Data Series Similarity Search via Deep Learning

Qitong Wang
supervised by Themis Palpanas
Université Paris Cité, LIPADE
qitong.wang@etu.u-paris.fr

ABSTRACT

A key operation for the (increasingly large) data series collection analysis is similarity search. According to recent studies, SAX-based indexes offer state-of-the-art performance for similarity search tasks. However, their performance lags under high-frequency, weakly correlated, excessively noisy, or other dataset-specific properties. In this work, we propose to facilitate data series similarity search with deep learning techniques, involving both data series approximation and data series indexing. Our preliminary study focuses on developing Deep Embedding Approximation (DEA), a novel family of data series summarization techniques based on deep neural networks. Moreover, we describe SEAnet, a novel architecture specially designed for learning DEA, that introduces the Sum of Squares preservation property into the deep network design. Finally, we propose a new sampling strategy, SEASam, that allows SEAnet to effectively train on massive datasets. Comprehensive experiments verify the advantages of DEA learned using SEAnet. These preliminary results can lead to further progress in this area, by developing more customized architectures and training strategies, better integrating DEA with index structures, learning novel data series indexes, and facilitating faster model training.

PVLDB Artifact Availability:
The source code, pretrained models, and datasets have been made available at https://helios.mi.parisdescartes.fr/~themisp/seanet/.

1 INTRODUCTION

With the rapid developments and deployments of modern sensors, massive data series[1] datasets are now being generated, collected and analyzed in almost every scientific domain [24]. Data series similarity search aims to find the closest series in a dataset to a given query series according to a distance measure, such as Euclidean distance, which is one of the most widely used [31]. Similarity search can be divided into exact search and approximate search [7]. Approximate similarity search may not always produce the exact answers, but in most cases, it produces answers that are very close to the exact ones [8]. Thus, it is very popular in practice, and widely used on massive series collections to enable interactive data exploration and other latency-bounded applications [12]. In this work, we focus on approximate similarity search under Euclidean distance.

Indexes are widely employed to speed up data series similarity search [7, 8]. Most indexes are based on summarized representations of the data series [31] of lower dimensionality. Symbolic Aggregate approXimation (SAX) [19] is a popular and effective discretized summarization. SAX-based indexes [23] are the state-of-the-art (SOTA) data series similarity search methods [7, 8].

Nevertheless, SAX-based indexes suffer from the problem that SAX fails in hard datasets with specific properties [17]. Since SAX is the symbolization of Piecewise Aggregate Approximation (PAA) [19], failure of PAA to correctly represent some data series directly translates to failure of the PAA-based SAX. For example, the high frequency of Deep1B series means that each PAA segment has to average many highly-varying points, leading to similar PAA values across different segments, and to indistinguishable SAX words across different series. Introducing more SAX words could alleviate the problem, but would lead to an undesirably long summarization that could not be effectively indexed.

To address the aforementioned problems, we propose to build a data series index based on ***D***eep ***E***mbedding ***A***pproximations (DEA), i.e., data series summarizations derived from embeddings learned using deep neural networks. Embedding techniques, or representation learning [1], is to learn vectors possessing necessary latent information for classification, clustering and other downstream applications. Embedding techniques have been proven to be capable of capturing frequency [29] and other latent properties. However, data series embedding has not been adapted to and evaluated for similarity search (and could also be applied to other tasks, e.g., anomaly detection [2, 3, 25]).

Specifically, we propose to replace traditional summarizations (e.g., PAA) with DEA, and then be symbolized and indexed by an iSAX index. DEA targets to preserve original pairwise distances in the lower-dimensional DEA space. Thus, it is naturally capable of being symbolized into SAX, on which an iSAX index can be built.

Our preliminary results show that compared to PAA and SAX (which is based on PAA), DEA better preserves pairwise distances, leading to a more effective index for data series similarity search. This can be used as a blueprint to facilitate further progress in this area. Promising further directions include developing more customized architectures and training strategies based on observations from preliminary results, better integrating DEA learning with index structure designs, and facilitating faster DEA learning or transferring [35] on massive datasets.

Note that existing studies on learned indexes [16] cannot be straightforwardly employed for facilitating data series similarity search with deep learning techniques. This is true, because most existing methods assume that the data are sortable in a natural order, which can then be captured by learned distribution functions [16]. However, such global orders for data series similarity search do not exist (since the order depends on the queries) [18]. Furthermore, existing methods suitable for similarity search are built upon grid indexes [5, 22], which do not scale to the high dimensionalities (i.e.,

Proceedings of the VLDB 2022 PhD Workshop, September 5, 2022. Sydney, Australia.

[1] A data series, or data sequence, is an ordered sequence of points. The most common type of data series is time series, where the dimension that imposes the sequence ordering is time; though, this dimension could also be the mass, angle, or position [24].

in the order of 100s-1000s) of data series. Hence, how to extend existing studies to resolve the aforementioned open problems remains a challenging research direction.

In this work we propose the following research directions:

(1) **[Architecture]** Design novel architectures that are specifically built to support high-quality DEA and similarity search. Our preliminary solution, SEAnet (cf. Section 3.1), introduces and formalizes the principle of Sum of Squares (SoS) preservation.

(2) **[Training Dataset]** Propose novel sampling strategies for massive data series collections, enabling effective training for the deep models. One such example is SEAsam (cf. Section 3.2), which demonstrates that intelligent sampling strategies can help improve the performance of the deep network models.

(3) **[Learned Indexes]** Integrate index structure building into DEA learning to fully exploit the edges of DEA. Pushing further in this direction, it would be interesting to learn a specifically designed index structure *together* with DEA learning.

(4) **[Model Training]** Address the problem of the long training times needed by the deep neural models (which can be significantly slower than traditional approaches), by introducing transfer learning and domain adaptation techniques in this context.

2 BACKGROUND

A **data series**, $S = \{p_1, ..., p_m\}$, is a sequence of points, where each point $p_i = (v_i, t_i)$, $1 \leq i \leq m$ is associated to a real value v_i and a position t_i. The position corresponds to the order of this value in the sequence. We call m the *length*, or *dimensionality* of the data series. $\mathcal{S}$ denotes a collection of data series, i.e., $\mathcal{S} = \{S_1, ..., S_n\}$. We call n the *size* of the data series collection. A **summarization** $E = \{e_1, ..., e_l\}$ of a series S is a lower, l-dimensional representation, which preserves some desired properties of $\mathcal{S}$. For similarity search, the target property is pairwise distance space structure of $\mathcal{S}$, i.e., $\forall S_i, S_j \in \mathcal{S}, d'(E_i, E_j) \approx d(S_i, S_j)$, where E_i, E_j are summarizations of S_i, S_j, $d(\cdot,\cdot)$, and $d'(\cdot,\cdot)$ are distance measures in series and summarization spaces, respectively. The **distance measure** d we use is Euclidean distance [31]. d' in the summarization space needs not be the same as d, e.g., for PAA, $d'(\cdot,\cdot) = \sqrt{m}/\sqrt{l} \times d(\cdot,\cdot)$. d' for DEA is the same as PAA if it's scaled for SoS preservation. Otherwise, $d'(\cdot,\cdot) = d(\cdot,\cdot)$. Given a query series S_q of length m, a series collection $\mathcal{S}$ of size n and length m, a distance measure d, **similarity search** targets to identify the series $S_c \in \mathcal{S}$ whose distance to S_q is the smallest, i.e., $\forall S_o \in \mathcal{S}, S_o \neq S_c, d(S_c, S_q) \leq d(S_o, S_q)$. Instead of finding the exact closest series S_c, **approximate similarity search** targets to find a series $S'_c \in \mathcal{S}$ such that $d(S'_c, S_q) \approx d(S_c, S_q)$. $d(S_c, S_q)/d(S'_c, S_q) \in (0, 1]$ is called S'_c's **tightness**.

The most prominent **data series indexing** techniques can be categorized into optimized scans [10], and tree-based indexes [27]. Recent studies [7, 8] have demonstrated that the SAX-based indexes [23] achieve SOTA performance under several conditions. In this work, we use MESSI as our iSAX index [26] , because its main-memory operation and parallel design lead to SOTA performance.

3 DEA-BASED SIMILARITY SEARCH

Figure 1 illustrates the proposed DEA-based data series similarity search framework, including the SEAnet architecture. Given a series collection, SEAsam first draws representative samples to train

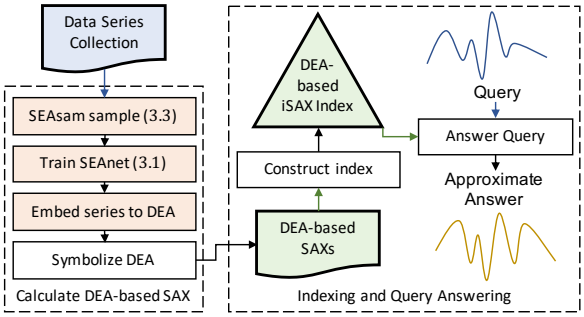


Figure 1: Workflow of DEA-based similarity search.

SEAnet. After SEAnet converges, it embeds all series into DEAs, which are further discretized into SAXs. Thus, DEA-based SAXs are structured into an iSAX index, where approximate similarity search can be efficiently conducted.

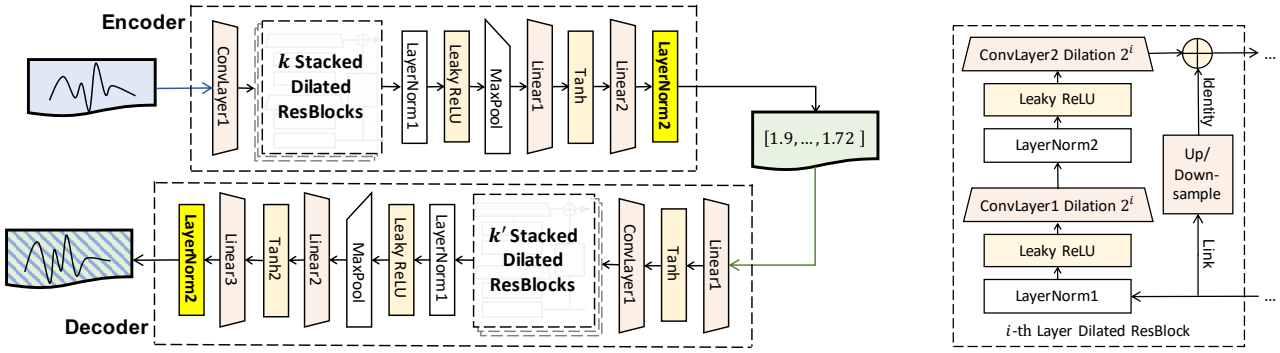
SEAnet is a novel autoencoder proposed to learn high-quality DEA (cf. Section 3.1). Moreover, it introduces the principle of SoS preservation for lower dimensionality representation learning (cf. Section 3.1.1). SEAsam makes use of the inverse iSAX sortable summarization [15] (cf. Section 3.2).

3.1 SEAnet Architecture

The SEAnet architecture is illustrated in Figure 2a. The first part of the SEAnet encoder, from ConvLayer1 to MaxPool, comprises k stacked dilated full-preactivation ResBlocks in Figure 2b for non-linear transformations. The second part of the SEAnet encoder, from Linear1 to LayerNorm2, comprises two linear layers for dimensionality reduction. Unlike most existing encoders with linear final layers [11], the SEAnet encoder is finalized by LayerNorm2, which is specifically designed using the SoS preservation principle.

SEAnet is trained in a pairwise manner by mini-batched Stochastic Gradient Descent (SGD). Its loss function is a linear combination of two components: (1) The Compression Error L_C (i.e., the average differences between the original distance of data series pairs (S_i, S_j) and their DEA distance) evaluates whether original distances are well preserved in the DEA space. (2) The Reconstruction Error L_R (i.e., the average distance between the original series S_i and the reconstructed series) L_R evaluates how well the original series can be reconstructed using SEAnet.

3.1.1 Sum of Squares Preservation. We propose a SoS preservation framework for effective DEA learning. SoS preservation has been observed before [32], but to the best of our knowledge, has never been formally introduced to representation learning. Given an $n \times m$ matrix M, where each row $M_{i,*}$ corresponds to a series and each column $M_{*,j}$ corresponds to a position, $\text{SoS} = \sum_{i,j} M_{i,j}^2$. Note that defining new axes based on the largest SoS is equivalent to selecting the largest eigenvalues in linear dimensionality reductions on z-normalized datasets, with the purpose of preserving information about the dataset through linear transformations [32]. Thus, SoS may be regarded as an indicator of transformation quality. By keeping SoS invariant, the quality of DEAs is upheld from this perspective, and the networks then focus on learning the nonlinear transformations.



(a) SEAnet Architecture (b) Dilated ResBlock in SEAnet

Figure 2: The SEAnet architecture and the details of a dilated full-preactivation ResBlock.

We now elaborate on the architecture design and model training under SoS preservation. Given the (z-normalized) input dataset, SoS preservation requires two steps: (1) z-normalizing the output of encoder (DEAs) and decoder (the reconstructed series); and (2) scaling the series by $1/\sqrt{m}$ and DEA by $1/\sqrt{l}$ in L_C and L_R.

Based on theoretical analysis [30], we observe that scaling series and DEA will not only keep the two distances to the same level, but will also largely stabilize the distance distributions. Thus, by z-normalizing DEA, and scaling series and DEA in L_C and L_R, SEAnet succeeds in providing high-quality DEAs by preserving SoS.

3.2 Sampling with SEAsam

The representativeness of the training set upper bounds for the quality of the deep models. Not only we need our sample to effectively cover the entire space of a given dataset, but also we need to efficiently select this sample without having to perform expensive computations on the full dataset.

To this end, we propose SEAsam (SEA Sampling), a novel data series sampling strategy based on the sortable data series representation, InvSAX [15]. Recall that SAX first transforms the data series into l real values , and then quantizes these real values, representing them using discrete symbols [27]. The core observation is that every subsequent bit in a SAX word contains a decreasing amount of information about the location of its corresponding data point, and simply increases the degree of precision. Interleaving SAX's bits such that all significant bits across each SAX word precede all less significant bits presents a value array with descending significance, i.e., InvSAX. SEAsam orders the series collection by their InvSAX representations, and draws samples at equal-intervals (e.g., every 1,000 series) from this sorted order. Thus, SEAsam samples are expected to preserve the distribution of the series collection by evenly covering its InvSAX space. Moreover, the time complexity of SEAsam is $O(nm)$, and the space complexity of SEAsam is $O(nl)$, rendering SEAsam an efficient strategy.

4 PRELIMINARY RESULTS

We present our experimental evaluation of SEAnet, DEA-based data series similarity search, and SEAsam using 7 diverse synthetic and real datasets. Totally, 5,040 deep models were trained to provide a thorough profile of DEA architectures. In summary, the results demonstrate that the SEAnet DEA is robust across various dataset properties and outperforms its competitors by better preserving original pairwise distances and nearest neighborhood structure, leading to better approximate similarity search results than traditional (PAA-based) and alternative deep learning (DEA-based using FDJNet [11], TimeNet [20], and InceptionTime [9]) approaches.

We evaluate the benefit of using DEA for similarity search, by reporting the 1st Best-So-Far (BSF) tightness, i.e., the 1st Nearest-Neighbor (NN) distance divided by the 1st BSF distance given a specific query, as a function of the number of series that the similarity search algorithm examines. The results on 100M datasets and 1K queries, are shown in Figure 3. SEAnet-nD is an encoder-only version of SEAnet. SEAnet improved the 1st BSF tightness, and thus the similarity search results, in 61 out of the 63 experiments. Its advantage was particularly obvious on the hard datasets, namely, Deep1B, Seismic, and Astro (detailed experimental results in [30]).

5 DISCUSSION AND CONCLUSIONS

In this paper, we introduce the use of deep learning embeddings, DEA, for data series similarity search. We propose a novel autoencoder, SEAnet, designed under the firstly introduced SoS preservation principle, for effectively learning DEA. A new sampling strategy, SEAsam, is introduced in order to facilitate SEAnet's training on massive collections. We demonstrate that the DEA learned by SEAnet more closely approximates the original data series distances, better preserves the true nearest neighbors in the summarized space, better reconstructs the original series, and leads to better similarity search results than the SOTA PAA-based iSAX (when examining either a small, or a large number of candidates). These preliminary results are very promising, they set the ground for further advancements in this area, and have the potential to also improve the performance of kNN classification, anomaly detection, and other similarity search-based applications.

Promising directions in our future studies include the following:

(1) Develop more customized architectures and training strategies. An interesting candidate would be to quantify with differentiability the nearest neighborhood preservation in the DEA space [28], which shows positive correlations with the qualities of query answers in our preliminary results.

(2) Investigate the lower bounding properties for DEA that will enable exact similarity search [13].

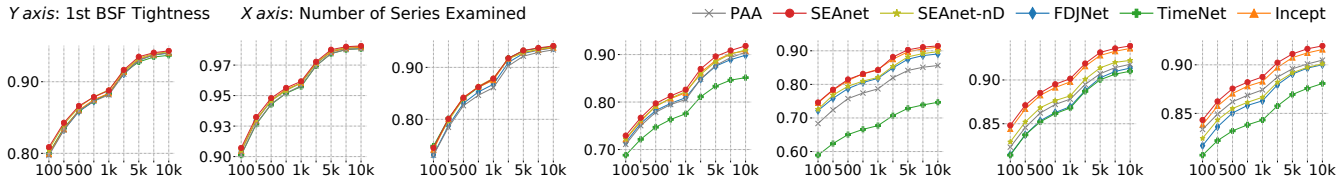


Figure 3: Approximate query answers quality: 1st BSF tightness vs number of series visited (higher is better); 100M series

(3) Integrate DEA learning with index structure building. Such an end-to-end framework will have more potential to reduce information loss during the DEA and indexing steps. Candidate index structures could be extended from trees[6] to clusters [14] and hash tables [18]. DEA and index structure could be learned together to fully exploit advantages from both sides.

(4) Design more powerful sampling strategies [33] to cover the large (pairwise distances) space, whose size is $O(n^2)$ (where n is the number of series in the collection). Ideally, a small sample should train models able to efficiently serve any *ad-hoc* query.

(5) Facilitate faster DEA learning on massive datasets. Promising techniques include incremental learning [34] and transfer learning [35]. How to identify the useful common information and how to best transfer this knowledge between massive datasets makes this a very challenging problem.

(6) Benchmark data series summarizations for similarity search [21, 25]. We will design a unified workflow and proper metrics to evaluate different summarization techniques, based on a set of representative data series collections for similarity search. The compatibility between different summarization techniques and indexing techniques [4, 6] will also need to be studied.

ACKNOWLEDGMENTS

Work supported by ANR-18-IDEX-000, Chinese Scholarship Council, HIPEAC 4, GENCI–IDRIS (Grant 2020-101471), and NVIDIA Corporation for the Titan Xp GPU donation used in this research.

REFERENCES

[1] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *PAMI* (2013).
[2] Paul Boniol, Mohammed Meftah, Emmanuel Remy, and Themis Palpanas. 2022. dCAM: Dimension-wise Activation Map for Explaining Multivariate Data Series Classification. In *SIGMOD*.
[3] Paul Boniol and Themis Palpanas. 2020. Series2Graph: Graph-based Subsequence Anomaly Detection for Time Series. *PVLDB* 13, 11 (2020).
[4] Georgios Chatzigeorgakidis, Dimitrios Skoutas, Kostas Patroumpas, Themis Palpanas, Spiros Athanasiou, and Spiros Skiadopoulos. 2022. Efficient Range and kNN Twin Subsequence Search in Time Series. *TKDE* (2022).
[5] Jialin Ding, Vikram Nathan, Mohammad Alizadeh, and Tim Kraska. 2020. Tsunami: A Learned Multi-dimensional Index for Correlated Data and Skewed Workloads. *PVLDB* (2020).
[6] Karima Echihabi, Panagiota Fatourou, Kostas Zoumpatianos, Themis Palpanas, and Houda Benbrahim. 2022. Hercules Against Data Series Similarity Search. *PVLDB* (2022).
[7] Karima Echihabi, Kostas Zoumpatianos, Themis Palpanas, and Houda Benbrahim. 2018. The lernaean hydra of data series similarity search: An experimental evaluation of the state of the art. *PVLDB* (2018).
[8] Karima Echihabi, Kostas Zoumpatianos, Themis Palpanas, and Houda Benbrahim. 2019. Return of the Lernaean Hydra: experimental evaluation of data series approximate similarity search. *PVLDB* (2019).
[9] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. 2020. InceptionTime: Finding AlexNet for time series classification. *DMKD* (2020).
[10] Hakan Ferhatosmanoglu, Ertem Tuncel, Divyakant Agrawal, and Amr El Abbadi. 2000. Vector Approximation based Indexing for Non-uniform High Dimensional Data Sets. In *CIKM*.
[11] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. 2019. Unsupervised scalable representation learning for multivariate time series. In *NeurIPS*.
[12] Anna Gogolou, Theophanis Tsandilas, Karima Echihabi, Anastasia Bezerianos, and Themis Palpanas. 2020. Data Series Progressive Similarity Search with Probabilistic Quality Guarantees. In *SIGMOD*.
[13] Piotr Indyk, Rajeev Motwani, Prabhakar Raghavan, and Santosh S. Vempala. 1997. Locality-Preserving Hashing in Multidimensional Spaces. In *SOTC*.
[14] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. 2011. Product Quantization for Nearest Neighbor Search. *PAMI* (2011).
[15] Haridimos Kondylakis, Niv Dayan, Kostas Zoumpatianos, and Themis Palpanas. 2018. Coconut: A Scalable Bottom-Up Approach for Building Data Series Indexes. *PVLDB* (2018).
[16] Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, and Neoklis Polyzotis. 2018. The Case for Learned Index Structures. In *SIGMOD*.
[17] Oleksandra Levchenko, Boyan Kolev, Djamel Edine Yagoubi, Reza Akbarinia, Florent Masseglia, Themis Palpanas, Dennis Shasha, and Patrick Valduriez. 2020. BestNeighbor: Efficient Evaluation of kNN Queries on Large Time Series Databases. *KAIS* (2020).
[18] Mingjie Li, Ying Zhang, Yifang Sun, Wei Wang, Ivor W. Tsang, and Xuemin Lin. 2020. I/O Efficient Approximate Nearest Neighbour Search based on Learned Functions. In *ICDE*.
[19] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. 2003. A symbolic representation of time series, with implications for streaming algorithms. In *SIGMOD*.
[20] Pankaj Malhotra, Vishnu TV, Lovekesh Vig, Puneet Agarwal, and Gautam M. Shroff. 2017. TimeNet: Pre-trained deep recurrent neural network for time series classification. In *ESANN*.
[21] Ryan Marcus, Andreas Kipf, Alexander van Renen, Mihail Stoian, Sanchit Misra, Alfons Kemper, Thomas Neumann, and Tim Kraska. 2020. Benchmarking Learned Indexes. *PVLDB* (2020).
[22] Vikram Nathan, Jialin Ding, Mohammad Alizadeh, and Tim Kraska. 2020. Learning Multi-Dimensional Indexes. In *SIGMOD*.
[23] Themis Palpanas. 2019. Evolution of a Data Series Index. In *ISIP*.
[24] Themis Palpanas and Volker Beckmann. 2019. Report on the first and second interdisciplinary time series analysis workshop (itisa). *SIGMOD Record* (2019).
[25] John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael J. Franklin. 2022. TSB-UAD: An End-to-End Benchmark Suite for Univariate Time-Series Anomaly Detection. *PVLDB* (2022).
[26] Botao Peng, Panagiota Fatourou, and Themis Palpanas. 2020. MESSI: In-Memory Data Series Indexing. ICDE.
[27] Jin Shieh and Eamonn Keogh. 2008. iSAX: indexing and mining terabyte sized time series. In *KDD*.
[28] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR* 9, 11 (2008).
[29] Jingyuan Wang, Ze Wang, Jianfeng Li, and Junjie Wu. 2018. Multilevel Wavelet Decomposition Network for Interpretable Time Series Analysis. In *KDD*.
[30] Qitong Wang and Themis Palpanas. 2021. Deep Learning Embeddings for Data Series Similarity Search. In *KDD*.
[31] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn J. Keogh. 2013. Experimental comparison of representation methods and distance measures for time series data. *DMKD* (2013).
[32] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* (1987).
[33] Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. 2017. Sampling Matters in Deep Embedding Learning. In *ICCV*.
[34] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. 2019. Large Scale Incremental Learning. In *CVPR*.
[35] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *PIEEE* (2020).