SPASE: SPAtial Saliency Explanation for time series models in Microsoft Azure products

Pranay Lohia (✉ pranaylohia@microsoft.com)  
Microsoft (India)

Badri Narayana Patro  
Microsoft (India)

Naveen Panwar  
Microsoft (India)

Vijay Agneeswaran  
Microsoft (India)

Research Article

Keywords: Spatial, Saliency, Time-Series, Explainability, Quantile Density

Posted Date: August 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3290148/v1

License: ☕️ 🔖 This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License

Additional Declarations: No competing interests reported.
SPASE: SPAtial Saliency Explanation for time series models in Microsoft Azure products

Pranay Lohia¹, Badri Narayana Patro¹, Naveen Panwar¹, Vijay Agneeswaran¹

¹Microsoft, Bengaluru, India.

Contributing authors: pranaylohia@microsoft.com; badripatro@microsoft.com; naveenpanwar@microsoft.com; vagneeswaran@microsoft.com;

Abstract

We have seen recent advances in the field 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. We also outline details of its implementation and integration into Microsoft Azure products.

Keywords: Spatial, Saliency, Time-Series, Explainability, Quantile Density

1 Introduction

As per our business use case, we have a Machine Learning / Deep Learning model built for different time series tasks i.e., time series forecasting and anomaly detection, etc. The nature of time series data is completely different from the image and textual data. Explainability or interpretability for time series model predictions involved challenges like identifying the sequence of time stamps contributing to the model prediction and existing techniques (LIME/SHAP/CAM) are under-explored for such time series use-case. There aren’t any specific methods for Time series explanation. Hence, we explored and tried to address the gaps and drawbacks of existing techniques for the time series explanation. Explainability for the AI models has been explored in different aspects of model prediction and choice of the model architecture.

Most model explanation techniques are post-hoc explanations, meaning these techniques will provide explanations on prediction after the model is trained on data and there isn’t any consideration for explainability while building the model. In the domain of post-hoc techniques, there are two model explanations 1) Black-box model explanation examples: LIME and SHAP 2) White box techniques model example: Class Activation Map (CAM) based techniques. A detailed overview of these is provided in Related Work section. Here we discuss the disadvantages or drawbacks of using these techniques directly for time series data.
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 quantile and its range. The Red color in the table shows less importance and blue indicates more importance feature range and its quantile.

- **Black Box or Token based (Example: Word important in a sentence):** In the time series data having few isolated points that are contributing to prediction are less intuitive for explanation from user perspective. It’s not much of an explanation for the user unlike text data where individual tokens are words. There are a few perturbation based black box techniques like DeepLIFT [1] and Integrated Gradient [2], which usage gradient map in image space to explain salience regions for CNNs' model only.

- **White Box/Gradient Map/Class Activation Map (Example: spatial region in Image):** We can use existing spatial saliency techniques from computer vision, but existing techniques are used for white box models where gradient maps overlap with regions in images.

As for time series data, having isolated data points(tokens) as explanation are less intuitive than having individual feature contribution in tabular data and models’ architecture also varies from problem to problem. Thomas et al. [3] has discussed various model explanation methods from vision and language domain, which can be applied to time-series data. But in this survey paper they haven’t shown any quantitative results or comparisons among different explanation methods applied on time-series corpus. Similarly, Theisler et al. [4] has discussed the existing explanation system (XAI) from different domain like: textual and image from time series data perspective and it has presented categorization and taxonomy of existing explanation techniques, not any new/novel technique for time series model explanations. To address the above-mentioned drawbacks, we proposed SPASE aka Spatial Saliency Explanation technique.

SPASE is built upon the black-box model explanation technique, and it provides spatial saliency over the quantile regions to better under the model prediction. A detailed overview of the technique is provided in the methodology section. In the paper, we have also provided details on results and experimentation on time series explanation for Anomaly Detection as a service (ADaaS) [5] for Azure workload. Prior art and existing techniques details can be found in the related work section. For ease of understanding, we used features and time stamps interchangeably, for time series data they both meant the same. We explain our method using this flow diagram as shown in Figure 1. Our SPASE model takes as input a sample time series from the data and provides important feature ranges and its spatial saliency score as shown in the figure.

2 **Related Work**

Explainability and interpretability for machine learning models have been explored in the domain of computer vision (CV), natural language processing (NLP) and tabular data, etc. In this section, we will discuss a few of the prior works or techniques to better understand the explainability of time series data.

LIME (Black-box model, [6])- Local Interpretable Model-agnostic Explanations) is a state of art local explainability method that identifies the importance of features for class prediction using the linear approximation method. For every data point, it creates multiple test points around the original data point in the data space by perturbing it. It then fits a linear model (e.g., linear regression model) on the data points and then
identifies the features’ ‘importance’ of this linear model as local explainability of the original data point. It is a model agnostic explainability method as it only needs the prediction function of the model and not its internal details of it.

SHAP (Black-box model, [7]) is a game theoretic approach to explain the output of a machine learning model. It connects optimal credit allocation with local explanations using the classic ‘Shapley’ values from various extensions of game theory to compute the importance. SHAP assigns each feature an ‘importance’ value for a particular prediction. The general idea is that the most relevant features are the ones that result in the largest change to the model outputs when we remove them. We apply SHAP-based analysis from both local and global perspectives to provide a comprehensive view. Its main advantages include 1) Identification of a new class of additive feature-importance measures. 2) Theoretical results depict a unique solution in each class with a set of desirable properties.

CAM: Class Activation Map [8] is a method to localize objects in an image based on its class label. This method is a white box explanation technique that uses the global average pooling layer on Convolution Neural Network to get the class contribution weights. The CAM method obtains a generic localized deep representation that captures the implicit attention of CNNs on an image. The class activation map for a particular category indicates that the discriminative image regions are used by CNN to identify the category. Grad-CAM [9] has proposed a gradient-based class activation map to explain visual objects based on the class label. Grad-CAM++ [10] based explanation method is used to capture and explain every instance of the object.

3 Method

In this section, we describe the proposed 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. Hence, in SPASE we proposed an explanations technique that is model agnostic and treat 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.

\[
M(X, y) = (R^{N \times D}, R^N)
\]

\[
X = \text{Training data}
\]

\[
N = \text{Number of data points (time series)}
\]

\[
D = \text{Number of features (time stamps)}
\]

\[
T = \{T^0, T^1, \ldots, T^R\}, \text{ where } T \text{ is a subset of } X
\]

\[
T^i = (R^1 \times D, R^i), \text{ a data-point from } T \subseteq X
\]

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 which 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. 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 filtered 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.

\[
L(T^i_D) = [l_0, l_1, \ldots, l_D]
\]

\[
L_K(T^i) = \text{Top } K \text{ salient token from } L(T^i_D)
\]
Fig. 3 LIME based SPASE explanation for Azure Core dataset. It’s provide global explanation for model and, top contributing time-stamps and their distribution. We can see quantiles for top contributing time-stamps are skewed toward prediction point(right-side) with mean density at 1465. First two quantile at 1980 and 1872 provide insights that most recent time-stamps are contributing most for prediction point(1993).

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.

$$Hist(L_K(T), N_b) = [h_1, h_2, \ldots h_{(N_b)}]$$

$$Q(L_K(T), N_Q) = [q^1, q^2, \ldots q^{(N_Q)}]$$

$$q^i = (q_{i,low}, q_{i,high})$$

$$H^q = \{\text{List of } h_i | i \in q^i\}$$

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

$$SS(q^i) = \frac{H_{q^i}}{R \cdot (q_{i,high} - q_{i,low})}$$

where $SS$ represents the spatial saliency for 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 spatial saliency explanation is provided in the Results section. Table 1 shows the quantile density as per equation (1)
and a more detailed overview of SPASE results is provided in the next section. Figure 1 articulates a detailed input-output system flow providing a snapshot of sample output highlighting saliency score explanation of different spatial region of input time-series.

4 Experiments and Results

4.1 Dataset

4.1.1 Azure Core Workload Insights data

We have used Azure Core workload insights data collected by their Mario service for our algorithm building and proof-of-concept testing. The data that Mario service collects from the customer tenant is from Azure Monitor insights. Data collected in Azure Monitor is stored in a time-series database that’s optimized for analyzing timestamped data. Metrics are numerical values that describe some aspect of a system at a particular point in time. They are collected at regular intervals and are identified with a timestamp, a name, a value, and one or more defining labels. Metrics can be aggregated using a variety of algorithms, compared to other metrics, and analyzed for trends over time. We trained our ADaaS model with 7-day data for ‘Availability’ metrics with around 16k time-series. Each series is of length 1992 time-series data-points. Inference data has 15k time-series, and we validated our results every 1 hr. Figure 2 provides a sample time-series from Azure Core workload insights data.

4.1.2 Electricity

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

4.2 Results

To showcase the explainability using SPASE, we have chosen two dataset and two different token based explainability techniques. We used one public dataset Electricity (4.1.2) and one internal dataset of Azure Core (4.1.1). To showcase the SPASE extensibility, we have used two popular token based black-box explanation technique i.e LIME and SHAP. In our experiments, we obtained the frequency of top-10 contributing features using both of these for a subset of input training data, for this have picked 50 random sample from training dataset. To provide the visual and quantitative explanation for time series data, we presented histogram based frequency distribution of Top contributing features in Figure 3 and Figure 4, we have also provided quantile wise density value as saliency in Table 1. For direct comparison with existing technique i.e. LIME, we have presented LIME explanation on same Azure Core data and model in Table 3. We can see LIME present a few isolated pointed and their importance as explanation. For time series data this token-based explanation is harder to understand as compare to words in textual data. In next subsections, we will be discussing results and explanations for both the dataset.
4.2.1 Azure Core Explanations

Figure 3 showcase frequency distribution of LIME based explanation for Azure Core workload data. We obtain the mean density of the features, and it shows that the mean density lies near 1405 features. The mean density values indicate important features towards the end of the series. We obtain five quantiles across all the features and their range and spatial saliency scores are shown in Table 1. In this table, we show the feature range in each quantile and its spatial saliency score. In the first quantile, the feature range starts in 1980 and ends in 1991 with a spatial silence score of 0.02364. The top salience score indicates a more important feature and vice versa. The feature ranges from quantiles 1 and 2 contribute more for future prediction of the series. The feature range and its salience score are also shown in Figure 4. Here we have shown the initial three quantile values and features in different colors. We observed that maximum density features are present at the end of the series that is in the first quantile. So, our explanation method SPASE can help to identify the important feature ranges and their salience scores for those features, whose contribution is more to the model’s prediction. Similarly, we obtained the quantile-based feature ranges and frequency density using SHAP. SHAP based spatial saliency exhibits similar results as LIME. SHAP based spatial saliency explanation as per Figure 4 showing top most contributing quantiles lie towards the end, means time stamps very close to the forecasting time stamps, contribute most for the model prediction. Mean density for SHAP based top contributing features is 1937 and it’s higher than LIME mean density 1405. SHAP based results also showing top contributing features are more skewed toward the end than LIME, as lower bound for 4th quantile is 1828 as compare to 670 of LIME.

4.2.2 Electricity Data Explanations

To better understand SPASE based explanation and to showcase it’s generalization and extensibility on different kinds of dataset, We have picked up publicly available Electricity dataset 4.1.2 and build the time series a forecasting using ADaAS. Spatial saliency based Explanation using histogram based quantile density for electricity dataset is shown in Figure 6 and Figure 7. Mean
density of top contributing features for electricity dataset using LIME is 91 and it’s near to the features mid points i.e. 81. For this dataset, we observe all the quantiles are equally distributed and having similar density values. Having equally contributing quantiles provides us the information that top contributing feature are recurrent in every quantile interval. This lead us to interesting insights about seasonality in dataset and how forecasting model is capturing it. As we can see in Figure 5, time series from electricity dataset are non-stationary in nature, meaning it’s changing value recurrently with in some intervals. This nature of time series is well captured by the model and it’s reflecting in model explanation as well, in terms of having seasonality among top contributing features/time stamps. Similar results showcased by SHAP based spatial saliency in Figure 7. Mean density for top contributing features is 90 and quantiles are equally distributed like LIME based results in Figure 6. Quantile wise spatial saliency is presented in Table 2. We can see 3rd quantile is having lower width and high saliency value for both LIME and SHAP based explanation, which signifies seasonality from this quantile contributing more towards the time series model forecasting than any other quantiles.

4.3 Validation and Benefit Analysis

In order to validate the outcome of SPASE, we perform human evaluation with 30 annotators (internal company employees) on the outcomes of Azure Core Workload Insights ADaaS model. For example, a study [11] mention how explainability methods like LIME and SHAP have been also been human evaluated. We achieved 93% Precision and 86% Recall value on task of identification of most important/explainable region in 5k time-series (a subset of inference data). In terms of benefit in consumption of models, Azure MLaaS reported 20% more sales in time-series model based products post SPASE explanation method integration. Azure Core Workload insights users reported better understanding of the outcomes when clubbed with explainability results. This resulted in around 25% more retention in Azure Core workload insights products. The user -interaction on Azure Core workload insights dashboard increased by 40%.
### 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

<table>
<thead>
<tr>
<th># QUANTILE</th>
<th>LIME RANGE</th>
<th>LIME SALIENCY</th>
<th>SHAP RANGE</th>
<th>SHAP SALIENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ST</td>
<td>(141.2, 163.0]</td>
<td>0.00936</td>
<td>(140.0, 163.0]</td>
<td>0.01479</td>
</tr>
<tr>
<td>2ND</td>
<td>(108.0, 141.2]</td>
<td>0.00645</td>
<td>(108.0, 140.0]</td>
<td>0.00825</td>
</tr>
<tr>
<td>3RD</td>
<td><strong>(83.6, 108.0]</strong></td>
<td><strong>0.00909</strong></td>
<td><strong>(91.0, 108.0]</strong></td>
<td><strong>0.01342</strong></td>
</tr>
<tr>
<td>4TH</td>
<td>(43.0, 83.6]</td>
<td>0.00626</td>
<td>(40.0, 91.0]</td>
<td>0.00687</td>
</tr>
<tr>
<td>5TH</td>
<td>(0, 43.0]</td>
<td>0.00609</td>
<td>(0, 40.0]</td>
<td>0.00505</td>
</tr>
</tbody>
</table>

### 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

<table>
<thead>
<tr>
<th>LIME Top-5</th>
<th>FEATURE</th>
<th>IMPORTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ST</td>
<td>1991</td>
<td>2.6706</td>
</tr>
<tr>
<td>2ND</td>
<td>1979</td>
<td>0.3905</td>
</tr>
<tr>
<td>3RD</td>
<td>1647</td>
<td>0.2995</td>
</tr>
<tr>
<td>4TH</td>
<td>288</td>
<td>0.2529</td>
</tr>
<tr>
<td>5TH</td>
<td>1087</td>
<td>0.2466</td>
</tr>
</tbody>
</table>

### 5 Business Application

Azure Machine Learning-as-a-Service (MLaaS) has multiple time-series based solutions like 1.) Anomaly detection-as-a-service (ADaaS) 2.) Azure Credit consumption early warning system. These systems provide various output corresponding to time-series data however lacks interpretability/explainability at their core. Typically, in business use-cases machine learning model output and prediction are consumed as a black-box decision. To provide better understanding on why model made certain prediction, it’s necessary to have an explainability system to increase confident on models’ prediction and reduce unforeseen biases. Therefore, a need for time-series explainability as a spatial saliency feature comes into the picture. We have integrated our solution with these two products. As part of our next application, we partnered with Azure Core to provide a time-series explainability solution for their workload insights time-series data anomaly detection service. Our ‘proof-of-concept’ has been widely tested on Azure Core workload insights data and put into their MARIO system.

### 6 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 into 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 is an important aspect needed to provide transparency of outcome especially in applications like healthcare, manufacturing, electronics etc., where use of time-series data is prevalent.

### References


