IM1102-232433M - Deep Neural Engingeering assignment 2

Modifying the attention mechanism of transformers for time series forecasting

Arne Lescrauwaet (TODO) - Joachim Verschelde (TODO) - Alexander Van Hecke (852631385)

March 23, 2024

Introduction

This report details the steps taken by Arne Lescrauwaet (TODO), Joachim Verschelde (TODO) and Alexander Van Hecke (852631385) for the second assignment of the 2023 Deep Neural Engineering course organised by the Open University (1).

For this assignment we look at different attention mechanisms in transformers (2) for use with time series data. The attention mechanism enables a transformer model to selectively focus on relevant parts of the input data. The goal is to be able to capture long range dependencies and relationships between items of the input data. This is particularly important for time series data containing recurring patterns, e.g. hourly traffic counts on busy highways and hourly power consumption of nations. We expect these types of data to contain clear and recurring patterns (i.e. traffic will typically be lower during weekends) and we want an attention mechanism to capture these patterns. In addition to capturing recurring patterns, we would also like to be able to capture the "local context" of a pattern to predict new values. That is, when encountering an event that is similar to a past event, we want to take the outcome of that past event into account in our prediction.

Different kinds of attention mechanisms exist. Convolutional self-attention is introduced in (3), which aims to capture the local context of input events, but does this using a symmetric convolution, thereby taking both input data leading to a particular event and the outcome of that event into account. A dual-stage attention mechanism is used for a Recurrent Neural Network (RNN) architecture in (4), using an input attention mechanism in an encoder step, and a temporal attention mechanism in a decoder step.

Even though transformers were originally designed in the field of natural language processing (NLP), a lot of work has been done to use transformers with time series data. An overview of different ways to adapt transformers to time series data is given in (5). The time2vec encoding mechanism is introduced in (6). The authors of this paper use transformer models to predict stock prices, and claim these models can be used both for short and long term predictions. The effectiveness of applying transformers to time series data is tested in (7).

TODO hier korte literature review TODO nog meer zoeken naar verschillende attention mechanismen.

Goal

In this paper, we focus on using transformers for time series forecasting. We aim to compare different attention mechanism and determine which mechanism best captures the outcome of past events.

> **RQ : When comparing regular self-attention, convoluted self-attention, xyz self-attention, xyz2 self-attention, which mechanism best predicts future values?**

TODO nagaan wat we juist willen onderzoeken? accuracy? best outcomes capteren (zonder naar resultaat te kijken zegt dit niet veel)?

Firstly, this report will look at the characteristics of the dataset used and discuss pre-processing steps. Then, we will consider several attention mechanisms, discuss design and implementation details and finally evaluate the performance of these attention mechanisms on the dataset.

TODO nagaan of we idd performance willen meten.

Data analysis

Dataset description

The layout of the dataset is fully described here (8). We recap the most important points here.

Table 1: Features captured per quarter-hour in (8)

feature	description	range
todo	todo	[0.0-0.0]

TODO outliers

The features are organized as a time series of quarter-hour values.

TODO welke feature of features gaan we bekijken? Indien solar pv, kijken we ook naar de forecasts van elia zelf?

Data general properties

TODO spreken over regularity. zon, dag / nacht, . . . correlation is hier allicht geen concern als we maar 1 feature zouden in beschouwing nemen

Data pre-processing

TODO omschrijven van de data pre-processing stappen die we gaan nemen. bv selectie van data (maanden), aggregeren tot uren / dagen / weken? concatenatie van waarden per maand? per zomer? op welke lengte opslitsen? nachtelijke uren eruit halen?

PCA analysis

TODO NVT?

Duplicates analysis

TODO NVT?

Outlier analysis

TODO NVT?

Methodology and Implementation

Research methodology

We started by examining the dataset provided (8).

Beschrijven van de verschillende stappen.

All analysis was done using the pytorch python package.

Design elaboration

We decided to evaluate the following attention mechanisms :

- regular self-attention
- convoluted self-attention as described in (3)
- XYZ
- XYZ

TODO experiment beschrijven naar dataset of datasets, parameters (bv kernel size) en hyperparameters (indien van toepassing)

Table 2: parameters used for the attention mechanisms

algo	parameter	range
regular self-attention	?	?
convoluted self-attention	kernel size	2-10
XYZ	?	?

TODO indien meerdere scenarios (bv meerdere datasets, of een scenario zomer/winter/...) hier opsommen van scenarios

This entire design is repeated for a number of different scenarios. We detail these in Table 3.

Table 3: learning scenarios

scenario	description
summer	TODO july - september
winter	TODO december - february

Implementation

All code is available in a github repository (9).

TODO eventueel speciale vermeldingen rond implementatie

Evaluation and Results

Evaluation

TODO beschrijven wat we exact willen meten en hoe dit te meten (loss) (accuracy?)

To evaluate whether ... TODO ... self-attention ... , we formulate the following H_0 hypothesis :

> **H_0 : A self-attention mechanism using XYZ is not better at predicting ... than regular self-attention .**

If the p-value is below $\alpha = 0.05$, we can reject H_0 and accept the alternative hypothesis, that there is indeed a difference between the TODO.

Results

Scenario 1 : Regular self-attention

This is the baseline scenario. Results are summarized in Table 4.

Table 4: one sample t-test to determine whether TODO

metric	metric mean	population mean	t-test value	p-value	H_0 rejected
accuracy	0.9580	0.9864	-1.77E+01	2.29E-32	yes
accuracy	0.9858	0.9864	-1.84E+01	1.17E-33	yes
accuracy	0.9782	0.9864	-2.58E+01	9.62E-46	yes

TODO resultaten beschrijven

Scenario 2 : Convoluted self-attention

TODO idem hierboven

Recap

TODO hier summary van scenarios

Table 5: Features importance summary

scenario	attention mechanism	TODO
summer	regular self-attention	?
summer	convoluted self-attention	?
winter	regular self-attention	?
winter	convoluted self-attention	?

Conclusions and Discussion

In this study, we have used xyz dataset and pre-processed thus and thus.

We evaluated x self-attention mechanisms, x, y and z in x different scenarios. Results were :

- result 1
- result 2

TODO some discussion

TODO future work

References

1.   Startpagina - IM1102-232433M - Deep Neural Engineering [Internet]. [cited 2024 Mar 21]. Available from: https://brightspace.ou.nl/d2l/home/8636

2.   Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2017 [cited 2024 Mar 6]. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

3.   Li S, Jin X, Xuan Y, Zhou X, Chen W, Wang YX, et al. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2019 [cited 2024 Mar 10]. Available from: https://proceedings.neurips.cc/paper/2019/hash/6775a0635c302542da2c32aa19d86be0-Abstract.html

4.   Qin Y, Song D, Chen H, Cheng W, Jiang G, Cottrell GW. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence [Internet]. Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization; 2017 [cited 2024 Mar 6]. p. 2627–33. Available from: https://www.ijcai.org/proceedings/2017/366

5.   Wen Q, Zhou T, Zhang C, Chen W, Ma Z, Yan J, et al. Transformers in Time Series: A Survey [Internet]. arXiv; 2023 [cited 2024 Mar 7]. Available from: http://arxiv.org/abs/2202.07125

6.   Muhammad T, Aftab AB, Ibrahim M, Ahsan MdM, Muhu MM, Khan SI, et al. Transformer-Based Deep Learning Model for Stock Price Prediction: A Case Study on Bangladesh Stock Market. International Journal of Computational Intelligence and Applications [Internet]. 2023 Sep [cited 2024 Mar 16];22(03):2350013. Available from: https://www.worldscientific.com/doi/full/10.1142/S146902682350013X

7.   Cholakov R, Kolev T. Transformers predicting the future. Applying attention in next-frame and time series forecasting [Internet]. arXiv; 2021 [cited 2024 Mar 16]. Available from: http://arxiv.org/abs/2108.08224

8.   Solar-PV power generation data [Internet]. [cited 2024 Mar 23]. Available from: https://www.elia.be/en/grid-data/power-generation/solar-pv-power-generation-data

9.   Hecke AV, Lescrauwaet A, Verschelde J. Avhou/dne [Internet]. 2024. Available from: https://github.com/avhou/dne