Paper Summaries for Master Thesis

October 20, 2023

Contents

1 Probabilistic Transformer for Time Series Analysis

1.1 Introduction and Objective

The paper titled "Probabilistic Transformer for Time Series Analysis" by Binh Tang and David S. Matteson from Cornell University aims to address the challenges in generative modeling of multivariate time series. The authors propose a novel approach that combines state-space models (SSMs) with transformer architectures. The focus is on creating models that can handle complex, non-deterministic dynamics across long-distance time steps, particularly for applications like demand forecasting, autonomous driving, and healthcare.

1.2 Methodology

The authors introduce the Probabilistic Transformer (ProTran), which incorporates attention mechanisms to model non-Markovian dynamics in the latent space. Unlike traditional SSMs, ProTran avoids the use of recurrent neural networks and instead employs a hierarchical structure of stochastic latent variables for increased expressiveness.

1. Single-Layered Model: The generative model uses a Gaussian distribution parameterized by two sequential steps of attention—one self-attention over previously inferred states and another attention over projected contexts. The inference model shares most parameters with the generative model and is trained end-to-end using stochastic variational inference.

2. Multi-Layered Extension: The authors extend ProTran to include multiple layers of latent variables. This hierarchical structure allows for more flexibility in modeling complex temporal dependencies. The multi-layered model also uses self-attention mechanisms and is trained using a similar variational inference framework.

3. Computational Aspects: The model has a time complexity of $O(T2d)O(T2d)$ and a memory cost of $O(T2d)O(T2d)$, where TT is the total sequence length and dd is the dimensionality of the latent space.

1.3 Main Findings

- ProTran is probabilistic, non-autoregressive, and capable of generating diverse long-term forecasts with uncertainty estimates.

- The model outperforms various state-of-the-art baselines in tasks like time series forecasting and human motion prediction.

- The multi-layered extension of ProTran offers increased flexibility and expressiveness, making it suitable for real-world applications with complex emissions or temporal dependencies.

1.4 Summary Bullet Points

- Introduction of Probabilistic Transformer (ProTran) that combines SSMs and transformer architectures for time series analysis.

- Utilization of attention mechanisms to model non-Markovian dynamics, avoiding the limitations of recurrent neural networks.

- Hierarchical structure of stochastic latent variables for increased model expressiveness.

- ProTran outperforms state-of-the-art baselines in various tasks, demonstrating its effectiveness and versatility.

2 Are Transformers Effective for Time Series Forecasting?

2.1 Introduction and Objective

The paper titled "Are Transformers Effective for Time Series Forecasting?" by Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu challenges the effectiveness of Transformer-based models for long-term time series forecasting (LTSF). The authors question whether the self-attention mechanism in Transformers, which is inherently permutation-invariant, is suitable for tasks that require understanding the temporal relations in a sequence of data points.

2.2 Methodology

The authors introduce a simple model called LTSF-Linear to serve as a baseline for comparison with Transformer-based models. This model uses a one-layer linear regression to directly forecast future time series based on historical data. The model is embarrassingly simple but serves to challenge the complexity and effectiveness of existing Transformer-based models.

1. **LTSF-Linear Model**: The model uses a single linear layer to directly regress historical time series data for future prediction. It employs a weighted sum operation, where $W \in \mathbb{R}^{T \times L}$ is a linear layer along the temporal axis. X_i and $\hat{X}_i$ are the input and prediction for each ith variate, respectively.

2. **Comparison Metrics**: The paper uses Mean Squared Error (MSE) and Mean Absolute Error (MAE) as core metrics for performance evaluation.

3. **Datasets**: The authors conduct experiments on nine widely-used real-world datasets that cover various applications like traffic, energy, economics, weather, and disease predictions.

2.3 Main Findings

- LTSF-Linear outperforms existing Transformer-based models in all cases, often by a large margin (20% - 50%).

- Most Transformer-based models fail to extract temporal relations effectively, as evidenced by their performance not improving (sometimes even deteriorating) with the increase of look-back window sizes.

- The paper suggests that the temporal modeling capabilities of Transformers for time series are exaggerated.

2.4 Summary Bullet Points

- Introduction of a simple one-layer linear model, LTSF-Linear, as a new baseline for long-term time series forecasting.

- Comprehensive empirical studies reveal that LTSF-Linear outperforms complex Transformer-based models.

- The paper challenges the effectiveness of Transformer-based models for time series forecasting, advocating for a reevaluation of their applicability in this domain.

3 Energy forecasting tools and services

The paper "Energy forecasting tools and services" provides a comprehensive overview of the various tools and techniques available for energy-related forecasting, particularly focusing on load and volatile renewable power. The authors aim to guide both researchers and practitioners through the labyrinth of existing tools and services, making it easier to understand what each offers.

3.1 Methodology

The paper is divided into four main sections:

1. Background information on time series forecasting.

2. Specific properties of energy time series and information regarding load and volatile renewable power forecasting.

3. Description and categorization of various energy forecasting tools and services.

4. Conclusion summarizing the findings.

The authors provide a detailed explanation of the most commonly used forecasting models, such as Exponential Smoothing and ARIMA, and their applications in energy forecasting. They also discuss the importance of considering weather parameters like solar irradiation and wind speed, which significantly influence renewable energy production.

3.2 Types of Forecasting Models

The paper categorizes forecasting models into three types:

- White-box models: Utilize known relations and expert knowledge.

- Black-box models: Rely solely on data mining techniques.

- Gray-box models: A combination of white and black-box models.

3.3 Point vs. Probabilistic Forecasting

The authors highlight the growing interest in probabilistic forecasting, which provides information about forecast uncertainty, in contrast to traditional point forecasting models.

3.4 Main Findings

- The paper serves as a comprehensive guide for understanding various energy forecasting tools and services.

- Exponential Smoothing and ARIMA models are commonly used for energy-related time series forecasting.

- Weather parameters like solar irradiation and wind speed are crucial for renewable energy forecasting.

- There is a growing interest in probabilistic forecasting to understand forecast uncertainty.

4 Make Transformer Great Again for Time Series Forecasting: Channel Aligned Robust Dual Transformer

4.1 Introduction

The paper introduces a specialized Transformer model, known as the Channel Aligned Robust Dual Transformer (CARD), designed to address the limitations of using standard Transformer models for time series forecasting. The authors argue that despite the Transformer's success in Natural Language Processing (NLP) and Computer Vision (CV), its application in time series forecasting has been less effective compared to Multi-Layer Perceptrons (MLPs).

4.2 Methodology

The CARD model introduces two key innovations:

- **Dual Transformer Structure**: Unlike typical transformers that only capture temporal dependencies, the dual structure in CARD captures both temporal correlations among signals and dynamic dependencies among multiple variables over time.

- **Robust Loss Function**: A new loss function is introduced to weigh the importance of forecasting over a finite horizon based on prediction uncertainties, aiming to mitigate the overfitting issue.

4.2.1 Tokenization

The authors adopt a tokenization strategy that converts the input time series into a token tensor. This is done to make the attention scheme more efficient compared to the vanilla point-wise counterpart.

4.2.2 Dual Attention Mechanism

The dual attention mechanism operates both over tokens and hidden dimensions. Exponential Moving Average (EMA) is applied to stabilize the training process.

4.2.3 Dynamic Projection

To handle high-dimensionality and reduce computational overhead, a dynamic projection technique is used in the attention mechanism over channels.

4.2.4 Signal Decay-based Loss Function

The authors introduce a new loss function that takes into account the variance of future observations, giving more weight to near-future predictions.

4.3 Results

CARD outperforms state-of-the-art models on multiple long-term and short-term forecasting datasets, confirming the effectiveness of the self-attention scheme.

4.4 Key Takeaways

- CARD efficiently and robustly aligns information among different channels.

- The model demonstrates superior performance in both long-term and short-term forecasting.

- The robust signal decay-based loss function effectively improves the model's performance.

5 Probabilistic Transformer for Time Series Analysis

5.1 Introduction

The paper "Probabilistic Transformer for Time Series Analysis" by Binh Tang and David S. Matteson addresses the challenges in generative modeling of multivariate time series. The authors propose a deep probabilistic method that combines state-space models (SSMs) with transformer architectures. Unlike traditional SSMs, this approach uses attention mechanisms to model non-Markovian dynamics in the latent space, thereby avoiding the use of recurrent neural networks.

5.2 Methodology

5.2.1 Single-Layered Probabilistic Transformer

The proposed model, termed as Probabilistic Transformer (ProTran), consists of a generative model and an inference model. The generative model uses attention mechanisms to model the latent dynamics. Specifically, it employs two sequential steps of attention: a self-attention over previously inferred states and another attention over the projected contexts. The model uses Gaussian sampling and approximating rectifier operators to sample from the distribution.

5.2.2 Multi-Layered Extension

The paper also introduces a multi-layered extension of the Probabilistic Transformer. This extension incorporates several layers of stochastic latent variables organized in a hierarchy, thereby increasing the model's flexibility for handling sequential data.

5.2.3 Training and Inference

The model is trained end-to-end using a stochastic variational inference objective. The inference model is parametrized in a similar fashion to the prior, sharing most parameters and undergoing the same recursive loop of operations.

5.3 Main Findings

- The proposed Probabilistic Transformer successfully combines the strengths of SSMs and transformer architectures, offering a new way to model complex, non-Markovian dynamics in time series data.

- The model is non-autoregressive and capable of generating diverse long-term forecasts with uncertainty estimates.

- The multi-layered extension further increases the model's expressiveness without compromising the tractable inference procedure.

- Extensive experiments show that the model consistently outperforms competitive baselines on various tasks and datasets.

6 A Review on Deep Learning Models for Forecasting Time Series Data of Solar Irradiance and Photovoltaic Power

6.1 Introduction

The paper reviews deep learning models for forecasting time series data related to solar irradiance and photovoltaic (PV) power. The focus is on comparing the performance of three standalone models—Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU)—and one hybrid model—Convolutional Neural Network-LSTM (CNN–LSTM).

6.2 Methodology

The authors compare the selected models based on various criteria such as accuracy, input data, forecasting horizon, type of season and weather, and training time. The performance metrics used for comparison include the root-mean-square error (RMSE) and relative RMSE.

6.3 Key Findings

- LSTM shows the best performance among standalone models in terms of RMSE.

- The hybrid model (CNN–LSTM) outperforms the standalone models but requires a longer training time.

- Deep learning models are more suitable for predicting solar irradiance and PV power compared to other conventional machine learning models.

- The authors recommend using relative RMSE as the representative evaluation metric for facilitating accuracy comparison between studies.

6.4 Implications

The paper provides valuable insights into the applicability and limitations of deep learning models in solar energy forecasting. It serves as a comprehensive guide for researchers and practitioners in selecting the most appropriate model based on specific requirements and conditions.

7 Scalable Reservoir Computing on Coherent Linear Photonic Processor

7.1 Introduction

The paper presents a novel approach to photonic neuromorphic computing using a scalable on-chip photonic implementation of a simplified recurrent neural network, known as a reservoir computer. The study leverages the ultrawide bandwidth and parallel processing capabilities of photonic computing hardware for ultrafast and energy-efficient computation.

7.2 Methodology

The authors employ a coherent linear photonic processor to implement the reservoir computer. Both the input and recurrent weights are encoded in the spatiotemporal domain by photonic linear processing. This is in contrast to previous approaches that rely on electrical components, which limit bandwidth and scalability. The architecture uses an array of Mach-Zehnder interferometers (MZIs) for various types of optical topologies. The system is driven by an optical pulse source that up-converts the input signals to a higher frequency region beyond the electrical bandwidth. The paper also introduces the concept of parallel processing based on wavelength division multiplexing (WDM), allowing the use of ultrawide optical bandwidth as a computational resource.

7.3 Results

Experiments were conducted for standard benchmarks, showing good performance for chaotic time-series forecasting and image classification. The device can perform 21.12 tera multiplication-accumulation operations per second (MAC/s) for each wavelength and can reach petascale computation speed on a single photonic chip using WDM.

7.4 Main Findings

- Demonstrated a scalable on-chip photonic implementation of a reservoir computer.

- Introduced spatiotemporal encoding of both input and recurrent weights, enabling ultrafast computing beyond electrical bandwidth.

- Showed the potential for petascale computation speed on a single photonic chip using WDM.

- Achieved good performance on standard benchmarks for chaotic time-series forecasting and image classification.

8 Temporal Convolutional Networks Applied to Energy-Related Time Series Forecasting

8.1 Introduction

The paper aims to address the challenges in energy demand forecasting by employing Temporal Convolutional Networks (TCNs). The authors argue that TCNs offer advantages over traditional Long Short-Term Memory (LSTM) networks, particularly in capturing long-term dependencies and computational efficiency. The study focuses on two energy-related time series datasets from Spain: national electric demand and power demand at electric vehicle charging stations.

8.2 Methodology

The authors employ TCNs, a specialized form of Convolutional Neural Networks (CNNs), designed to handle time series data. TCNs use causal dilated convolutions to capture long-term dependencies without information leakage from the future to the past. The paper involves an extensive parameter search, experimenting with more than 1900 models with different architectures and parameter settings.

8.2.1 Data Preprocessing

The data is preprocessed using min-max normalization and transformed into a format suitable for multi-step forecasting. A moving window scheme is used to create input-output pairs for the neural network.

8.2.2 Model Architecture

The TCN model is designed to output a complete forecasting window, adhering to the Multi-Input Multi-Output (MIMO) strategy. This approach allows the model to capture dependencies between the predicted values and avoids the accumulation of errors over predictions.

8.3 Main Findings

- TCNs outperform LSTMs in forecasting accuracy for both national electric demand and electric vehicle power demand.

- TCNs are computationally more efficient, offering advantages in real-time applications.

- The extensive parameter search validates the robustness of the TCN architecture for energy-related time series forecasting.

9 Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting

9.1 Introduction

The paper introduces Temporal Fusion Transformer (TFT), a deep learning architecture designed for multi-horizon time series forecasting. Unlike traditional models, TFT aims to provide both high-performance forecasting and interpretability. It is designed to handle a variety of inputs, including static covariates, known future inputs, and other exogenous time series, offering a more flexible and robust approach to time series forecasting.

9.2 Methodology

TFT employs a combination of specialized components to achieve its goals:

- **Gating Mechanisms:** These allow the model to adaptively control its complexity, enabling it to handle a wide range of datasets and scenarios.

- **Variable Selection Networks:** These are used to select relevant input variables at each time step.

- **Static Covariate Encoders:** These integrate static features into the network, providing context for temporal dynamics.

- **Temporal Processing:** A sequence-to-sequence layer is used for local processing, while a multi-head attention block captures long-term dependencies.

- **Quantile Forecasts:** TFT uses quantile regression to provide prediction intervals, offering a measure of uncertainty in its forecasts.

9.3 Main Findings

- TFT outperforms existing benchmarks in multi-horizon time series forecasting.

- The architecture provides valuable insights into the importance of different features and their temporal dynamics.

- The model is highly adaptable and can be applied to a wide range of real-world scenarios.

10 A Worrying Analysis of Probabilistic Time-series Models for Sales Forecasting

10.1 Introduction

The paper aims to rigorously evaluate the performance of three popular probabilistic time-series models—DeepAR, DeepState, and Prophet—in the context of sales forecasting. The authors employ a large-scale dataset, named EC dataset, and standardized training and hyperparameter selection to ensure a fair comparison.

10.2 Methodology

The authors set up two experimental principles:

1. Utilization of a large-scale dataset with various cross-validation sets.

2. Standardized training and hyperparameter selection for all models.

Three probabilistic models—DeepAR, DeepState, and Prophet—are compared against four baseline models: Moving Average (MA), Linear Regression (LR), Multi-layer Perceptron (MLP), and Seasonal ARIMA (SARIMA).

10.3 Results

The paper finds that simple models like MLP and Linear Regression outperform the more complex probabilistic models on RMSE (Root Mean Squared Error) and MAPE (Mean Absolute Percentage Error) metrics. Specifically, MLP had the lowest RMSE and MAPE, followed by Linear Regression.

10.4 Discussion

The paper suggests that while probabilistic models like DeepAR are good for capturing uncertainty, they may not be the best for point estimation tasks such as RMSE and MAPE. The authors also highlight the importance of feature scaling and the limitations of DeepState in capturing complex system dynamics.

10.5 Conclusions

- Simple models like MLP and LR outperform probabilistic models in RMSE and MAPE.

- Probabilistic models are not necessarily better for all types of forecasting tasks.

- Feature scaling plays a critical role in the performance of these models.