Related Work

Many real-world problems happen to be time series foresting problems. Currently, there has been a plethora of research using transformers for time series forecasting. This establishes the universality of transformers for being fit to use for univariate and multivariate, as well as time series embeddings (Neo et al. 1).

In 2020, Farsani and Pazouki attempted to use a “transformer neural network based on the self-attention” on “two benchmarks real-world TSF data sets from the different domains” (1). The result was a “20 per cent improvement in estimation accuracy, compare to other well-known methods,  along with the significant reduction in computation complexity” (1).

Wu et al. leveraged the self-attentive transformers on Influenza Like Illness (ILI) forecasting problem which achieved a 42.4% reduction in RMSE and a 20.7% increase in Pearson Correlation compared to the baseline model. The model also performed equivalently to the state-of-the-art ILI forecasting model, ARGONet (8).

In the field of speech recognition, Dong et al.worked with the “Speech-Transformer, a no-recurrence sequence-to-sequence model entirely relies on attention mechanisms to learn the positional dependencies” and compared the performance with the current “Recurrent sequence-to-sequence models using encoder-decoder architecture”. They found out their Speech Transformer takes only 1.2 days on 1 GPU to train and produces a “competitive 10.9% word error rate” which happens to be significantly faster than the known results of recurrent sequence to sequence models.

Transformers have been used to predict the motion of cars, an integral part of autonomous car driving, and have achieved state-of-the-art results while “substantially improving the diversity and the accuracy of predicted trajectories”(Liu et al. 1). They have also been used to predict the traffic flow which is a key for any good transportation system. Xu et al. have proposed a “Spatial-Temporal Transformer Networks (STTNs) which combines “...dynamical directed spatial dependencies and long-range temporal dependencies to improve the accuracy of long-term traffic flow forecasting (1)”. The key point to note is that the STTNs can achieve 10-40% and 40-60% reduction in computational costs compared to GraphWaveNet and DCRNN which are two of the eight baseline models (8).

They have also been used in the e-commerce industry to predict future sales to implement better stocking, promotions, and marketing practices. In this regard, Qi et al. proposed “Aliformer” based on the bi-directional Transformer which employs “historical information, current factor, and future knowledge” to predict future sales. The experiment resulted in “MSE improvement of 52% (0.321 → 0.154) for Informer, 53% (0.327 →0.154) for LogTrans, whereas, Informer and LogTrans are state-of-the-art models (5). Interestingly, their model greatly outperforms the above-mentioned models on the real-world product-sales data (T-Mall Sales) provided by Alibaba.

Transformer architecture has also been combined with other models such as state-space models (SSMs). This technique has consistently outperformed competitive baselines on time series forecasting and human motion prediction (Tand and Matteson 1). Their model, ProTran, has achieved the lowest average displacement error and final displacement error in predicting human motion (9).

Furthermore, the limitation of transformers, which is their inability to “process long sequences due to their self-attention operation, which scales quadratically with the sequence length”, has been addressed upon introducing “Longformer” that “that scales linearly with sequence length, making it easy to process documents of thousands of tokens or longer”. This allows the model to achieve, on a character level, the same result as *text8* and *enwik8*. Moreover, the “pre-trained Longformer” also outperforms  RoBERTa on long document tasks and sets new state-of-the-art results on WikiHop and TriviaQA” (Beltagy et al. 1).

More on extending the capabilities of transformers, Wu et al. have introduced an “Autoformer” with an “Auto-Correlation” mechanism. It deals with the traditional transformers' inability to find reliable dependencies due to “...intricate temporal patterns of long term future…”, and also lifts the “information utilization bottleneck” due to the transformer’s adaptation of “...sparse versions of point-wise self-attentions for long series efficiency…” (1).

Their model’s novelty lies in breaking away from “...the pre-processing convention of series decomposition, and making it an inner block of deep models…”. They also leverage stochastic process theory to design the Auto-correlation mechanism based on the series periodicity which results in “...dependencies discovery and representation aggregation at sub-series level…”. This technique has yielded a 38% improvement in relative performance on six benchmarks while covering five areas of application names; 1. Energy, 2. Traffic, 3. Economics, 4. Weather, and finally, 5. Disease.

Furthermore, on improving the transformers, Zhu et al. have proposed the model named “Informer” to increase the efficiency of transformers for long sequence time series forecasting. Their model also targets the limitation of transformers, namely “...quadratic time complexity, high memory usage, and the inherent limitation of the encoder-decoder architecture…” (1). Informer has three distinctive features, “...(i) a ProbSparse self-attention mechanism, which

achieves O(Llog L) in time complexity and memory usage and has comparable performance on sequences’ dependency alignment. (ii) the self-attention distilling highlights dominating attention by halving cascading layer input, and efficiently handles extreme long input sequences. (iii) the generative style decoder, while conceptually simple, predicts the long time-series sequences at one forward operation rather than a step-by-step way, which drastically improves the inference

speed of long-sequence predictions…” (1).  The model, running on five distinct datasets on univariate settings,”... outperforms DeepAR, ARIMA and Prophet on MSE by decreasing 49.3% (at 168), 61.1% (at 336), and 65.1% (at 720) in average…” (6). The brackets denotes prediction length. While on the multivariate settings, the model produces better results than “... RNN-based LSTMa and CNN-based LSTnet, and the MSE decreases 26.6% (at 168), 28.2% (at 336), 34.3% (at 720) in average…”. The authors do note that the substantial decrease in the performance on multivariate settings is due to the “...the anisotropy of feature dimensions’ prediction capacity…”(7).

In the area of predicting price movements, Transformers have shown great potential. Wallbridge has developed the model named, TransLOB, which leveraged the time series data of Limit Order Books and outperformed models ranging from SVM to CNN on prediction horizon starting from 10 till 100 (9).

Many forecasting techniques suffer from two limitations: 1. The inability to capture the stochasticity of the data. 2. The danger of accumulating error in inference. To solve this, Wu et al. leveraged “Adversarial Sparse Transformer” (AST) based on “Generative Adversarial Networks” (GANs) which works as a “generator to learn a sparse attention map for time series forecasting, and uses a discriminator to improve the prediction performance at a sequence level” (1). Their work has led to much greater effectiveness and efficiency in the field of time series forecasting.

Zerveas et al. have also showed the effectiveness of transformers based framework over the traditional supervised models such as regression and classification (1). Their work on multivariate regression problems has yielded “...on average, our models attain 30% lower RMSE than the mean RMSE among all models, and approx. 16% lower RMSE than the overall second best model (XGBoost), with absolute improvements varying among datasets from approx. 4% to 36%” (7). On classification problems, they used a total of 11 datasets and compared their model with five other models which included XGBoost, Rocket, LSTM, etc. Their model has able to achieve an average rank of 1.7, out performing all the models on 7 out of 11 datasets. It is interesting to note that the author’s model’s performance drops when it deals with very low dimensional time series.

On the front of multivariate time series problems, another classe of transformers namely, Gated Transformer Networks (GTN), has been used by Liu et al. They propose that since gating allows for modeling channel-wise and step-wise correlations, GTN are suitable for multivariate time series classification tasks. To test the hypothesis, they ran GTN on 13 different datasets and compared the results with 9 other models. GTN had outperformed all the models on 7 out of 13 datasets which is worthy of remarkable. However, the deeper analysis of the model has yield the propensity towards overfitting which is a cause of concern (Liu et al. 4).

Cryptocurrency, after the inception of Bitcoin in 2009, has been gaining popularity in all spheres of the society. It is evident in the fact that during the first month of 2021, the crypto market cap reached 1 Trillion Dollars . It is no wonder that predicting the price of different cryptocurrencies at varied time interval with significant accuracy is a valuable problem to solve.

Much research has been done in the above-mentioned problem sphere. In 2021, Livieris et al. attempted to use the “CNN-LSTM Model’ on three different cryptocurrencies, namely; 1. Bitcoin (BTC), 2. Ethereum (ETH), and 3. Ripple (XRP). Their data consisted of three consecutive years. The empirical study conducted on the data led to the finding that the model is well-versed in dealing with, in fact exploiting, the mixed cryptocurrency data, reducing the danger of over-fitting, and decreasing the computational costs in comparison to   “... traditional fully-connected deep neural networks…” (1).

Furthermore, Akyildirim et al. have also investigated the validity of Machine Learning in the business of predicting cryptocurrencies’ prices either at daily or minute level frequencies. The study included the data of the 12 most traded cryptocurrencies, based on volume and popularity, and paired with the following models: Support Vector Machines, Logistic Regression (A baseline from traditional techniques), Artificial Neural Networks, and Random Forest. All the models included past price information and technical indicators as their model features. They found that there exists a possibility of using Machine Learning successfully, for the algorithms do reach the predictive accuracy ranging from 55%-65% on average at daily or last-minute intervals. They also found the SVM performs consistently compared to Logistics Regression, Artificial Neural Networks, and Random Forest Classification (1).

Interestingly, in the study conducted by McNally where RNNs did not perform well when pitted against the LSTM, the experiment conducted by Christoforou et. al points otherwise. However, one must note that the feature selection by  Christoforou et. al was vastly different from McNally’s. They did take a holistic approach by including the data of financial information which includes market capitalization, historical price information, blockchain activity which encompasses the demand and supply of any cryptocurrency, and lastly the data from GitHub activity to extrapolate the speed of software development. The result of their comparative study with their distinctive feature selection yielded RNNs to deliver the promising result (1).

More specifically, Bitcoin’s price is of more importance, for it has much greater acceptance relative to other cryptocurrencies and enjoys the “grand-father” status. To build a better model which can incorporate the daily volatility of Bitcoin and achieve better results in predicting the price movement, Awoke et al. implemented the “...efficient deep learning-based prediction models specifically long short-term memory (LSTM) and gated recurrent unit (GRU)... (1)”. The study involved using the above-mentioned model to forecast the daily price of Bitcoin. It is worth noting that the identification and selection of features were done by the models themselves (5). The study concluded that the GRU-based forecasting model is more suited to forecast time series data of the highest price volatility. One should also note that GRU did not perform well when the forecasting window was 7 and 12 days in comparison to LSTM (9).

More on predicting the price of Bitcoin, Roy et al. proposed the use of “...autoregressive integrated moving average (ARIMA) model…” which was able to achieve  “... an accuracy of 90% for deciding volatility in weighted costs of bitcoin in the short run… (1)”. More specifically, running ARIMA on the dataset comprising from 2013 to 2018, they achieved 90.31% accuracy in predicting the price of Bitcoin for the next consecutive 10 days. While comparing the performance with the baselines used in the study, AR and MA achieved 89.25% and  87.58%  respectively.

Wirawan et al. have also contributed to the area of predicting the price of Bitcoin by using ARIMA with the hypothesis of predicting the short-term price of Bitcoin with high accuracy. By using the data ranging from 2013 to 2017, and basing the performance on generating the least Mean Absolute Percentage Error, the authors concluded that the model works best to predict the price one to seven days ahead. The MAPE was .87 and 5.98 for the next day's prediction and the next 7 day’s prediction respectively.

Many other researchers have been conducting studies to investigate the performance between LSTM and ARIMA when it comes to predicting Bitcoin Prices for varying lengths of time. Hua conducted a comparative study between ARIMA and LSTM to predict Bitcoin prices for varied time lengths. The findings were in line with other studies which show that for a shorter time period horizon, ARIMA proves to be an efficient model to predict the price, while LSTM could reach a better performance if given some more time to train via CPU (1).

However, on the opposite end, Mendes did find LSTM’s better performance to predict Bitcoin’s price when compared with the ARIMA model. The author’s data ranged from 2017 to 2019. The comparison resulted in the following finding that LSTM forecasts Bitcoin’s, on average, 92% and 94% better than ARIMA as per RMSE and MAE (1). More on the efficacy of LSTM when compared to other models including RNNs and ARIMA, McNally has found that LSTM achieves the highest classification accuracy (52%) and an RMSE of 8%. Interestingly, the findings of Hua which recommended the use of CPU to train the LSTM model are challenged by McNally’s finding which stipulates that the training on GPU has outperformed by 67.7% when compared with CPU implementation (1).

In 2019, a comparative study on predicting the price of Bitcoin was conducted by  Felizardo et al. The investigation included WaveNets (WNs), Recurrent Neural Networks (RNN), Random Forest, etc. The goal was to compare the results of recent advancements in machine learning algorithms with the traditional methods of price forecasting such as univariate Autoregressive (AR), univariate Moving Average (MA), Simple Exponential Smoothing (SES), and especially Autoregressive Integrated Moving Average (ARIMA). The authors ran the models on the data which had the following dimensions on a per-day basis: close price, open, high, low, percentage of change and the volume of transactions. The data comprised from 2012, 2013, 2014, 2015 and 2016. Although the years 2017 and 2018 were very happening for the price of Bitcoin, the authors decided to leave them out of the dataset due to the following reason: “...they are outliers and undergo great variance and random aspects…” (3). The study aimed to yield prediction results for the following time horizons: D1, D5, D10, and D30. The authors found that on a one-day time horizon (D1), ARIMA and SVR are the best performing models among the five models while optimizing on these metrics: 1. MAE, 2. MSE, 3. RMSE, 4. MAPE, 5. MPE. The results remain consistent up until D10. However, D30 reduces all models’ performance.

More on the topic of predicting Bitcoin prices using machine learning techniques on time series data, KARASU et al. have used Support Vector Machines to predict the price for different time window lengths and compared the performance against Logistic Regression. They used “...Filters with different weight coefficients are used for different window lengths. For different window lengths, Bitcoin price prediction is made using filters with different weight coefficients…”(1). The performance was measured with the traditional statistical metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Pearson Correlation. The finding is consistent with the overall findings of other researchers that SVM performs better than LR.

Kavitha et al have also evaluated the performance of machine learning algorithms in predicting the Bitcoin price. They evaluated RNN and LSTM and used LR as their baseline model. The hypothesis of the study was based on Big Data (the study used 30,00,000 rows of data) will affect the accuracy of prediction. The study concluded with the following finding: “...RNN model with LSTM is evidently effective for forecasting and prediction of Bitcoin prices than linear regression model because of its capability to recognize longer term dependencies…”(5). They did note that both models require significant computing time and more specifically, RNN loses its efficacy in training if trained on a small dataset.

Another area of research is stemming from using Transformers to predict cryptocurrencies’ prices. Zhao et al. conducted the study to investigate the Transformer model’s efficacy in forecasting Bitcoin and Ethereum prices. The baseline model to compare the performance was LSTM. Their work illustrated an interesting finding: when comparing the performance of LSTM and Transformer head to head, LSTM beats down Transformer. However, when Transformers are combined with sentiment analysis, which in this case was based on Twitter data and Valence Aware Dictionary and sEntiment Reasoner (VADER), Transformers outperform LSTM on Bitcoin price forecasting, but not on ETH’s price. It’s worth noting that with the same amalgamation of sentiment analysis with LSTM, there was no improvement in the performance of LSTM. Furthermore, they showed how transferred learnings can improve upon the prediction of ETH price by the Transformer Model.

**\cite{9538640} works on the price forecasting of the doge coin using Transfomer but modifies the default positional encoding with the Time2vec \cite{time2vec} that includes the periodic and non-periodic representation.Moreover, it uses both encoder and decoder with no attention module in decoder unlike default architecture.**

Works Cited

Akyildirim, Erdinc, Ahmet Goncu, and Ahmet Sensoy. "Prediction of cryptocurrency returns using machine learning." *Annals of Operations Research 297.1* (2021): 3-36.

Awoke, Temesgen, et al. "Bitcoin price prediction and analysis using deep learning models." *Communication software and networks. Springer, Singapore*, 2021. 631-640.

**Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." *arXiv preprint arXiv:2004.05150* (2020).**

Christoforou, Emmanouil, Ioannis Z. Emiris, and Apostolos Florakis. "Neural networks for cryptocurrency evaluation and price fluctuation forecasting." *Mathematical research for blockchain economy*. Springer, Cham, 2020. 133-149.

**Dong, Linhao, Shuang Xu, and Bo Xu. "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2018.**

Felizardo, Leonardo, et al. "Comparative study of bitcoin price prediction using WaveNets, recurrent neural networks and other machine learning methods." *2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC). IEEE*, 2019.

Hua, Yiqing. "Bitcoin price prediction using ARIMA and LSTM." *E3S Web of Conferences. Vol. 218. EDP Sciences*, 2020.

Karasu, Seçkin, et al. "Prediction of Bitcoin prices with machine learning methods using time series data." *2018 26th signal processing and communications applications conference* (SIU). IEEE, 2018.

Kavitha, H., Uttam Kumar Sinha, and Surbhi S. Jain. "Performance evaluation of machine learning algorithms for bitcoin price prediction." *2020 Fourth International Conference on Inventive Systems and Control (ICISC)*. IEEE, 2020.

Livieris, Ioannis E., et al. "An advanced CNN-LSTM model for cryptocurrency forecasting." *Electronics 10.3 (2021)*: 287.

**Liu, Yicheng, et al. "Multimodal motion prediction with stacked transformers." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2021**

Liu, Minghao, et al. "Gated transformer networks for multivariate time series classification." *arXiv preprint arXiv:2103.14438* (2021).

McNally, Sean. *Predicting the price of Bitcoin using Machine Learning.* Diss. Dublin, National College of Ireland, 2016.

Mendes, João Filipe Batista. Forecasting bitcoin prices: ARIMA vs LSTM. Diss. 2019.

**Mohammadi Farsani, R., and Ehsan Pazouki. "A transformer self-attention model for time series forecasting." *Journal of Electrical and Computer Engineering Innovations (JECEI)* 9.1 (2020): 1-10.**

**Qi, Xinyuan, et al. "From known to unknown: Knowledge-guided transformer for time-series sales forecasting in Alibaba." *arXiv preprint arXiv:2109.08381* (2021).**

Roy, Shaily, Samiha Nanjiba, and Amitabha Chakrabarty. "Bitcoin price forecasting using time series analysis." 2*018 21st International Conference of Computer and Information Technology (ICCIT). IEEE*, 2018.

**Tang, Binh, and David S. Matteson. "Probabilistic transformer for time series analysis." *Advances in Neural Information Processing Systems 34* (2021): 23592-23608.**

Wallbridge, James. "Transformers for limit order books." *arXiv preprint arXiv:2003.00130* (2020).

Wirawan, I. Made, Triyanna Widiyaningtyas, and Muchammad Maulana Hasan. "Short term prediction on bitcoin price using ARIMA method." *2019 International Seminar on Application for Technology of Information and Communication (iSemantic). IEEE*, 2019.

**Wu, Neo, et al. "Deep transformer models for time series forecasting: The influenza prevalence case." *arXiv preprint arXiv:2001.08317* (2020).**

Wu, Sifan, et al. "Adversarial sparse transformer for time series forecasting." *Advances in neural information processing systems* 33 (2020): 17105-17115.

Wu, Haixu, et al. "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting." *Advances in Neural Information Processing Systems 34* (2021): 22419-22430.

Xu, Mingxing, et al. "Spatial-temporal transformer networks for traffic flow forecasting." *arXiv preprint arXiv:2001.02908* (2020).

Zerveas, George, et al. "A transformer-based framework for multivariate time series representation learning." *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021.

Zhou, Haoyi, et al. "Informer: Beyond efficient transformer for long sequence time-series forecasting." *Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 12*. 2021.

Zhao, Huali, Martin Crane, and Marija Bezbradica. "Attention! Transformer with sentiment on cryptocurrencies price prediction." (2022).