Exploration of CGAN Pix2Pix Model with Image Data for S&P Stock Price Prediction

By Cindy Su

Author Bio

Cindy is a high school junior at Phillips Exeter Academy in Exeter, New Hampshire. She is interested in education technology, business, and the usage of artificial intelligence in our society. She is a self-starter and entrepreneur, her most recent project being the co-founder and Curriculum Director for the DreamAI, an organization intended to introduce AI and its various applications to middle schoolers. She is a constant learner of coding and computer science, picking up Java on her own through Khan Academy and studying Python through various elective classes at her school and summer programs such as AI4ALL. Currently, she is working on Speciify, a business and app aiming to help special educators better organize their classrooms and manage their students.

Abstract

With the incredibly volatile stock market in 2020, many professional investors won big, but almost all average citizen investors lacked the resources to participate, such as investment knowledge, time, attention, and buying capital,, leaving them at a great disadvantage. There have been AI and prediction algorithms that have been developed that are making “smart” investing available to more people, but the algorithms are still difficult to understand as there is a code language barrier. The algorithms are also difficult to trust as they are opaque in how they process strictly numerical data into predictions. Thus, this experiment tested to see if an image data-based algorithm like the Pix2Pix Conditional Generative Adversarial Network (CGAN) would be able to predict the next week’s stock prices. By focusing on image data, it promotes more understanding of the algorithm as common investors can more easily see patterns through visual representations of stock prices. It also explores the Gramian Angular Summation/Difference Fields (GAF) image representation method that could encode more pertinent relationships between data points than a direct candlestick image. Experiments ended up showing that the model’s prediction GAF image for the next week’s prices had a 74% similarity rate with the GAF image of the next week’s actual prices, suggesting that the image-based model and GAFs could be considered for future stock prediction.

Keywords: Generative Adversarial Networks, Stock Market, Time-Series, Optimization, Prediction, Artificial Intelligence, Gramian Angular Fields, Image Data, CGAN, Candlestick Graphs

Introduction

Millions of Americans with money staked in the stock market lose precious savings in the stock market every day. In fact, more than 50% of all US families own publicly traded stock in some form, often drawn in by promises of money and fortune (1). The COVID-19 pandemic has only increased public interest in the stock market, especially with Congress representatives making more than $290 million through trading in 2020 (2). However, most of the time, investments in the stock market made by common citizens lose money, with over $415 billion lost in 2020 as they do not have the same access to resources and knowledge as professional investors (3). Thus, this research aims to address how technology may be able to help more citizens win in the stock market, the computer algorithm acting as a stock market "advisor" by predicting the next week's stock prices. This isn't a new concept, as there are already existing AI models that operate by predicting future events through the analysis of historical data and trends. The most common method used to predict stock price change is based on LSTM (Long Short-Term Memory), a deep learning framework for time-series, to build a predictive model (4, 5). This method uses numerical stock values (open, close, and high, low, and volume) directly to train the models (6). Although this method has seen success so far, it is impossible for a human to look at the same numerical data and process the information immediately and draw useful conclusions (7). If the algorithm used visual data, particularly starting with graphs that humans can easily read, it could increase the trust and transparency of these predictive model to the average consumer. Therefore, we also wanted to focus on trying to use image data in our stock prediction algorithm so it would be easier to understand from an untrained eye. We turned towards Generative Adversarial Networks to create an understandable and accurate model.

Generative Adversarial Networks, or GANs, are a class of machine learning frameworks designed by Ian Goodfellow in June 2014 (8). The adversarial process involves using two neural networks that compete against each other in a zero-sum game; the generator tries to create similar images from an existing data set that will confuse the other network, the discriminator, into classifying the generated images as real data. Since the "indirect" training happens through the discriminator, it also updates itself and the generator constantly, allowing the model to learn in an unsupervised manner. This powerful architecture has researched in the past for stock prediction, using networks like Multi-Layer Perception (MLP) and LSTM, and showed promising results that show the potential of image-based prediction (9). Specifically, the Pix2Pix CGAN, typically used for image-to-image translation, works off of the previous data entered. Thus, the conditional aspect of the network ensures that the newly generated image is a time-series prediction, not just a random candlestick graph, hypothetically making the predicted images more accurate. The Pix2Pix CGAN is unique in the way that it does not use the popular LSTM algorithm and instead uses a generator with a U-Net-based architecture and a discriminator represented by a convolutional PatchCGAN classifier (10). U-Net is an architecture for semantic segmentation. PatchCGAN is a type of discriminator that runs convolutionally across the image, trying to classify if each patch in an image is real or fake, and then averaging all responses to provide the ultimate decision

We hypothesized that image data and the Pix2Pix CGAN model would be able to predict stock prices, and conducted two main experiments, which will be referenced throughout this paper. Both experiments were meant as an iterative experimental process to further discover the potential and use of CGANs in time series prediction. The main difference between the two experiments is the type of data that was used. The same stock market index was used, the Standard and Poor's 500, otherwise known as SPY, because it gives a general sense of how the market is doing and the specific stock is not particularly important as we are testing the overall effectiveness of the prediction method. For the first experiment, I used images of candlestick graphs from 1993 to 2003, but in the second experiment, we used more recent data spanning from 2009 to 2019 and eliminated the graph formatting distractions by directly encoding the numbers into GAF images. Candlestick charts cleanly display the open, high, low, and close of the day, and are easy to interpret at a glance, making them our first choice as image data to train the CGAN on. However, because the stock market's prices have such a large range over each 10-year period, the candlestick charts automatically scale themselves to display each week's candles most optimally (Fig. 1). This raises the problem of standardizing each image's scale, which is incredibly difficult to do as many of the smaller candlesticks become invisible when the scale becomes too big. After reading about Gramian Angular Fields (GAFs), which represent numerical time-series data in a non-Cartesian coordinates system, we realized it would eliminate the issue of scaling for all pieces of data (11). The black and white encoded images do lose transparency for the general reader of the data images (Fig. 2). Nonetheless, using the GAF package is a clean way to ensure all images are equal (5x5 pixels) and can speak to the promise of time-series prediction using image-based data (12, 13). After both experiments, we saw that our results supported our hypothesis: the conditional generative adversarial network was able to use image data to predict the price of the S&P 500 with relative accuracy.

Results

In the first experiment, images of candlestick graphs of SPY data from 1993 to 2003 were analysed to see if the Pix2Pix CGAN on Google Colabratory could be used to predict the next week's stocks. Unfortunately, we were unable to come up with conclusive results as the generated images looked completely different from the real graphs. They could also not be interpreted because each candlestick would have multiple ending points and spindles, which indicate the prices at which the stock closes and opens at, crucial to determining a prediction for the next day. However, all the fake images were able to pick up the general candlestick shape, with a rectangle and spindles sticking out on the ends, as well as the separation between the five candlesticks, which can speak to the effectiveness of the CGAN itself (Fig. 3).

In the second experiment, GAFs were analyzed to encode the highest price that the S&P 500 hit each day from 2009 to 2019. We also wanted to quantify our results and accuracy of the method, so we decided to use a cosine similarity function to determine the similarity between the fake images that were generated as predictions and the real images that were used as tests (Fig. 4). After calculating the similarity percentage of each pair of images using a pixel-by-pixel comparison, we got a final average of 0.7383, or around a 74% similarity rate of the entire model.

Discussion

When analyzing our first experiment, it was identified that there were some fundamental things amiss with the images we trained the Pix2Pix CGAN on, so it wasn't unexpected that we didn't have clear and conclusive answers. The CGAN was unable to pick up the significance of white and black candlesticks, which represented bullish and bearish patterns, and tended to make all the generated images white candlesticks. There were also often many spindles and cutoff points on one candlestick, which was inaccurate and didn't provide any actual interpretable value. One of the most fundamental problems with these graphs in experiment 1 is that the images included the axis numbers, axis titles, and white space around each image. The axis labels were unreadable and only added noise to the image. A deeper problem discovered was that the images did not have the same scaling for each photo, which meant some photos would display more range of monetary value than others. In the second experiment, although an average of 74% accuracy may not seem very high, if we consider the context of CGANs as well as the danger of overfitting, this proves that the CGAN model has great potential in being used as a trustworthy way to predict stock market data.

There are benefits and drawbacks of both types of usage of the CGAN model proposed. The model that uses the candlestick graphs makes the data easier to understand for the public eye, but they are more difficult to train with because they have more components within the image the CGAN needs to filter out and account for when trying to make a prediction. The second method with the GAF images is harder to understand, but it can streamline the process of training the CGAN and prediction, which would likely provide more accurate results.

There are two novel things about the research completed in this paper: the usage of image rather than raw numerical data, and the encoding of time-series data using GAFs. The goal of exploring and implementing this new method of predicting stocks is to increase awareness and promote more engagement with this intersection of financial and AI literacy. Future work could involve a series of "mix and matches". This means changing variables such as the kind of GAN used (implementing LSTM, MLP, or other networks), the specific stock or index used, using a different data point out of the OHLC of a stock, running multiple experiments with OHLC and then evaluating all of them together to get a better overall sense of the stock, etc. (14). We believe there is a point where these two approaches explored in our research can come together so that not only can we include visual data to increase transparency with consumers, but also preserve the accuracy rate achieved when using GAFs (15). This method could be applicable to different kinds of time-series datasets and can be further developed so that the model is easier for the public to use, adapt to, and understand.

Materials And Methods

Experiment 1: Candlestick Graphs

In the first experiment, I used SPY (SPDR S&P 500 trust) data from 2/22/1993 to 4/17/2003, 10 years, to create 512 images in total. Each image featured 5 candlesticks on one graph since each candlestick represents one day, the entire image represents one week of trading. In addition, each image is at the size of 800 x 575 pixels, including the extra margin of white space. Each graph is lined with a black outline, with axis labels including the written month and date. The background of the graph is the same for each one, a light blue with white lines, and each candlestick is of the same width, though they are either black or white depending on if the stock fell or rose from its opening price that day, respectively (Fig. 1). They also have needles on each side of each candlestick as it represents the range the price hit that day.

Experiment 2: GAF Images

In the second experiment, the data was shifted to be more recent SPY data in a time range from 01/02/2009 to 12/30/2019, 10 years with 552 images in total. This time frame was chosen because it avoids any extremely dramatic movements in the overall market. The data starts by recovering from the 2008 recession and ends right before the COVID-19 pandemic, which resulted in drastic aftereffects, including the huge crash in March of 2020. Considering the goal of this experiment was to test the effectiveness of the image prediction method, this particular time frame avoided potentially “confusing” the CGAN with such dramatic stock differences in short amounts of time, as they were “black swan” events and don’t accurately represent what the algorithm would be predicting if applied to the real world.

The most important difference in the second experiment was that the use of Gramian Angular Fields, or GAFs, was implemented.. My original reason to choose GAFs is that it was difficult to manipulate the graphs do not include the white margins and different words on the axis changes a numerical data point to a pixel, it can only do one numerical data point at a time, so the high data point was chosen since we were curious about possible predicting the “potential” of a stock. However, after studying the mathematics meaning of GAFs, it was concluded that a GAFs image provides more pertinent information of the times series data: GAFs essentially are relationships between every point and every other point in the time series. It also allows for the exploration of a more effective image-based presentation than a traditional candlestick chart for time series data.

To turn the numerical stock data into images, I used two slightly different methods for the two experiments. In the first experiment, I used the Python package Matplotlib Finance to create candlestick graphs from the numerical data. I looped through all the data in 5-day increments and saved them to my Drive through a mounted folder on Google Colab. In the second experiment, I followed the same steps, only creating GAF images using the code package from the Chen and Tsai experiment. Then, I split my image data in both experiments into A and B sets, with subsections of training and testing sets. For experiment 1, I split the 512 images into two folders- A and B. 256 images went into each, and within each A and B folder were 51 images placed into testing and 205 in the training folder that met a 20/80 split, respectively. Folder A contained the first week and every other week’s (even weeks) data after that, while the B folder included the second week and every other week’s (odd weeks) data afterward. This allowed for the Pix2Pix CGAN to place every two weeks’ images together and hypothetically find a correlation from one week to the next. In the second experiment, I followed the same steps, the only difference being that there were 552 GAF images in total, with 276 in A and B each and 54 test and 222 training images, maintaining the 20/80 split from the first experiment.

Acknowledgements

To Dr. Ganesh Mani at the Institute for Software Research in Carnegie Mellon University, thank you so much for your constant guidance and support. Thanks to Grace Su, Computer Science Student and AI Researcher at Columbia University, who provided me with initial research feedback and coding guidance.

References

1. Federal Reserve. (2021, March 2). USA Facts. https://usafacts.org/articles/what-percentage-of-americans-own-stock/
2. Unusual Whales. (2022, January 10). Congressional trading in 2021. Unusual Whales. https://unusualwhales.com/i_am_the_senate/full
3. Dream Exchange. (2021, March 12). Cision PR Newswire. https://www.prnewswire.com/news-releases/3-financial-literacy-statistics-that-need-to-change-in-2021-301246352.html
4. Loukas, S. (2020, July 10). Time-Series forecasting: Predicting stock prices using an LSTM model. Towards Data Science. https://towardsdatascience.com/lstm-time-series-forecasting-predicting-stock-prices-using-an-lstm-model-6223e9644a2f
5. P, S. (2021, May 19). Stock price prediction and forecasting using stacked LSTM. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/stock-price-prediction-and-forecasting-using-stacked-lstm/
6. Kumar, V. (2020, March 27). Hands-On guide to LSTM recurrent neural network for stock market prediction. Analytics India Magazine. https://analyticsindiamag.com/hands-on-guide-to-lstm-recurrent-neural-network-for-stock-market-prediction/
7. Lee, Y. H. (2021, February 16). Analysis of stock price predictions using LSTM models. Analytics Vidhya. https://medium.com/analytics-vidhya/analysis-of-stock-price-predictions-using-lstm-models-f993faa524c4
8. Goodfellow, I., Pouget-abadie, J., Mirza, M., Xu, B., Warde-farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144. https://doi.org/10.1145/3422622
9. Zhang, K., Zhong, G., Dong, J., Wang, S., & Wang, Y. (2019). Stock market prediction based on generative adversarial network [Paper presentation]. International Conference on Identification, Information and Knowledge in the Internet of Things. https://www.sciencedirect.com/science/article/pii/S1877050919302789
10. Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (n.d.). Image-to-Image translation with conditional adversarial networks. arXiv. https://doi.org/10.48550/arXiv.1611.07004
11. Chen, J.-H., & Tsai, Y.-C. (2020). Encoding candlesticks as images for pattern classification using convolutional neural networks. Financial Innovation, 6(1). https://doi.org/10.1186/s40854-020-00187-0
12. Mazzoni, C. (2021, February 19). How to encode time-series into images for financial forecasting using convolutional neural networks. Towards Data Science. https://www.google.com/url?q=https://towardsdatascience.com/how-to-encode-time-series-into-images-for-financial-forecasting-using-convolutional-neural-networks-5683eb5c53d9&sa=D&source=docs&ust=16409628781-286&usg=AOvVaw1IxgRloBawDANaJeaV2vEV
13. de Vitry, L. (2018, October 14). Encoding time series as images gramian angular field imaging. Analytics Vidhya. https://medium.com/analytics-vidhya/encoding-time-series-as-images-b043becbdbf3#:~:text=Gramian%20Angular%20Field%20Imaging%20%7C%20by,de%20Vitry%20%7C%20Analytics%20Vidhya%20%7C%20Medium
14. Ganegedara, T. (2020, January 1). Stock market predictions with LSTM in python. Datacamp. https://www.datacamp.com/community/tutorials/lstm-python-stock-market
15. Wang, Z., & Oates, T. (n.d.). Imaging time-series to improve classification and imputation. ArXiv.org. https://arxiv.org/pdf/1506.00327.pdf

Figures and Figure Captions

Figure 1. Image #430 from the training data of Set A in Experiment 1. 5 candlesticks are displayed with 3 white (bullish) and 2 black (bearish) candles. Created using Matplotlib graphs in Google Colaboratory.

Figure 2. Real GAF image #442 from the training data of Set A in Experiment 2. Originally 5 by 5 pixels, it is a blown up black and white image, and was created in Google Colabratory using the GAF package from the original "Encoding Candlesticks as Images" paper (8).

Figure 3. Fake Generated Image #2270 in Experiment 1. There are 5 white rectangles on a light blue square background, which was generated by the Pix2Pix CGAN in Google Colabratory as the next week prediction of stock prices.

Figure 4. Screenshot of my code in Google Colabratory of the cosine similarity function I used to calculate the accuracy rate displayed on the last line of output.

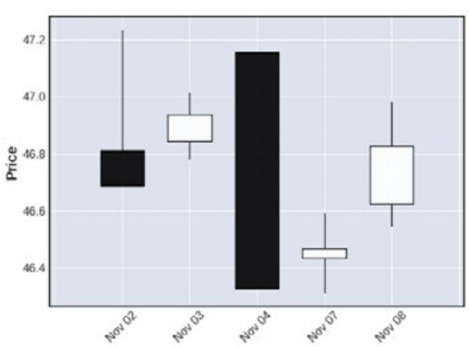

Figure 1

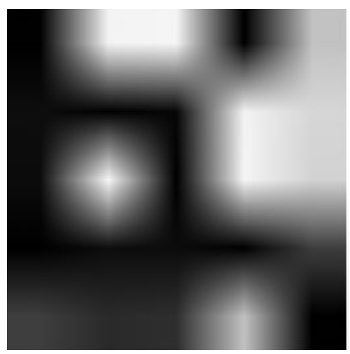

Figure 2

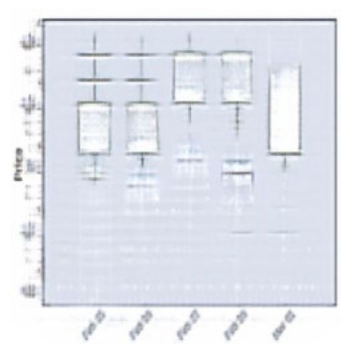

Figure 3

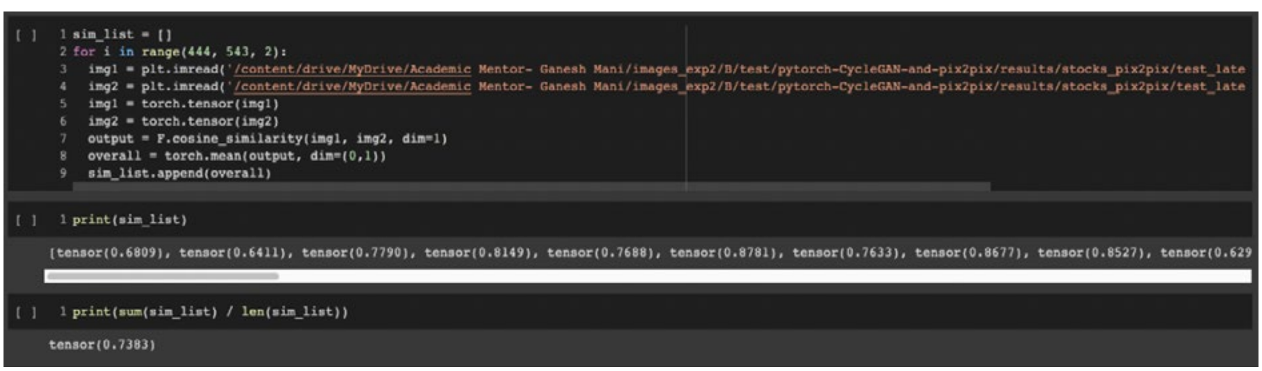

Figure 4