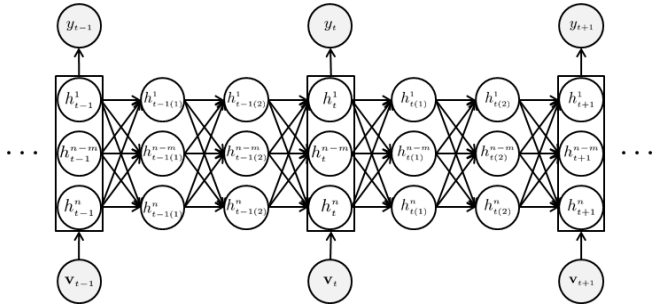
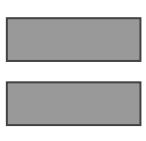
Implementation Project: SHAP for Visual Explanations of Time Series Models

Group 18

Introduction

Introduction





Related Work

- Model Explainer
  - Gradient Information: CAM, Grad-CAM
  - Back Propagation: DeepLIFT

- Time Series Model
  - Univariate Time Series: ARIMA, LSTM, GRU
  - Multivariate Time Series: VARMA, MT-GNN

SHAP (SHapley Additive exPlanations)

- Unified framework for interpreting predictions, proposed in 2017.
- The goal of SHAP is to explain the prediction for any instance x_i as a sum of contributions from its individual feature values.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)]$$

Notation: |M| is the total number of features. S represents any subset of features that doesn't include the i-th feature and |S| is the size of that subset. $f_x()$ represents the prediction function for the model.

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems. 2017.

SHAP (SHapley Additive exPlanations)

- Deep Explainer

Time Series Forecasting

Time Series Forecasting

- Datasets
- Metrics
- Models

Datasets

Dataset	Number of features (n)	TS length (T)	Time interval
Solar	137	52560	10 minutes
Traffic	862	17544	1 hour
Electricity	321	26304	1 hour
Exchange rate	8	7588	1 day

Table 1: Summary of datasets: the number of features in the multivariate time series, the total number of time steps, and the time interval between time steps.

Metrics

Root Relative Square Error (RSE):

$$RSE = \frac{\sqrt{\sum_{(i,t)\in\Omega_{\text{Test}}} \left(Y_{it} - \hat{Y}_{it}\right)^2}}{\sqrt{\sum_{(i,t)\in\Omega_{\text{Test}}} \left(Y_{it} - \text{mean}(Y)\right)^2}}$$

Relative Absolute Error (RAE):

$$RAE = \frac{\sum_{(i,t)\in\Omega_{\text{Test}}} \left|Y_{it} - \hat{Y}_{it}\right|}{\sum_{(i,t)\in\Omega_{\text{Test}}} \left|Y_{it} - \text{mean}(Y)\right|}$$

Empirical Correlation Coefficient (CORR):

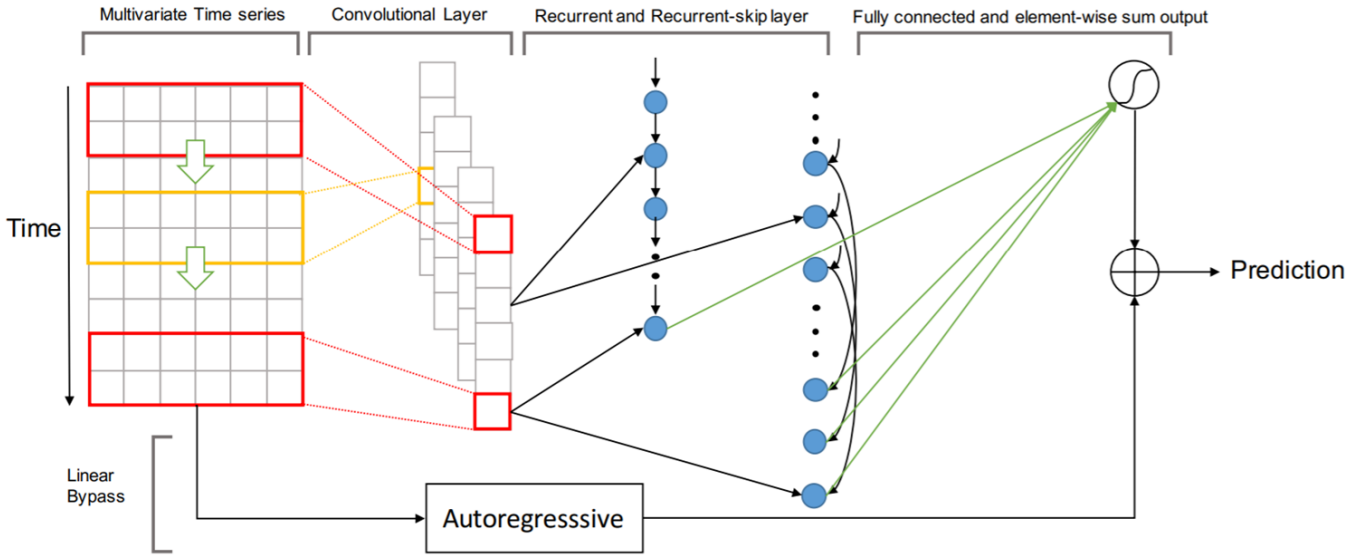
$$CORR = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_t \left(Y_{it} - \text{mean}(Y_i)\right)\left(\hat{Y}_{it} - \text{mean}(\hat{Y}_i)\right)}{\sqrt{\sum_t \left(Y_{it} - \text{mean}(Y_i)\right)^2 \left(\hat{Y}_{it} - \text{mean}(\hat{Y}_i)\right)^2}}$$

where Ω_{Test} denotes the set of indices in the test set, and $Y, \hat{Y} \in \mathbb{R}^{n\times T}$ are the ground truth and predicted time series, respectively.

Time Series Models

- LSTNet
- TPA-LSTM
- mWDN

LSTNet


Multivariate Time series
Convolutional Layer
Recurrent and Recurrent-skip layer
Fully connected and element-wise sum output
Time
Prediction
Linear
Bypass
Autoregresssive

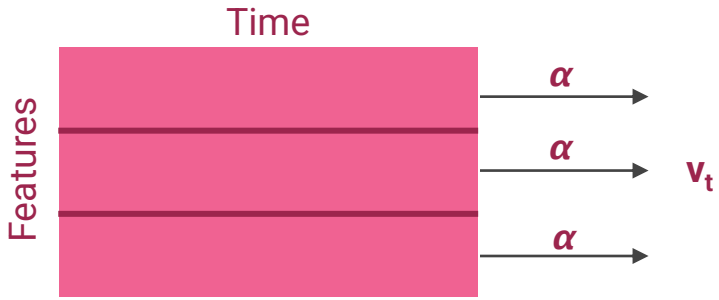
LSTNet

Dataset	Implementation	*RSE*	*RAE*	*CORR*
Solar	Ours	0.4239	0.2844	0.9094
	Original	**0.3254**	/	**0.9467**
Traffic	Ours	0.5071	0.3404	0.8598
	Original	**0.4950**	/	**0.8614**
Electricity	Ours	**0.0995**	0.0542	0.9049
	Original	0.1007	/	**0.9119**
Exchange rate	Ours	0.0357	0.0296	**0.9538**
	Original	**0.0356**	/	0.9511

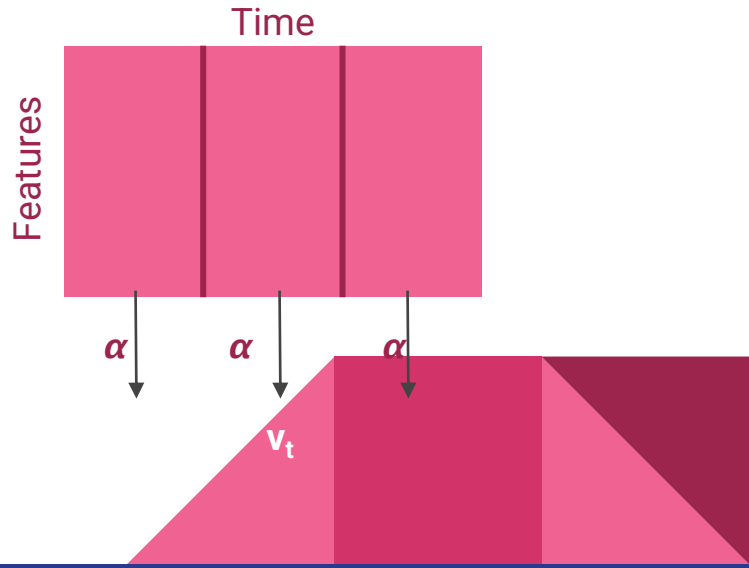
Performance of LSTNet implementation: our implementation vs reported by original paper

TPA-LSTM

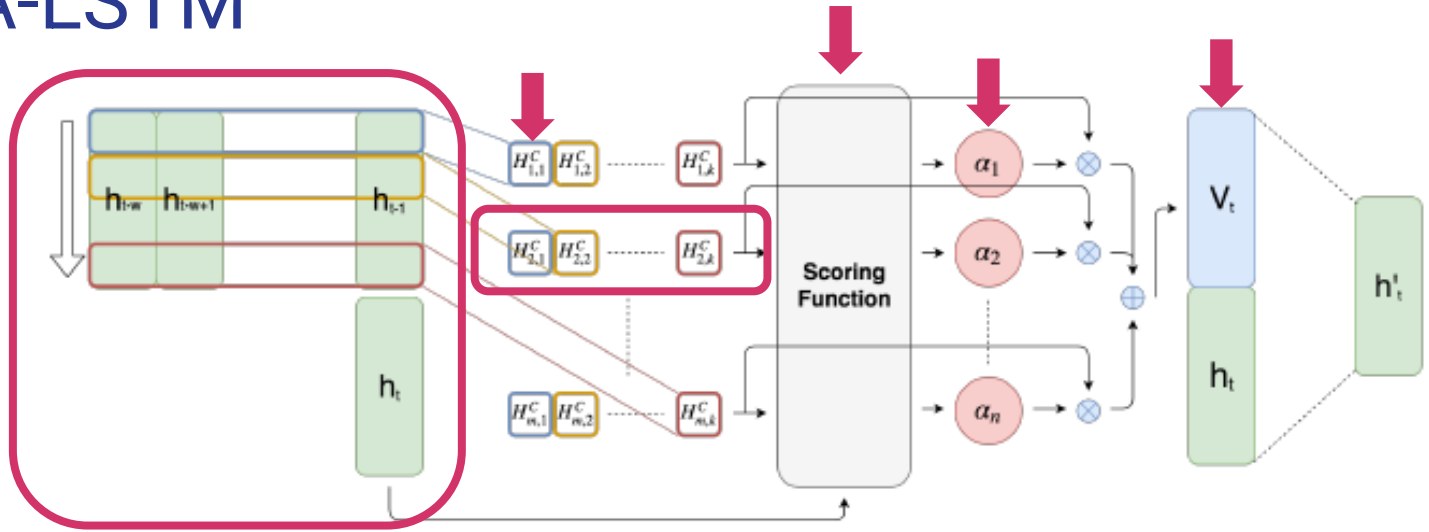
TPA-LSTM:
Capture relevent variables



Conventional (e.g. LSTNet):
Capture relevant time steps



TPA-LSTM



- RNN(H): Long term dependency of TS
- CNN(C): k 1-D filters for frequency domain information of TS
- $H^C = HC^T$: Convolution values
- $H^C_{i,j}$: Convolution value of TS i and filter j
- Scoring function: Relevency between H^C_i and the target

Results

Dataset	Model	*RSE*	*RAE*	*CORR*
Solar	mWDN	**0.3923**	**0.2439**	**0.9250**
	LSTNet	0.4239	0.2844	0.9094
	TPA-LSTM	0.4160	0.2798	/
Traffic	mWDN	0.8800	0.8685	0.7687
	LSTNet	**0.5071**	**0.3404**	**0.8598**
	TPA-LSTM	0.4843	0.3310	/
Electricity	mWDN	0.1668	0.1267	0.8655
	LSTNet	**0.0995**	**0.0542**	**0.9049**
	TPA-LSTM	0.1017	0.0621	/
Exchange rate	mWDN	0.1060	0.1018	0.8923
	LSTNet	**0.0357**	**0.0296**	**0.9538**
	TPA-LSTM	0.0317	0.0258	/

Table 2: Performance of time series models estimated on four datasets. *CORR for TPA-LSTM is not reported since TPA-LSTM was implemented with a univariate assumption and the comparison between CORR of multivariate and univariate models may not be fair.*

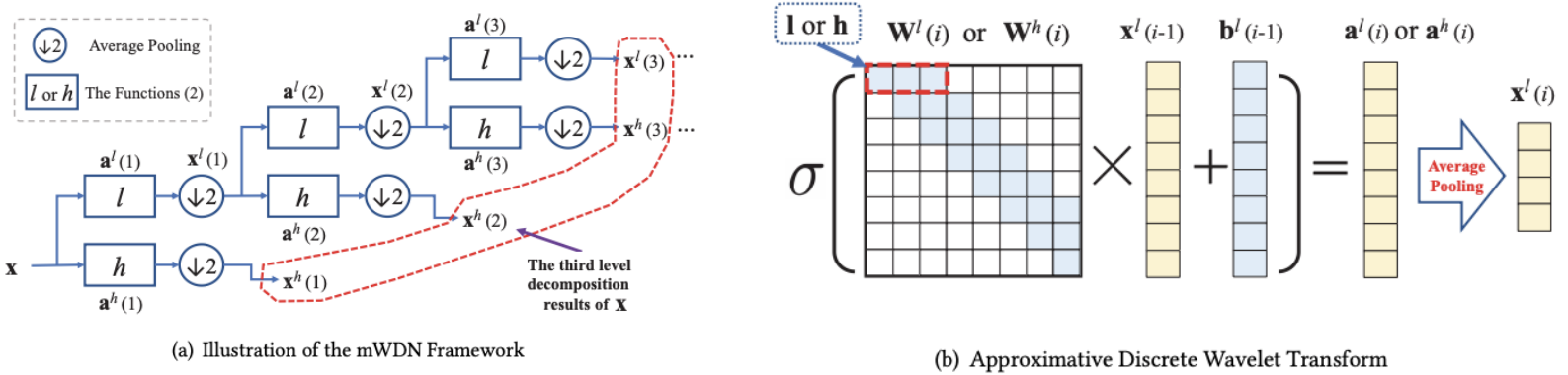
mWDN[1] - Intuition

An overview of current time series neural networks:

- Time domain methods
  - e.g. RNN, LSTM
  - taking a time series of data points as input
    -> better performance by modeling correlations
- Time + frequency domain methods
  - e.g. mWDN
  - decomposition into high-frequency and low frequency sub-series

[1] Wang, J., Wang, Z., Li, J., & Wu, J. (2018, July). Multilevel wavelet decomposition network for interpretable time series analysis. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2437-2446).

mWDN[1] - Architecture



(a) Illustration of the mWDN Framework

(b) Approximative Discrete Wavelet Transform

Figure 1. The mWDN framework.

[1] Wang, J., Wang, Z., Li, J., & Wu, J. (2018, July). Multilevel wavelet decomposition network for interpretable time series analysis. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2437-2446).

mWDN[1] - Performance

Dataset	Model	*RSE*	*RAE*	*CORR*
Solar	mWDN	**0.3923**	**0.2439**	**0.9250**
	LSTNet	0.4239	0.2844	0.9094
	TPA-LSTM	0.8408	5.7878	/
Traffic	mWDN	0.8800	0.8685	0.7687
	LSTNet	**0.5071**	**0.3404**	**0.8598**
	TPA-LSTM	1.6133	4.3277	/
Electricity	mWDN	0.1668	0.1267	0.8655
	LSTNet	**0.0995**	**0.0542**	**0.9049**
	TPA-LSTM	0.1487	1.8715	/
Exchange rate	mWDN	0.1060	0.1018	0.8923
	LSTNet	**0.0357**	**0.0296**	**0.9538**
	TPA-LSTM	0.0982	0.3127	/

Table 2: Performance of time series models estimated on four datasets. *CORR for TPA-LSTM is not reported since TPA-LSTM was implemented with a univariate assumption and the comparison between CORR of multivariate and univariate models may not be fair.*

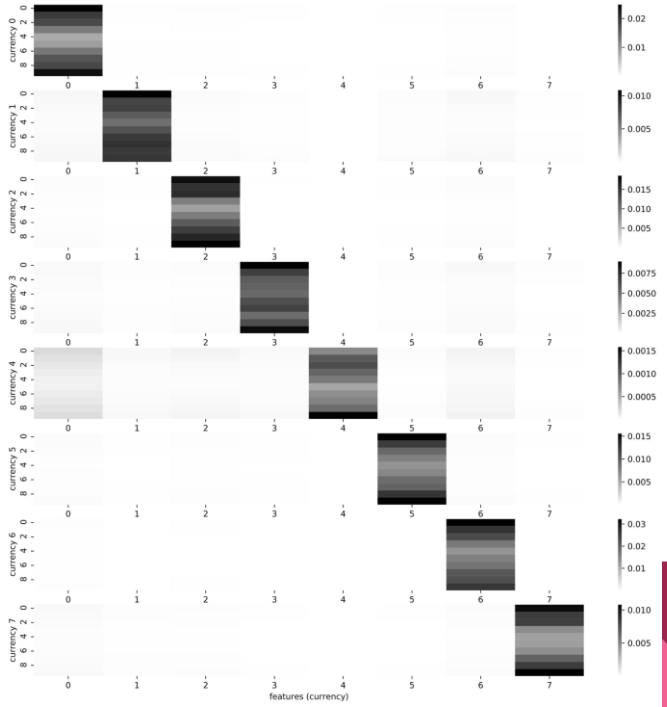
[1] Wang, J., Wang, Z., Li, J., & Wu, J. (2018, July). Multilevel wavelet decomposition network for interpretable time series analysis. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2437-2446).

SHAP Explanations

SHAP on LSTNet (Exchange Rate Dataset)

SHAP values are generated for each feature (currency) in the first 10 samples (timepoint).

The heatmap shows the sum of absolute SHAP values in all data points used in the model, for each feature. Darker color in corresponding feature (x axis) means this feature are more important in predicting the corresponding label (y axis).

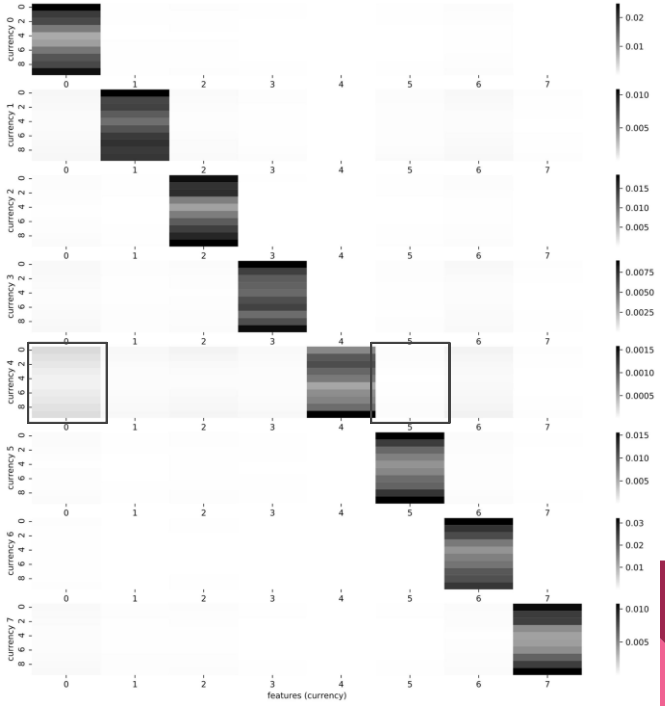


Evaluation of Explainable Model on LSTNet

Two features are chosen to evaluate the explainable model. The most important one (except the label itself) is currency 0 and the less important one is currency 5.

Evaluation Dataset 1:
All values in currency 0 <- the mean of currency 0
Resulting CORR: 0.8950 -> 0.8905

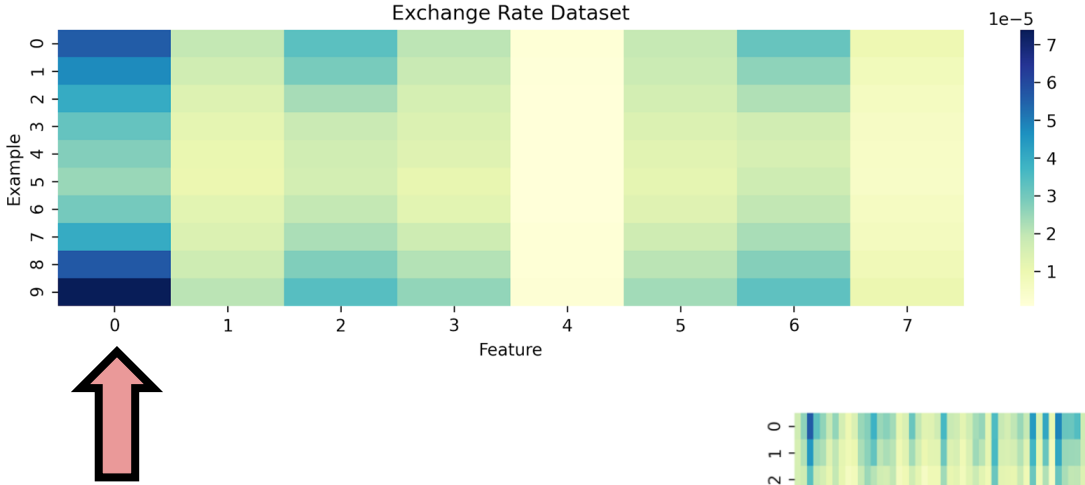
Evaluation Dataset 2:
All values in currency 5 <- the mean of currency 5
Resulting CORR: 0.8950 -> 0.8961

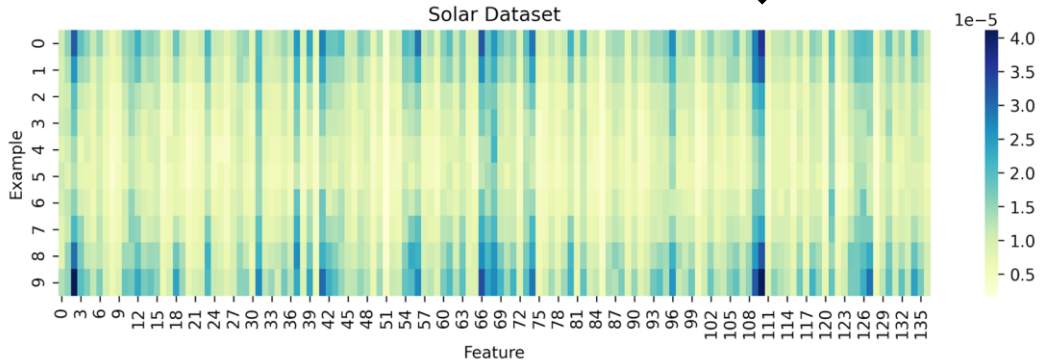


SHAP on mWDN (1. Features)

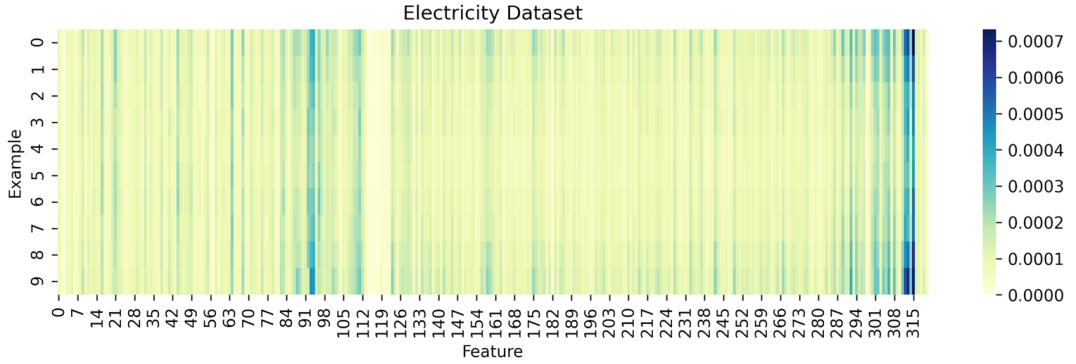
- Objective: Identify the most influential features for each example
- Methodology:
  - Each time point of each feature of each example has an individual SHAP value (168 x n x 10 SHAP values)
  - Sum up all SHAP values of each time point of a given feature of a given example (n x 10 Sums of SHAP values)
  - Visualize and analyze using a heatmap

SHAP on mWDN (1. Features)





SHAP on mWDN (1. Features)



SHAP on mWDN (1. Features)

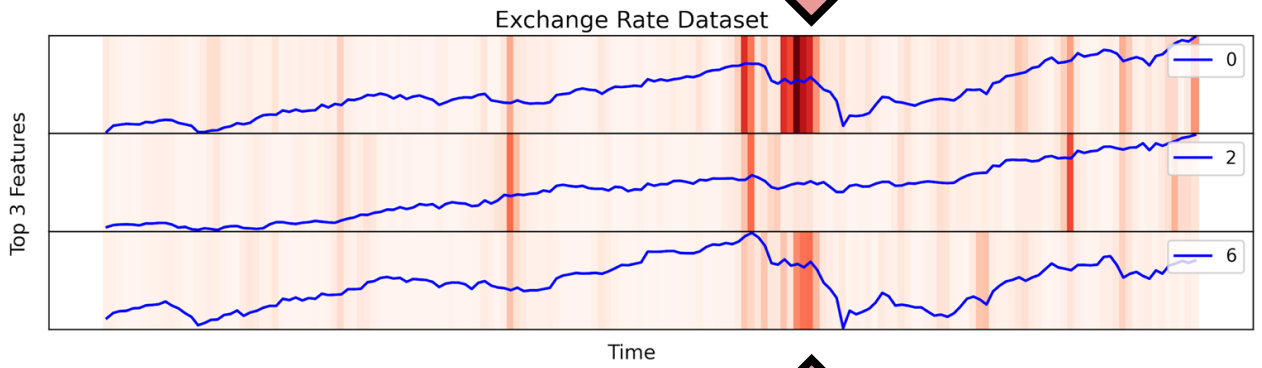
- Main Observation

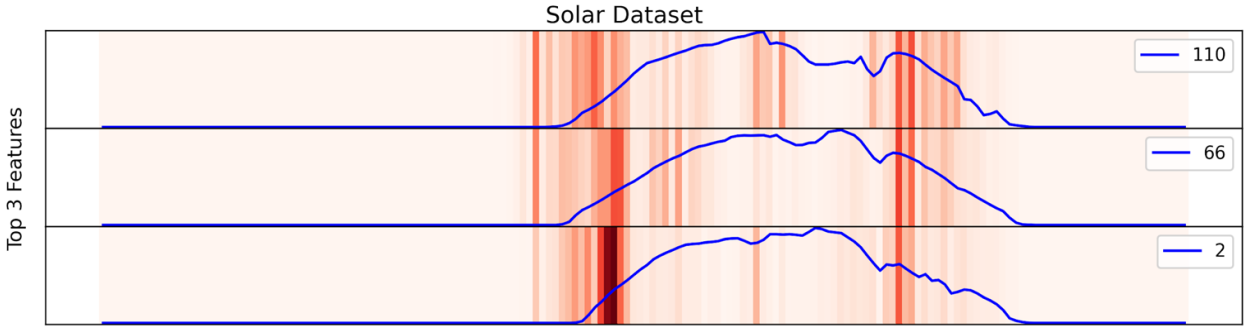
Only a small number of features
are truly influential for a given prediction

SHAP on mWDN (2. Time Periods)

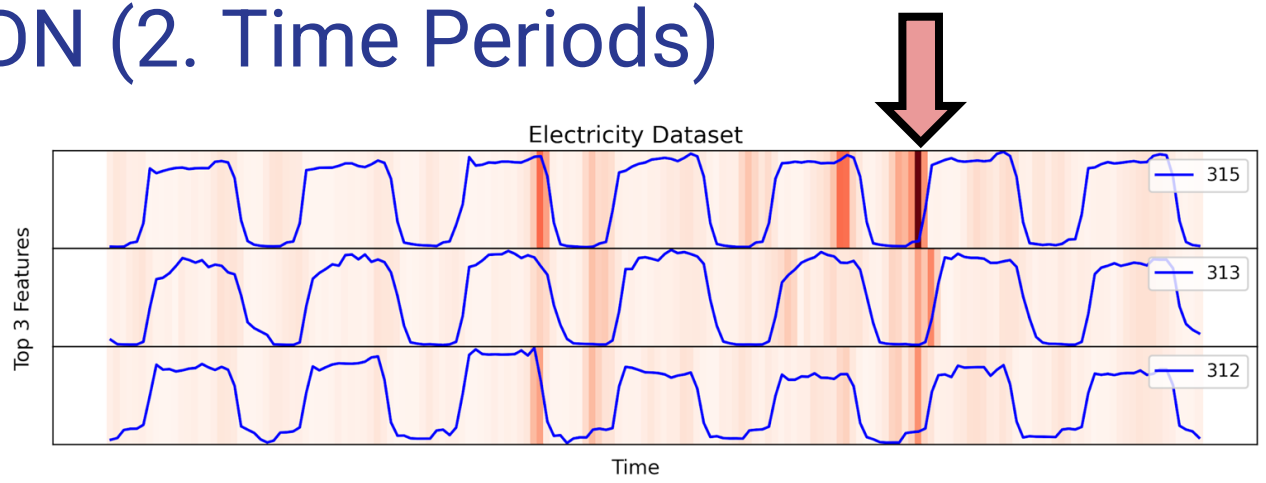
- Objective: Identify the most influential time periods for an example
- Methodology:
  - Each time point of each feature of one example has an individual SHAP value (168 x n x 1 SHAP values)
  - Select only the top 3 most influential features using the previous analysis (168 x 3 x 1 SHAP values)
  - Visualize and analyze these SHAP values using a heatmap-line-graph hybrid

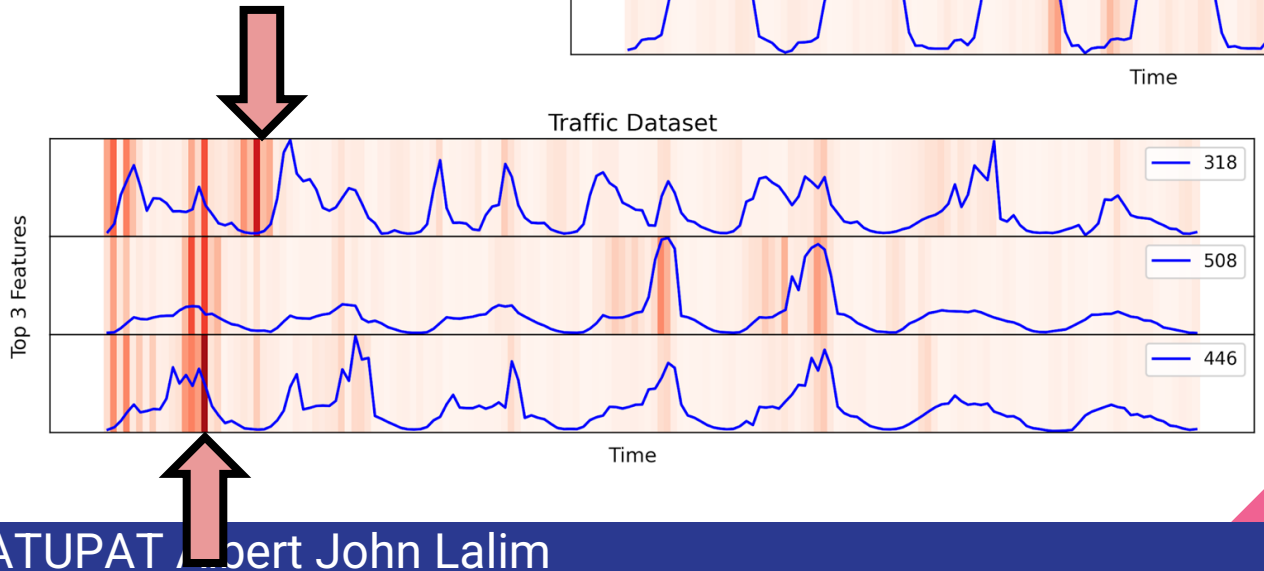
SHAP on mWDN (2. Time Periods)





SHAP on mWDN (2. Time Periods)





SHAP on mWDN (2. Time Periods)

- Main Observation

Only a small number of time intervals
are truly influential for a given prediction

SHAP on mWDN (3. SHAP for Prediction)

- Objective: Analyze sensitivity of the SHAP values
- Methodology:
  - Each time point of each feature of one example has an individual SHAP value (168 x n x 1 SHAP values)
  - Select the top 10% most influential data points and replace each with the mean of its 25-length window (smoothened input)
  - Compare error of original and smoothened input

SHAP on mWDN (3. SHAP for Prediction)

Dataset	Exchange Rate	Solar	Electricity	Traffic
Average % Difference	1.8195%	-0.1079%	0.8653%	7.5525%

- Assumptions of SHAP are incompatible with time series data
- Implemented mWDNs performed poorly due to size limits
- Smoothing method for destroying patterns is not effective

SHAP on mWDN (3. SHAP for Prediction)

- Main Observation

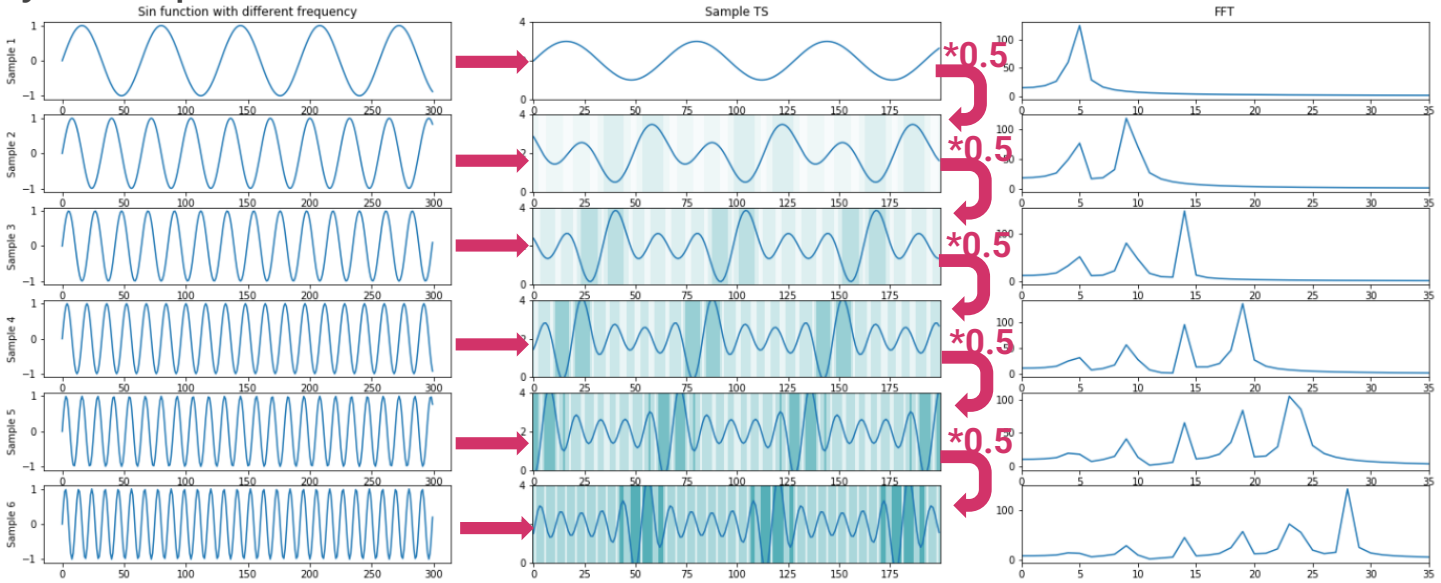
SHAP values are not always feasible
for time series forecasting models

SHAP on TPA-LSTM

- SHAP does not support Tensorflow 1.9 without Keras
- To implement SHAP
  - Rewrite the current implementation for Pytorch or Keras
  - Write an entirely new implementation, which is more compatible

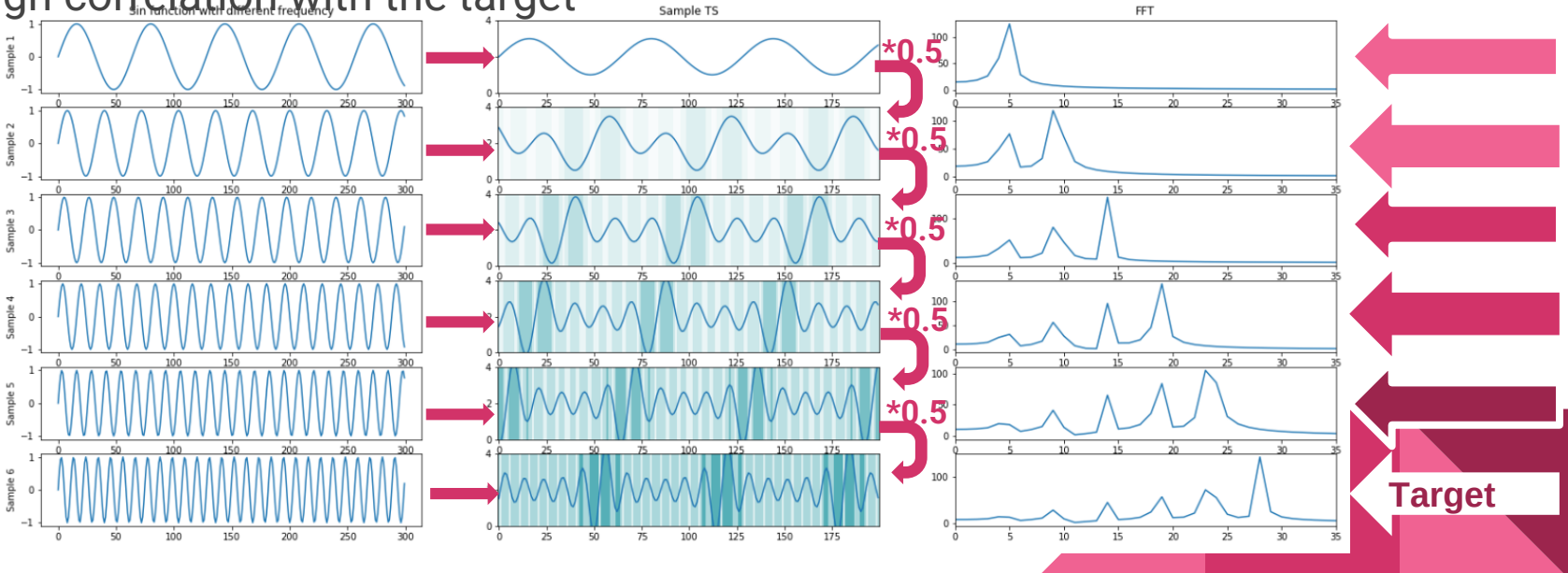
SHAP on TPA-LSTM (Expected)

- Toy Example



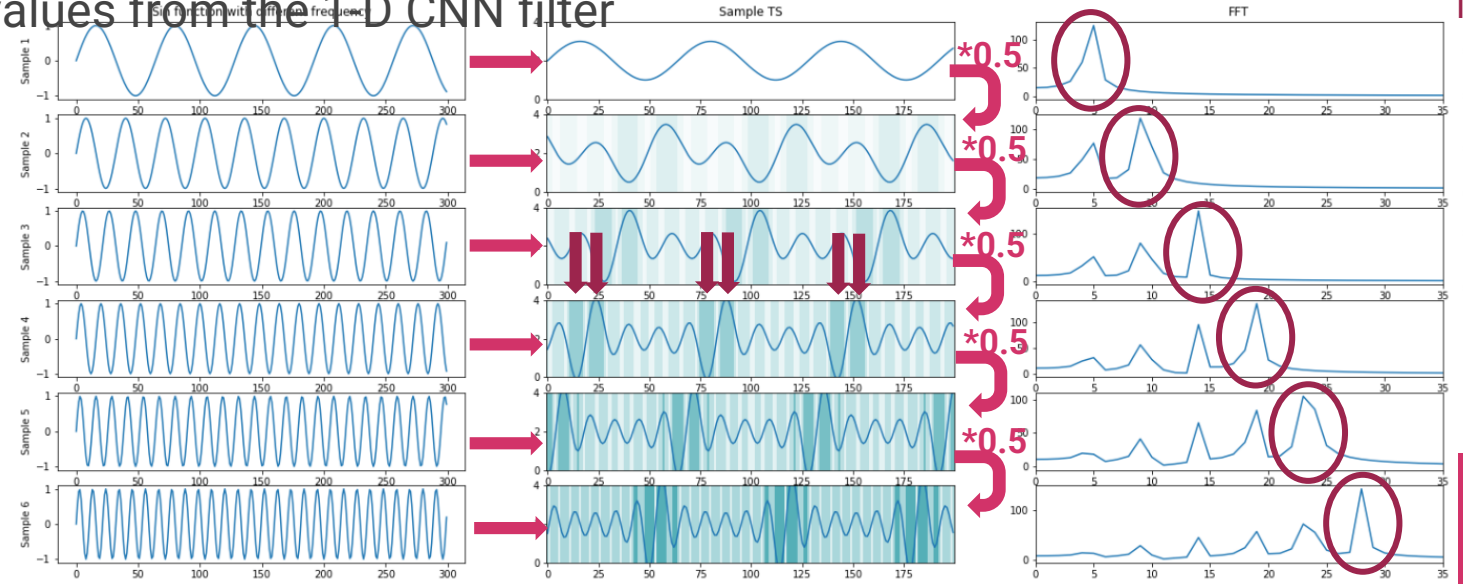
SHAP on TPA-LSTM (Expected)

- To select a feature: TPA-LSTM captures the frequencies of features having high correlation with the target



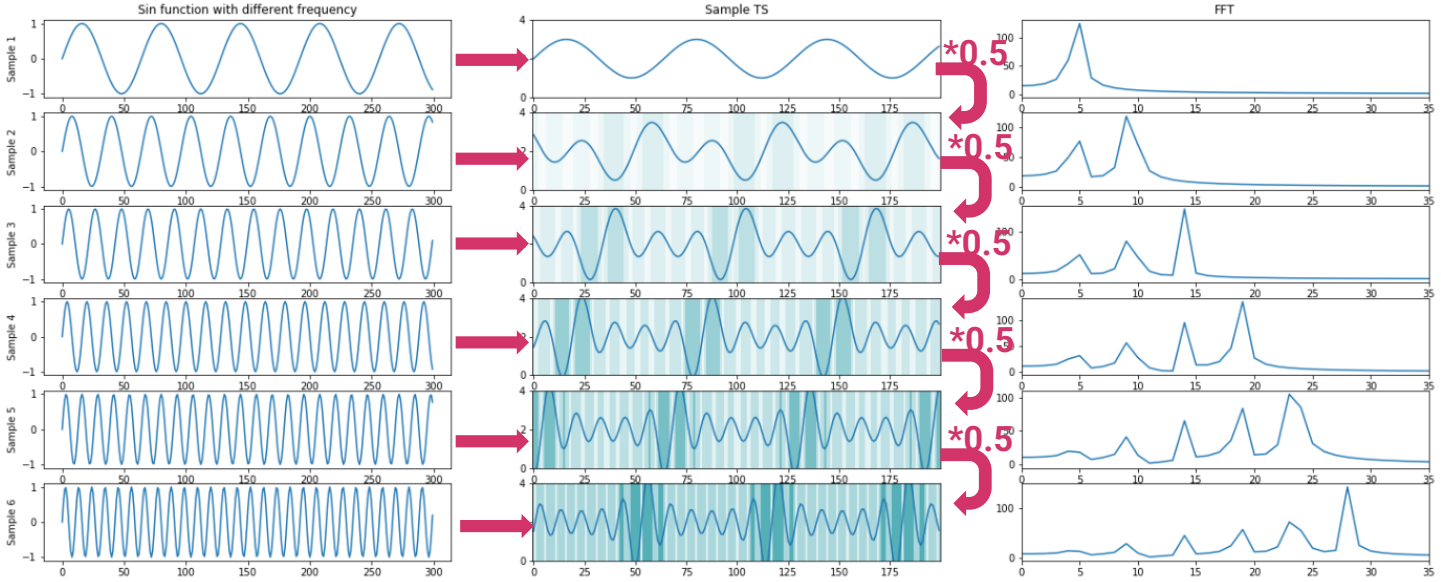
SHAP on TPA-LSTM (Expected)

- To select a time period for each selected feature: Consider the convolution values from the 1-D CNN filter



SHAP on TPA-LSTM (Expected)

- Shifted time frame would not affect



Conclusion

- Implemented, trained, analyzed three multivariate time series model
- Implemented SHAP on the models, analyzed its ability
- Results suggest a research gap in explainable time series models