Partial Correlation-based Attention for Multivariate Time Series Forecasting

Won Kyung Lee
Department of Information & Industrial Engineering, Yonsei University, Seoul, Korea
wk.lee@yonsei.ac.kr

Abstract

A multivariate time-series forecasting has great potentials in various domains. However, it is challenging to find dependency structure among the time-series variables and appropriate time-lags for each variable, which change dynamically over time. In this study, I suggest partial correlation-based attention mechanism which overcomes the shortcomings of existing pair-wise comparisons-based attention mechanisms. Moreover, I propose data-driven series-wise multi-resolution convolutional layers to represent the input time-series data for domain agnostic learning.

Introduction

A multivariate time-series forecasting has great potentials for making decisions in various domains—transportation, finance, etc. For accurate forecasting, it is crucial to find dependency structure among the time-series variables and appropriate time-lags for each variable. This can be a challenging task as appropriate time-lags change dynamically over time, and there are long-term as well as short-term dependency. Furthermore, the dependency also needs to cover non-linear relationships.

In recent years, many artificial intelligence (AI) algorithms have been proposed, which can consider not only linear relationship but also non-linear relationships among the time series variables. To be specific, many neural network-based algorithms incorporating recurrent neural networks (RNN), convolutional neural networks (CNN), or graph neural networks (GNN), have been suggested to forecast time series by virtue of recent developments of machine learning. However, the existing RNN or CNN-based algorithms are limited to consider long-term dependency. Moreover, the GNN-based algorithms require much prior knowledge about the dependency structure between the time-series variables. To be more, results from these existing algorithms are hard to be interpreted.

In this thesis, I suggest attention-based multivariate time-series forecasting algorithm. The algorithm is inspired by Transformer (Vaswani et al., 2017), but several components are modified by my work for time-series forecasting. First of them is the attention mechanism which plays a key component in the Transformer algorithm. The Transformer has the encoder-decoder architecture and each component of the algorithm consists of dot-product self-attention layers. However, the dot-product attention can show a myopic view resulting in erroneous dependency structure for forecasting as it calculates only pair-wise similarity neglecting the other variables. Some researchers proposed additive attention (Bahdanau, Cho, and Bengio, 2015), but it is also based on pair-wise comparison. I suggest partial correlation-based attention mechanism. The partial cross-correlation is obtained for attention score.

Another component that I try to improve is to embed continuous time-series data. As the canonical Transfomer and its some variants (Vaswani et al., 2017; Huang et al., 2019) deal with only a discrete as well as replicable input data, such as word or musical sign, how to embed continuous time-series input data for the Transformer algorithm has been underdeveloped. There are some preprocessing methods to represent the time-series, such as Fourier or wavelet transformations. However, optimal filters are different from domain to domain so that finding appropriate filters is heavily dependent on the domain knowledge. In this study, I propose data-driven series-wise multi-resolution convolutional layers to represent the input time-series data for domain agnostic learning.

Method

Let me assume that there are S kinds of regular time-series over time T, $\boldsymbol{\mathcal{X}} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T)' \in \mathbb{R}^{S \times T}$, where $\mathbf{X}_t$ indicates S dimensional multivariate data at time t. The mul-

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

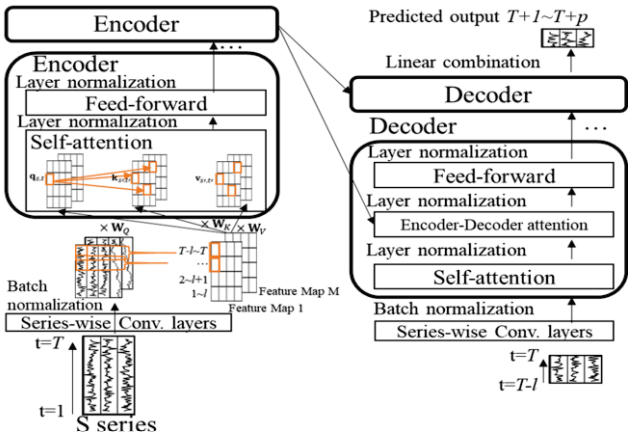

Figure 1: Proposed model architecture which forms encoder-decoder architecture.

tivariate time-series forecasting predicts future time-series over next p periods, $\mathbf{X}_{T+1}, \dots, \mathbf{X}_{T+p}$. As shown in Figure 1, the proposed model architecture forms encoder-decoder architecture. The encoder makes a representation of input time-series, and the decoder generates an output sequences one time period at a time. The architecture consists of a few self-attention layers as well as convolutional layers. Each component is presented in the following sections.

Partial correlation-based self-attention

The self-attention mechanism discovers the dependency structure among features and obtains representations of input in order to optimize the prediction loss function $\sum_{i=T+1}^{T+p}(\mathbf{X}_i - \widehat{\mathbf{X}}_i)^2$. The self-attention only utilizes its input feature map of the previous layer to find the dependency structure. Each feature map consists of the features which are enumerated by series, and time. In addition, each feature is designed to be l-length of subsequence in order to consider local contexts, and the length of the subsequence, l, is a hyperparameter. In detail, the input feature map is linearly transformed into query, key, and value features by learnable weights matrix. For each query feature, all of the key features are compared in terms of similarity. While the existing self-attention mechanism calculates dot-product similarity between each query-key pair without considering the other query-key pairs, these are conditioned in the proposed partial correlation-based self-attention. The partial correlation is obtained by fitting multiple linear regression of all of the key features on the query feature. To be specific, one of the penalized regressions, ridge regression could be applied to deal with the correlated key features. It means that $\mathbf{q}_{s,t} = \mathbf{K}_{:,:}\boldsymbol{\beta} + \mathbf{b}_s + \lambda\|\boldsymbol{\beta}\|_2^2,\ \lambda \geq 0$, where $\mathbf{q}_{s,t}$ represents query feature of s-th series at time t, $\mathbf{K}_{:,:}$ does all of the key features $[\mathbf{k}_{1,1}, \mathbf{k}_{2,1}, \dots, \mathbf{k}_{S,T-l}]$ covering time T over S series, and $\boldsymbol{\beta}$ is $S(T-l)\times 1$ dimensional coefficient vector. The $\mathbf{b}_s$ is a bias, and the λ is ridge hyperparameter for regularization. $||\cdot||_2$ means L2-norm. Then, we compute attention score as:

$$\text{Attention}(\mathbf{q}_{s,t}, \mathbf{k}_{s\prime,t\prime}) = \text{softmax}(\hat{\beta}_{s',t'}).$$

The attention score is then linearly combined with value features, $\mathbf{V}_{:,:}$, $[\mathbf{v}_{1,1}, \mathbf{v}_{2,1}, \dots, \mathbf{v}_{S,T-l}]$. Finally, representation for the s-th series at time t, $\mathbf{h}_{s,t}$ can be described as:

$$\mathbf{h}_{s,t} = \sum_{s\prime=1,t\prime=1}^{S,T-l} \text{Attention}(\mathbf{q}_{s,t}, \mathbf{k}_{s\prime,t\prime})\, \mathbf{v}_{s\prime,t\prime}$$

In case of encoder-decoder attention at the decoder part, this is quite similar with the self-attention. Only the difference is that key and value features come from the results at the last encoder part, but the query comes from the previous layer at decoder part. The encoder-decoder attention plays a role to look up the representations of the input time-series in order to generate an output sequence.

Series-wise multi-resolution convolution

As various smoothing methods are useful to extract features from the raw time-series data, convolutional layers are proposed to embed the time-series data. Inspired by wavelet transformation, various lengths of filters are employed in this paper, expecting to capture various lengths of temporal patterns. The filters are applied for each univariate time-series. From the convolution layers, the feature map can be obtained and each channel of the map is fed into the self-attention layers. In detail, as the different self-attention weights are learned for each channel, the proposed model learns various kinds of dependency structure between time-series.

Future Works

The proposed method will be evaluated with state-of-the-art baseline models in prior literature. The real-world open dataset for traffic forecasting, and stock market prediction will be used for the evaluation. Moreover, I plan to extend our framework to incorporate existing domain knowledge about dependency structure among time series. For example, road network structure can be harmonized in our algorithms to discover dependency structure among the multivariate time series variables.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998-6008.

Huang, C. Z. A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.; Simon, I.; Hawthorne, C.; Dai, A. M.; Hoffman, M. D.; Dinculescu, M.; and Eck, D. 2019. Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations.*