What plays the role of CNNs for sequential data?

Tropical quasisymmetric functions in time-series analysis

J. Diehl (Universität Greifswald)

joint with K. Ebrahimi-Fard (NTNU), N. Tapia (TU Berlin)

August 26th, RSS

Slides at: `https://diehlj.github.io`

Convolutional Neural Networks

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 4 & 3 & 4 & 1 \\ 1 & 2 & 4 & \boxed{3} & 3 \\ 1 & 2 & 3 & 4 & 1 \\ 1 & 3 & 3 & 1 & 1 \\ 3 & 3 & 1 & 1 & 0 \end{pmatrix}$$

Why they work so well (probably ...)

1. Weight sharing.
2. Structure compatible with image data (“receptive field”, approximate translation invariance).

CNNs can, of course, be applied to sequential data.

$$\underset{I}{\begin{pmatrix} 0 & 1 & 1 & \boxed{1 \;\; 0 \;\; 3} & 0 \end{pmatrix}} * \underset{K}{\begin{pmatrix} \boxed{1 \;\; 0 \;\; 1} \end{pmatrix}} = \underset{I * K}{\begin{pmatrix} 1 & 2 & 1 & \boxed{4} & 0 \end{pmatrix}}$$

Does it make sense?

1. Weight sharing. ✓
2. Structure compatible with time-series data ?

Using a CNN to answer:
"Did a person visit Rome directly before visiting London?"

Hamburg	Berlin	Rome	London	Amsterdam

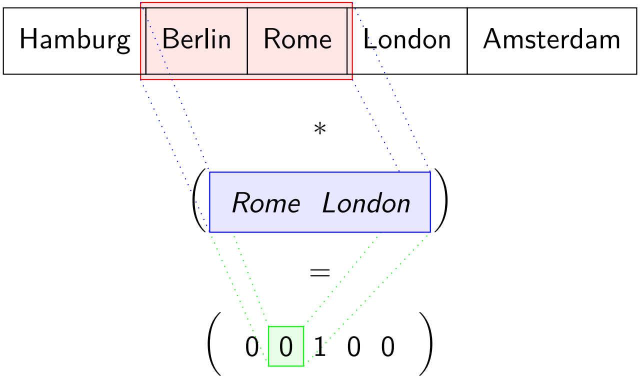
$$*$$

$$\begin{pmatrix} \boxed{\textit{Rome} \quad \textit{London}} \end{pmatrix}$$

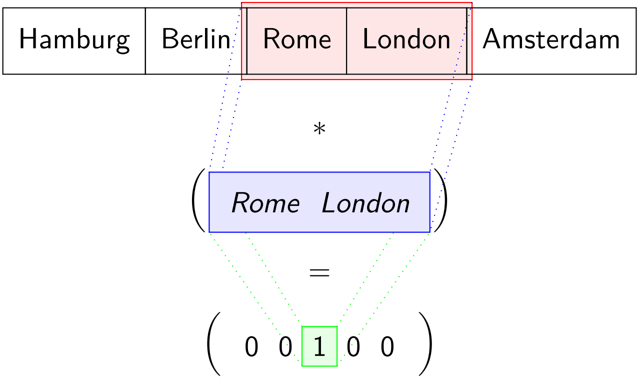
$$=$$

$$\begin{pmatrix} \boxed{0} & 0 & 1 & 0 & 0 \end{pmatrix}$$

Using a CNN to answer:
"Did a person visit Rome directly before visiting London?"



Using a CNN to answer:
“Did a person visit Rome directly before visiting London?”



But what if the person visits Rome some time before visiting London?

Hamburg	Rome	Berlin	Amsterdam	London

A (one-layer) CNN has difficulties detecting this (unless the kernel is large enough).

Chronological question:

"Did a person visit Rome some time before visiting London?"

Hamburg	Rome	Berlin	Amsterdam	London

$$*$$

$$\begin{pmatrix} \boxed{\mathit{Rome} \quad \mathit{London}} \end{pmatrix}$$

$$=$$

$$\begin{pmatrix} \boxed{0} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Chronological question:

"Did a person visit Rome some time before visiting London?"

Hamburg	Rome	Berlin	Amsterdam	London

$$*$$

$$\begin{pmatrix} \boxed{\textit{Rome} \quad \textit{London}} \end{pmatrix}$$

$$=$$

$$\begin{pmatrix} 0 & \boxed{0} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Chronological question:

"Did a person visit Rome some time before visiting London?"

Hamburg	Rome	Berlin	Amsterdam	London

$$*$$

$$\left(\boxed{\mathit{Rome} \quad \mathit{London}} \right)$$

$$=$$

$$\begin{pmatrix} 0 & 0 & \boxed{0} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Chronological question:

“Did a person visit Rome some time before visiting London?”

Hamburg	Rome	Berlin	Amsterdam	London

$$*$$

$$\left(\boxed{\mathit{Rome}\quad \mathit{London}}\right)$$

$$=$$

$$\left(\begin{matrix} 0 & 0 & 0 & \boxed{0} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}\right)$$

Chronological question:

"Did a person visit Rome some time before visiting London?"

Hamburg	Rome	Berlin	Amsterdam	London

$$*$$

$$\left(\boxed{\textit{Rome} \quad \textit{London}} \right)$$

$$=$$

$$\left(\begin{array}{cccccccccccccccccccc} 0 & 0 & 0 & 0 & \boxed{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right)$$

More formal

Let

$$K : \text{Cities} \times \text{Cities} \to \{\texttt{true}, \texttt{false}\}$$

$(\text{cityA}, \text{cityB}) \mapsto \left(\text{cityA} =\right.$  $\left.\right) \wedge \left(\text{cityB} =\right.$  $\left.\right)$

$$\text{pool} : \{\texttt{true}, \texttt{false}\}^{\binom{n_{\text{in}}}{2}} \to \{\texttt{true}, \texttt{false}\}$$

$$z \mapsto z_1 \ \vee \ z_2 \ \vee \ \ldots \ \vee \ z_{\binom{n_{\text{in}}}{2}}.$$

Then

$\text{pool}\left(K(x_I) : I \in \binom{[n_{\text{in}}]}{2}\right) = \bigvee_{0 < i_1 < i_2 \leq n_{\text{in}}} \left(x_{i_1} =\right.$  $\left.\right) \wedge \left(x_{i_2} =\right.$  $\left.\right),$

is `true` <u>if and only if</u> Rome was visited some time before London.

(There is nothing "learnable" here yet, we'll come to this later.)

First, we want to deal with a problem: $\binom{n_{\mathrm{in}}}{2}$ gets large real quick !

(There is nothing "learnable" here yet, we'll come to this later.)

First, we want to deal with a problem: $\binom{n_{\text{in}}}{2}$ gets large real quick!

To clarify, let us do 3 cities whose ordered visit we want to detect:

$K(\dots) := (\text{cityA} =$  $) \wedge (\text{cityB} =$  $) \wedge (\text{cityC} =$  $)$

$$\operatorname{pool}\left(K(x_I) : I \in \binom{[n_{\text{in}}]}{3}\right) := \bigvee_{I \in \binom{[n_{\text{in}}]}{3}} K(x_I).$$

This needs $O(n_{\text{in}}^3)$ evaluations of K. $\lightning$

But! There is a better way.

$$\bigvee_{I \in \binom{[n_{\mathrm{in}}]}{3}} K(x_I) = \bigvee_{i_1 < i_2 < i_3} (x_{i_1} = \quad) \wedge (x_{i_2} = \quad) \wedge (x_{i_3} = \quad)$$

$$= \bigvee_{i_3} \left(\bigvee_{i_1 < i_2 < i_3} (x_{i_1} = \quad) \wedge (x_{i_2} = \quad) \right) \wedge (x_{i_3} = \quad)$$

$$=: \bigvee_{i_3} \mathrm{pool}'_{i_3} \wedge (x_{i_3} = \quad).$$

Only n_{in} evaluations!

Further

$\mathrm{pool}'_{i_3} = \bigvee_{i_1<i_2<i_3} (x_{i_1} =$  $) \wedge (x_{i_2} =$  $)$

$= \bigvee_{i_2<i_3} \left(\bigvee_{i_1<i_2} (x_{i_1} =$  $) \right) \wedge (x_{i_2} =$  $)$

$=: \bigvee_{i_2<i_3} \mathrm{pool}''_{i_2} \wedge (x_{i_2} =$  $).$

Only n_{in} evaluations (to calculate all of $\mathrm{pool}'_{\bullet}$)!

Finally,

$\mathrm{pool}''_{i_2} = \bigvee_{i_1<i_2} (x_{i_1} =$  $)$

Only n_{in} evaluations (to calculate all of $\mathrm{pool}''_{\bullet}$)!

total amount of evaluations: $O(3n_{\mathrm{in}}) = O(n_{\mathrm{in}})$

What have we achieved?

We calulated

$$\operatorname{pool}\left(K(x_I) : I \in \binom{[n_{\text{in}}]}{3}\right)$$

$$= \bigvee_{I \in \binom{[n_{\text{in}}]}{3}} K(x_I)$$

$= \bigvee_{i_1 < i_2 < i_3} (x_{i_1} =$  $) \wedge (x_{i_2} =$  $) \wedge (x_{i_3} =$  $),$

which, on paper, costs $O(n_{\text{in}}^3)$, in only $O(n_{\text{in}})$ time !

What did we use?

- $\wedge$ distributes over $\vee$
- $\wedge$ and $\vee$ are associative

And that's it.

What have we achieved?

We calulated

$$\mathrm{pool}\left(K(x_I) : I \in \binom{[n_{\mathrm{in}}]}{3}\right) = \bigvee_{I \in \binom{[n_{\mathrm{in}}]}{3}} K(x_I)$$

$= \bigvee_{i_1<i_2<i_3} (x_{i_1} =$  $) \wedge (x_{i_2} =$  $) \wedge (x_{i_3} =$  $),$

which, on paper, costs $O(n_{\mathrm{in}}^3)$, in only $O(n_{\mathrm{in}})$ time !

What did we use?

- $\wedge$ distributes over $\vee$
- $\wedge$ and $\vee$ are associative

And that's it.

Definition

The tuple $(\mathbb{S}, \oplus_{\mathbb{S}}, \odot_{\mathbb{S}}, \mathbf{0}_{\mathbb{S}}, \mathbf{1}_{\mathbb{S}})$ is a commutative **semiring** if

- $(\mathbb{S}, \oplus_{\mathbb{S}}, \mathbf{0}_{\mathbb{S}})$ is a commutative monoid with unit $\mathbf{0}_{\mathbb{S}}$
- $(\mathbb{S}, \odot_{\mathbb{S}}, \mathbf{1}_{\mathbb{S}})$ is a commutative monoid with unit $\mathbf{1}_{\mathbb{S}}$
- $\mathbf{0}_{\mathbb{S}} \odot_{\mathbb{S}} \mathbb{S} = \{\mathbf{0}_{\mathbb{S}}\}$
- multiplication distributes over addition, i.e.

$$a \odot_{\mathbb{S}} (b \oplus_{\mathbb{S}} c) = (a \odot_{\mathbb{S}} b) \oplus_{\mathbb{S}} (a \odot_{\mathbb{S}} c)$$

Examples of semirings

- any commutative ring
- boolean semiring $(\{\mathtt{false}, \mathtt{true}\}, \vee, \wedge, \mathtt{false}, \mathtt{true})$
- min-plus (“tropical”) semiring $(\mathbb{R} \cup \{+\infty\}, \min, +, +\infty, 0)$
- possibilistic (or Viterbi or Bayesian) semiring $([0, 1], \max, \cdot, 0, 1)$

Definition

The tuple $(\mathbb{S}, \oplus_{\mathbb{s}}, \odot_{\mathbb{s}}, \mathbf{0}_{\mathbb{s}}, \mathbf{1}_{\mathbb{s}})$ is a commutative **semiring** if

- $(\mathbb{S}, \oplus_{\mathbb{s}}, \mathbf{0}_{\mathbb{s}})$ is a commutative monoid with unit $\mathbf{0}_{\mathbb{s}}$
- $(\mathbb{S}, \odot_{\mathbb{s}}, \mathbf{1}_{\mathbb{s}})$ is a commutative monoid with unit $\mathbf{1}_{\mathbb{s}}$
- $\mathbf{0}_{\mathbb{s}} \odot_{\mathbb{s}} \mathbb{S} = \{\mathbf{0}_{\mathbb{s}}\}$
- multiplication distributes over addition, i.e.

$$a \odot_{\mathbb{s}} (b \oplus_{\mathbb{s}} c) = (a \odot_{\mathbb{s}} b) \oplus_{\mathbb{s}} (a \odot_{\mathbb{s}} c)$$

Examples of semirings

- any commutative ring
- boolean semiring
  $(\{\mathtt{false}, \mathtt{true}\}, \vee, \wedge, \mathtt{false}, \mathtt{true})$
- min-plus ("tropical") semiring
  $(\mathbb{R} \cup \{+\infty\}, \min, +, +\infty, 0)$
- possibilistic (or Viterbi or Bayesian) semiring
  $([0, 1], \max, \cdot, 0, 1)$

Examples of semirings $(\mathbb{S}, \oplus_{\mathbb{S}}, \odot_{\mathbb{S}}, \mathbf{0}_{\mathbb{S}}, \mathbf{1}_{\mathbb{S}})$

- semiring of subsets of a set M
  $(2^M, \cup, \cap, \emptyset, M)$
- any distributive lattice (with minimal and maximal element)
- ...

They are of huge interest in computer science / automata theory.

Corollary (DEFT '20)

Let $(\mathbb{S}, \oplus_{\mathbb{S}}, \odot_{\mathbb{S}}, \mathbf{0}_{\mathbb{S}}, \mathbf{1}_{\mathbb{S}})$ *be a commutative semiring. Then*

$$\operatorname{pool}\left(z_I : I \in \binom{[n_{\mathrm{in}}]}{k}\right) := \bigoplus_{\mathbb{S}\; i_1<\cdots<i_k\leq n_{\mathrm{in}}} z_{i_1}^{\odot_{\mathbb{S}}\alpha_1} \odot_{\mathbb{S}} \cdots \odot_{\mathbb{S}} z_{i_k}^{\odot_{\mathbb{S}}\alpha_k},$$

is calculable in $O(n_{\mathrm{in}})$*-time.*

Examples

- Over the ring $\mathbb{R}$

$$\sum_{i_1 < \cdots < i_k} z_{i_1}^{\alpha_1} \ldots z_{i_k}^{\alpha_k},$$

$\rightsquigarrow$ iterated-sums signature (quasisymmetric functions)

This has a long history.

- Graham '13 "Sparse arrays of signatures for ...".
- Lyons, Ni, Oberhauser '14 "A feature set for streams ..."
- various works by L Jin et al '15 on Chinese character recognition.
- Kiraly, Oberhauser '16 "Kernels for sequentially ordered data".
- Lyons, Oberhauser '17 "Sketching the order of events".
- D '13, D,Reizenstein '19 on invariant features.
- D,Ebrahimi-Fard,Tapia '19 "Time warping invariants".
- Kidger, Bonnier, Arribas, Salvi, Lyons '19 "Deep Signature Transforms".
- Toth, Bonnier, Oberhauser '20 "Seq2Tens".

In these works it progressively emerged that it is helpful to learn the signature-type features.

Paraphrasing

$$\rightsquigarrow \sum_{i_1 < \cdots < i_k} f_{\theta_1}(z_{i_1}) \cdot \cdots \cdot f_{\theta_k}(z_{i_k}).$$

with $f_\theta : \mathbb{R}^d \to \mathbb{R}$.

We propose to boil this down to the bare minimum needed, namely

distributivity and associativity,

to arrive at a richer set of features.

$$\rightsquigarrow \bigoplus_{i_1 < \cdots < i_k}{}_{\mathbb{S}} f_{\theta_1}(z_{i_1}) \odot_{\mathbb{S}} \cdots \odot_{\mathbb{S}} f_{\theta_k}(z_{i_k}),$$

with $f_\theta : \mathbb{R}^d \to \mathbb{S}$.

In these works it progressively emerged that it is helpful to learn the signature-type features.

Paraphrasing

$$\rightsquigarrow \sum_{i_1<\cdots<i_k} f_{\theta_1}(z_{i_1}) \cdot \dots \cdot f_{\theta_k}(z_{i_k}).$$

with $f_\theta : \mathbb{R}^d \to \mathbb{R}$.

We propose to boil this down to the bare minimum needed, namely

distributivity and associativity,

to arrive at a richer set of features.

$$\rightsquigarrow \bigoplus_{i_1<\cdots<i_k}{}_{\mathbb{S}} f_{\theta_1}(z_{i_1}) \odot_{\mathbb{S}} \cdots \odot_{\mathbb{S}} f_{\theta_k}(z_{i_k}),$$

with $f_\theta : \mathbb{R}^d \to \mathbb{S}$.

Examples

- Over the tropical semiring

$$\min_{i_1<\cdots<i_k} \{\alpha_1 \cdot z_{i_1} + \cdots + \alpha_k \cdot z_{i_k}\}$$

$\rightsquigarrow$ tropical-sums signature
(tropical quasisymmetric functions [DEFT '20])

By leaving the strict setting of tropical-sums (as in the previous slide), we can do a learnable version of the visiting-cities example:

- Fix some embedding z_i of the visited cities in $\mathbb{R}^d$ (e.g. one-hot-encoding).
- Fix parametrized functions $f_\theta : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$.

$$\rightsquigarrow \max_{i_1<i_2} \left\{ f_{\theta_1}(z_{i_1}) + f_{\theta_2}(z_{i_2}) \right\}.$$

Non-example

Not all type of sums work. For example

$$\sum_{i_1 < \cdots < i_k} \sigma(x_{i_1} + .. + x_{i_k}),$$

for a general nonlinear σ cannot be efficiently computed (since one can frame NP-complete problems in this form ..).

The algebraic setting

For $z_1, z_2, \dots \in \mathbb{S}$, $s < t$,

$$\left\langle \mathrm{ISS}^{\mathbb{S}}_{s,t}(z), w \right\rangle := \bigoplus_{s<i_1<\dots<i_k<t+1}{}_{\mathbb{S}} \; z_{i_1}^{\odot_{\mathbb{S}} w_1} \odot_{\mathbb{S}} \cdots \odot_{\mathbb{S}} z_{i_k}^{\odot_{\mathbb{S}} w_k}$$

Theorem (DEFT '20)

1. *(Quasi-shuffle identity)*

$$\left\langle \mathrm{ISS}^{\mathbb{S}}_{s,t}(z), w \right\rangle \odot_{\mathbb{S}} \left\langle \mathrm{ISS}^{\mathbb{S}}_{s,t}(z), u \right\rangle = \left\langle \mathrm{ISS}^{\mathbb{S}}_{s,t}(z), w \star u \right\rangle$$

2. *(Chen's identity) For $s < t < u$,*

$$\left\langle \mathrm{ISS}^{\mathbb{S}}_{s,u}(z), w \right\rangle = \bigoplus_{w'w''=w}{}_{\mathbb{S}} \left\langle \mathrm{ISS}^{\mathbb{S}}_{s,t}(z), w' \right\rangle \odot_{\mathbb{S}} \left\langle \mathrm{ISS}^{\mathbb{S}}_{t,u}(z), w'' \right\rangle$$

3. $\mathrm{ISS}^{\mathbb{S}}_{0,\infty}(z)$ *is invariant to inserting* $\mathbf{0}_{\mathbb{S}}$ *into* z.

Summary

- Expressions of the from

$$\mathrm{pool}\left(K(x_I) : I \subset \binom{n_{\mathrm{in}}}{k}\right)$$

extract meaningful, chronological information of time series. In this generality they are computationally untractable.

- Semirings provide a large class of examples that are tractable, namely

$$\bigoplus_{i_1<\cdots<i_k}^{\mathbb{S}} f_{\theta_1}(x_{i_1}) \odot \ldots f_{\theta_k}(x_{i_k}).$$

- In the special case of monomial f, we are led to the iterated-sums signature over a semiring

$$\left\langle \mathrm{ISS}^{\mathbb{S}}_{s,t}(z), w \right\rangle = \bigoplus_{s<i_1<\cdots<i_k<t+1}^{\mathbb{S}} z_{i_1}^{\odot_{\mathbb{S}} w_1} \odot_{\mathbb{S}} \cdots \odot_{\mathbb{S}} z_{i_k}^{\odot_{\mathbb{S}} w_k}.$$

Thank you!