Deep Learning Embeddings for Data Series Similarity Search

Qitong Wang
Université de Paris, LIPADE

Themis Palpanas
Université de Paris, LIPADE
French University Institute (IUF)

ACM SIGKDD 2021, Singapore
Data Series

- Sequence of points ordered along some dimension
Data Series from Various Domains

Electroencephalography (EEG) Data

Seismic Data

Data Series Similarity Search

Query series $S_q$

Series collection $S$

(Approximate) Similarity Search

Similar series

Classification

Clustering

Outlier Detection

Frequent Patterns

...
Data Series Approximate Similarity Search

• Similarity search
  • given a series set \( S \), a query series \( s_q \) and a similarity measure \( d(\cdot,\cdot) \)
    • \( d \) is commonly the Euclidean distance
  • find the closest series in \( S \) to \( s_q \), i.e.,
    \[
    s_a = \arg \min_{s_i \in S} d(s_q, s_i)
    \]

• Approximate similarity search
  • (efficiently) find \( s_{a'} \), \( d(s_q, s_{a'}) \approx d(s_q, s_a) \)
State-of-the-art: iSAX Family of Indexes

Raw series → PAA approximation → SAX symbolization → iSAX index

1. Backgrounds


Limitations of (PAA-based) iSAX

Depends on whether PAA successfully profiles the dataset

→ Need for better summarizations

PAA (and DFT) works to approximate and reconstruct a RandomWalk series

PAA (and DFT) fails to approximate and reconstruct a Deep1B series
DEA: Deep Embedding Approximation

Replace PAA by DEA for SAX symbolization and iSAX index
(PAA-based) iSAX $\rightarrow$ DEA-based iSAX

Depends on whether PAA successfully profiles the dataset

$\rightarrow$ Need for better summarizations

PAA (and DFT) works to approximate and reconstruct a RandomWalk series

PAA (and DFT) fails to approximate and reconstruct a Deep1B series
(PAA-based) iSAX $\rightarrow$ DEA-based iSAX

- DEA better profiles diversified dataset than PAA

✓ Fulfill the need for better summarizations

DEA works to approximate and reconstruct a RandomWalk series

DEA works to approximate and reconstruct a Deep1B series
DEA-based iSAX

Replace PAA by DEA for SAX symbolization and iSAX index
How to generate high-quality DEA on massive data series collections for approximate similarity search?

Challenges

1. Effective architecture for similarity search?

Solutions

✓ SEAnet: SEries Approximation network
  - exponentially dilated ResNet + Sum of Squares (SoS) regularization

2. Efficient learning on massive datasets?
  - 100 million 256-length series \( \approx 100 \text{GB} \)
SEAnet Architecture
SEAnet Architecture

(a) SEAnet Architecture

(b) Dilated ResBlock in SEAnet
SEAnet Architecture

(a) SEAnet Architecture

(b) Dilated ResBlock in SEAnet

SoS Preservation
SEAnet Architecture

Exponentially-increasing dilations
SEAnet Training

• Loss \( L = L_C + \alpha L_R \)

  • Compression error (pairwise) \( L_C \)
    \[
    L_C = \frac{1}{N_p} \sum_{(S_i, S_j) \in S \times S} \left| \frac{1}{\sqrt{m}} d(S_i, S_j) - \frac{1}{\sqrt{l}} d(\phi(S_i), \phi(S_j)) \right|
    \]

  • Reconstruction error \( L_R \)
    \[
    L_R = \frac{1}{N_s} \sum_{S_i \in S} \frac{1}{\sqrt{m}} d(S_i, \psi \cdot \phi(S_i))
    \]

  • \( \frac{1}{\sqrt{m}} \) and \( \frac{1}{\sqrt{l}} \): scaling coefficients under SoS regularization
    • \( m \): series length, \( l \): DEA length, \( \phi/\psi \): en-/decoder mapping
Sum of Squares Preservation

• Sum of Squares (SoS)
  • $\sum_{i,j} M_{i,j}^2$
    • $M_{i,*}$ denotes series, $M_{*,j}$ denotes position

⇒fix SoS, to learn the transformation
  • i.e., nonlinear encoder mapping
  • SoS works as a regularizer

⇒measures preserved information
  • in linear dimensionality reductions on z-normalized datasets
    • where SoS $\iff$ total variances
  • by selecting the largest eigenvalues

⇒fix transformation (linear), to preserve SoS
SoS-Preservation Regularization

• Regularize SEAnet by SoS preservations:
  1. *z-normalize embeddings* ([LayerNorm2](#))
  2. scale series/DEA by its length in loss functions

• Benefits
  - regularize by preserving SoS → higher-quality embeddings
  - stabilize gradients and latent values (by decreasing Var) → better model convergence

<table>
<thead>
<tr>
<th>Length m</th>
<th>Before Scaling</th>
<th>Scaled by (\sqrt{256/m})</th>
<th>Scaled by (\sqrt{1/m})</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Var</td>
<td>Mean</td>
</tr>
<tr>
<td>256</td>
<td>22.605</td>
<td>0.999</td>
<td>22.605</td>
</tr>
<tr>
<td>128</td>
<td>15.969</td>
<td>0.998</td>
<td>22.583</td>
</tr>
<tr>
<td>96</td>
<td>13.820</td>
<td>0.997</td>
<td>22.569</td>
</tr>
<tr>
<td>16</td>
<td>5.5692</td>
<td>0.984</td>
<td>22.277</td>
</tr>
<tr>
<td>8</td>
<td>3.8772</td>
<td>0.967</td>
<td>21.933</td>
</tr>
</tbody>
</table>
How to generate high-quality DEA on massive data series collections for approximate similarity search?

### Challenges

1. Effective architecture for similarity search?
   - Efficient learning on massive datasets?
     - 100 million 256-length series ≈ 100GB

### Solutions

- **✓ SEAnet: SEries Approximation network**
  - exponentially dilated ResNet + Sum of Squares (SoS) regularization

- **✓ SEAsam: SEA-sampling**
  - sampling based on a sortable series summarization
SEAsam

• Intuition
  • Sampling by dataset’s intrinsic distribution
    ✓ Draw samples by equal-intervals from the ordered set

• How to order series in a dataset?

→ Observation
  • every subsequent bit in one SAX symbol contains a decreasing amount of information about the location
    • ≈ space-filling curves

✓ Order by InvSAX
  • interleaving SAX’s bits
    ⇒ all significant bits across each SAX symbol precede all less significant bits

InvSAX transformation

• interleaving SAX’s bits

⇒ all **significant bits** across each SAX symbol precede all **less significant bits**
InvSAX transformation

- interleaving SAX’s bits

⇒ all significant bits across each SAX symbol precede all less significant bits
InvSAX transformation

- interleaving SAX's bits

⇒ all **significant bits** across each SAX symbol precede all **less significant bits**
InvSAX transformation

• interleaving SAX’s bits

⇒ all significant bits across each SAX symbol precede all less significant bits
InvSAX transformation

• interleaving SAX’s bits

⇒ all **significant bits** across each SAX symbol precede all **less significant bits**

<table>
<thead>
<tr>
<th>SAX</th>
<th>100</th>
<th>101</th>
<th>001</th>
<th>011</th>
</tr>
</thead>
<tbody>
<tr>
<td>InvSAX</td>
<td>110</td>
<td>000</td>
<td>010</td>
<td>111</td>
</tr>
</tbody>
</table>

**Most Significant** | **Least Significant**
Experimental Setup

• Datasets
  • 3 synthetic (RandWalk, F5, F10)
  • 4 real (Seismic, SALD, Deep1B, Astro)

• Comparison methods
  • PAA
  • SEAnet-nD (SEAnet without decoder)
  • TimeNet\textsuperscript{5}, FDJNet\textsuperscript{6}, InceptionTime\textsuperscript{7}
    • Adapted for similarity search

• Hyper-parameter tuning
  • \textasciitilde5,000 models trained

\begin{center}
\begin{tabular}{|c|c|c|}
\hline
Dataset & Length & Dataset Size & Training Size \\
\hline
RandWalk & 256 & 100 million series & 200k SEAsam samples \\
F5 & 256 & & \\
F10 & 256 & & \\
Seismic & 256 & & \\
SALD & 128 & & \\
Deep1B & 96 & & \\
Astro & 256 & & \\
\hline
\end{tabular}
\end{center}

SoS Preservation and SEAsam

SoS Preservation improves tightness of approximate answers
- vs. without SoS

SEAsam improves tightness of approximate answers
- vs. uniformly random sampling
DEA distances vs. Original distances

✓ SEAnet better preserve original distances in the DEA spaces than PAA
Preserving neighborhood in DEA space

✓ SEAnet well preserves the original neighbors in the embedding space
Approximate query answers’ tightness

✓ SEAnet provides tighter approximate answers
Conclusions

1. Proposed learned embeddings (DEA) as a replacement to traditional data series summarizations

2. Developed SEAnet to effectively learn DEA
   • designed using the novel Sum of Square (SoS) preservation regularization

3. Described SEAsam to efficiently train SEAnet on massive datasets
   • DEA by SEAnet outperforms PAA and other SOTA deep embeddings for data series approximate similarity search
   • DEA and SEAnet lead to faster and more accurate data series processing/analytics
   • several promising open research directions
Thanks!

Deep Learning Embeddings for Data Series Similarity Search

*Qitong Wang* and Themis Palpanas

ACM SIGKDD 2021, Singapore