SPASE: SPATIAL SALIENCY EXPLANATION FOR TIME SERIES MODELS

Pranay Lohia[*], *Badri Narayana Patro*[*], *Naveen Panwar*[*], *Vijay Agneeswaran*

Microsoft

ABSTRACT

We have seen recent advances in the fields of Machine Learning (ML), Deep Learning (DL), and Artificial intelligence (AI) that the models are becoming increasingly complex and large in terms of architecture and parameter size. These complex ML/DL models have beaten the state of the art in most fields of computer science like computer vision, NLP, tabular data prediction and time series forecasting, etc. With the increase in models' performance, model explainability and interpretability has become essential to explain/justify model outcome, especially for business use cases. There has been significant improvement in the domain of model explainability for Computer Vision and Natural Language Processing (NLP) tasks with fundamental research for both black-box and white-box techniques. In this paper, we proposed novel time series explainability techniques SPASE for black-box time series model forecasting and anomaly detection problems.

Index Terms**—** Spatial, Saliency, Explainability, Time-Series, Quantile Density

1. INTRODUCTION

Machine learning models for time series tasks, such as forecasting and anomaly detection, present unique explainability challenges due to the distinct nature of time series data. Existing techniques like LIME, SHAP, and CAM are not specifically designed for time series data, and their direct application may not provide intuitive explanations. While these post-hoc techniques offer black-box (LIME, SHAP) and white-box (CAM) explanations, they neglect considerations for explainability during model building. Additionally, techniques like DeepLIFT [1] and Integrated Gradient [2], which utilize gradient maps to explain salience regions, are limited to CNN models. Existing spatial saliency techniques from computer vision have their limitations when applied to time series data.

Time series data explanations can be less intuitive due to isolated data points and varying model architectures. Thomas et al. [3] discussed vision and language model explanation methods applicable to time-series data, but didn't provide quantitative results or comparisons. Theissler et al. [4] presented a categorization of existing explanation techniques without introducing new ones. To overcome these limitations, we propose SPASE (Spatial Saliency Explanation) technique, which provides spatial saliency over quantile regions. Its detailed overview is in the methodology section. The experiments on time series explanation for ADaaS [5]/Azure Core Workload Insights data and Electricity data are also included. Details on prior art are in the related work section. The term 'features' is used interchangeably with 'time stamps' in this context. The SPASE model's operation is demonstrated in Figure 1.

2. RELATED WORK

Interpretability and explainability of machine learning models are essential, especially in CV, NLP, and tabular data fields. This section discusses key techniques for improving time series data explainability. LIME is a local explainability method that highlights feature importance for class prediction using linear approximation [6]. Being model-agnostic, LIME only needs the prediction function, not the model's internal details. SHAP assigns 'importance' to features based on their impact on model outputs [7]. It uses local and global perspectives to detail feature importance. CAM localizes image objects based on class labels using the global average pooling layer on Convolution Neural Networks [8]. Grad-CAM and Grad-CAM++ are CAM extensions that provide detailed explanations for visual objects based on class labels [9, 10, 11].

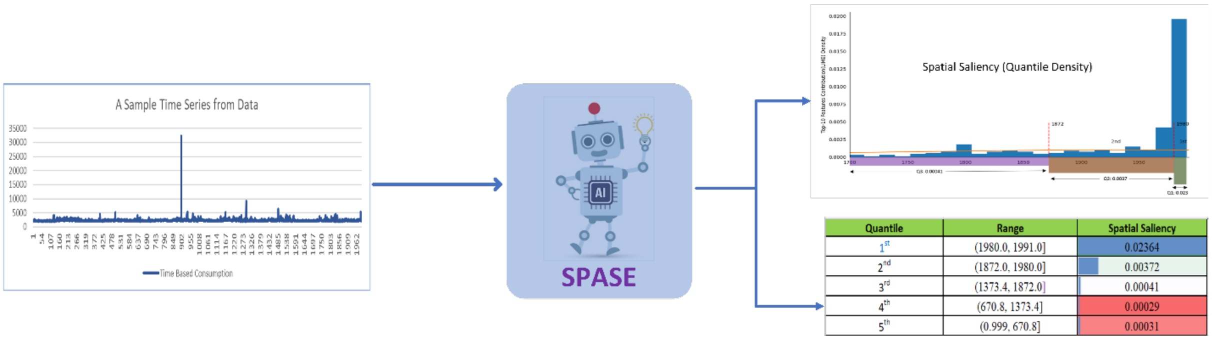
3. METHOD

In this section, we describe our approach Spatial Saliency Explanation aka SPASE for time series data. We specifically focused on black box models for an explanation as time series models can be built upon multiple architectures like RNN, LSTM, CNN, or transformer-based networks. In SPASE, we propose an explanations technique that is model agnostic and treats model prediction as an output from the black box.

3.1. SPASE: Spatial Saliency Explanation

SPASE is a post-hoc, black-box, and global model explanation technique, where it generates a global explanation for model-trained time series data for forecasting. Below are a few nomenclatures or notations, which we will be using to describe the SPASE technique.

[*] Equal Contribution



Quantile	Range	Spatial Saliency
1st	(1980.0, 1991.0]	0.02364
2nd	(1872.0, 1980.0]	0.00372
3rd	(1373.4, 1872.0]	0.00041
4th	(670.8, 1373.4]	0.00029
5th	(0.999, 670.8]	0.00031

Fig. 1. For a given input time series sample the SPASE model shows important features/ time stamps (Ranges: 1980-1991), which are the end of the series. It also shows the density score for different quantiles and its range. The Red color in the table shows less importance and blue indicates more importance feature range and its quantile

$$
\begin{aligned}
M(X, y) &= (R^{N\times D}, R^{N}) \\
X &= \text{Training data} \\
N &= \text{Number of data points(time series)} \\
D &= \text{Number of feature(time stamps)} \\
T &= [T^{0}, T^{1}, \dots T^{R}], \text{ where } T \text{ is subset of } X \\
T^{i} &= (R^{1\times D}, R^{1}), \text{ a data-point from } T \subseteq X
\end{aligned}
\tag{1}
$$

SPASE backbone is two important components Spatial and Saliency, and both these are important aspects of time series explanation.

1. Spatial: it refers to a set of features which are in proximity of each other and typically mean continuous regions/ranges in time series.
2. Saliency: it refers to the contribution of a certain range or points toward the model prediction i.e., time series forecasting.

Spatial saliency for time-series data provides us with important insights about feature ranges that are contributing most towards the time series forecasting in the given model. For SPASE explanation, we first get the individual token-based saliency for all the time series/data points in. To get the individual token-based saliency, we can leverage any black box token-based saliency method from prior work like LIME or SHAP. In this paper, we used LIME [6] to achieve token-based saliency. Token-based saliency time stamps for time series can be denoted as $L_K(T^i)$. We get each time series from lime and filter out the top-k salient feature/time stamp, we chose K = 10 for our experiment, this can be a user-defined parameter as well. contains the list of all top contributing features for the model prediction.

$$
\begin{aligned}
L(T_D^{i}) &= [l_0, l_1, \dots l_D] \\
L_K(T^{i}) &= \text{Top K salient token from } L(T_D^{i})
\end{aligned}
\tag{2}
$$

3.2. Quantile Density

We derived the spatial saliency from a sampled dataset (T) using the quantile-based density estimation for top contributing features $L_K(T)$. To get the quantile density for each quantile range of time series features, we first generate the frequency distribution histogram and the quantile range for $L_K(T)$. Then for each quantile, we calculate the density of feature contributions towards the model prediction. We consider quantile ranges and their density as Spatial regions in time series and their saliency.

$$
\begin{aligned}
Hist(L_K(T), N_b) &= [h_1, h_2, \dots h_{(N_b)}] \\
Q(L_K(T), N_Q) &= [q^1, q^2, \dots q^{(N_Q)}] \\
q^{i} &= (q^{i}_{low}, q^{i}_{high}) \\
H^{q^{i}} &= [\text{List} \quad \text{of} \quad \text{h}]_i \in q^{i}
\end{aligned}
\tag{3}
$$

N_b and N_Q are a number of bins and quantiles, these two can be user-defined parameters depending upon how wide and deep an explanation they are looking for. In our experiments, we choose N_b=150 and N_Q=5. q^i here is a tuple denoting quantile or spatial range of time stamps or contributing features. The spatial saliency explanation $SS(q^i)$ for q^i quantile is derived from the equation below.

$$
\begin{aligned}
SS(q^{i}) &= \frac{H^{q^{i}}}{R * (q^{i}_{high} - q^{i}_{low})} \\
SS &= [SS(q^{1}), SS(q^{2}) \dots\dots SS(q^{N_q})]
\end{aligned}
\tag{4}
$$

Here, SS represents the spatial saliency for a given dataset and R denotes the total number of data points/time series in Set T. Quantile density as saliency acts as a good explanation for users about what continuous ranges of time stamps/features contribute and by how much to the model prediction. An example of a spatial saliency explanation is provided in the Results section. Table 1 shows the quantile density as per

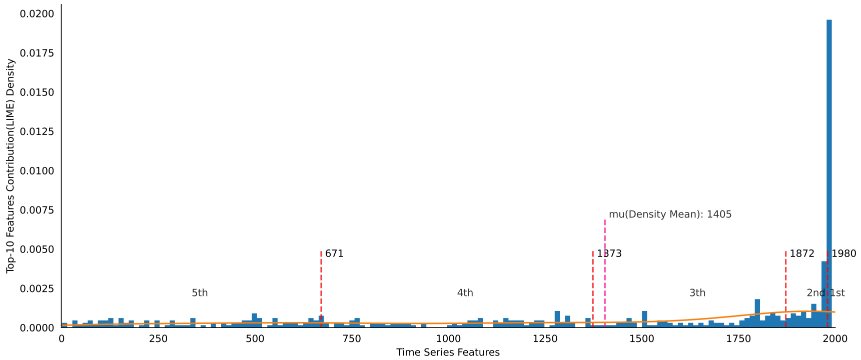


Fig. 2. LIME-based SPASE explanation for Azure Core dataset. It provides a global explanation for the model and, the top contributing time-stamps and their distribution. We can see quantiles for top contributing time-stamps are skewed toward the prediction point(right-side) with mean density at 1405. The first two quantiles at 1980 and 1872 provide insights that most recent time-stamps are contributing most for prediction point(1993)

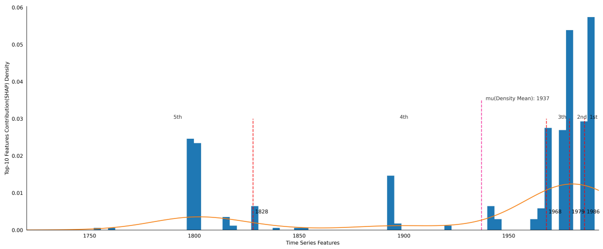


Fig. 3. SHAP-based SPASE explanation for Azure Core dataset. The top contributing feature and their distribution provide similar insights as LIME. They are skewed toward the prediction point(right-side) with mean density at 1937 more than LIME and top contributing features are more isolated and only occur after timestamps 1750. A more detailed analysis is provided in the results section.

equation (1) and a more detailed overview of SPASE results is provided in the next section.

4. EXPERIMENTS AND RESULTS

4.1. Dataset

4.1.1. Azure Core Workload Insights Data

We utilized Azure Core workload insights data collected by the Mario service from Azure Monitor insights for our algorithm development and proof-of-concept testing. The time-series data, stored in an optimized database, includes metrics describing system aspects at specific time points. We trained our anomaly detection model on a week's 'Availability' metrics data consisting of around 16k time series, each with 1992 data points. We ran validation of our results on the 15k time-series inference data hourly.

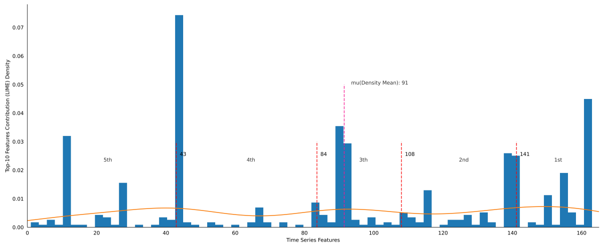


Fig. 4. LIME-based SPASE explanation for Electricity dataset. The top contributing features mean density is 91 and they are homogeneously distributed across all timestamps, unlike the explanation for the Azure Core dataset where the top contributing features were skewed towards the right side.

4.1.2. Electricity Data

The UCI Electricity Load Diagrams Dataset [12], containing the electricity consumption of 370 customers – aggregated on an hourly level. We use the past week (i.e., 168 hours) to find anomalies over the last 24 hours.

4.2. Results

We demonstrated SPASE's explainability using two datasets, Electricity and Azure Core, and two token-based black-box techniques, LIME and SHAP. We identified the top-10 contributing features from 50 random training data samples. Histogram-based frequency distributions of these features are visualized in Figures 2 and 3, and quantile wise density values are presented in Table 1. A direct comparison of LIME's explanation on Azure Core data is in Table 3. Further subsections will discuss these results and explanations.

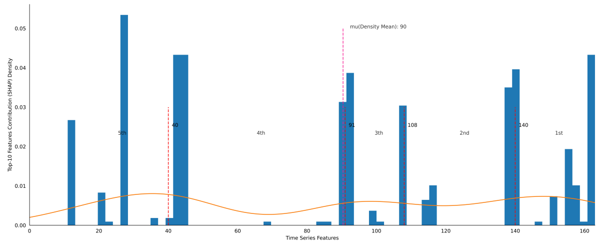


Fig. 5. SHAP-based SPASE explanation for Electricity dataset. Top contributing features mean density is 90 and show similar distribution as LIME but they are more sparse and isolated. A more detailed analysis is provided in the results section.

# Quantile	LIME		SHAP	
	Range	Saliency	Range	Saliency
1st	(1980.0, 1991.0]	0.02364	(1986.2, 1991.0]	0.06167
2nd	(1872.0, 1980.0]	0.00372	(1979.0, 1986.2]	0.03889
3rd	(1373.4, 1872.0]	0.00041	(1968.0, 1979.0]	0.03363
4th	(670.8, 1373.4]	0.00029	(1828.0, 1968.0]	0.00174
5th	(0, 670.8]	0.00031	(0, 1828.0]	0.00012

Table 1. Spatial Saliency explanation for Azure Core dataset as time series quantile (feature ranges) and their importance values as estimated density for token-based techniques from LIME and SHAP

4.2.1. Azure Core Workload Insights Data Explanations

Figure 2 displays the frequency distribution of LIME-based explanation for Azure Core workload data, with the mean density situated around 1405 features, indicating crucial end-of-series features. Table 1 presents five quantiles, their feature ranges, and spatial saliency scores, revealing that feature ranges in quantiles 1 and 2 are more crucial for future predictions while 4 and 5 contribute less. Figure 3 illustrates these findings and notes a maximum density of features at the series end, further supported by SHAP. SHAP's mean density for top contributing features is 1937, higher than LIME's 1405, showing a skew towards the end with the 4th quantile's lower bound at 1828 compared to LIME's 670.

4.2.2. Electricity Data Explanations

We used the publicly available Electricity dataset 4.1.2 to demonstrate SPASE's generalization and extensibility. Figures 4 and 5 present spatial saliency-based explanations. The mean density of top contributing features using LIME is 91, indicating recurrent top contributors in each quantile interval and reflecting the non-stationary nature of the series. This seasonality among top contributing features/time stamps is similarly exhibited by SHAP-based spatial saliency, with a mean density of top contributing features at 90. The 3rd quantile, with a lower width and high saliency value, contributes more to the forecasting than other quantiles as per Table 2.

# Quantile	LIME		SHAP	
	Range	Saliency	Range	Saliency
1st	(141.2, 163.0]	0.00936	(140.0, 163.0]	0.01479
2nd	(108.0, 141.2]	0.00645	(108.0, 140.0]	0.00825
3rd	(83.6, 108.0]	0.00909	(91.0, 108.0]	0.01342
4th	(43.0, 83.6]	0.00626	(40.0, 91.0]	0.00687
5th	(0, 43.0]	0.00609	(0, 40.0]	0.00505

Table 2. Spatial Saliency explanation for Electricity dataset as time series quantile (feature ranges) and their importance values as estimated density for token-based techniques from LIME and SHAP

LIME Top-5	Feature	Importance
1st	1991	2.6706
2nd	1979	0.3905
3rd	1647	0.2995
4th	288	0.2529
5th	1087	0.2466

Table 3. Example of LIME-based feature importance, as we can see token-based (individual features) importance for time series explanation does not provide much information than isolated points and their importance value

4.3. Validation and Benefit Analysis

We validated SPASE using a human evaluation study with 30 annotators, achieving 93% Precision and 86% Recall for identifying important regions in a 5k time-series. Post-SPASE integration, Azure MLaaS reported a 20% sales increase in time-series model-based products, Azure Core Workload insights users reported better outcome understanding, and user interaction on the Azure Core workload insights dashboard increased by 40%. This led to a 25% retention increase in Azure Core workload insights products.

5. CONCLUSION

Using our approach SPASE, we can provide regions/ranges in the time-series data and reasons for model prediction in Anomaly detection as a service and Azure Core workload insights fault detection service. We are providing time-series explainability as a spatial saliency feature in production for multiple Azure MLaaS. Apart from these services, time-series explainability has wider implications to any model service based on time-series data. Interpretability/Explainability is an important aspect needed to provide transparency of outcomes, especially in applications like healthcare, manufacturing, electronics, etc., where the use of time-series data is prevalent.

6. REFERENCES

[1] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje, “Learning important features through propagating activation differences,” in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.

[2] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, “Axiomatic attribution for deep networks,” in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[3] Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez, “Explainable artificial intelligence (xai) on timeseries data: A survey,” *arXiv preprint arXiv:2104.00950*, 2021.

[4] Andreas Theissler, Francesco Spinnato, Udo Schlegel, and Riccardo Guidotti, “Explainable ai for time series classification: A review, taxonomy and research directions,” *IEEE Access*, 2022.

[5] C S Krishna, Swarnim Narayan, Sourav Khemka, Ivan Barrientos, Vijay Agneeswaran, and R Kiran, “Probabilistic time-series forecasting with deep autoregressive flow models,” in *In Machine learning, AI and Data Science conference*. MSJAR, 2022.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, “” why should i trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[7] Scott M Lundberg and Su-In Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.

[8] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[9] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[10] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[11] Badri N. Patro, Mayank Lunayach, Shivansh Patel, and Vinay P. Namboodiri, “U-cam: Visual explanation using uncertainty based class activation maps,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[12] Artur Trindade, “ElectricityLoadDiagrams20112014,” UCI Machine Learning Repository, 2015, DOI: https://doi.org/10.24432/C58C86.