You’re your own best teacher: A Self-Supervised Learning Approach For Expressive Representations

Johan Vik Mathisen

June 3, 2024

The overarching theme of this thesis is machine learning, with a focus on two specialized areas: generative modeling and representation learning.

Machine learning algorithms are fundamentally pattern finders designed to pick up on patterns in data and utilize these patterns for various tasks. These tasks may include distinguishing images of dogs from cats, identifying fraudulent bank transactions from legitimate ones, or predicting tomorrow's weather. This process is known as *predictive modeling*, which aims to use patterns in existing data to make predictions about unseen data.

In *generative modeling*, the objective shifts to recognizing patterns that enable the creation of data that resembles the original training data. A familiar example of this is ChatGPT, a generative model capable of producing text similar to its training input.

The second major component of this thesis is representation learning, which involves viewing information from different perspectives. Just as diverse perspectives on an issue can highlight various aspects and be useful in different contexts, different data representations can reveal distinct patterns. In machine learning, representation learning is about finding perspectives that are computationally useful. For instance, consider the sequence of numbers $1, 3, 7, 15, 31, \ldots$. One approach to finding a generating formula might involve examining the differences between consecutive numbers. However, a change in perspective, such as expressing the numbers in binary (i.e., $1, 11, 111, 1111, 11111$), makes the pattern more apparent—simply add another 1 to each subsequent number. While this is a simple example, it illustrates how changing perspectives can clarify patterns, a process that is crucial in complex scenarios commonly encountered in modern machine learning.

In our research, we are interested in generative modelling of time series. Time series, as the name kinda spoils, is data with a time component. Time series data encompasses diverse measurements such as weather conditions, patient heart rates, economic indicators (e.g., GDP, inflation, unemployment rates), energy usage, social media activity, and sales figures, all recorded at regular intervals. With the ever forward pointing arrow of time, it is really no surprise that these types of data are everywhere.

In this thesis we attempt, and in many ways succeed, to find a better perspective on time series, such that we can create new ones that resemble, but not completely mimic, the training data. We do this by squishing the data into a smaller space, in a way that preserves the most important information. In this compressed space, we too push similar looking time series to the same regions, which in a sense organizes it. Finally, using a clever method, similar to the one used by ChatGPT, we create new data in the compressed space, and decompress them to

get new time series.

The primary contribution of this thesis lies in this compressed but structured perspective on time series data.

Technically speaking, we investigate possible enhancements to the TimeVQVAE model presented in [**TimeVQVAE**] by introcuding a non contrastive self-supervised loss to the tokenization model. We specifically examine if the representations learned are more informative, in the sense that they improve the downstream classification accuracy, while simultaneously enable high quality reconstruction, and investigate how the learned representations affect the quality of the synthetic samples.

0.1 Acknowledgements

This thesis is the culmination of 6 years of studies in Trondheim. There are many people that deserve big thank you for making these years so memorable. If you are reading this, chances are quite high that your name should be here, but the margins are sadly too narrow.

I would firstly like to thank my supervisors, Erlend Aune and Daesoo Lee, for introducing me to this fascinating research field, and being open minded and supportive throughout the process. Erlend Lokna, a dear friend and great collaborator, for your indispensable contributions on this project. Developing advanced machine learning models is both challenging and time consuming, and synchronously banging our heads made it all a little easier. From the many late nights of writing, experimentation and discussion, to trash talking over the ping pong table, this project would not be the same without you. It was sweet to end on a 9-1 victory the final day, which once and for all cemented me as the top dog, and nothing (except a rematch) could change that.

I would too like to thank my family, mom, dad, Otto, Andreas and Tiril for being a continuous source of inspiration and motivation. You move so graciously through life, sharing both wisdom and skills, paving the way for the youngest in the flock. My roommates Pernille and Eldrun for being devilishly funny and handling me as i go into goblin mode. Preben and Elias, despite placing bets on whether i would last one or two weeks into my bachelors, for their endless curiosity and fearlessness in face of difficult mathematics, pushing me to learn way more than i aught to have.

Finally, Tindegruppa, the university climbing group, my crew. You have such a special place in my heart. You reintroduced the joy of movement in my life, after abruptly retiring as professional handball player. You provided an arena to play, connect with nature and go on adventure. Someone once told me, to have

a great time as a student in Trondheim, you need to find your little cult, be it on Samfundet, in a band or some part of NTNUI. I can wholeheartedly say i found mine, and to all the members, such loving, crazy and interesting people, thank you. See you on top of some remote peak very soon.

0.2 Motivation

- The role and importance of time series. - The need for models that capture complex structures for which traditional statistical models fail. Real world time series data is often incomplete (missing datapoints), irregular (datapoints not evenly spaced in time) and noisy. ML4ITS. - Why do people care about time series generation (TSG)? - Applications - Why is (unsupervised) representation learning for time series interesting?

Distributions of time series in their original temporal representation are complex and difficult to model. One would like to translate time series to a space where modelling is easier. This is one of the reasons to investigate representation learning for time series. Time series are recorded at record speed from sensors of various kinds (IoT, wearable devices). Unfortunately many of these do not have easily recognizable patterns for human observers, which makes labeling of such data quite difficult. In order to take advantage of this vast amount of unlabeled data we need techniques that can extract useful patterns without supervision. This is one of the reasons for investigating possible unsupervised models. A subcategory of unsupervised learning called self-supervised learning has in recent times shown great potential for learning informative and useful representations without the need of labeled data in the fields of computer vision and natural language processing. Most notably the GPT models from OpenAI which utilizes masked language modelling for pre-training.

0.3 Overview/structure

- Main inspirations [**TimeVQVAE**]
- Structure of the thesis

0.4 Research questions

Stage1

RQ1: Will self a supervised learning approach enhance downstream classification while simultaneously reconstruct well?

RQ2: How does augmentations influence reconstruction and downstream classification?

Stage 2

RQ3: Will more expressive representations improve synthetic sample quality

RQ4: How does augmentations influence synthetic sample quality?

0.5 Sustainability impact

Ethical and environmental impact consideration with basis in UN sustainability goals.

The field of AI and machine learning is in rapid development and the fear of being left in the dust in the gold rush makes many actors scrape all data they can find to train ever larger models.

One downside, not considering the massive copyright disputes in LLMs, questionable privacy and the spread of misinformation by hallucinations and use of deepfakes, is their colossal environmental impact. Large models are power hungry.

AI and machine learning are not universal tools though the modern LLMs make it seem so.

0.6 Collaboration and AI Declaration

During the work on our theses, Erlend Lokna and myself have collaborated extensively. This has been declared from the beginning and talked about continuously with our supervisors. Erlend, with his developer experience, has taken lead on code development for our model, and should get credit for his high level of programming skills. I have, in addition to code contribution and being a source of ideas, taken the lead on data processing and visualization. We have exchanges ideas, possible paths forward and the overall structure of the theses. Additionally have conducted the same experiments, hence our works will have many similarities. All writing though, is done independently. The collaboration has been very fruitful, and working closely with Erlend has made the process quite enjoyable, despite him being lousy at ping pong.

Use of LLMs in this thesis. Grammar, assistance with code for visualization, Latex help,