International Conference on Data Mining
SIAM 2021

# Attention-Based Autoregression for Accurate and Efficient Time Series Forecasting

**Jaemin Yoo** and **U Kang**

Computer Science & Engineering

Seoul National University

**SDM 2021**

# Outline

- **<u>Introduction</u>**
- Previous Works
- Proposed Method
- Experiments
- Conclusion

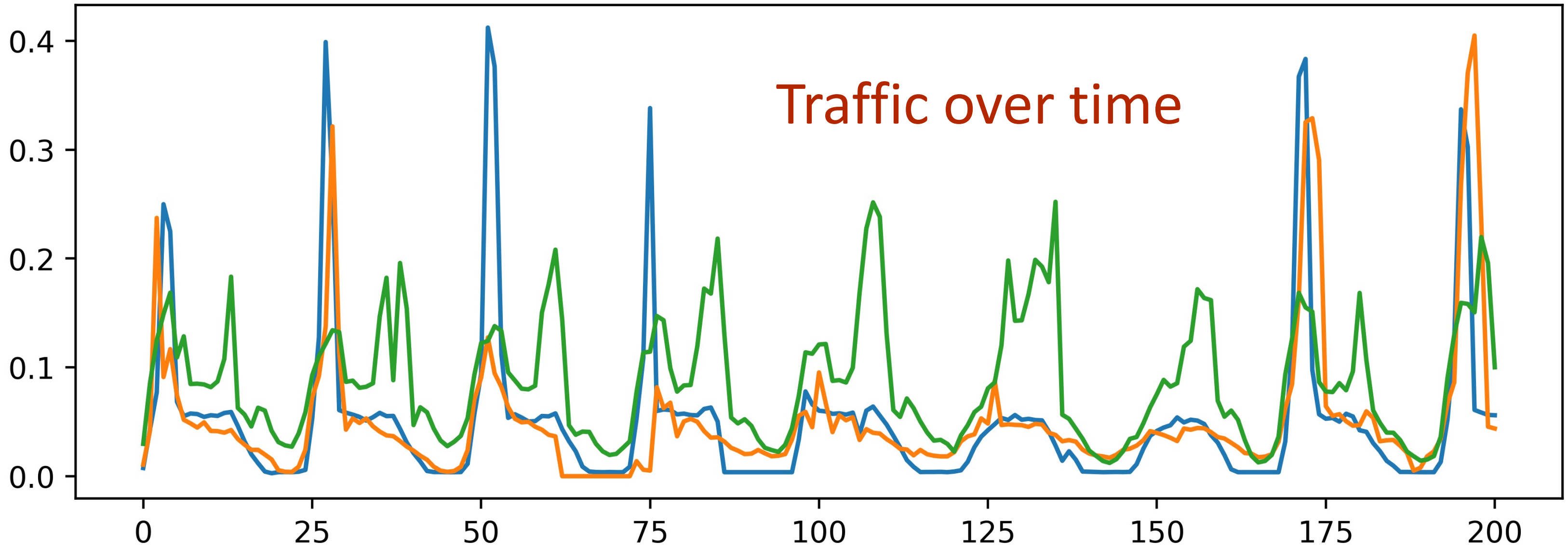
# Time Series Forecasting

- Core problem that has numerous applications
  - Stock price prediction
  - Product sales forecasting
  - Weather forecast

https://www.simplilearn.com/tutorials/data-science-tutorial/time-series-forecasting-in-r

# Multivariate Time Series

- Most time series data are **multivariate**
  - Such variables have *correlations* to each other
  - Prices of stocks, sales of products, …



# Problem Definition

- Multivariate time series forecasting
  - **Given**
    - Multivariate time series $\mathbf{X} \in \mathbb{R}^{d \times w}$
      - $d$ is the number of variables
      - $w$ is the number of recent observations
    - Prediction horizon $h$
      - Larger $h$ makes the problem more difficult
  - **Predict**
    - The observation $\mathbf{y} \in \mathbb{R}^{d}$ after $h$ time steps

# Outline

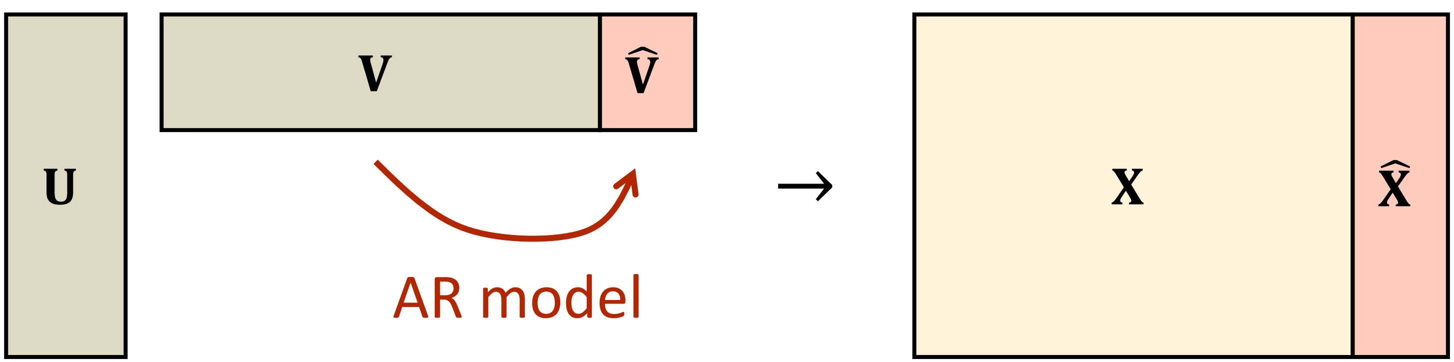
- Introduction
- **Previous Works**
- Proposed Method
- Experiments
- Conclusion

# AR Models

- **Autoregressive (AR)** models have been used widely for time series forecasting
  - **AR** is the simplest model for univariate forecasting
  - **VAR** extends AR to multivariate settings
- **Limitations**
  - They learn *linear* relationships between $\mathbf{X}$ and $\mathbf{y}$
  - VAR requires too many parameters: $O(d^2w)$
    - $d$ is the number of variables, and $w$ is the window size

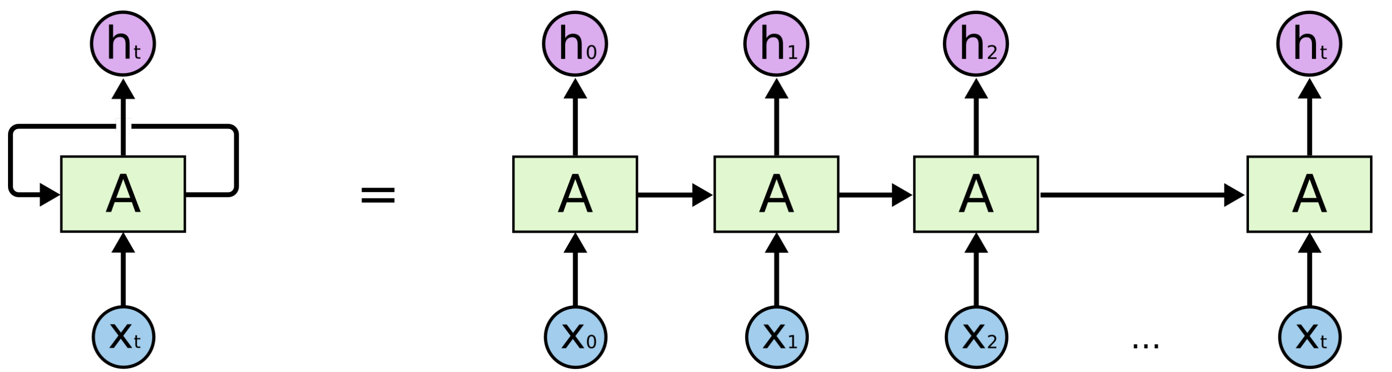
# TRMF

- **TRMF** (Yu et al., 2016) improves AR models based on matrix factorization (MF)
  - Applies an AR model to the time embedding matrix
- **Limitations**
  - Burdensome generation of the time embeddings



# RNN

- **Recurrent neural networks (RNN)** have been used widely for modeling sequential data
  - **GRU** and **LSTM** are popular variants of RNNs
- How to apply RNNs to multivariate forecasting
  - The $d$ observations at each step become an input




http://colah.github.io/posts/2015-08-Understanding-LSTMs/

# LSTNet

- **LSTNet** (Lai et al., 2018) is a recent approach based on recurrent neural networks (RNN)
  - Improves RNNs by applying temporal attention and skip connections between distant cells
- **Limitations**
  - Large number of parameters
    - The length of state vectors should be larger than $d$
    - This makes the model contain at least $d^2$ parameters
  - High sensitivity to its many hyperparameters

# Summary

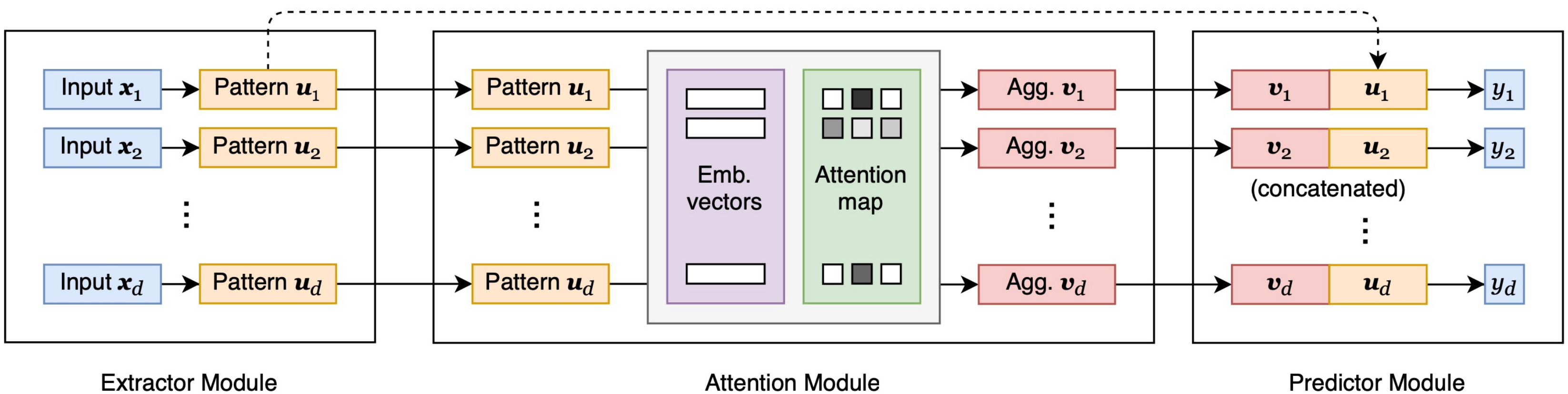
- Existing models are either too simple or have too many parameters
  - AR and TRMF
    - Cannot capture complex patterns in time series
  - VAR, LSTM, GRU, and LSTNet
    - Contain too many parameters and easily overfit
- **Research motivation:**
  - *To correlate variables with minimal parameters*

# Outline

- Introduction
- Previous Works
- **Proposed Method**
- Experiments
- Conclusion

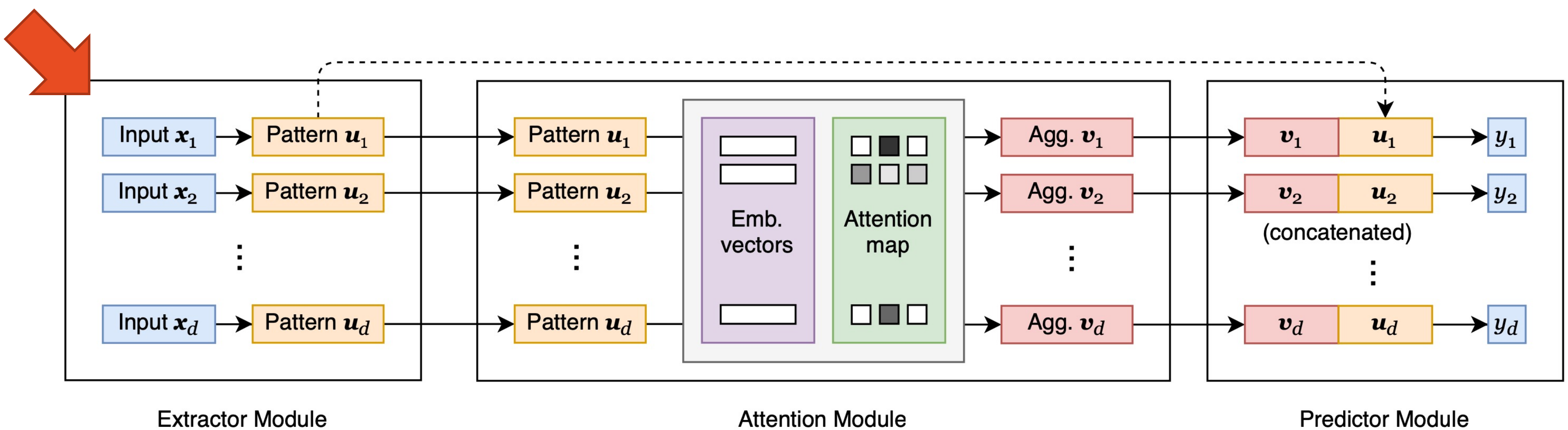
# Overview

- **AttnAR** (attention-based autoregression)
  - Our approach for efficient multivariate forecasting
  - End-to-end framework of three separable modules
    - **Extractor**, **attention** and **predictor** modules
    - Module structure is our key idea for high efficiency



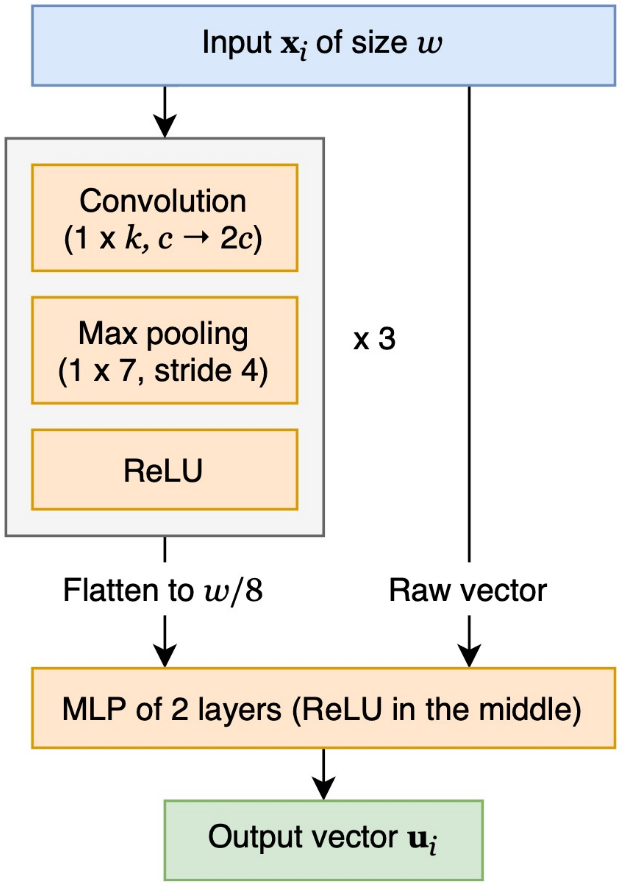
# Extractor Module (1)

- **Extractor module** captures univariate patterns
  - Transforms a raw observation $\mathbf{x}_i$ of each variable $i$ into a pattern vector $\mathbf{u}_i$ by a neural network
    - $\mathbf{u}_i$ is fed into both attention and predictor modules
  - *Which network should we use for efficiency?*



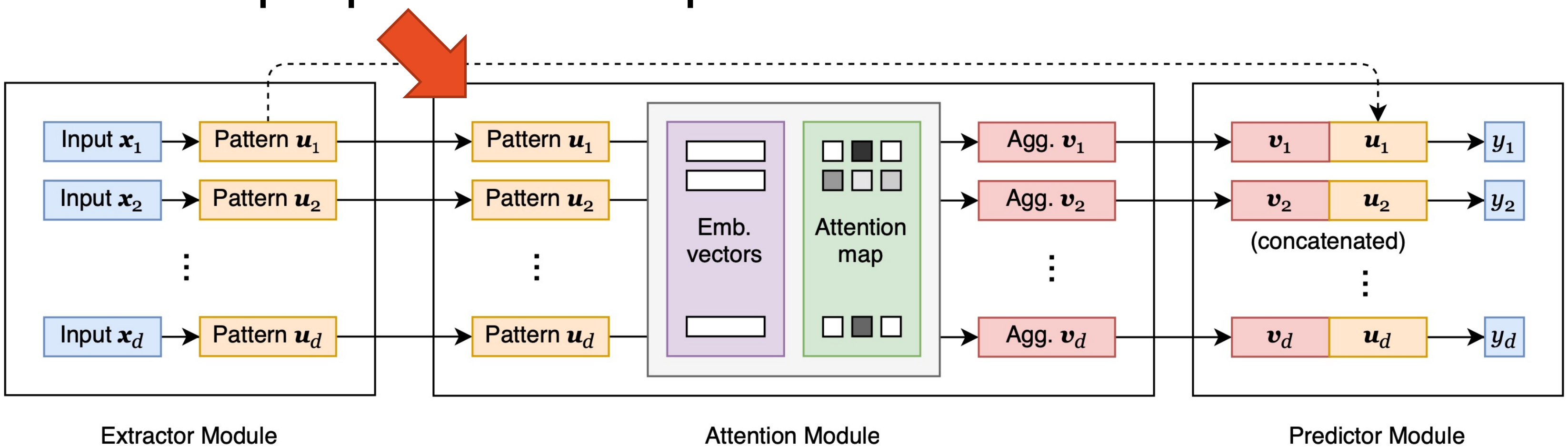
# Extractor Module (2)

- **MCE (mixed-convolution extractor)**
  - Our proposed model for efficient pattern extraction
  - **Shallow dense layers**
    - Connect distant time steps
    - Low degree of abstraction
  - **Deep convolution layers**
    - Focus on adjacent time steps
    - High degree of abstraction



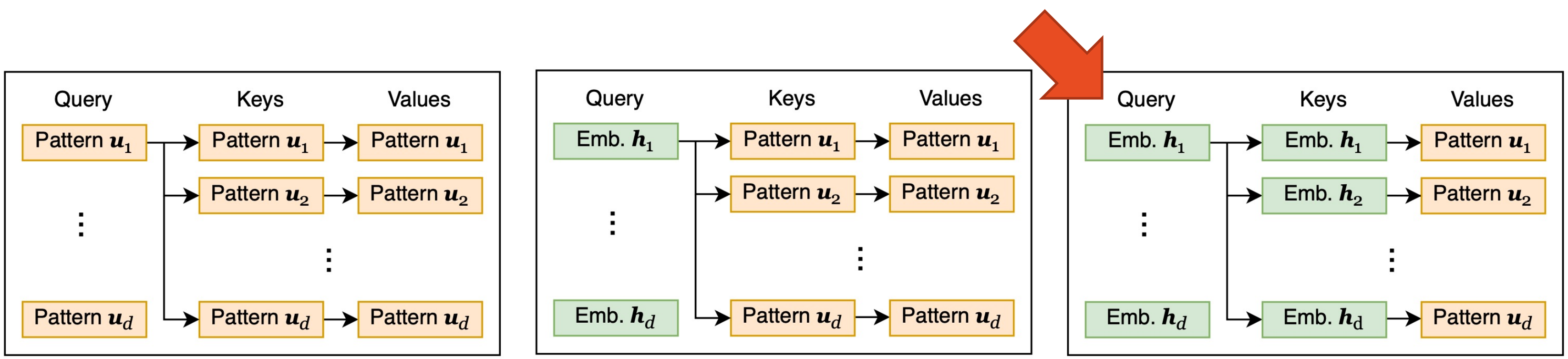
# Attention Module (1)

- **Attention module** correlates given variables
  - The main component of our AttnAR
  - Correlates the pattern vectors $\{\mathbf{u}_i\}$ of variables by an attention map $\boldsymbol{S} \in \mathbb{R}^{d \times d}$ and returns $\{\mathbf{v}_i\}$
  - We propose three options as the attention function



# Attention Module (2)

- We use the **time-invariant attention (TIA)** as our attention function
  - Learns a static embedding $\mathbf{h}_i$ for each variable $i$
  - Generates the attention map from $\mathbf{h}_i$, excluding $\mathbf{u}_i$
- The attention becomes robust and consistent

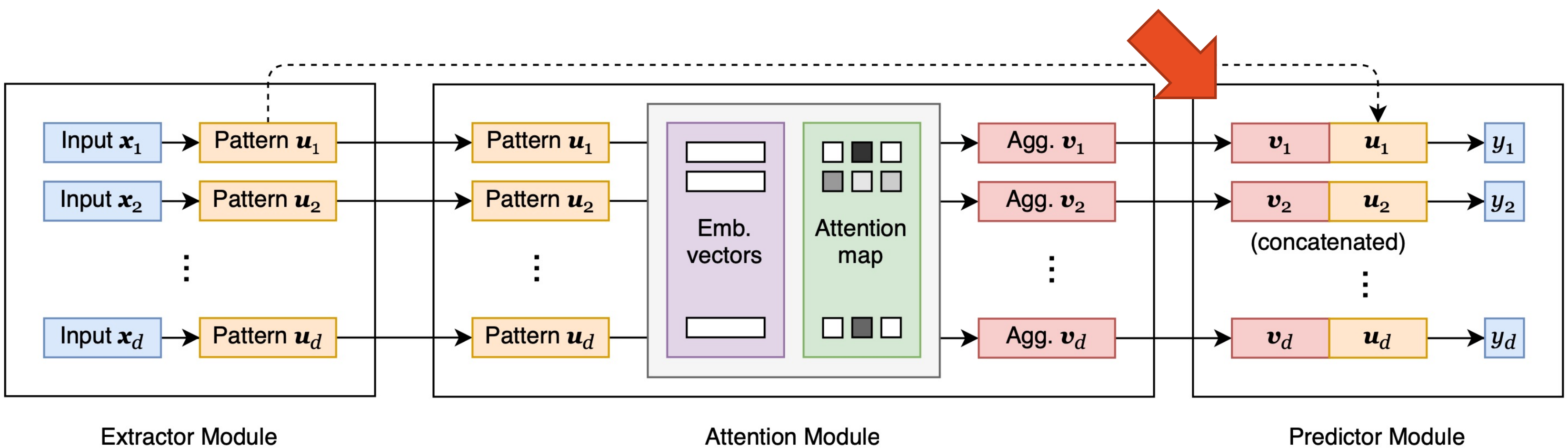



(a) Basic attention (Section 3.3.1). (b) Hybrid attention (Section 3.3.2). (c) Time-invariant attention (S. 3.3.3).

# Predictor Module

- Lastly, the **predictor module** simply produces the final prediction given the pattern vectors:

$$\hat{y}_i = f_{\text{mlp}}(\mathbf{u}_i \parallel \mathbf{v}_i)$$



# Outline

- Introduction
- Previous Works
- Proposed Method
- **Experiments**
- Conclusion

# Experimental Setup

- We use four multivariate time series datasets

| Dataset | Length | Dim. | Granularity |
|---|---|---|---|
| Traffic | 17,544 | 862 | 1 hour |
| Electricity | 26,304 | 321 | 1 hour |
| Solar-Energy | 52,560 | 137 | 10 minutes |
| Exchange-Rate | 7,587 | 8 | 1 day |

- The prediction horizon $h$ varies in $\{6, 12, 24\}$
- **Evaluation:** Root relative squared error (RSE)
  - RMSE divided by the standard deviation of $\mathbf{Y}$

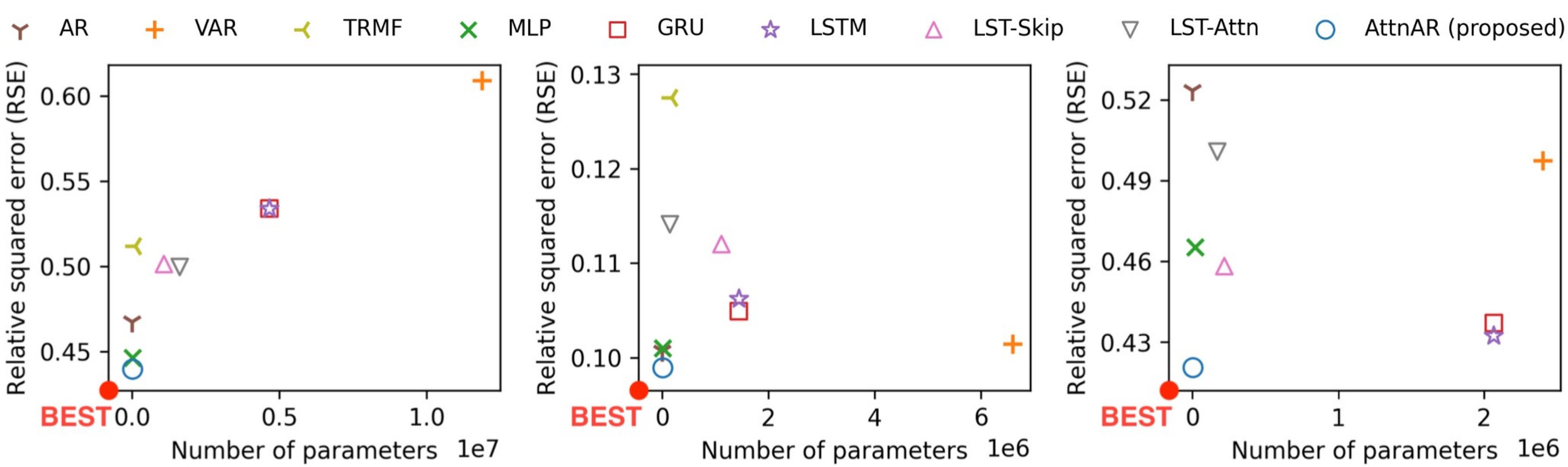
# Forecasting Accuracy

- AttnAR makes the most accurate predictions in nine of the twelve cases
  - Exchange-Rate is very noisy, and AR does the best
  - The improvement is significant in Solar-Energy

| Method | Traffic | | | Electricity | | | Solar-Energy | | | Exchange-Rate | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $h$=6 | $h$=12 | $h$=24 | $h$=6 | $h$=12 | $h$=24 | $h$=6 | $h$=12 | $h$=24 | $h$=6 | $h$=12 | $h$=24 |
| AR | .4647 | .4659 | .4675 | .0930 | .0983 | .1007 | .3120 | .4195 | .5235 | **.0238** | .0329 | **.0433** |
| VAR | .5909 | .6008 | .6088 | .0964 | .1010 | .1014 | .2965 | .4112 | .4974 | .0496 | .0652 | .0872 |
| TRMF | .4871 | .4909 | .5120 | .1050 | .1062 | .1275 | .6001 | .7112 | .8434 | .0425 | .0466 | .0542 |
| MLP | .4368 | .4436 | .4464 | **.0871** | .0965 | .1010 | .2747 | .3592 | .4652 | **.0238** | **.0328** | .0436 |
| GRU | .5158 | .5225 | .5340 | .1088 | .0974 | .1049 | .2485 | .3229 | .4370 | .0322 | .0465 | .0639 |
| LSTM | .5195 | .5268 | .5337 | .1043 | .1008 | .1062 | .2539 | .3328 | .4323 | .0412 | .0503 | .0658 |
| LST-Skip | .4811 | .4900 | .5013 | .0993 | .0959 | .1120 | .2537 | .3448 | .4582 | .0279 | .0425 | .0553 |
| LST-Attn | .4780 | .4895 | .4996 | .0936 | .0990 | .1141 | .2552 | .3528 | .5007 | .0379 | .0473 | .0590 |
| **AttnAR** | **.4287** | **.4370** | **.4396** | **.0871** | **.0942** | **.0989** | **.2272** | **.3057** | **.4205** | .0240 | .0336 | .0448 |

# Parameter-Efficiency (1)

- AttnAR makes the best parameter-efficiency
  - The error often increases with the model size
  - Overfitting is common in multivariate forecasting




(a) Traffic (b) Electricity (c) Solar-Energy

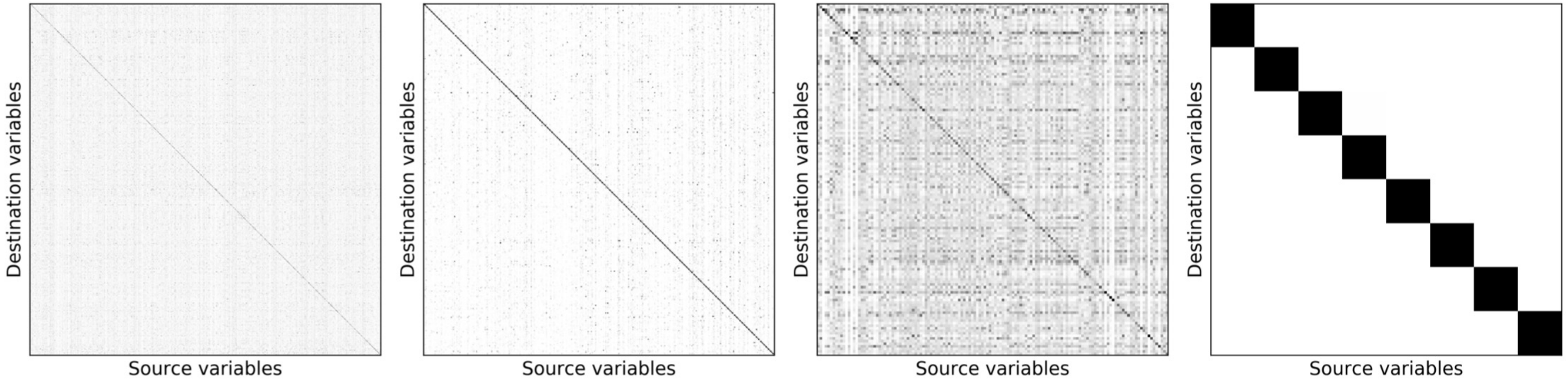
# Parameter-Efficiency (2)

- RNN-based models requires many parameters, especially in a dataset with many variables
- AttnAR has up to 42.6× fewer parameters

| Method | Traffic | Elec. | Solar | Exchange |
|---|---|---|---|---|
| GRU | 4665.3K | 1445.4K | 2066.9K | 14.5K |
| LSTM | 4665.3K | 1445.4K | 2066.9K | 804.4K |
| LST-Skip | 1086.1K | 1114.1K | 218.7K | 65.4K |
| LST-Attn | 1621.5K | 144.1K | 170.5K | 18.6K |
| AttnAR | **25.5**K | **9.5**K | **10.7**K | **0.9**K |

# Attention Map

- AttnAR generates interpretable attention maps
  - Strong correlations in Solar-Energy
  - Weak correlations in Traffic and Electricity
  - No correlations in Exchange-Rate




(a) Traffic (b) Electricity (c) Solar-Energy (d) Exchange-Rate

# Outline

- Introduction
- Previous Works
- Proposed Method
- Experiments
- **<u>Conclusion</u>**

# Conclusion

- **AttnAR (attention-based autoregression)**
  - Our proposed model for multivariate forecasting
- Main ideas of AttnAR
  - End-to-end learning of three separable modules
  - **MCE** for efficient extraction of univariate patterns
  - **TIA** for consistent and robust attention maps
- Experimental results
  - AttnAR consistently outperforms existing models

# Thank you!

Jaemin Yoo (jaeminyoo@snu.ac.kr)

Supported by **SIAM Travel Awards**