Schulich School of Business

FINE 6901 WF19: Individual Study

Deep Reinforcement Learning Architecture for Portfolio Optimization

Oct 26, 2019

Daniel Fudge
215868904

Executive Summary

Table of Contents

Background

This report is a component of a larger 3-term Independent Study with Professor Yelena as part of a concurrent Master of Business Administration (MBA) and Diploma in Financial Engineering from Schulich School of Business.

The first term was a general investigation into applications of machine learning to portfolio optimization. Here we reviewed the different aspects of machine learning and their possible applications to Portfolio Optimization. During this investigation we highlighted Deep Reinforcement Learning (DRL) as an especially promising area to research and proposed the development of a DRL framework to better understand its applications. The report capturing this research may be found at the following link.

https://github.com/daniel-fudge/DRL-Portfolio-Optimization/blob/master/docs/report1.pdf

The second term, which this report summarizes, pursues the future work identified in the first term. To get a better understanding of DRL, the Udacity DRL Nanodegree was completed. A description of this nanodegree can be found at the following link.

https://www.udacity.com/course/deep-reinforcement-learning-nanodegree--nd893

This Nanodegree involved the developing DRL networks to solve 3 different Unity ML-Agents environments. The solutions to these environments can be found in the following GitHub repositories.

1. Banana Collector - https://github.com/daniel-fudge/banana_hunter
2. Reacher - https://github.com/daniel-fudge/reinforcement-learning-reacher
3. Tennis - https://github.com/daniel-fudge/reinforcement-learning-tennis

With the Udacity Nanodegree complete, the next step is formally defining the problem statement and associated solution architecture, as captured in this report. With the problem statement and architecture defined, the task of the third term will be to implement the architecture. Note that the intent of this implementation is not to advance the state-of-the-art in DRL or its application to portfolio optimization. Instead it is to get a functioning architecture operational. From this conventional implementation, we can experiment with more advanced techniques.

Reinforcement Learning (RL) Review

At a high level, RL is the process of learning the optimal strategy (***Policy***) that defines ***Actions*** that an ***Agent*** should take given the current ***State***, ***Reward*** function, and future reward ***Discount Factor***. The figure below illustrates the classic RL feedback loop used to determine the optimal policy. Here, the environment is modeled as a function that predicts the next reward and state given the current state and action.

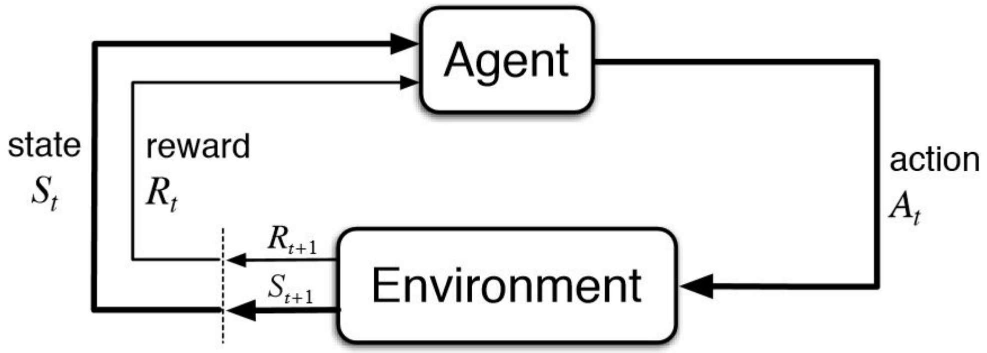


Figure 1. Reinforcement Learning feedback loop.
Image source: https://i.stack.imgur.com/eoeSq.png

Definitions

Before we begin the discussion we need to make the following definitions.[1]

1. Agent: The algorithm or person that takes the given action.
2. Environment: The world which the agents exists. It converts the agent's current state and actions into the associated rewards and next state.
3. Action (***a***): All the possible moves that the agent can take.
4. State (***s***): Current situation returned by the environment.
5. Reward (***R***): An immediate return sent back from the environment to evaluate the last action.
6. Policy (π): The strategy that the agent employs to determine the next action based on the current state.
7. Value (***V***): The expected long-term return with discount, as opposed to the short-term reward ***R***. ***V***π(***s***) is defined as the expected long-term return of the current state ***s*** under policy π.
8. Q-value or action-value (***Q***): Similar to ***V***, except that it takes an extra parameter, the current action ***a***. ***Q***π(***s***, ***a***) refers to the long-term return of the current state ***s***, taking action ***a*** under policy π.
9. Discount Factor (γ): The discount factor applied to future rewards.

[1]Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. 2nd edition, Cambridge, Massachusetts: *The MIT Press*, 2018.

Adding "Deep" to RL

Reinforcement learning can be implemented traditionally as an extension of dynamic programming, which requires a model of the environment. For instance, a classic model-based approach for option pricing uses a stochastic differential equation such as the Black-Scholes-Merton (BSM) model as the environment.[2] It then uses standard dynamic programming techniques and the Bellman optimality equation as an action-value function to identify the optimal policy.

In a model-free approach we do not assume a model of the environment, instead we deploy one or more deep neural networks to learn from the environment and build the optimal policy directly. As shown below a neural network becomes deep as the number of hidden layers increase, which allows more complex relationships to be modeled. In later sections we will discuss special networks such as CNN, RNN and LSTM that can designed to extract spatial and sequential structure from the input space.

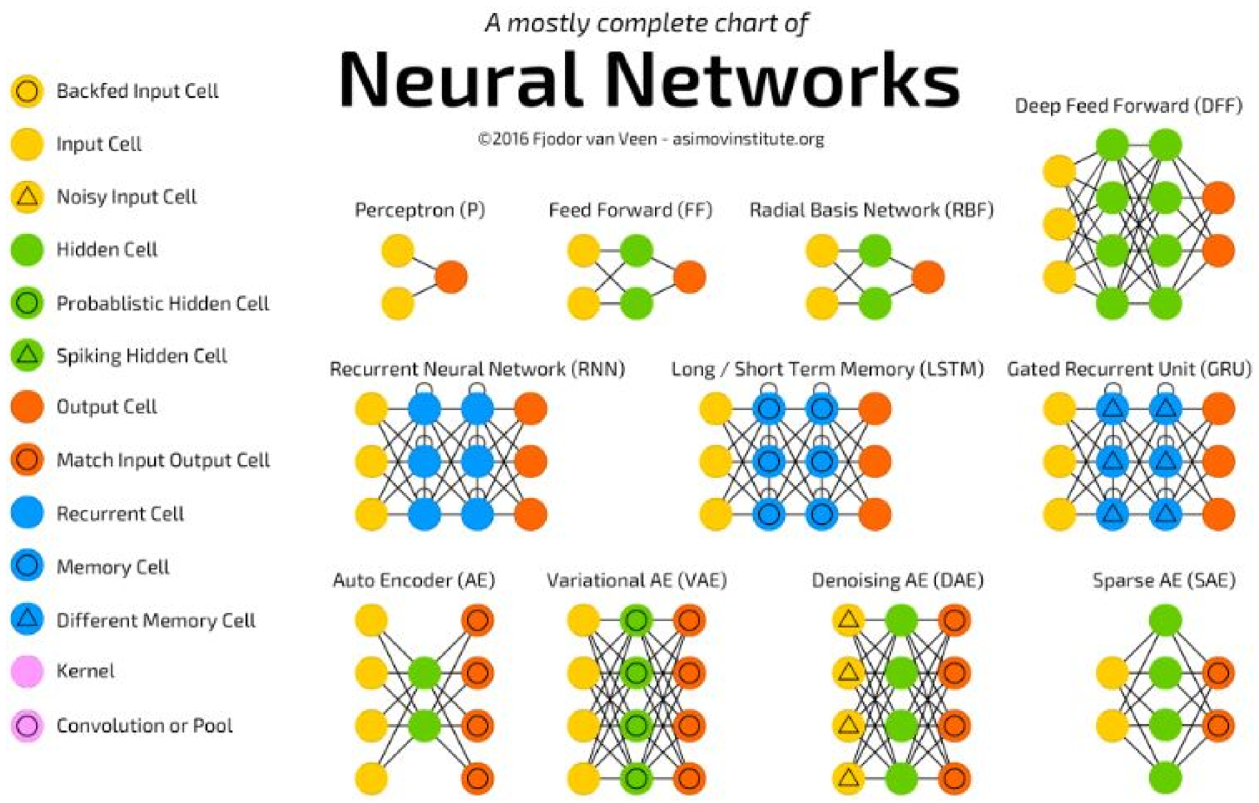


Figure 2. Neural network architectures.[3]
Image source: https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/neural_networks.html

[2] Igor Halperin. "QLBS: Q-Learner in the Black-Scholes(-Merton) Worlds." *SSRN Electronic Journal*, December 2017, DOI: 10.2139/ssrn.3087076, p2.
[3] Leonardo Araujo dos Santos. "Artificial Intelligence." *GitBook*, accessed October 6, 2019. https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/

Problem Statement

The problem statement is the state space, rewards, action space and objective.

State Space

Since we are performing portfolio optimization, the first obvious state selection will be the prices of the universe of assets we may place in our portfolio. For simplicity we will randomly choose *n* assets from the S&P500 that have data for the last 12 years; 10 years for training and 2 years for testing.

The next decision is what time scale we wish to model. On one extreme is high frequency trading, which operates on the sub-second time scale. A major advantage of this scale is the quantity of data available to train our model. More data allows use to train more complex networks that can extract extremely complex interactions. However, to capitalize at this time scale we require the ability to access and react to the data as fast as any other actor. The other end of the spectrum is monthly data, which gives us plenty of time to react, but doesn't provide enough data to model complex interactions before the environment shifts and reduces the accuracy of the data. For instance, in 5 years there are only 60 monthly data points, which is completely insufficient to build an accurate deep neural net. Therefore we will select daily prices as a compromise.

To capture some of the volatility in the daily data, we will also select the closing, high and low prices ($\mathbf{P}, \mathbf{P^H}, \mathbf{P^L}$). We will also assume the markets are not perfect and react to momentum, therefore we will include *k* days of prices. To add a signal for the market, we will add the S&P 500 index to the *n* assets. This gives a [3, *n*+1, *k*] price tensor $\mathbf{Y_t}$.

$\mathbf{Y_t} = [\mathbf{P}, \mathbf{P^H}, \mathbf{P^L}]$ with $p_{i,j}$ for asset $i \in [0, n]$ and time embedding $j \in [t-k, t]$

Note that this state tensor could easily be expanded to include a wide assortment of signals such as GDP, inflation rates, interest rates, company financial ratios or even twitter sentiment. This purpose of this study is to develop the architecture, the signal selection is a never ending area of research.

We also need to add the portfolio weighting $\mathbf{w}$ as defined below to define the full state space $\mathbf{S_t}$. Note asset 0 is set to cash for simplicity

$w_{i,t}$ *for asset* $i \in [0, n]$, $\sum_{i=0}^{n} w_{i,t} = 1$ *for all* t and $\mathbf{w_0} = [1, 0, \ldots 0]^\top$

$$\mathbf{S_t} \doteq [\mathbf{Y_t}, \mathbf{w_t}]$$

Reward Function

The reward function r_t has three components. The first component is the change in portfolio value from t to t+1, which we will denote $\Delta\rho_t$. To the n S&P assets we will also carry cash to make a portfolio of n+1 assets. To calculate the change in portfolio value we must consider the transaction cost of changing the portfolio weighting $\mathbf{w}$ as defined below. Note c is a fee for selling or purchasing assets. A more accurate representation of these fees is left for future work.

Due to the changes in asset prices between t-1 and t, we must define the weighting $\mathbf{w'}_t$ before the trades at time t as follows. Note $\odot$ and $\oslash$ are element-wise multiplication and division respectively.

$$\mathbf{u_t} \doteq \mathbf{P_t} \oslash \mathbf{P_{t-1}} = [1, p_{1,t}/p_{1,t-1}, \ldots, p_{n,t}/p_{n,t-1}]^{\mathsf{T}}$$

$$\mathbf{w'_t} = (\mathbf{u_t} \odot \mathbf{w_{t-1}})/(\mathbf{u_t} \cdot \mathbf{w_{t-1}})$$

$$\mathbf{\Delta w_t} = \mathbf{w_t} - \mathbf{w'_t}$$

$$\overline{c_t} \doteq 1 - c\sum_{i=0}^{n} |\Delta w_{i,t}|$$

$$\Delta\rho_t \doteq \ln(\overline{c_t}\mathbf{u_t} \cdot \mathbf{w_{t-1}})$$

The second reward component is added to penalize risk as indicated by large drops in the portfolio value over a given number of time steps m, denoted as d_t. Setting m > 1, prevents the penalization of short term volatility as long as the drops are recovered in the m time steps. Note this also requires the recording of the portfolio value ρ_t for the previous m time steps, where ρ_0 is set to 1 for simplicity.

$$\rho_t \doteq \rho_{t-1}\overline{c_t}\mathbf{u_t} \cdot \mathbf{w_{t-1}}$$

$$d_t \doteq -ln(min(1, \rho_t/\rho_{t-m}))^2$$

The third component is simply a binary gain g_t applied to the reward that is set to zero if the portfolio value hits zero, indicating a default. In practice, the code will simply end the episode and set the remaining rewards to zero. Therefore the reward is the sum below.

$$r_t \doteq \Delta\rho_t + d_t = ln(\overline{c_t}\mathbf{u_t} \cdot \mathbf{w_{t-1}}) - ln(min(1, \rho_t/\rho_{t-m}))^2 \quad [1]$$

Action Space

From the above reward function it can be seen that the action space is simply $\Delta \mathbf{w_t}$.

Objective

To define the objective we also need a discount factor γ, an investment time horizon *T* and portfolio value ϱ_t. Therefore the objective is to maximize the discounted expected portfolio value at time T as defined below.

$$MAX : \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t\right] \qquad [2]$$

Methodology Review

Now that the problem statement is defined our task is to build a network that learns the optimal policy (π) that select the actions ($\mathbf{a}$) for a given state ($\mathbf{s}$) to maximize the objective. There are a wide variety of network architectures to solve this problem. The list below is a short summary of the popular methods with features that will be incorporated into the chosen network.

Deep Q-Learning (DQN)[4]

A Deep Neural Network is generated that learns the **Q**-value of all possible actions. Therefore the optimal policy can simply take the greedy action that maximizes the **Q**-value for the given state. Often an ϵ-greedy action is taken where there is a probability ϵ that a random action will be taken. This allows for learning and exploration and the value of ϵ decays over time as the network becomes smarter.

Double DQN[5]

Google's DeepMind developed Double DQN to overcome DQN's overestimation of the **Q**-values by creating two identical networks; one for action selection (local) and one for evaluation (target). The local network that selects the actions is constantly learning from experience and uses the difference form the more stable target as loss function. The target is then updated periodically from the local network with an update factor τ; target = τ*local + (1-τ)*target. A common value of τ is 0.001 but often tuned for a specific application. This type of update is referred to as a ***soft update***.

[4] Volodymyr Mnih, et al.. "Human level control through deep reinforcement learning." *Nature*, 518(7540): 529–533, 2015.

[5] Hado van Hasselt, et al. "Deep Reinforcement Learning with Double Q-Learning." *Google DeepMind*, 13th Conference on Artificial Intelligence (AAAI-16), 2016.

Prioritized Experience Replay (PER)[6]

Google's DeepMind also improved on DQN by prioritizing which experiences are used to train the model. An experience is a $\langle \mathbf{s}_t, \mathbf{a}_t, \mathbf{r}_{t+1}, \mathbf{s}_{t+1} \rangle$ tuple that records the current state and action and the resulting reward and state. In a classical reinforcement learning, the network learns from the experience as it happens and then the experience is forgotten. Two major issues with this approach is the correlation between the experience impairs gradient-based algorithms used to train the network and possibly rare and valuable experiences are forgotten. To overcome these issues, a memory buffer is used to store these experiences, which can then be randomly sampled to break the temporal correlations and boost learning through repetition. Google's PER takes this a step further by extending the tuple to include a priority term that captures the value of the experience and then weighting the memory sampling by this priority value. The exact calculation and implementation of this priority value for this implementation will be discussed in a later section.

An excellent comparison of 7 DQN variants and their combination (Rainbow) on 57 Atari games is shown below.

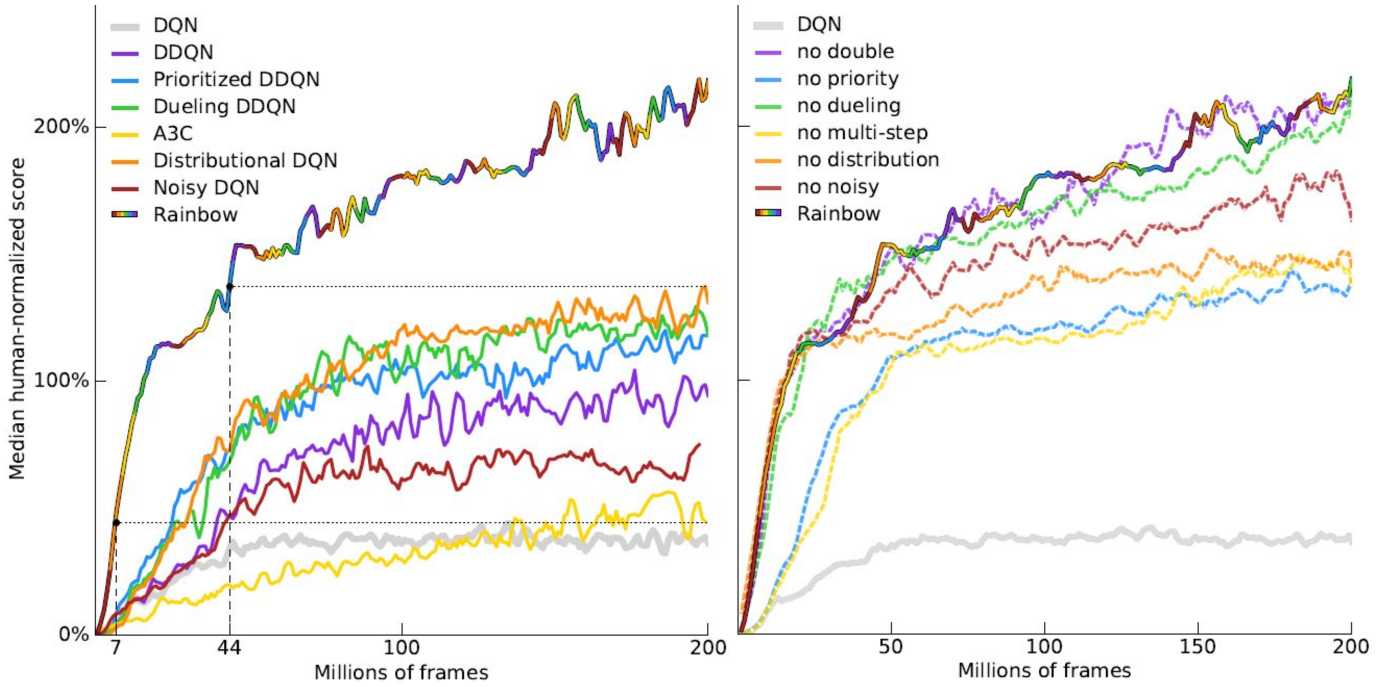


Figure 3. Google DeepMind's DQN variant comparison over 57 Atari games.[7]
Image source: https://arxiv.org/abs/1710.02298

[6] Tom Schaul, et al. "Prioritized Experience Replay." *Google DeepMind*, ICLR 2016.

[7] Matteo Hessel, at al.. "Rainbow: Combining Improvements in Deep Reinforcement Learning." *Google DeepMind*, Association for the Advancement of Artificial Intelligence (AAAI), 2018.

The image on the right on the previous page is especially telling. Removing the Prioritized Experience Replay (blue 'no priority') shows a considerable drop in performance, illustrating the power of PER. Note that this performance difference may not be replicated in all applications but it is still suggestive.

Policy-Based Methods

Up to this point we have been discussing value-based methods that depend on creating a network to estimate the **Q**-value of the state-action pairs and then select the optimal action to maximize the **Q**-Value. The following policy-based algorithms skip the estimation of the **Q**-Value and directly learn the optimal policy (actions). A value-based method also requires discrete actions to perform the max value operation. In our application the action space $\Delta \mathbf{w_t}$ is continuous. We could discretize the action space but policy-based methods give us the ability to directly predict the optimal $\Delta \mathbf{w_t}$ in a continuous space.

The policy-based methods often incorporate a gradient-ascent algorithm to find the network weights that maximize the objective. This ranges from the classic gradient-based REINFORCE[8] method to Trust Region Policy Optimization (TRPO)[9] and more advanced Proximal Policy Optimization (PPO)[10], which overcomes some of the complexity of implementing other policy-based methods.

Deterministic Policy Gradient (DPG)[11]

Policy gradient algorithms are often stochastic with a parametric probability policy distribution $\pi_\theta(\mathbf{a}|\mathbf{s}) = \mathbb{P}[\mathbf{a}|\mathbf{s};\theta]$, which provides the probability of selecting action ($\mathbf{a}$) for a given state ($\mathbf{s}$) according to the parameter vector $\boldsymbol{\theta}$. In a deterministic policy gradient algorithm we train a network directly determine $\mathbf{a} = \mu_\theta(\mathbf{s})$. Stochastic methods must integrate over both state and action spaces, whereas deterministic methods only integrate over the state space. This reduces the data requirement for training a deterministic method especially in an action space with high dimensionality.

[8] Ronald J. Williams. "Simple statistical gradient-following algorithms for connectionist reinforcement learning." *Reinforcement Learning*. Springer, Boston, MA, 1992.5-32.

[9] John Schulman, et al. "Trust Region Policy Optimization." *University of California*, 31st ICM L, 2015.

[10] John Schulman, at al. "Proximal Policy Optimization Algorithms." *OpenAI*, August 28, 2017.

[11] David Silver, et al. "Deterministic Policy Gradient Algorithms." *31st ICML*, Beijing, China, 2014.

Actor-Critic Methods[12]

Actor-Critic methods combine Value-based and Policy-based methods by creating two separate neural networks; the actor and critic. The actor is like a policy-based network, which selects which actions to take. The critic is similar to the value-based network and estimate the value of the actions chosen by the actor and is used as a .

Deep Deterministic Policy Gradient (DDPG)

The authors of DDPG classified it as a special form of Actor-Critic method.[13] However Miguel Morales argued in the Udacity DRL Nanodegree that DDPG is best classified as a DQN method for continuous action spaces with normalized advantage functions.[14] Regardless of the classification, it is a relatively simply architecture to implement and is very effective in a wide range of applications including portfolio management.[15,16] To prepare for the application to portfolio management, a DDPG architecture was created and successfully solved 3 unity ML-Agent environments to illustrate its flexibility and effectiveness.[17]

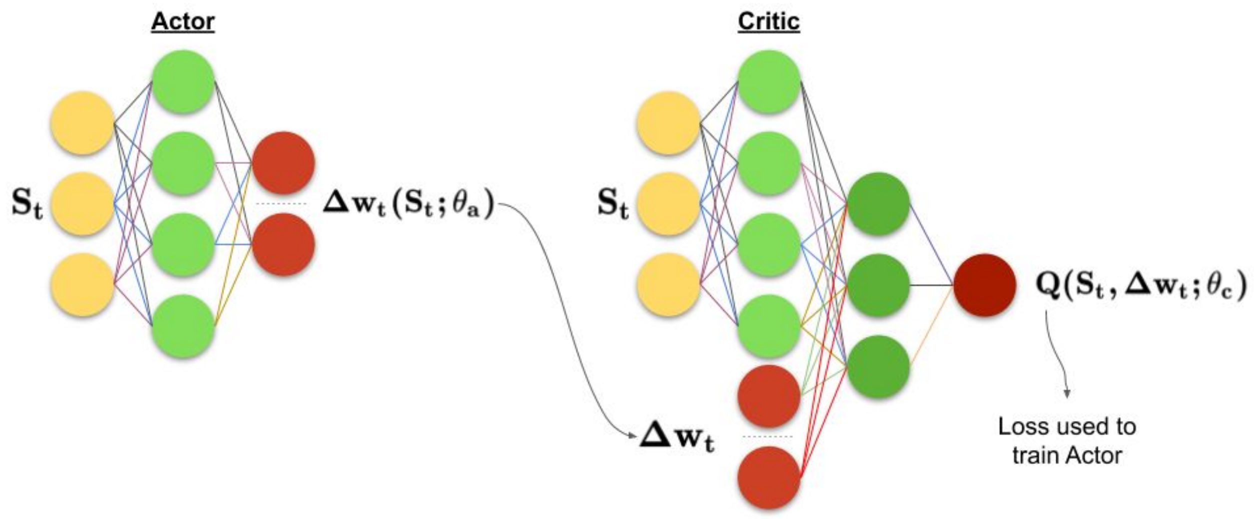


Figure 4. Generic DDPG network diagram.

The figure above illustrates the basic concept of the DDPG network as applied to our problem statement. The actor selects what action $\mathbf{\Delta w_t}$ to take based on the given state $\mathbf{S_t}$ and the learned actor network weights $\theta_\mathbf{a}$. The critic then determines the

[12] Vijay R. Konda and John N. Tsitsiklis. “Actor-Critic Algorithms.” *MIT*, Cambridge, 2000.
[13] Timothy Lillicrap, et al. “Continuous Control With Deep Reinforcement Learning.” *ICLR* 2016.
[14] Shixiang Gu, et al. “Continuous Deep Q-Learning with Model-based Acceleration.” *Google*, 2016.
[15] Pengqian Yu, et al. “Model-based Deep RL for Financial Portfolio Optimization.” *ICML*, 2019.
[16] Shashank Hegde, et al. "Risk aware portfolio construction using DDPG," *IEEE SSCI*, 2018
[17] Daniel Fudge. “DRL-Portfolio-Optimization.” *GitHub Project*, 2019.

$\mathbf{Q}$-value based on the state $\mathbf{S_t}$, the action selected be the actor $\Delta \mathbf{w_t}$ and the learned critic network weights $\theta_{\mathbf{c}}$. This $\mathbf{Q}$-value is then used to train the actor. Note the simple single layer fully connected network shown in the image does not reflect the actual network.

Ornstein-Uhlenbeck Noise Process[18]

All RL algorithms must balance the need to explore and exploit the action-state space to avoid getting caught in a local minimum while still learning the optimal policy as quickly as possible. A standard approach is to randomly perturb the predicted optimal action by applying noise to the selected actions. Another approach that appears very promising is to add the noise directly to the actor network weights $\theta_{\mathbf{a}}$.[19,20] A common noise process used in both cases is the Ornstein-Uhlenbeck (OU) noise process, which is a Gaussian process, a Markov process and is temporally homogeneous.[21]

Recurrent Neural Network (RNN)

Recurrent neural networks were developed to exploit temporal dependencies in sequence data such our price $p_{i,j}$ for asset $i \in [0, n]$ over the time embedding $j \in [t-k, t]$ as shown in the figure below.

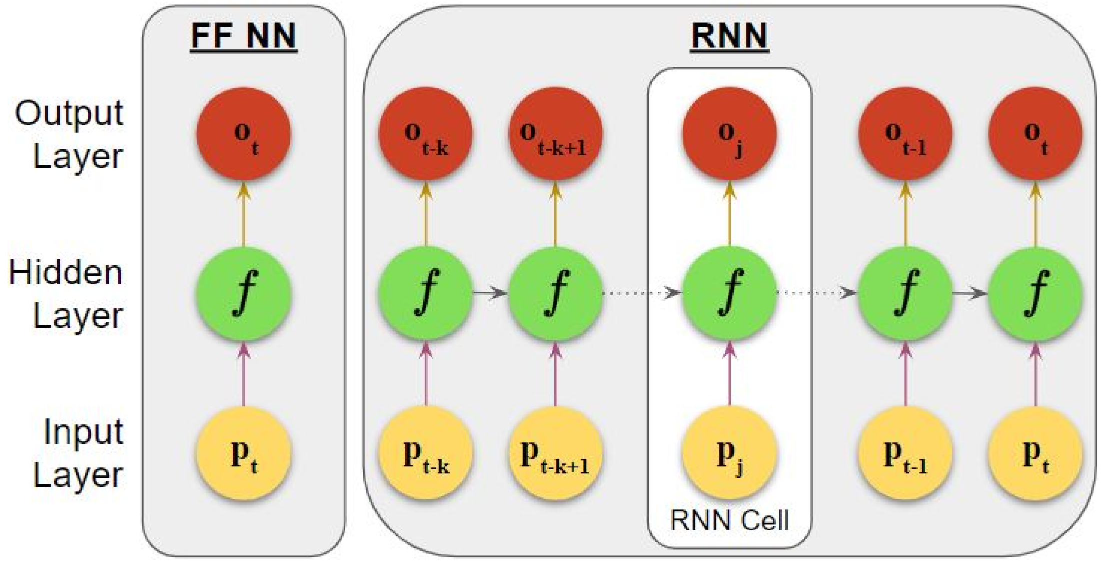


Figure 5. Feed-Forward (FF) vs. Recurrent Neural Network (RNN).

[18] George Uhlenbeck and Leonard Ornstein. “On the Theory of the Brownian Motion.” *Physical review*, 1930.
[19] Matthias Plappert,et al. “Parameter Space Noise for Exploration.” *OpenAI*, ICLR, 2016.
[20] Matthias Plappert,et al. “Parameter Space Noise for Exploration.” *OpenAI*, Blog, 2017.
[21] J.T. Doob. “The Brownian Movement and Stochastic Equations.” *Annals of Mathematics*, 1942.

The basic concept of a RNN is the output of the hidden layer is a function of not only the input layer, such as in a standard feed-forward network, but also of the hidden layer of the previous time step. The final cell may generate the desired output or the full RNN output may be passed to additional RNN layers to capture higher levels of complexity. Note for simplicity the above figure illustrates a 1D input (single asset). Much of the RNN literature also only addresses 1D input such as text or a single asset however this concept can be extended to multiple dimensions.[22]

Long Short-Term Memory (LSTM)[23]

A major short-coming of RNNs is the vanishing gradients experienced during the back-propagation through time while training the network. Effectively the impact of past events decay exponentially with time so important events are quickly forgotten. LSTM was introduced to overcome this issue and also add the ability to forget the past, which is important if an event occurs that makes previous information irrelevant.[24]

An excellent description of LSTM and RNN can be found in colah's blog, which I will condense and adapt for our application.[25] A LSTM cell which feeds information from one time step to the next like the RNN has three new features that generate memory that is both flexible and stable. The first is a cell state ($\mathbf{C_j}$) that represents the memory. The second is a sigmoid ($\boldsymbol{\sigma}$) function shown below, which being [0, 1], acts as a gate when multiplied to a signal. The last is a ***tanh*** function, which being [-1, 1], prevents a signal from exploding in either the positive or negative direction.

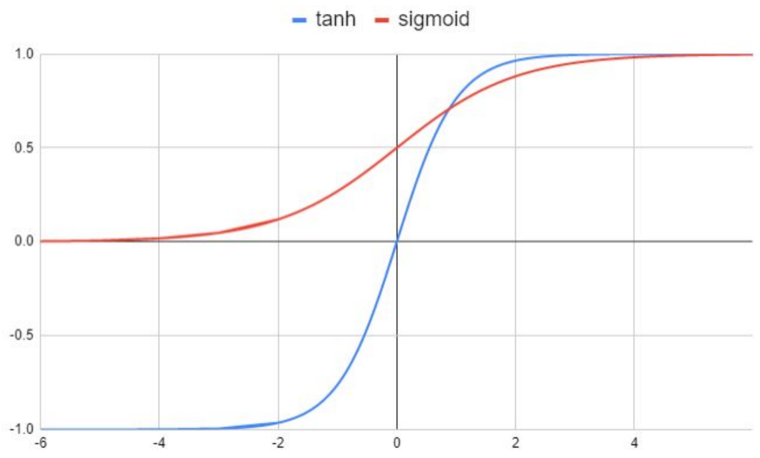


Figure 6. Sigmoid and tanh functions.

[22] Alex Graves, et al. "Multi-Dimensional Recurrent Neural Networks." *IDSIA*, 2013.
[23] Sepp Hochreiter and Jurgen Schmidhuber "Long short-term memory." *Neural computation*, 1997.
[24] Klaus Greff, et al. "LSTM: A Search Space Odyssey." *IEEE TNNLS*, Vol. 28, No. 10, pp. 2222-2232, 2017.
[25] Christopher Olah. "Understanding LSTM Networks." *colah's blog*, August 2015.

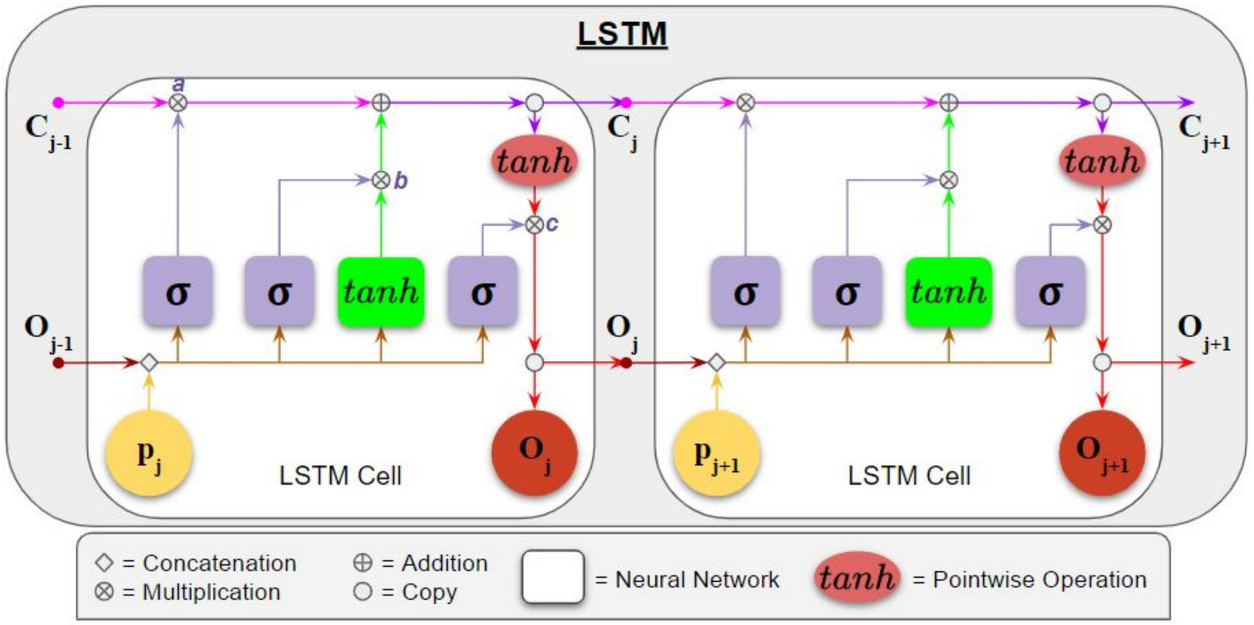


Figure 7. LSTM Network Architecture.

It is worth taking the time to study the above figure and compare it to the previous RNN figure. Both cells pass the output $\mathbf{O}_j$ from one time step to the next however the LSTM also passes a hidden cell state $\mathbf{C}_j$. In addition, if the sigmoid feeding into the "***a***" multiplication is zero, the cell state is "forgotten". If the sigmoid feeding into the "***b***" multiplication is zero, the price signal $\mathbf{p}_j$ is ignored and the previous cell state is passed to the next cell. And finally if the sigmoid feeding into the "***c***" multiplication is zero, the output is completely zeroed. Note the three sigmoids and the green ***tanh*** function all have weights and biases that need to be learned during training. This makes the LSTM more data intensive but also very powerful.

DDPG Neural Network Architecture

Now that all of the pieces have been discussed it is time to add them all together in the proposed DDPG network. Let's start by restating the state $\mathbf{S}_t$ as combination of the [3, n+1, k] price tensor $\mathbf{Y}_t$ and the [n+1] portfolio weighting $\mathbf{w}_t$. Recall that all n assets plus the S&P 500 index in $\mathbf{Y}_t$ has a low, high and close price for all k time embedding steps making a 3(n+1) x k tensor. In addition, $\mathbf{w}_t$ contains the portfolio weights of cash plus the n assets making a (n+1) vector. These are indicated by yellow in the figure below.

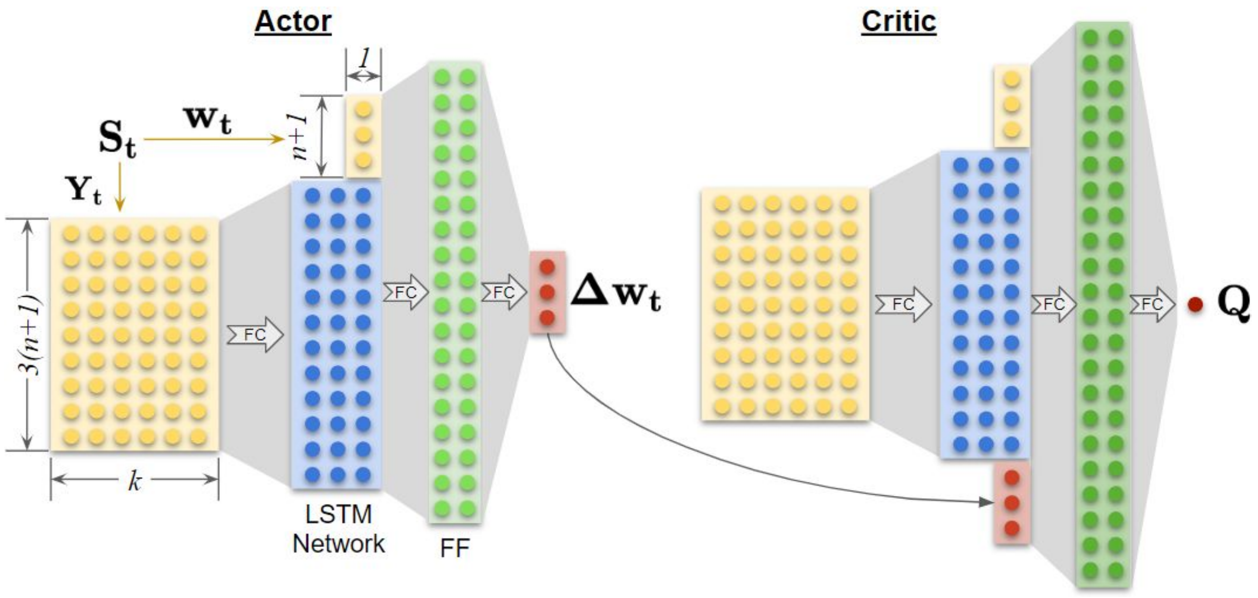


Figure 8. Proposed DDPG network diagram.

In the image above we see the price tensor $\mathbf{Y}_t$ is first passed to a LSTM network in both the actor and critic network. Note the depth and number of nodes in the LSTM network will be free parameters that we will need to experiment with in the next phase of this research. In the actor network, the output layer of the LSTM network is concatenated with the weight vector $\mathbf{w}_t$ and then passed to a standard feed-forward (FF) neural network with the final layer representing the predicted optimal action $\Delta\mathbf{w}_t$.

The critic network has a similar structure except the optimal action $\Delta\mathbf{w}_t$ generated by the actor is also concatenated with the LSTM output and the weight vector $\mathbf{w}_t$ before entering the FF network, which now generates the **Q**-value of the state-action pair. Note the "FC" arrow in the diagram represents "fully-connected", which indicates that each node on the left is connected to every node on the right.

The actor and critic networks are trained independently and will have different weights and biases. For simplicity and considering the actor and critic LSTM networks see the same $\mathbf{Y_t}$ tensor, they will have the same structure. However the two FF networks see different input and generate an output layer with a different number of nodes, therefore may have different structures. Testing will have to determine an optimal number of layers and number of nodes on each layer.

Both the actor and critic have both a local and target network giving a total of four networks. The local and target networks have the same structure as discussed in the Double DQN section and will implement a soft update from the local to the target network.

Prioritized replay will be implemented in the architecture. It is believed that this will be very effective in training the network since particular market events may have a significant impact on the portfolio value; for both the good and bad. Therefore it is important that these events are emphasized in the training.

Finally the network will incorporate an Ornstein-Uhlenbeck noise process directly on the network weighting as discussed in the OpenAI study.[26] This is a departure from the methodology covered in the Udacity Nanodegree but it appears to be a convenient way to boost performance.

[26] Matthias Plappert,et al. "Parameter Space Noise for Exploration." *OpenAI*, ICLR, 2016.

Training and Evaluation Process

We will take twelve years of daily price data and dividing it into ten years of training and two years of evaluation. The length of these periods are relatively arbitrary and can be easily modified. However, periods need to be long enough to smooth out random motion and provide sufficient data for training. Ten years provide approximately 2530-k days of training. This is not a lot of data considering the complexity of the network illustrated above.

Data Preparation

Prior to training we must condition the price data. The first filter will be to only consider assets with a full history over the desired period. We must also alter the high and low price data to compensate for splits, dividends and distributions.[27] Dividends are an interesting issue. A more realistic option would be to keep the dividend drops in the price data but also include the dividends in the reward function. However, the simpler price adjustment method will be used for this implementation.

Training

With clean data, we can begin training and pass through the training period, called an episode, many times. An episode continues from the beginning to the end unless the portfolio reaches zero, in which case a large negative reward is returned. Each of these episodes is analogous to a game if we were training the algorithm to play chess. The network retains both the learned weights and biases from episode to episode allows it to become "smarter" as it experiences more episodes. The memories defined by the $\langle \mathbf{s}_t, \mathbf{a}_t, \mathbf{r}_{t+1}, \mathbf{s}_{t+1}, \mathbf{v}_t \rangle$ tuples are also cumulative across all episodes, where $\mathbf{v}_t$ in the tuple is the priority of the memory. This allows us to accelerate learning by periodically resampling the memory to remember important events as the learning progresses.

Since each state includes the previous k-1 prices, each episode must begin after k-1 time steps. The network will also begin without any learning so the weights and bias will be initialized with the Xavier Initialization.[28] This initialization prevents the problem of vanishing or exploding gradients during the backpropagation in the training operation.

One of the major assumptions in this training process is that our actions do not impact the market. This allows us to reuse the asset prices across all episodes regardless of our actions. Including market impacts will be left for future work. However, if the network performs well in the remaining two-year evaluation period and

[27] StockCharts.com. "Historical Price Data is Adjusted for Splits, Dividends and Distributions."
[28] Kian Katanforoosh and Daniel Kunin, "Initializing Neural Networks", *deeplearning.ai*, 2019.

someone was willing to risk real money, the training could continue with real actions. This would incorporate real market movements and transaction costs into the network.

Evaluation

With a trained network, we will then apply it to the final two years and compare its performance with respect to the overall return and Sharp ratio to the iShares Core S&P 500 ETF (IVV) including its 0.04% management fees. Note that during training the network never sees the two-year price data however it can still learn as the two years progress. If the evaluation is repeated, the network is reset to the original trained state before it saw the evaluation data.

Implementation Environment

Now that the architecture and process are defined we need to evaluate which implementation environment to select. The first part of that question is the software development environment. We have selected Python as the based software language due to its extreme popularity, ease of use and extensive machine learning packages.[29,30] From the many excellent Interactive Development Environments (IDE) available for Python we have selected PyCharm, which includes plug-ins for both GitHub and Amazon Web Service (AWS).[31]

The Python distribution we are using is Anaconda Python 3.7.[32] This is an industry standard that allows user to tailor their environment with the Conda, an open source package management system.[33]

We are using the Git version control system and the GitHub code hosting platform. This not only provides a means of securing and controlling the code versions but also provides a collaboration platform to share this report and code and gather feedback from the community. GitHub also allows anyone to download (clone) the repository to any system including a local PC or AWS to test the code and provide updates.

[29] Nick Heath. "GitHub: The top 10 programming languages for machine learning." *TechRepublic*, Jan 2019.
[30] Mahesh Chand. "Best Programming Language for Machine Learning." C# Corner, Feb 2019.
[31] Jet Brains. "PyCharm: The Python IDE for Professional Developers." accessed Oct 26, 2019. https://www.jetbrains.com/pycharm/
[32] Anaconda. "Anaconda Distribution: The World's Most Popular Python/R Data Science Platform." accessed Oct 26, 2019. https://www.anaconda.com/distribution/
[33] Anaconda. "Conda: Package, dependency and environment management for any language." accessed Oct 26, 2019. https://conda.io/

PyTorch

With a general software development environment established we need to select the Machine Learning framework. The two main contenders currently are TensorFlow released by Google in 2015 and PyTorch by Facebook in 2017.[34,35] This is a very rapidly evolving space and both packages are very capable and easily deployed with Python. Generally TensorFlow's TensorBoard visualization and production ready serving are advantages over PyTorch. PyTorch's advantages are a more pythonic style and dynamics graphing that make it easier to debug and develop.[36] For these reasons, we select PyTorch for development but the general strategy could be deployed in either framework.

Both PyTorch and TensorFlow also have built-in Graphics Processing Unit (GPU) support. GPUs make large deep learning networks practical to train.[37,38] When training a neural network there is first a forward pass that takes the input and propagates the values through the network to generate the output. The error is then calculated and back propagated through the network by calculating the gradient of the error with respect to each node weight, which is effectively a chain rule application. As the network becomes larger, these become extremely large matrix multiplication operations. These highly parallel matrix (tensor) operations are perfectly suited for GPUs and to automatically switch between CPU and GPU execution all we need to add is the following line in our code.

```
device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")
```

For a full PyTorch implementation, please review the DDPG agent module in the tennis GitHub repository.[39] In this code the "`import torch`" line is importing the PyTorch package into the Python code.

[34] TensorFlow. "An end-to-end open source machine learning platform." accessed Oct 26, 2019. https://www.tensorflow.org/

[35] PyTorch. "From Research To Production." accessed Oct 26, 2019. https://pytorch.org/

[36] Vihar Kurama. "Pytorch Vs. Tensorflow: Which Framework Is Best For Your Deep Learning Project?" *Builtin*, Sept 2019.

[37] Bernard Fraenkel. "For Machine Learning, It's All About GPUs." *Forbes*, Dec 2017.

[38] Shachi Shah. "Do we really need GPU for Deep Learning? - CPU vs GPU." *Medium.com*, Mar 2018.

[39] Daniel Fudge. "ddpg.py: DRL applied to Tennis Environment." accessed Oct 26, 2019. https://github.com/daniel-fudge/reinforcement-learning-tennis/blob/master/tennis/ddpg_agent.py

Amazon Web Service (AWS)

The second major component of the environment is the hardware. As discussed above, deploying this algorithm on GPUs is a major speed boost that makes testing and debugging much faster. Aside from buying GPU hardware, there are many cloud computing environments available that provide GPU support. The environment we have selected is AWS. It's Deep Learning Amazon Machine Images (DLAMIs) with Conda allows us to rapidly launch the latest GPU enabled hardware with PyCharm and all necessary software pre-installed.[40] This greatly simplifies the setup and management of the environment and allows new users to rapidly access the exact same environment that has been proven to work.

During the Udacity DRL Nanodegree the p3.2xlarge Ubuntu 16.04 DLAMI with one GPU was used to train and evaluate the tennis DDPG network.[41,42] Due to the limited training data size and small network, the switch from CPU to GPU wasn't significant, however we expect this to be more important for the much larger network we are training.

[40] AWS. "AWS Deep Learning AMIs." accessed Oct 26, 2019. https://aws.amazon.com/machine-learning/amis/

[41] AWS Marketplace. "Deep Learning AMI (Ubuntu 16.04)." accessed Oct 26, 2019. https://aws.amazon.com/marketplace/pp/B077GCH38C

[42] Daniel Fudge. "DRL applied to Tennis Environment." accessed Oct 26, 2019. https://github.com/daniel-fudge/reinforcement-learning-tennis

Future Research Prior to Implementation

AWS Development Environment

Testing in AWS environment included local laptop development then transfer to a DLAMI that cost $3.06/hr for on-demand usage. Another option is the AWS Cloud9 development environment.[43] Developing on AWS allows live debugging with GPUs and presents a fixed development environment for multiple developers. However, we do not want to be charged $3.06/hr for development that does not include training. One option is to split the development and training environments. For instance if you spend 90 hours per month developing, the monthly charge for a t2.micro instance with 10 GB of storage would be $2.05, which is 40 minutes on the DLAMI.[44] However, you would not want to train in this instance, so you need to dispatch training to the DLAMI. To reduce cost upto 90%, spot pricing may be used for the DLAMI.[45] This basic idea of spot pricing is you submit a request for a resource at a given price. Your job will not start until the "Spot Price", which is set by AWS based on supply and demand for that specific resource, drops below your requested price. If the price rises above your price while your are running, the instance will be terminated. This is acceptable in our workflow as long as we periodically save our network weights. Training can restart from the saved weights once the price drops again.

A different approach is using the AWS SageMaker.[46] However it isn't clear if this would provide much benefit for custom PyTorch implementations designed for research instead of production workflows. The Pluralsight "Using PyTorch in the Cloud: PyTorch Playbook" course includes a section devoted to deploying PyTorch deep learning models on AWS SageMaker.[47] The "AWS Certified Machine Learning - Specialty 2019" by A Cloud Guru also prepares covers the broader implementation of SageMaker.[48]

Understanding the optimal development environment within AWS is future research that must be performed before implementation begins. The documentation of this development environment would be captured in the next term report and will lower the barrier for new researchers.

[43] AWS, "AWS Cloud9." accessed Oct 26, 2019. https://aws.amazon.com/cloud9/
[44] AWS, "Cloud9 Pricing." accessed Oct 26, 2019. https://aws.amazon.com/cloud9/pricing/
[45] AWS, "Amazon EC2 Spot Instances." accessed Oct 26, 2019. https://aws.amazon.com/ec2/spot/
[46] AWS, "Reinforcement Learning with Amazon SageMaker RL." accessed Oct 26, 2019. https://docs.aws.amazon.com/sagemaker/latest/dg/reinforcement-learning.html
[47] Pluralsight. "Using PyTorch in the Cloud: PyTorch Playbook." accessed Oct 26, 2019. https://www.pluralsight.com/courses/using-pytorch-in-the-cloud-playbook
[48] A Cloud Guru. "AWS Certified Machine Learning - Specialty 2019." accessed Oct 26, 2019. https://acloud.guru/learn/aws-certified-machine-learning-specialty

LSTM Network Architectures

The DDPG networks developed for this research did not include a LSTM network. The theory and basic implementation of a LSTM network are described in the preceding sections, however a true understanding of how to size and connect a LSTM network can only be obtained through implementing such a network. The Udacity "Intro to Deep Learning with PyTorch" course includes building a LSTM network in PyTorch and could provide the necessary experience.[49] This course could also be extended to the Udacity "Deep Learning" Nanodegree.[50] Coursera also has a "Advanced Machine Learning with TensorFlow on Google Cloud Platform Specialization" that appears very useful but it uses TensorFlow instead of PyTorch and Google Cloud instead of AWS as is used in Udacity.

Future Work Beyond Implementation

Increasing State Definition

The state definition we have included here is purely dependent on the prices of the assets in the portfolio and a market index. Clearly this could be expanded to incorporate a multitude of signals. This could include macroeconomic signals such as Gross Domestic Product (GPD), inflation rates, interest rates and unemployment rates. It could also include firm specific financial indicators such as liquidity, efficiency, profitability and leverage ratios. More creative signals could be added such as sentiment analysis of twitter feeds or news reports or image processing of satellite imagery. Selecting which signals to include and how to pre-process this data is a critical area of research.

The Udacity Artificial Intelligence (AI) for Trading Nanodegree may be relevant to this research.[51] It focuses on using AI to generate trading signals from a wide variety of sources. These signals could be added to the state definition for the DRL algorithm.

Increasing Transaction Cost Accuracy

Our representation of the transaction cost is simply a constant times the change in portfolio weighting. Clearly this could be replaced by a more accurate cost model.

---

[49] Udacity. "Free Course: Intro to Deep Learning with PyTorch." accessed Oct 26, 2019. https://www.udacity.com/course/deep-learning-pytorch--ud188.

[50] Udacity. "Nanodegree Program: Deep Learning." accessed on Oct 26, 2019. https://www.udacity.com/course/deep-learning-nanodegree--nd101

[51] Udacity. "Nanodegree Program: Artificial Intelligence for Trading." accessed Oct 26, 2019. https://www.udacity.com/course/ai-for-trading--nd880

Including Market Impact of Trades

Our implementation is completely ignoring how our trades would impact the asset prices. This is a reasonable assumption as long as our trades are a small fraction of the market. However, adding these market impacts would be a reasonable upgrade to the algorithm.

Implementing more Advanced DRL Architectures

The field of DRL is constantly evolving and future work should compare our DDPG algorithm to many of the competing algorithms such as OpenAI's Proximal Policy Optimization (PPO).[52] This effort should be expanded as more algorithms are developed and this framework could also be used to experiment with new algorithms.

In our architecture we employ a LSTM to capture the time history of the asset prices, however this is also a very active field that needs to be explored. Many of the methods try to incorporate the concepts of Convolutional Neural Networks (CNN), which are most often deployed in image recognition.[53,54] In this context they are called Temporal Convolutional Networks (TCN). Another extension to CNNs adds dilation, where the output of a CNN is fed into a subsequent CNN that samples at a lower frequency as shown below with a fixed stride and dilation factor of one.[55] This should be much more memory efficient with minimal loss of information and can also be applied to LSTM.[56]

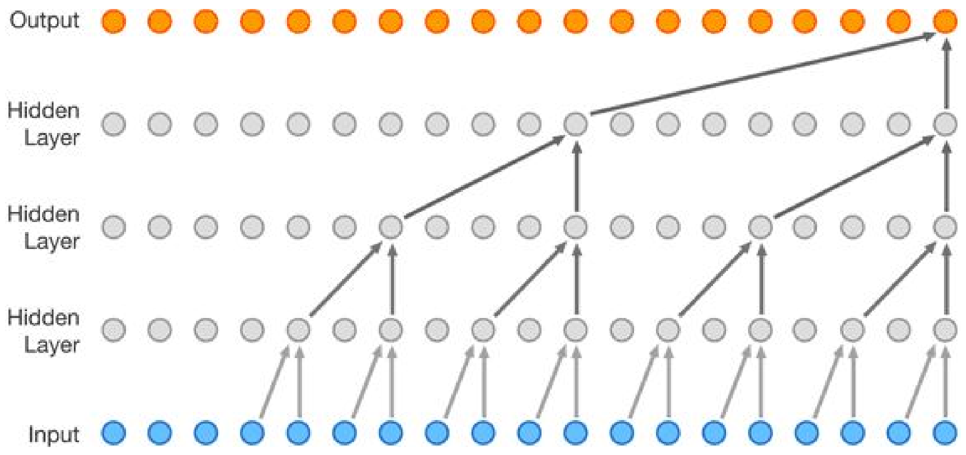


Figure 9. Dilated convolutional layers.
Image Source: https://i.stack.imgur.com/OCyiq.gif

[52] John Schulman et al.. "Proximal Policy Optimization Algorithms." OpenAI, August 28, 2017.

[53] Taewook Kim et al.. "Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data." *PLoS One*, Feb 2019.

[54] Shaojie Bai et al.. "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling." *ArXiv*, April, 2018.

[55] Fisher Yu and Vladlen Koltun. "Multi-scale Context Aggregation By Dilated Convolutions." ICLR, 2016.

[56] Shiyu Chang et al.. "Dilated Recurrent Neural Networks." *NIPS 2017*, April 19, 2018.

Conclusion

Bibliography

Bai, Shaojie, J. Zico Kolter and Vladlen Koltun. “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling.” *ArXiv*, April 19, 2018. https://arxiv.org/abs/1803.01271

Chand, Mahesh. “Best Programming Language for Machine Learning.” *C# Corner*, February 2, 2019. https://www.c-sharpcorner.com/article/best-programming-language-for-machine-learning/

Chang, Shiyu, Yang Zhang, Wei Han, Mo Yu, Xiaoxiao Guo, Wei Tan, Xiaodong Cui, Michael Witbrock, Mark Hasegawa-Johnson, Thomas S. Huang. “Dilated Recurrent Neural Networks.” *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, April 19, 2018. https://arxiv.org/abs/1803.01271

Doob, J.T.. “The Brownian Movement and Stochastic Equations.” *Annals of Mathematics*, Second Series, Vol. 43, No. 2, pp. 351-369, April 1942. https://www.jstor.org/stable/1968873

Fraenkel, Bernard. “For Machine Learning, It's All About GPUs.” *Forbes*, December 1, 2017. https://www.forbes.com/sites/forbestechcouncil/2017/12/01/for-machine-learning-its-all-about-gpus/#45e138417699

Fudge, Daniel. “DRL-Portfolio-Optimization.” *GitHub Project*. https://github.com/daniel-fudge/DRL-Portfolio-Optimization

Fudge, Daniel and Yelena Larkin. “Application of Machine Learning to Portfolio Optimization.” Independent Study, *Schulich School of Business*, April 2019. https://github.com/daniel-fudge/DRL-Portfolio-Optimization/blob/master/docs/report1.pdf

Graves, Alex, Santiago Fernandez and Jurgen Schmidhuber. “Multi-Dimensional Recurrent Neural Networks.” *IDSIA*, 2013. https://arxiv.org/abs/0705.2011

Greff, Klaus, Rupesh K. Srivastava, Jan Koutnık, Bas R. Steunebrink and Jurgen Schmidhuber. “LSTM: A Search Space Odyssey.” *IEEE Transactions on Neural Networks and Learning Systems,* Vol. 28, No. 10, pp. 2222-2232, October 2017. https://arxiv.org/abs/1503.04069

Gu, Shixiang, Timothy Lillicrap, Ilya Sutskever and Sergey Levine. “Continuous Deep Q-Learning with Model-based Acceleration.” *Google Brain and Google DeepMind*, March 2016. https://arxiv.org/abs/1603.00748

Halperin, Igor. “QLBS: Q-Learner in the Black-Scholes(-Merton) Worlds.” *SSRN Electronic Journal*, (December 2017) DOI: 10.2139/ssrn.3087076. https://arxiv.org/abs/1712.04609

Hasselt, Hado van, Arthur Guez and David Silver. “Deep Reinforcement Learning with Double Q-Learning.” *Google DeepMind*, 13th Conference on Artificial Intelligence (AAAI-16), 2016. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewPaper/12389

Heath, Nick. “GitHub: The top 10 programming languages for machine learning.” *TechRepublic*, January 25, 2019. https://www.techrepublic.com/article/github-the-top-10-programming-languages-for-machine-learning/

Hegde, Shashank, Vishal Kumar and Atul Singh, "Risk aware portfolio construction using deep deterministic policy gradients," *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018, pp. 1861-1867. doi: 10.1109/SSCI.2018.8628791. https://ieeexplore.ieee.org/document/8628791

Hessel, Matteo, Will Dabney, Joseph Modayil, Dan Horgan, Hado van Hasselt, Bilal Piot, Tom Schaul, Mohammad Azar, Georg Ostrovski, David Silver. “Rainbow: Combining Improvements in Deep Reinforcement Learning.” *Google DeepMind*, Association for the Advancement of Artificial Intelligence (AAAI), 2018. https://arxiv.org/abs/1710.02298

Hochreiter, Sepp and Jurgen Schmidhuber “Long short-term memory.” *Neural computation*, 9(8):1735–1780, 1997. https://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735

Hui, Jonathan. “RL-Value Learning.” *Medium*, October 11, 2018. https://medium.com/@jonathan_hui/rl-value-learning-24f52b49c36d

Jiang, Zhengyao, Dixing Xu and Jinjun Liang. “A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem.” *Journal of Machine Learning Research (JMLR)*, July 2017. https://arxiv.org/abs/1706.10059

Katanforoosh, Kian and Daniel Kunin, "Initializing Neural Networks", *deeplearning.ai*, 2019. https://www.deeplearning.ai/ai-notes/initialization/

Kim, Taewook, Ha Young Kim and Raul Harnandez Montoya. “Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data.” *PLoS One*, February 15, 2019. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6377125/

Konda, Vijay R. and John N. Tsitsiklis. “Actor-Critic Algorithms.” *Massachusetts Institute of Technology*, Cambridge, 2000. http://papers.nips.cc/paper/1786-actor-critic-algorithms.pdf

Kurama, Vihar. “Pytorch Vs. Tensorflow: Which Framework Is Best For Your Deep Learning Project?” *Builtin*, September 22, 2019. https://builtin.com/data-science/pytorch-vs-tensorflow

Liang, Zhipeng, Hao Chen, Junhao Zhu, Kangkang Jiang, Yanran Li. “Adversarial Deep Reinforcement Learning in Portfolio Management.” *Sun Yat-sen University*, November 2018. https://arxiv.org/abs/1808.09940

Lillicrap, Timothy, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver and Daan Wierstra. “Continuous Control With Deep Reinforcement Learning.” *Proceedings from the International Conference on Learning representations (ICLR)*, 2016. https://arxiv.org/pdf/1509.02971.pdf

Olah, Christopher. “Understanding LSTM Networks.” *colah's blog*, August 2015. http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Peters, Jan, Sethu Vijayakumar and Stefan Schaal. “Natural Actor Critic.” *Neurocomputing*, Volume 71, Issues 7-9, March 200 https://doi.org/10.1016/j.neucom.2007.11.026

Plappert, Matthias, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y. Chen, Xi Chen, Tamim Asfour, Pieter Abbeel and Marcin Andrychowicz. “Parameter Space Noise for Exploration.” *OpenAI*, Proceedings from the International Conference on Learning representations (ICLR), 2016. https://arxiv.org/abs/1706.01905

Plappert, Matthias, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y. Chen, Xi Chen, Tamim Asfour, Pieter Abbeel and Marcin Andrychowicz. “Parameter Space Noise for Exploration.” *OpenAI (Blog)*, 2017. https://openai.com/blog/better-exploration-with-parameter-noise/

Santos, Leonardo Araujo dos. “Artificial Intelligence.” GitBook, accessed October 6, 2019. https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/

Schaul, Tom, John Quan, Ioannis Antonoglou and David Silver. “Prioritized Experience Replay.” *Google DeepMind*, International Conference on Learning Representations (ICLR) 2016. https://arxiv.org/abs/1511.05952

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford and Oleg Klimov. “Proximal Policy Optimization Algorithms.” *OpenAI*, August 28, 2017. https://arxiv.org/abs/1707.06347

Schulman, John, Sergey Levine, Phillipp Moritz, Michael Jordon and Pieter Abbeel. "Trust Region Policy Optimization." *University of California*, Berkeley, 31st International Conference on Machine Learning, 2015. https://arxiv.org/abs/1502.05477

Shah, Shachi. "Do we really need GPU for Deep Learning? - CPU vs GPU." *Medium.com*, March 26, 2018. https://medium.com/@shachishah.ce/do-we-really-need-gpu-for-deep-learning-47042c02efe2

Silver, David, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra and Martin Riedmiller. "Deterministic Policy Gradient Algorithms." *Proceedings of the 31 st International Conference on Machine Learning*, Beijing, China, 2014. http://www.jmlr.org/proceedings/papers/v32/silver14.pdf

StockCharts.com. "Historical Price Data is Adjusted for Splits, Dividends and Distributions." Accessed Oct 25, 2019. https://support.stockcharts.com/doku.php?id=policies:adjusted_data

Sutton, Richard S. and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd edition, Cambridge, Massachusetts: The MIT Press, A Bradford Book, (November 5 2018). http://incompleteideas.net/book/RLbook2018.pdf

Udacity. "Deep Reinforcement Learning." Nanodegree program. https://www.udacity.com/course/deep-reinforcement-learning-nanodegree--nd893

Uhlenbeck, George E. and Leonard S. Ornstein. "On the Theory of the Brownian Motion." *Physical review*, 36(5): 823, 1930. https://doi-org.ezproxy.library.yorku.ca/10.1103/PhysRev.36.823

Wawrzynski, Pawel. "Real-time reinforcement learning by sequential Actor-Critics and experience replay." *Neural networks*: the official journal of the International Neural Network Society. 22. 1484-97 (June 2019). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.904.8961&rep=rep1&type=pdf

Williams, Ronald J. "Simple statistical gradient-following algorithms for connectionist reinforcement learning." Reinforcement Learning. Springer, Boston, MA, 1992.5-32.

Yang, Qing, Tingting Ye and Liangliang Zhang. "A General Framework of Optimal Investment." *SSRN* (February 2, 2019). https://ssrn.com/abstract=3136708

Yu, Fisher and Vladlen Koltun. "Multi-scale Context Aggregation By Dilated Convolutions." International Conference on Learning Representations (ICLR), April 30, 2016. https://arxiv.org/abs/1511.07122

Yu, Pengqian, Joon Sern Lee, Ilya Kulyatin, Zekun Shi and Sakyasingha Dasgupta. "Model-based Deep Reinforcement Learning for Financial Portfolio Optimization." *arXiv:1901.08740* (January 2019). https://arxiv.org/abs/1901.08740

Appendix A - Related Training

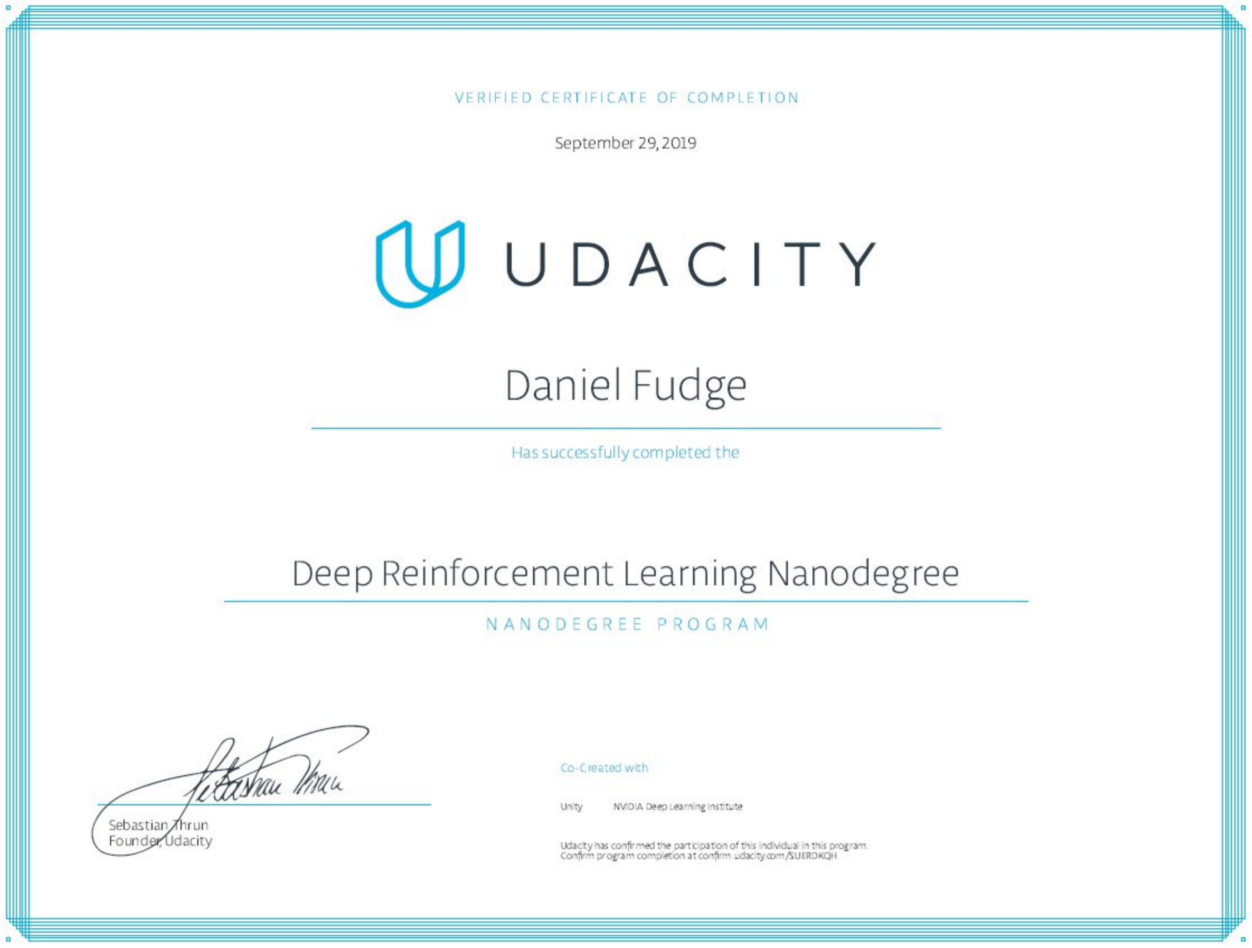
A.1 Udacity Nanodegree: Deep Reinforcement Learning (DRL)

https://www.udacity.com/course/deep-reinforcement-learning-nanodegree--nd893

This Nanodegree involved developing DRL networks to solve 3 different Unity ML-Agents environments (https://unity3d.com/machine-learning/).

The solutions to these environment can be found in separate GitHub repositories:

1. Banana Collector: https://github.com/daniel-fudge/banana_hunter
2. Reacher: https://github.com/daniel-fudge/reinforcement-learning-reacher
3. Tennis: https://github.com/daniel-fudge/reinforcement-learning-tennis



VERIFIED CERTIFICATE OF COMPLETION

September 29, 2019

UDACITY

Daniel Fudge

Has successfully completed the

Deep Reinforcement Learning Nanodegree

NANODEGREE PROGRAM

[illegible]

Sebastian Thrun
Founder, Udacity

Co-Created with

Unity NVIDIA Deep Learning Institute

Udacity has confirmed the participation of this individual in this program.
Confirm program completion at confirm.udacity.com/SUERDKQH