50) 50 periods

Table 2: High-priced 2D table

$v1^{(high)}_{t-n+1}/v1^{(close)}_{t}$	$v1^{(high)}_{t-n+2}/v1^{(close)}_{t}$	…	$v1^{(high)}_{t-1}/v1^{(close)}_{t}$	$v1^{(high)}_{t}/v1^{(close)}_{t}$
$v2^{(high)}_{t-n+1}/v2^{(close)}_{t}$	$v2^{(high)}_{t-n+2}/v2^{(close)}_{t}$		$v2^{(high)}_{t-1}/v2^{(close)}_{t}$	$v2^{(high)}_{t}/v2^{(close)}_{t}$
…				
$v11^{(high)}_{t-n+1}/v11^{(close)}_{t}$	$v11^{(high)}_{t-n+2}/v11^{(close)}_{t}$		$v11^{(high)}_{t-1}/v11^{(close)}_{t}$	$v11^{(high)}_{t}/v11^{(close)}_{t}$

Table 3: High-priced 2D table

$v1^{(low)}_{t-n+1}/v1^{(close)}_{t}$	$v1^{(low)}_{t-n+2}/v1^{(close)}_{t}$	…	$v1^{(low)}_{t-1}/v1^{(close)}_{t}$	$v1^{(low)}_{t}/v1^{(close)}_{t}$
$v2^{(low)}_{t-n+1}/v2^{(close)}_{t}$	$v2^{(low)}_{t-n+2}/v2^{(close)}_{t}$		$v2^{(low)}_{t-1}/v2^{(close)}_{t}$	$v2^{(low)}_{t}/v2^{(close)}_{t}$
…				
$v11^{(low)}_{t-n+1}/v11^{(close)}_{t}$	$v11^{(low)}_{t-n+2}/v11^{(close)}_{t}$		$v11^{(low)}_{t-1}/v11^{(close)}_{t}$	$v11^{(low)}_{t}/v11^{(close)}_{t}$

These three two-dimensional tables are used as one input data.

Figure 3 shows the convolutional neural network used as the policy function in the method of [15]. 1x3 convolution for the input data after obtaining two feature maps, one-dimensional convolution is performed in the column direction. The softmax function is a 12-dimensional vector obtained by convolutioning the 11-dimensional feature map with the transaction fee. Enter and output the optimal portfolio for the next step relative to the current portfolio as a continuous value vector.

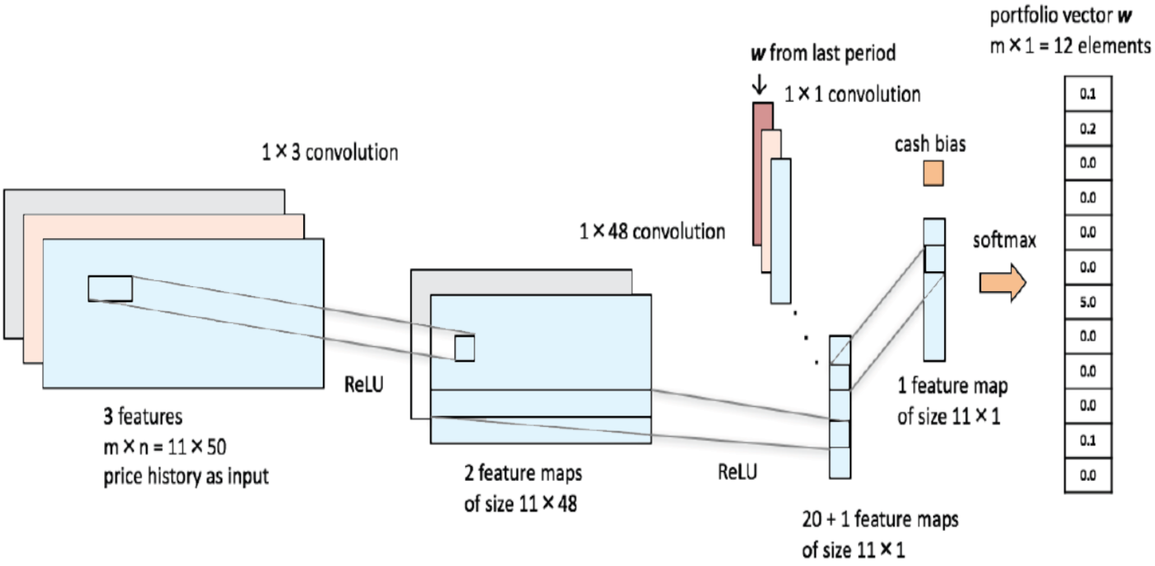


Figure 3: Deep Portolio Management DNN Design

5.3 Summary

Table 4 summarises the related studies described in Sections 5.1 and 5.2. They improved Q-learning to maximise the compound interest effect. Because the action space can only handle discrete values, the buying and selling options are limited. So the operation is separate from trading, allowing you to control how much you buy or sell multiple stocks. Using the method of Zhengyao et al. [15], multiple stocks can be managed simultaneously, and the portfolio can be managed with continuous values, allowing for flexible buying and selling.

Table 4: Summary of related research

	Proposal method	Advantage	question
Matsui [14]	Compound interest reinforcement learning using Deep Q-network	• Maximum compound interest effect Reward design to become Because you are In line with the purpose of investment Designed	• Handle only one brand • Limited action choices.
Zhengyao [15]	Deep Deterministic Portfolio management with Policy Gradient	• You can handle multiple brands at the same time. • Continuous action selection it can be handled by value.	• There is room for improvement in DNN's architecture.

6. Methodology

In this section, we use the method of [15] to recognise multiple input parallel time series data and generate an optimal asset distribution portfolio. It is the foundation of the proposed network design method. After describing the Casual dilated convolution [15].

6.1Casual Dilated Convolution

Casual dilated convolution is a type of convolutional layer in a convolutional neural network proposed by Yu et al. [17] in 2015. Figure 4 shows the operation of the convolutional filter used in a conventional convolutional neural network.

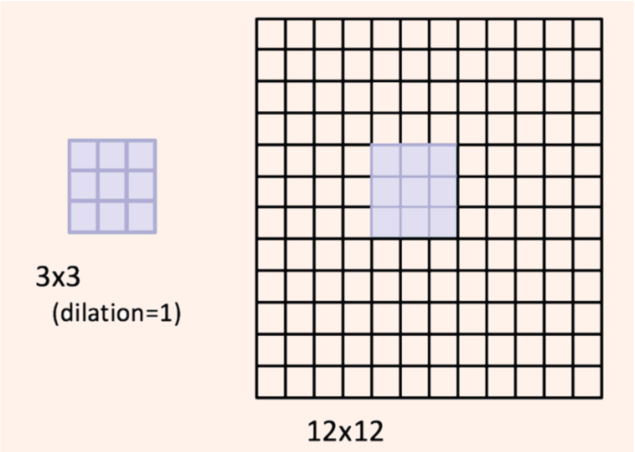


Figure 4: Convolution of 12x12 input data with a 3x3 filter (dilation = 1)

Figure 4 shows 12x12 data convolved with a 3x3 size filter. As shown in Figure 4, a conventional convolution layer convolves adjacent values. Dilation-1 convolution is the process of convolutioning adjacent values without a gap.

However, Yu et al. proposed convolution with a dilation of 2 or more and confirmed that the conventional method improved image division accuracy. About dilation-2 convolution. Figure 5 explains this.

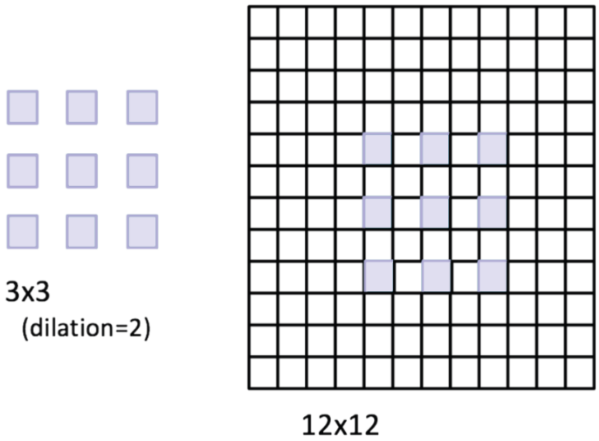


Figure 5: Convolution of 12x12 input data with a 23x3 filter (dilation = 2)

In Figure 5, when dilation = 2, the filter shape is a convolution with a gap composed of the values skipped by one. Similarly, for dilation = 3, a filter composed of two skip values is used.

The term casual dilated convolution dilation refers to two or more convolutional genera, and it has recently been applied to the method of text content speech synthesis [18]. Compared to conventional convolution dilation = 1, the visual receptive field is expanded to allow recognition of the entire region's background.

6.2 Application of Casual Dilated Convolution

Using the [15] method's casual dilated convolution so I want to improve it.As shown in Figure 6, the first layer input data is 1x3.The convolution filter convolved it according to the brand's price.The method proposed. As shown in Figure 6, a convolution filter with dilation = 2

We suggest a DNN design.

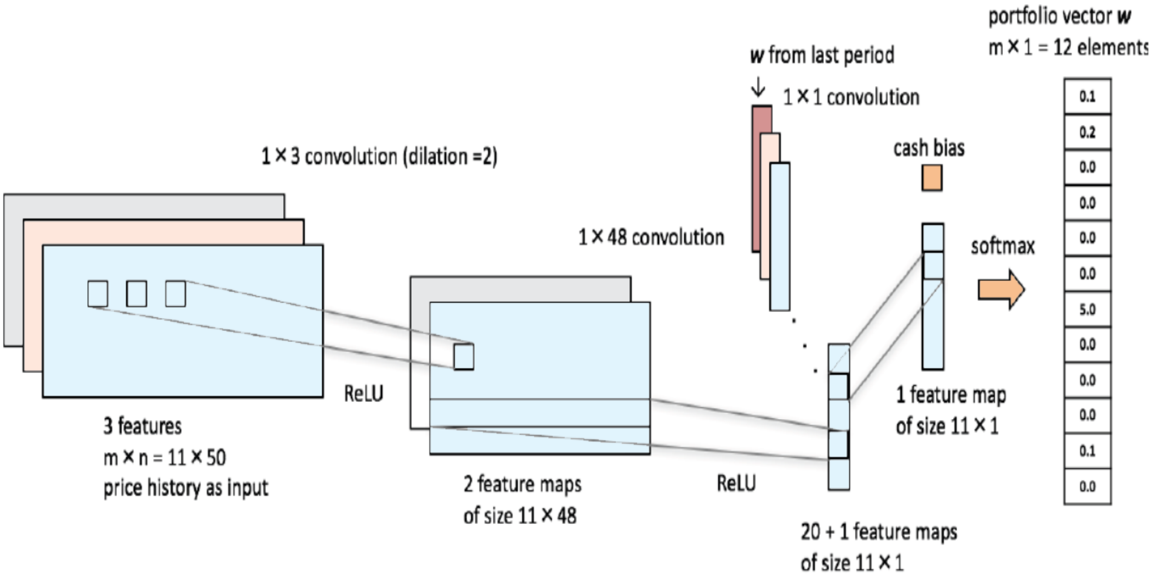


Figure 6: DNN design of the proposed method

By doing this, the longer-term correlation of the time-series price data of financial instruments is recognized. By doing so, we aim to improve performance.

7. Evaluation Experiment

7.1 Dataset

Similar to the [17] paper. Bitcoin (BTC) was the basis. 10 types of virtual currencies (ETH / BTC.LTC / BTC XRP / BTC.USD / BTC ETC / BTC, DASH / BTC XMR / BTC XEM / BTC FCT / BTC GNT / BT CZEC / BTC) and US dollar (USD / BTC) Experimenting a total of 11 types of financial products as a portfolio Use the past actual prices of 11 types of financial products obtained from the API of Poloniex of the virtual currency exchange as a data seminar. Poloniex, https://poloniex.com/

7.2 Experimental Method

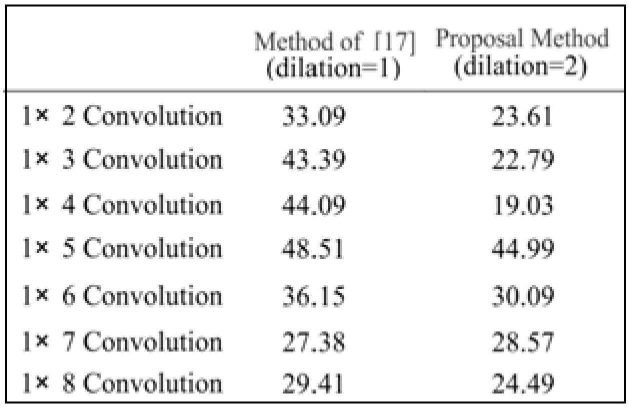
The portfolio is also reassembled for each step of 30 minutes, and the training period's data is used to learn 80,000 steps. The evaluation index is fAPV expressed by the following formula.

$$assets\ APV = Total: \frac{Total\ amout\ of\ final}{amount\ of\ initial\ assets} \quad (4)$$

7.3 Experimental Results

To compare the methods of [17] and [18], the proposed method uses the conventional convolution (dilation = 1) and the proposed method uses the convolution (dilation = 2) However, experiments were conducted in each of the following cases shown in Table 5.

Table 5: FAPV values of experimental results of the method and proposed method in [17]



	Method of [17] (dilation=1)	Proposal Method (dilation=2)
1× 2 Convolution	33.09	23.61
1× 3 Convolution	43.39	22.79
1× 4 Convolution	44.09	19.03
1× 5 Convolution	48.51	44.99
1× 6 Convolution	36.15	30.09
1× 7 Convolution	27.38	28.57
1× 8 Convolution	29.41	24.49

7.4 Results

Table 5 summarises the experimental results for dilation -1. The model convolves with 1x5 Convolution in both the [15] and dilation-2 methods. That is why it is critical to select the correct filter size for each model. The method of [15] used the 1x5 Convolution convolution in the paper of the method of [15] 1x3 we got better performance than Convolution convolution. No other filter sizes were outperformed by the proposed method in comparison to the method in [15]. Thus, the conventional convolution of dilation-1 is considered suitable for the method of [15]. The proposed method is much lower than the method of [15] for filter sizes 1 2.1 3.1 4. The proposed method outperforms the method [15] by 1 5.1 61 7.1 8 in the filter size. As a result, using Casual dilated convolution for a small size filter in this method will significantly degrade performance. This study modified the embedded filter and tested it. However, future experiments will be needed to determine how changing data size and stock combination affects the results. In this study, we used two convolution filters with the same filter size and dilation value. This one's a bit different. Using the same convolution layer for each filter may help recognise time series of different short-term and long-term periods and contribute to performance. It will also be tested for other financial products such as stocks and legal tender in the future.

8. Conclusion

In this paper, we will discuss how to automate financial trading strategies to maximise long-term profits. We proposed an online deep-strengthening learning method that outputs the optimal multi-stock portfolio. The existing method's convolutional neural network architecture has been improved, enabling long-term background recognition of time-series price data and improving performance. So we tested the casual dilated convolution, which expands the visual receptive field of the convolutional layer. The proposed method did not outperform the conventional method, but the 1x5 Convolution was the optimum filter size for the embedded layer in both methods. The next step is to see if the performance can be improved by convolving the same convolution layer with multiple filters of varying sizes and dilations is done. There is also room to verify differences in results due to changes in input data size and brand count.

References

[1] Farabet, Clement, et al. "Learning hierarchical features for scene labeling." IEEE transactions on pattern analysis and machine intelligence 35.8 (2012): 1915-1929.

[2] Sainath, T. N., et al. "Acoustics, Speech and Signal Processing (ICASSP)." 2013 IEEE International Conference on. IEEE. 2013.

[3] Jangmin, O., et al. "Adaptive stock trading with dynamic asset allocation" using reinforcement learning. "Information Sciences 176.15 (2006): 2121-2147. APA

[4] Bertoluzzo, Francesco, and Marco Corazza. "Making financial trading by recurrent reinforcement learning. "International Conference on Knowledge- Based and Intelligent Information and Engineering Systems. Springer, Berlin, Heidelberg, 2007.

[5] JJMurphy, Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Metods and Applications, New York, NY, USA: New York Institute of Finance

[6] Lee, Jay, and Linxia Liao. "Methods for prognosing mechanical systems." US Patent No. 8,301,406. 30 Oct. 2012.

[7] Collobert, Ronan, Christian Puhrsch, and Gabriel Synnaeve. "Wav2letter: an end-to-end convnet-based

speech recognition system. "ArXiv preprint arXiv: 1609.03193 (2016).

[8] Blunsom, Phil, Edward Grefenstette, and Nal Kalchbrenner. "A convolutional neural network for modeling sentences. "Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014.

[9] Srivastava, Nitish, et al. "Dropout: A simple way to prevent neural networks from overfitting. "The Journal of Machine Learning Research 15.1 (2014): 1929-1958.

[10] Silver, David, et al. "Deterministic policy gradient algorithms." ICML. 2014.

[11] Mnih, Volodymyr, et al. "Playing atari with deep reinforcement learning. "ArXiv preprint arXiv: 1312.5602 (2013).

[12] Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning. "Nature 518.7540 (2015): 529.

[13] Watkins, Christopher JCH, and Peter Dayan. "Q-learning." Machine learning 8.3-4 (1992): 279-292.

[14] Togoro Matsui, Masahiro Katagiri: "Compound Interest Deep Reinforcement Learning for Financial Transaction Strategy". Artificial Knowledge Noh Society Financial Informatics Study Group (SIG-FIN-016-01), pp.1-7, 2016.

[15] Zhengyao Jiang, et al. "A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem. "ArXiv: 1706.10059 (2017)

[16] Lillicrap, Timothy P., et al. "Continuous control with deep reinforcement learning. "ArXiv preprint arXiv: 1509.02971 (2015).

[17] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions. "ArXiv preprint arXiv: 1511.07122 (2015).

[18] Van Den Oord, Aaron, et al. "Wavenet: A generative model for raw audio. "arXiv preprint arXiv: 1609.03499 (2016).

[19] Sainath, T. N., et al. "Acoustics, Speech and Signal Processing (ICASSP)." 2013 IEEE International Conference on. IEEE. 2013.

[20] Lawrence, S., Giles, C. L., Tsoi, A. C. & Back, A. D. Face recognition: a convolutional neural-network approach. IEEE Trans. Neural Networks 8, 98–113 (1997).

[21] Waibel, A., Hanazawa, T., Hinton, G. E., Shikano, K. & Lang, K. Phoneme recognition using time-delay neural networks. IEEE Trans. Acoustics Speech Signal Process. 37, 328–339 (1989).