uc3m | Universidad Carlos III de Madrid

Master Degree in Big Data Analytics

Academic Year 2020-2021

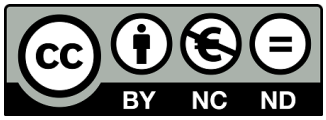
Master Thesis

"Time-Series Forecasting with Transformers"

Andrés Carrillo López

Pablo Martínez Olmos

Madrid, June 2021



This work is licensed under Creative Commons **Attribution – Non Commercial – Non Derivatives**

SUMMARY

Keywords:

DEDICATION

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

"There are two kinds of forecasters: those who don't know, and those who don't know they don't know."

John K. Galbraith

1.1. Motivation

In the history of mankind, an extensive amount of social and scientific science fields have benefited from the analysis and forecasting of time-dependent data. From disease monitoring and control [1], to financial and economic assets forecasting [2] and even supply chain logistics optimization [3].

Fueled by the continuously growing amount of data sources available thanks to the Internet as well as the proliferation of IoT ("*Internet of Things*" [4]) devices during the last decade, *Machine Learning* and other advanced statistical techniques have greatly benefited from this event. In particular, time-series forecasting, which targets the prediction or future behavior of data points of any kind based on past or historical data. Those methods takes advantage (in general) of having increasingly more and more historical data available to tune and fit their used models.

However, one should not be fooled by novel and technologically advanced methods for time series forecasting, given the fact that this field has been studied for decades, if not centuries. Some of the most studied and applied forecasting methods in history are still being applied nowadays, such as the *Exponential Smoothing* (ETS) family of methods [5], or *ARIMA* models [6], to name just a few. Even though these methods are widely applied on many different scenarios, their core assumptions on which are based, as well as basic limitations due to their formulation make them unsuited for certain temporal series or scenarios.

Over the last decade, models based on *Deep Neural Networks* (DNNs) have been popularized for time-series forecasting through known architectures such as *Recurrent Neural Networks* (RNNs), *Convolutional Neural Networks* (CNNs), and more recently, *Transformer* architectures. Deep neural networks benefits from the aforementioned big amounts of data available in many situations; as their training procedures are enhanced by having more historical data available, and are able to take local context into account and keep some long-term "memory" on recurring events. However, these recent approaches are far from being suited-for-all cases: one of their main downsides is that, in general, these models have a very complex inner structure and their produced forecasts are often

difficult to interpret (black-box nature) [7] and debug.

Attempting to solve the latter issues, *Transformer* architectures work by introducing a new element into their complex structure: an *attention* mechanism. In this way, the model forecast based on prior data is weighted on certain past moments based on a given influence learnt by such models. The attention mechanisms work as human subconscious would: by taking special attention and emphasis on past data which occurred at a certain key past moment when trying to forecast future behavior. These models have been focused since their recent introduction in 2017 [8] for *Natural Language Understanding* (NLU) tasks such as conversational and translation models, given that languages and their inherent syntactic properties are very well suited and benefited from this attention mechanism [9] [10].

However, the interesting attention mechanism characterizing *Transformers* can be further generalised to work on more general sequential data apart from text, such as traditional time-series data gathered from sensors or any other kind of historical continuous data. This work is thus focused on the usage of *Transformer* architectures when being applied to forecast general sequential datasets, such as road occupation with traffic, or electricity consumption in different households. Following this motivation, *Transformer* models will be tested and evaluated in comparison with classical forecasting methods as well as a popular DNN model.

1.2. Work Objectives

After introducing the main motivations of this work, here are briefly presented the dominant objectives that have been pursued to that end. Namely:

1. To explore the usage of Transformer network architectures for time-series forecasting, focusing on its main features and advantages compared to classical forecasting methods as well as other recent deep learning approaches on forecasting such series.

2. To survey the scarcely available code implementations of such Transformer architectures, and evaluate their usability and applicability to our target benchmark datasets.

3. To implement and tests different scenarios on forecasting time-series applied on real benchmark datasets, and perform a fair comparison of some of the most popular time-series forecasting models.

In order to achieve these objectives, a series of tests have been carried out and implemented in code notebooks hosted in the thesis repository; as well as the comparison analysis and results interpretations that will be carried out throughout this written work. More details about the technical development of these tests, work organization and environment used will be explained in chapter 3 and section 3.1.

1.3. Thesis Structure

Once the motivations and main objectives driving this work have been laid out, this written work follows the next structure from here on:

- **State of the art**: In this chapter, first the problem definition will be set, introducing the time-series forecasting scenarios. Then, after introducing the classical forecasting models that have been traditionally used in the literature, a small survey on deep learning models for time-series forecasting will be described, laying special emphasis on *Transformer* architectures, by first exploring their traditional use in *Natural Language Processing* tasks and finally introducing their usage for time-series analysis.
- **Technical implementation and tests**: Once the theoretical basis has been laid out in the previous chapter, the work and tests done for this thesis will be described. First, the environment as well as repository and thesis project organization will be set. Afterwards, the datasets used, as well as the retrieval, aggregation and overall processing applied to them will be described. Finally, all the different models, tests executed and followed workflows will be described.
- **Comparison and overall results interpretation**: After describing the tests carried out, a final comparison as well as results will be analyzed: from resulting error metrics as well as forecasts, to model's architectures explanatory plots, with the goal of interpreting how the different models have reacted to the fitted datasets.
- **Future work**: Finally, some additional tests, data processing and models worth noted to explore in future developments extending this work will be also mentioned.

2. STATE OF THE ART

2.1. Problem definition - Time-series forecasting

2.2. Classical Forecasting Methods

2.3. Transformers

2.3.1. Transformers - Natural Language Processing approaches

2.3.2. Transformers - Time-series Forecasting approaches

3. TECHNICAL IMPLEMENTATION AND TESTS

3.1. Environment Specifications, Thesis Organization and Repository

3.2. Survey of available and used libraries

3.3. Data: Definition, Retrieval, Aggregation and Processing

3.3.1. Electricity Benchmark Dataset

3.3.2. Traffic Benchmark Dataset

3.4. Models Definition, Training and Testing

3.4.1. Classical Models: (S)ARIMA, SES, Holt-Winters

3.4.2. DeepAR

3.4.3. Temporal Fusion Transformer (TFT)

3.4.4. Informer

4. COMPARISON AND OVERALL RESULTS INTERPRETATION

5. FUTURE WORK

BIBLIOGRAPHY

[1] N. Wu, B. Green, X. Ben, and S. O'Banion, "Deep transformer models for time series forecasting: The influenza prevalence case," *arXiv preprint arXiv:2001.08317*, 2020.

[2] I. Kaastra and M. Boyd, "Designing a neural network for forecasting financial and economic time series," *Neurocomputing*, vol. 10, no. 3, pp. 215–236, 1996.

[3] G. Wang, A. Gunasekaran, E. W. Ngai, and T. Papadopoulos, "Big data analytics in logistics and supply chain management: Certain investigations for research and applications," *International Journal of Production Economics*, vol. 176, pp. 98–110, 2016. DOI: https://doi.org/10.1016/j.ijpe.2016.03.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925527316300056.

[4] J. Clark, "What is the internet of things?" *IBM Blog*, 2016. [Online]. Available: https://www.ibm.com/blogs/internet-of-things/what-is-the-iot/ (visited on 04/30/2021).

[5] R. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder, *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media, 2008.

[6] G. E. P. Box and G. M. Jenkins, "Some Recent Advances in Forecasting and Control," *Journal of the Royal Statistical Society Series C*, vol. 17, no. 2, pp. 91–109, Jun. 1968. DOI: 10.2307/2985674.

[7] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PloS one*, vol. 13, no. 3, e0194889, 2018.

[8] A. Vaswani *et al.*, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/pdf/1706.03762.pdf.

[9] T. B. Brown *et al.*, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020. arXiv: 2005.14165. [Online]. Available: https://arxiv.org/abs/2005.14165.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.