

Exploring Scalability in Large-Scale Time Series in DeepVATS framework

Inmaculada Santamaria-Valenzuela¹[0000–0002–7497–8795], Victor
Rodriguez-Fernandez¹[0000–0002–8589–6621], and David
Camacho¹[0000–0002–5051–3475]

Dept. of Computer Systems, Universidad Politécnica de Madrid
{mi.santamaria,victor.rfernandez,david.camacho}@upm.es

Abstract. Visual analytics is essential for studying large time series due to its ability to reveal trends, anomalies, and insights. DeepVATS is a tool that merges Deep Learning (Deep) with Visual Analytics (VA) for the analysis of large time series data (TS). It has three interconnected modules. The Deep Learning module, developed in R, manages the load of datasets and Deep Learning models from and to the Storage module. This module also supports models training and the acquisition of the embeddings from the latent space of the trained model. The Storage module operates using the Weights and Biases system. Subsequently, these embeddings can be analyzed in the Visual Analytics module. This module, based on an R Shiny application, allows the adjustment of the parameters related to the projection and clustering of the embeddings space. Once these parameters are set, interactive plots representing both the embeddings, and the time series are shown. This paper introduces the tool and examines its scalability through log analytics. The execution time evolution is examined while the length of the time series is varied. This is achieved by resampling a large data series into smaller subsets and logging the main execution and rendering times for later analysis of scalability.

Keywords: DeepVATS · Visual Analytics · Scalability Analysis · Shiny.

1 Introduction

Visual analytics in Deep Learning refers to the use of interactive visual interfaces and plots to identify emerging patterns and trends within the data, supporting its understanding. Dimensional reduction techniques are essential for the analysis and interaction with large time series data [1].

DeepVATS (Deep Visual Analytics for Time Series) [16] provides a tool inspired by TimeCluster [1] that integrates Visual Analytics (VA) and Deep Learning (DL) to extract valuable insights from Time Series (TS) through interactive projections and TS data plots. It uses dimensionality reduction (DR) techniques for efficient analysis and interaction with large time series. As Fig. 1 shows, DeepVATS has three modules: DL, Storage and VA. These modules interact

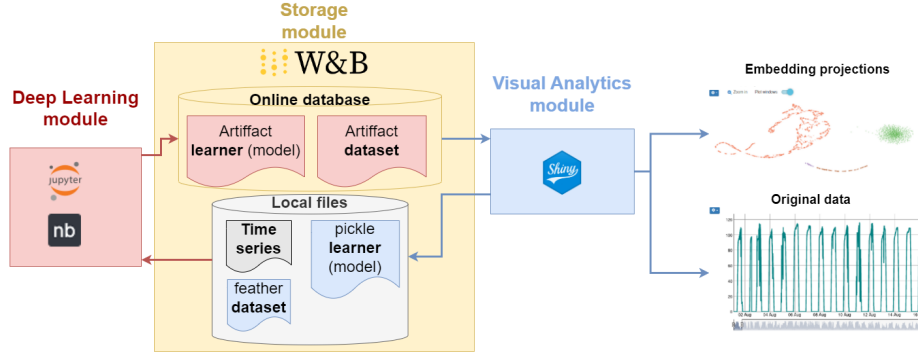


Fig. 1. DeepVATS' global workflow.

to perform three main tasks. These tasks include training neural networks to obtain meaningful data representations (embeddings); projecting the content of the trained neural network's latent space in two dimensions (2D) to detect clusters and anomalies; and finally, providing interactive plots to explore different perspectives of the projected embeddings.

The main goal of DeepVATS is to analyze large TS datasets. This article presents a scalability analysis performed using the Monash benchmark [9] to analyse future improvements that can be carried out. As “Solar Power dataset (4 Seconds Observations)” [7] is the dataset that contains the largest TS in the benchmark, it has been selected as the case for use for the study. This TS dataset represents the solar power production in MW recorded every 4 seconds over a year ($7397222 \approx 7.4\text{M}$ elements in total). To get representative subsets for scalability analysis, this TS has been resampled into datasets from 4 seconds to 10 minutes frequencies. This analysis has shown some performance problems. First, there is an issue of stability related to the use of `cuml` UMAP implementation [15] that is unstable due to an internal issue [14]. Two alternatives are proposed for future analysis. Second, the use of `reticulate` may add some time because of the communication between R and python. Thus, some operations have been joined or moved to be used from a pickled file instead of R variables to get better performance. Finally, some extra execution steps have been detected due to Shiny's reactivity. This can be easily managed by a better use of the cache in the R shiny up, thus some alternatives are proposed within the analysis.

The remainder of the paper is organized as follows. Section 2-DeepVATS description: provides a description of DeepVATS including a step-by-step case of use. Section 3-Scalability Analysis: examines the most representative steps within the execution. It includes the load of the dataset, obtainment of the embeddings, the calculus of the projections, the clustering, and the interactions with the interactive plots. Finally, section 4-Conclusions and future work summarize the

conclusions of the scalability analysis and future lines and expectations on the development of the tool.

2 DeepVATS description

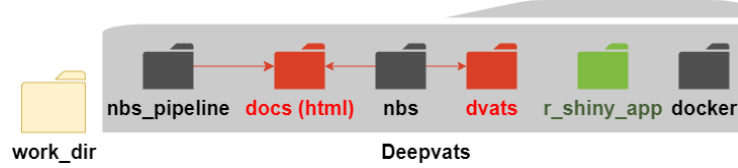


Fig. 2. DeepVATS main folders schema and main jupyter notebooks for the executions.

The DeepVATS code base¹ integrates three main modules: Deep Learning (DL) Module, Storage Module and Visual Analytics (VA) Module (See Fig. 1). The DL Module is a Python library implemented using Jupyter Notebooks [10] and nbdev [5]. The Jupyter notebooks contained in the folder **nbs** are used for the generation of **dvats** library. This library is the base for both the **nbs_pipeline** notebooks and the **r_shiny_app** (see Fig. 2). As detailed in the github repository, the docker containers must be preconfigured, assuming that another folder **work_dir** already exists in the same path as the global “Deepvats” for the use of the tool functionalities.

2.1 Deep Learning (DL) Module

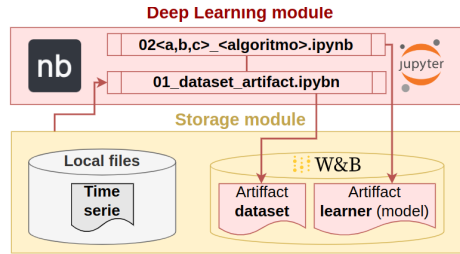


Fig. 3. DeepVATS’ load dataset workflow.

The DL module gives tools for loading datasets from/to the storage module and use them into the **nbs_pipeline** folder notebooks. As Fig. 3 shows, the first

¹ The source code and the dockers used for the execution of the tool are hosted in the GitHub repository <https://github.com/vrodriguezf/deepvats>.

step is to execute `01_dataset_artifact`. This notebook assumes preconfigured dataset options defined on the `config/*.yaml` files. Some examples with free accessible datasets are included via direct python execution. This execution will load a dataset from local to Weight & Biases (wandb.ai) as an artifact. It can be originally a 'csv', 'txt' or 'tsf' file. Now, we can use the `02_<a,b>_encoder_<DCAE, MVP>` notebooks to train the associated fastai learners' models [6]. There are currently implemented two algorithms for data encoding: Deep Convolutional auto-encoder (DCAE) [1] and Masked Value Prediction [16]. This three jupyter notebooks includes its equivalent using Sliding Window View for better performance. Once we got the model trained and loaded to W&B, the visual application should be used. However, the functionalities can be tested in `03a_embeddings`, which is based on TSAI's `get_acts_and_grads()` for extracting the embeddings from the trained model [12] and `04a_dimensionality_reduction` can be used for getting the embeddings from the learner and applying UMAP [11], TSNE or PCA (rapidsai's cuml implementations [15]). Finally, `hdbscan` is used for clustering [3] and the plots are shown.

In order to get dynamic plots depending on the dataset, the encoder, projection algorithms parameters and clustering parameters, the VA module is implemented. It includes their parameters so that different configurations can be tested and time series can be easily analyzed. This module is based on R and uses shiny for building up an app where a projection plot (with the projection of the dimensionally reduced embeddings) and a time series plots are shown. As it will be described in the next sections, both plots are interactive, so insightful details can be obtained from them.

2.2 Visual Analytics (VA) module

This section presents the VA module through images. As Fig. 4 show, the visual app does the following main steps. First, it downloads all artifacts from the W&Biases system. Second, from a selected dataset and an associated encoder, it obtains the embeddings and projects them into a projection plot (PP) using `dvats`. At the same time, it generates the TS plot associated to the dataset artifact. The application also includes the option of clustering the projections for better analysis. To open the app we must run the r-studio-server docker and run `shiny::runApp("app")`. This will open a new window (permissions may be requested by the browser).

This window contains the following main elements (see Fig. 5: On the left, selectors for getting specific dataset, encoders and associated configurations:

- Encoder selector. It gives a list with all W&B encoder artifacts trained for the loaded dataset. Default: last uploaded artifact for the dataset.
- Slider window size selector. Default: used for `02b prepare_forecasting_data()`. Maximum and minimum are proposed in the artifact configuration file.
- DR parameters slider selectors: `n_neighbors`, `min_dist` and `random_state`. Default: configured in the artifact configuration file.

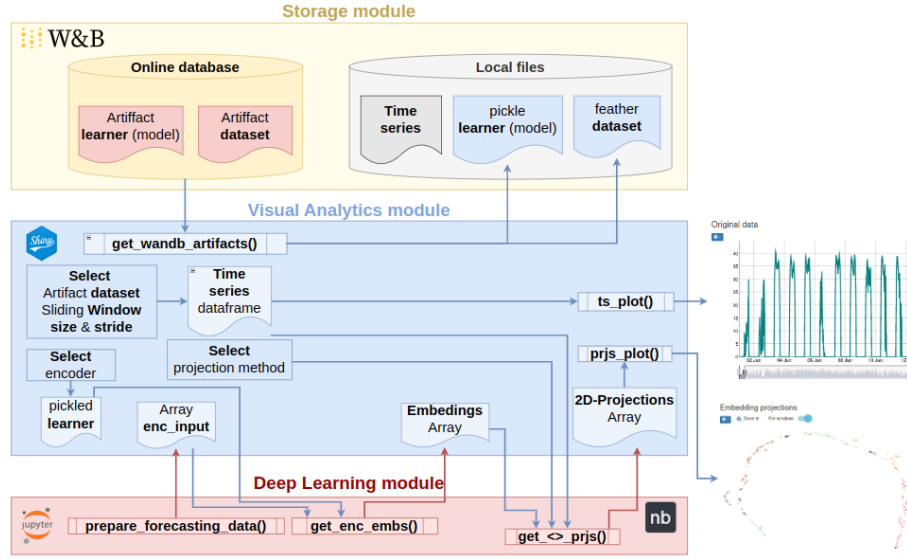


Fig. 4. DeepVATS Visual Analytics (VA) module workflow and interaction with Storage module schema

- Projections algorithm selector: UMAP, TSNE, PCA
- GPU/CPU selector. Default: GPU.
- Clustering option: Calculate and show clusters. For clustering the embedding projections dataset. When calculate and show cluster is selected, hdbscan parameters sliders are shown.

On the right side, there are three tabs. First, **projections** tab, which shows the interactive plots. Second, the **information** tab, which shows the metadata of the selected artifacts. Finally, the **Logs** tab, which is under development and summarises the main logs including the execution times.

Focusing on the **Projections** tab, we can interact with the plots in different manners. First, we can change the aesthetics through the embedding projections options menu for better visualisation (see Fig. 6). This includes different parameters for the aesthetics and a download button for saving the embedding projections plot.

Once we have the embeddings projections configured, we can get more insights on the relation between the embeddings and the TS data. By selecting points in the embeddings projections, their corresponding windows are displayed in the time series and viceversa. To avoid the application from crashing, the number of timepoints to display has been reduced to 10k. This also gives better rendering performance. The under-development lines below approximately show the windows in the same position as in the graphic timeline. The time selector can be used so that different windows can be analyzed. Also, zoom can be

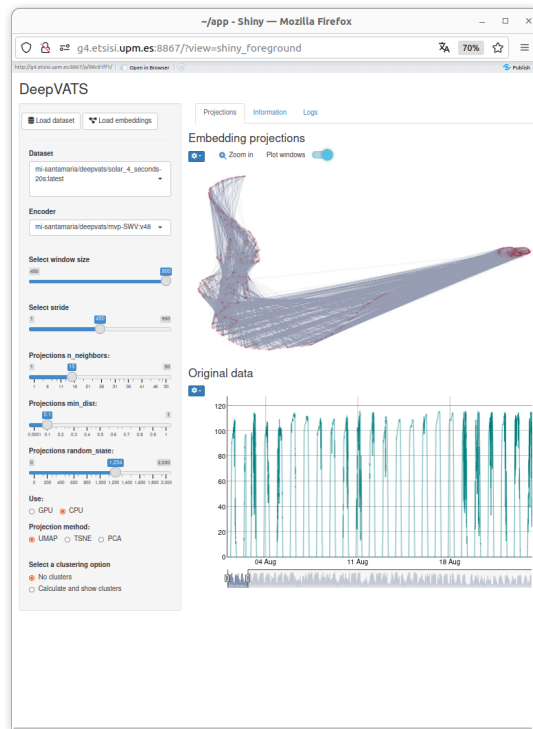


Fig. 5. DeepVATS' Visual Analytics (VA) Module first view example.

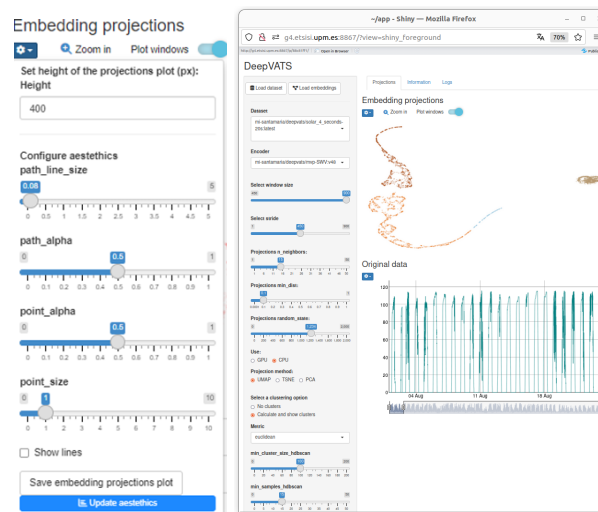


Fig. 6. DeepVATS' Visual Analytics (VA) Module example of projections plot aesthetics configuration

used for better comprehension of the embeddings structure by pushing the zoom in-zoom out button (see Fig. 7).

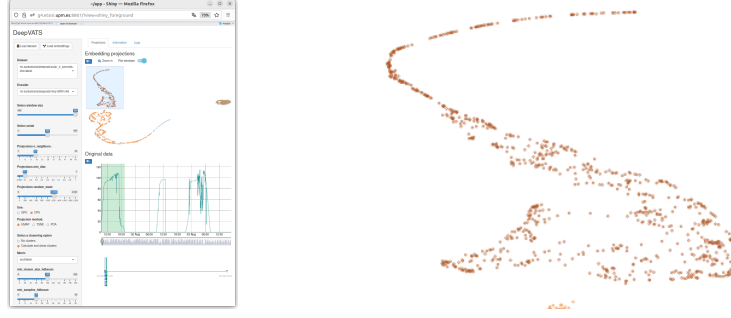


Fig. 7. DeepVATS' Visual Analytics (VA) Module projections plot zoom in example

The information tab shows the dataset and the encoder metadata and the Logs tab, which is under development, saves a datatable with basics logs of the app.

3 Scalability Analysis

Table 1. Resamples of **Solar Power Dataset (4 Seconds Observations)** used for the scalability analysis. Seconds is the frequency generated by using the proposed frequency factor. The last column shows the number of elements of each dataset.

Seconds	Frequency factor	Number of elements
4	1	7,397,222
20	5	3,698,611
60	15	493,149
300 (5m)	75	98,630
600 (10m)	150	49,315

The main goal of deepVATS is to visually analyse large time series. The Monash benchmark [9] presents more than 20 datasets with different number of series and lengths, thus providing for good data for testing the app's performance. To test the most challenging case, the dataset with the longest time series has been used: Solar Power Dataset (4 Seconds Observations) [7]. This dataset contains a single time series containing the solar power production in MW (megawatts) recorded with a frequency of 4 seconds along one year, obtaining more than 7M elements (see Table 1).

To check the app’s scalability for ensuring its performance in large time series, this dataset has been resampled into smaller datasets by selecting the dataset rows based in a frequency factor, defining the frequency factor ($f_{factor} = 150$) as the number we multiply frequency by ($f = 4s$) so that desired frequency is obtained ($f = 4s \cdot 150 = 600s = 10m$). For simplicity, after checking that 10 minutes frequency had also been previously used in the Monash benchmark [8], factors 5, 15, 75 and 150 have been selected to obtain divisions of 4 seconds to 10 minutes frequencies. Table 1 shows the evolution in terms of number of elements. To get the execution times and logs, the following steps have been executed for each dataset:

- Load the dataset from the feather file.** Select the desired dataset with the last associated encoder logged and check the time required to load the dataset from the binary file. The app will automatically get the embeddings from the encoder and apply UMAP using GPU for dimensionally reduction (DR) and will generate the associated projections and time series plots.
- Change to PCA.** As will be detailed in the following sections, it is interesting to check the use of PCA and other DR algorithms.
- Change to CPU.** It is interesting to check the app performance according to this selection.
- Clustering.** To gain insight into the structure of the embeddings.
- Selection of projection points.** The TS points associated to the selected points will be shaded in the TS plot.
- Plots interactions.** To check the plots rendering after interactions, basic steps have been added: projections plot (PP) zoom in and zoom out, select a point on the time series plot so its associated projections points are remarked.
- PP aesthetics updates.** To check rendering times for aesthetics updates, minimal changes have been selected: update point alpha aesthetic in PP and removing lines in PP for clearer visualisation.

The following sections analyses the different steps that the app executes when those steps are done: load dataset, get encoder embeddings, compute projections, compute and visualize plots and compute clusters.

3.1 Load dataset

TS plot is generated using `dygraph` [17]. This function expects to receive a dataframe with timestamps as rownames. Thus, a rownames conversion was initially added to move the time index column to rownames. This step rapidly increases the execution time when increasing the TS length. After taking insights on the execution progress, a dataframe to `xts` conversion was detected when calling `dygraph`. Fig. 8 shows how explicitly making the conversion via `xts(TS_dataframe, order.by=TS_dataframe$timeindex)` reduces the execution time up to 27 seconds for the largest case, thus notably enhancing the app performance.

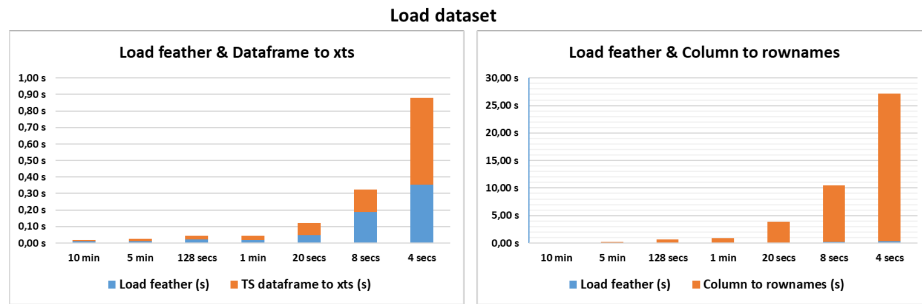


Fig. 8. Load dataset aggregated time plot. In blue, read feather operation. In orange, on the left, time series dataframe conversion to xts. In orange, on the right, move the timeindex column to rownames operation.

3.2 Get embeddings

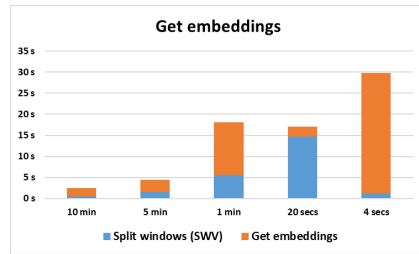


Fig. 9. Aggregated plot for `get embeddings` step. In blue, the execution time consumed by the splitting window operation using Sliding Window View (in seconds). In orange, the execution time consumption for the get encoder embeddings operation (in seconds).

The time and memory consumption increases rapidly when the embeddings are obtained from the trained models. For memory handling and better R-python communication the implementation has been modified. Now, the the dataset input is stored as a `feather` file, enabling data loading and manipulation within python. Furthermore, the embeddings obtainment is chunked so that GPU memory errors are avoided.

3.3 Compute projections

The computation of projections is known to be much faster on a GPU than on a CPU. However, there exists a known bug in the GPU version that makes UMAP fail when using `cuml UMAP (GPU)` function that makes projections unstable and with low quality cluster distribution [13, 14]. This quality can be checked in

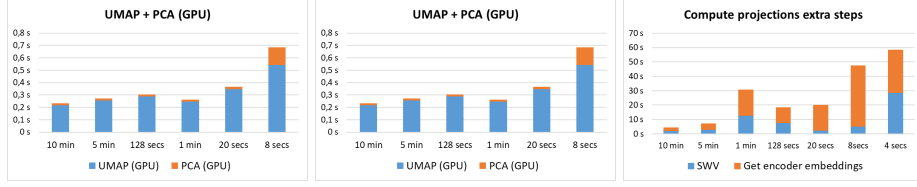


Fig. 10. Compute projections. On the left UMAP + PCA (GPU) execution time aggregated plot. On the middle, the comparison with UMAP (CPU) executions. On the right, the execution time associated to extra executed steps.

both the app logs and the 04 jupyter notebook, where `cluster_score(prjs, clusters_labels) = silhouette_score(prjs, clusters_labels)` is used to obtain a value in $[-1, 1]$, with -1 being the worst score.

To avoid this problem, the CPU version can be selected, but this results in an excessive execution time for a visual dynamic application. Thus, there exist two possible work lines. The first option is to change the GPU UMAP library that we use, for example, using Peter Eisenmann's parallel UMAP function [4]. The second option is to explore the solution proposed by Corey J. Nolet: Execute UMAP followed by PCA in order to obtain better results [14]. This has not been implemented yet, but, as Fig. 10 shows, the time would still be much better with this option than using the CPU.

3.4 Compute and visualise plots

Table 2. TS plot and PP plot rendering time after computing projections

	Time (seconds)				
	10m	5m	1m	20s	4s
TSP rendering time	4.11	5.25	20.48	79.19	228.05 \approx 4m
PP rendering time	5.55	7.24	21.16	64.80	228.10 \approx 4m

Table 3. TS plot and PP plot computation time media and extra detected steps execution time in seconds

	Time (seconds)					Time (% Relative to total)				
	10m	5m	1m	20s	4s	%10m	%5m	%1m	%20s	%4s
TSP computation time	0.03	0.03	0.08	5.45	25.37	0.63	0.39	0.26	23.23	46.27
PP computation time	2.16	3.28	12.46	0.93	0.05	45.92	42.65	40.71	3.97	0.09
Extra steps	2.51	4.38	17.06	29.4	29.83	53.45	56.95	59.03	72.78	53.63

A great reduction in the performance was detected when trying to load the Solar Power Dataset (4 Seconds Observations) and interacting with the plots. This is reasonable as it is the largest one. This is not a huge problem for small

datasets, but causes the app to crash with large datasets. But our goal is to analyse these large datasets with the provided tool. Thus, this scalability analysis was started, and we detected that some reactive operations are being executed without need (see Table 3). This adds a lot of time to the rendering of the plots. Also, a notable difference was noticed between execution and rendering time (see Table 2), allowing nearly 4 minutes to render the plots for the 4 second frequency dataset. To avoid this problem, cache can be used in shiny to avoid plot recalculations. Also, the reduction of the number of showed points have reduced the renderization time. Therefore, the next step of the work is to analyse the use of more `reactiveVal`, and cache plots [2], and to check more options for the reduction of the plotted points.

3.5 Compute clusters

The clusters computation is really fast even for large time series. However, some reactive functions have also been detected that affects to the performance of clustering step. Thus, the cache analysis is useful for enhancing this step too.

4 Conclusions and future work

DeepVATS is a powerful tool designed for the easy visual analysis of both univariate and multivariate time series. It allows users to interact with the embeddings projection plot and the original time series data plot, facilitating the detection of patterns within the data. The scalability analysis has demonstrated an excellent performance for small datasets (ranging from 49.3K to 98.6K elements) and medium-sized datasets (up to 493.1K elements). However, performance issues appeared when processing larger datasets, with noticeable degradation at 3.7 million elements and application crashes at 7.4 million elements.

To enhance the application’s usability for large datasets, we propose the following development work lines. First, to eliminate redundant processes, the `reactive` variables should be checked to determine if they can be converted to `reactiveVal` to ensure the use of cached values. Also, `renderCachedPlot` is proposed as an alternative to `renderPlot` for the projections plot [2]. Second, to ensure high-quality dimensionality reduction, two strategies are suggested: adopting an alternative UMAP implementation [4], and exploring the application of PCA followed by UMAP, rather than UMAP alone [14].

These modifications are anticipated to give the application a substantial improvement in performance (e.g., the elimination of the additional 28 seconds required for compute projections for the 4 seconds frequency dataset). Also, greater stability in the GPU clustering is expected, making DeepVATS an more robust tool for the visual and interactive analysis of time series.

Acknowledgements. This work has been funded by Grant PLEC2021-007681 (XAI-DisInfodemics) and PID2020-117263GB-I00 (FightDIS) funded by MCIN/AEI/10.13039/501100011033 and by “ERDF A way of making Europe”, by the

“European Union” or by the “European Union NextGenerationEU/PRTR”, by Calouste Gulbenkian Foundation, under the project MuseAI - Detecting and matching suspicious claims with AI, and by "Convenio Plurianual with the Universidad Politécnica de Madrid in the actuation line of Programa de Excelencia para el Profesorado Universitario”.

References

1. Ali, M., Jones, M.W., Xie, X., Williams, M.: Timecluster: dimension reduction applied to temporal data for visual analytics. *The Visual Computer* **35**(6-8), 1013–1026 (2019)
2. Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., Borges, B.: shiny: Web Application Framework for R (2022), <https://shiny.rstudio.com/>, r package version 1.7.4
3. scikit-learn contrib: scikit-learn-contrib/hdbscan. <https://github.com/scikit-learn-contrib/hdbscan> (2024), accessed: Jan. 04, 2024
4. Eisenmann, P.: p3732/gpumap. <https://github.com/p3732/gpumap> (2023), accessed: Jan. 04, 2024
5. fast.ai: nbdev – create delightful software with jupyter notebooks. <https://nbdev.fast.ai/>, accessed: Jan. 04, 2024
6. fast.ai: fastai/fastai. <https://github.com/fastai/fastai> (2024), accessed: Jan. 04, 2024
7. Godahewa, R., Bergmeir, C., Webb, G., Abolghasemi, M., Hyndman, R., Montero-Manso, P.: Solar power dataset (4 seconds observations) (Apr 2021). <https://doi.org/10.5281/zenodo.4656027>
8. Godahewa, R., Bergmeir, C., Webb, G., Hyndman, R., Montero-Manso, P.: Solar dataset (10 minutes observations) (Apr 2021). <https://doi.org/10.5281/zenodo.4656144>
9. Godahewa, R., Bergmeir, C., Webb, G.I., Hyndman, R.J., Montero-Manso, P.: Monash time series forecasting archive. arXiv preprint arXiv:2105.06643 (2021)
10. Jupyter, P.: Project jupyter. <https://jupyter.org>, accessed: Jan. 04, 2024
11. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
12. Oguiza, I.: tsai - a state-of-the-art deep learning library for time series and sequential data. Github (2023), <https://github.com/timeseriesAI/tsai>
13. rapidsai: [bug] accuracy of lanczos solver. · issue #313 · rapidsai/raft. <https://github.com/rapidsai/raft/issues/313>, accessed: Jan. 04, 2024
14. rapidsai: [bug] different outputs for umap on cpu vs. gpu · issue #5474 · rapidsai/cuml. <https://github.com/rapidsai/cuml/issues/5474>, accessed: Jan. 04, 2024
15. rapidsai: Rapids cuml api documentation. <https://docs.rapids.ai/api/cuml/stable/api/>, accessed: Jan. 04, 2024
16. Rodriguez-Fernandez, V., Montalvo-Garcia, D., Piccialli, F., Nalepa, G.J., Camacho, D.: Deepvats: Deep visual analytics for time series. *Knowledge-Based Systems* **277**, 110793 (2023)
17. Vanderkam, D., Allaire, J., Owen, J., Gromer, D., Thieurmél, B.: dygraphs: Interface to 'Dygraphs' Interactive Time Series Charting Library (2018), <https://github.com/rstudio/dygraphs>, r package version 1.1.1.6