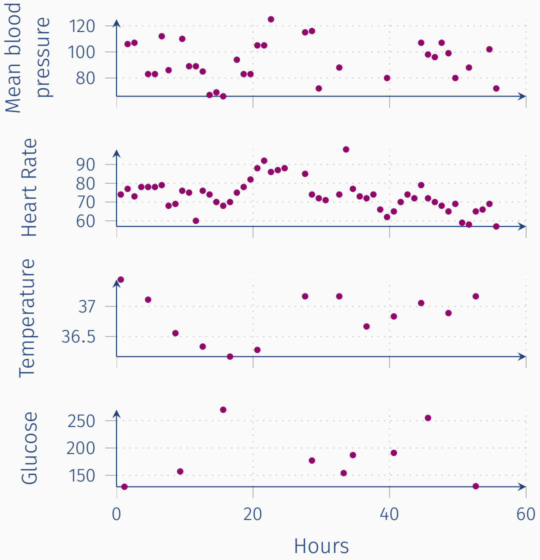
Set Functions for Time Series

ICML 2020

Max Horn, Michael Moor, Christian Bock, Bastian Rieck and Karsten Borgwardt

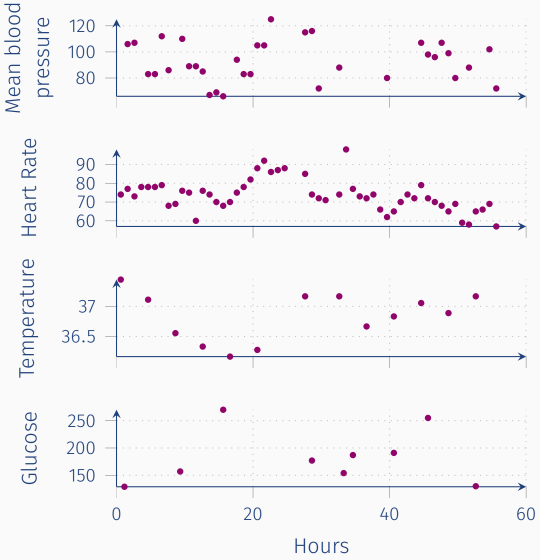
Machine Learning and Computational Biology Group, ETH Zurich

Motivation - Medical time series



Challenges

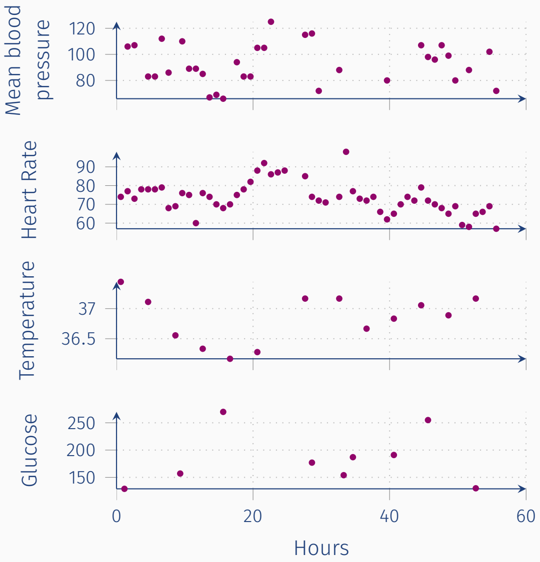
Motivation - Medical time series



Challenges

- Irregular sampling of data

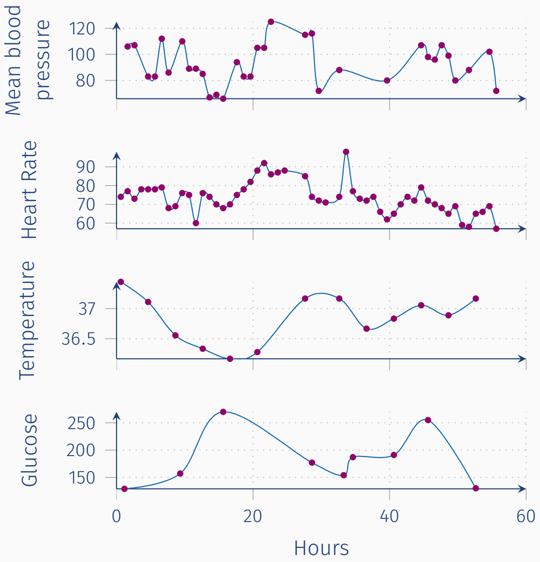
Motivation - Medical time series



Challenges

- Irregular sampling of data
- High demands on interpretability

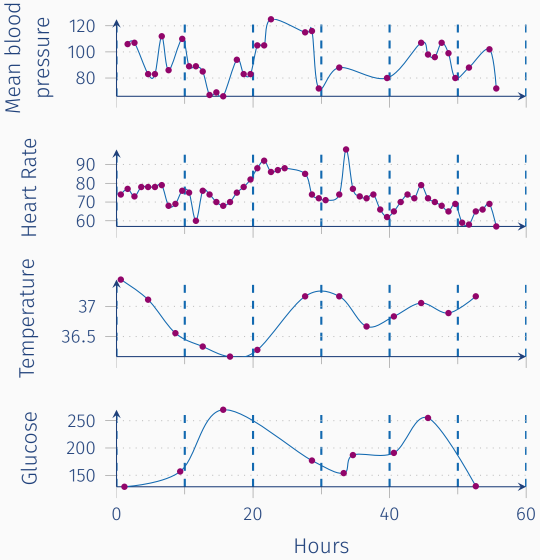
Motivation - Medical time series



Challenges

- Irregular sampling of data
- High demands on interpretability

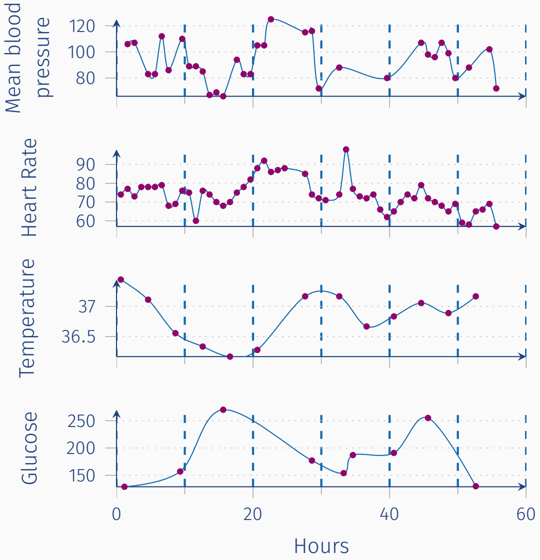
Motivation - Medical time series



Challenges

- Irregular sampling of data
- High demands on interpretability

Motivation - Medical time series



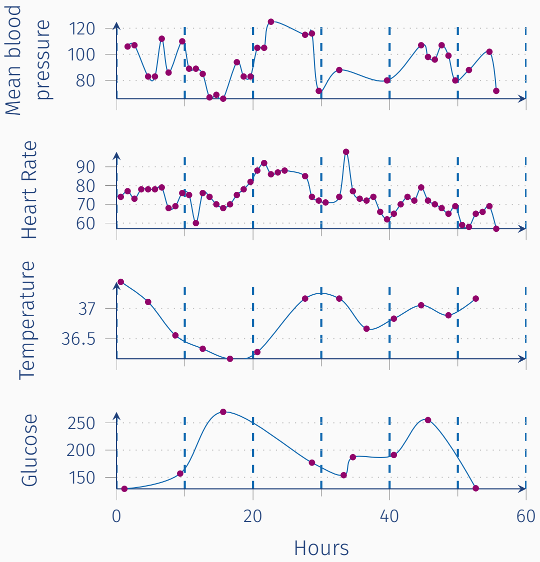
Challenges

- Irregular sampling of data
- High demands on interpretability

Problem statement

Learning classification models on irregularly-sampled time series without prior imputation.

Motivation - Medical time series



Challenges

- Irregular sampling of data
- High demands on interpretability

Problem statement

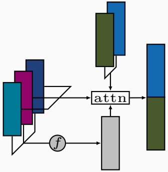
Learning classification models on irregularly-sampled time series without prior imputation.

Set **F**unctions for **T**ime Series

→ Time series classification as set classification

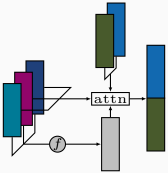
Contributions

Contributions

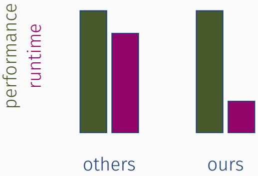


New approach for Irregularly-sampled Time Series

Contributions

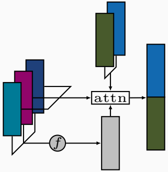


New approach for Irregularly-sampled Time Series

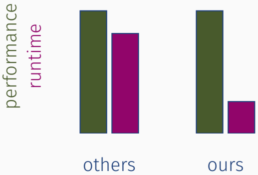


Competitive Performance with Lower Runtime

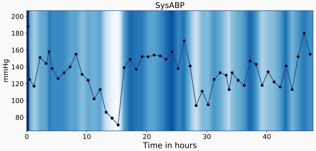
Contributions



New approach for Irregularly-sampled Time Series

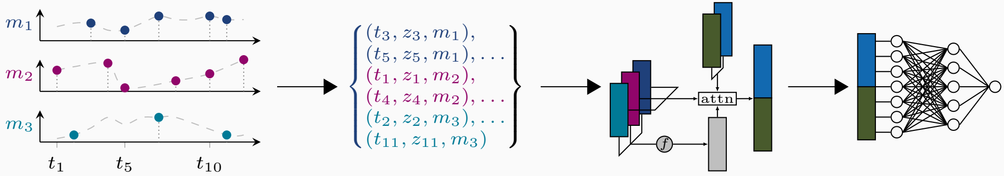


Competitive Performance with Lower Runtime



Per Observation Contributions

Architecture Overview



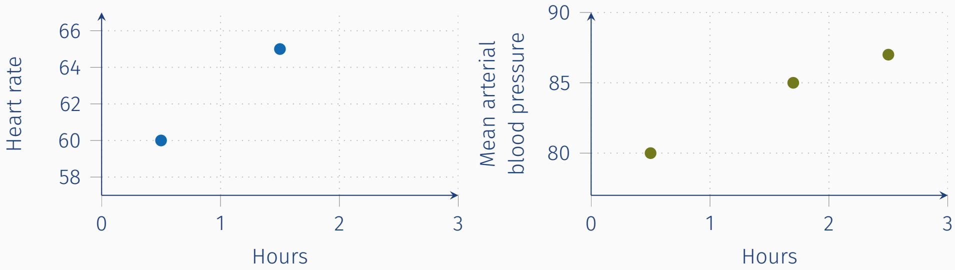
Irregularly sampled Multivariate Time Series

Set Encoding

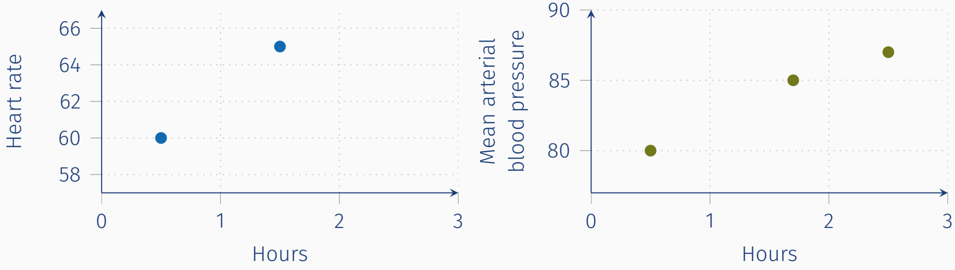
Embedding, Attention, and Aggregation

Classification

Key idea - Time Series as Sets of Observations

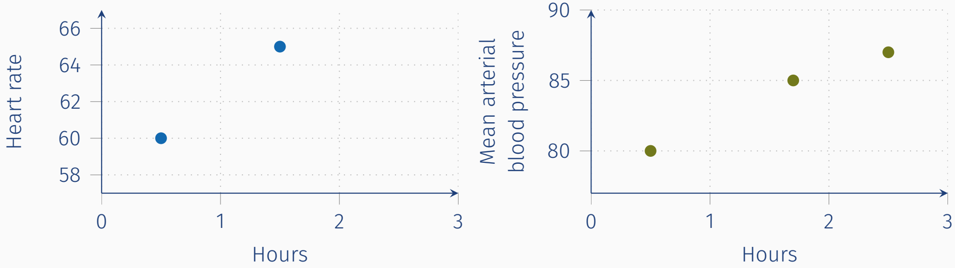


Key idea - Time Series as Sets of Observations



Each observation s_j is represented as a tuple (t_j, z_j, m_j)

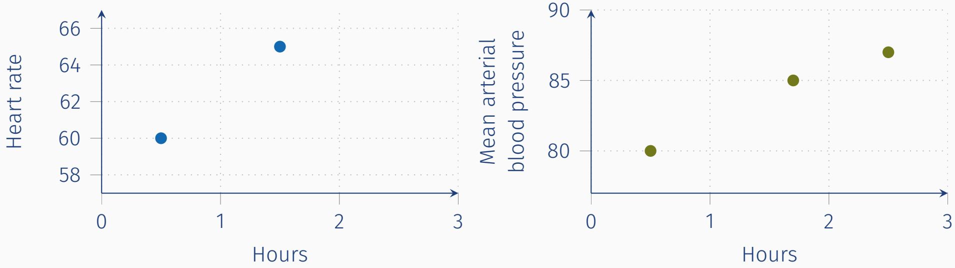
Key idea - Time Series as Sets of Observations



Each observation s_j is represented as a tuple (t_j, z_j, m_j)

$$\mathcal{S} = \{(0.5, 60, 1), (1.5, 65, 1),$$

Key idea - Time Series as Sets of Observations



Each observation s_j is represented as a tuple (t_j, z_j, m_j)

$$\mathcal{S} = \{(0.5, 60, 1), (1.5, 65, 1), (0.5, 80, 2), (1.7, 85, 2), (3, 87, 2)\}$$

Deep Sets [1]

$$f(\mathcal{S}) = g\left(\frac{1}{|\mathcal{S}|}\sum_{s_j \in \mathcal{S}} h\left(s_j\right)\right)$$

where $h\colon \Omega \to \mathbb{R}^d$ and $g\colon \mathbb{R}^d \to \mathbb{R}^C$ are neural networks

[1] Zaheer et al., NeurIPS 2017

Deep Sets [1]

$$f(\mathcal{S}) = g\left(\frac{1}{|\mathcal{S}|}\sum_{s_j \in \mathcal{S}} h(s_j)\right)$$

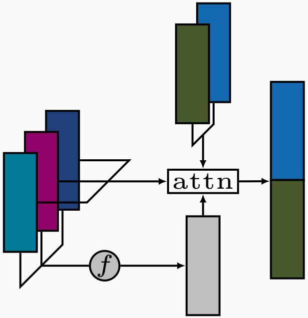
where $h\colon \Omega \to \mathbb{R}^d$ and $g\colon \mathbb{R}^d \to \mathbb{R}^C$ are neural networks

Problem

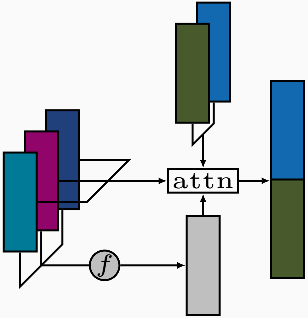
Influence of an element shrinks as $|\mathcal{S}|$ grows!

[1]Zaheer et al., NeurIPS 2017

Set Attention Mechanism

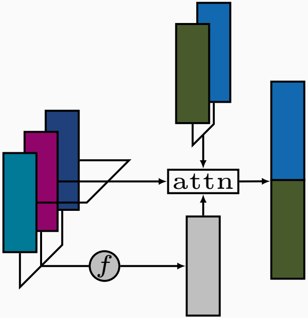


Set Attention Mechanism



Keys: $K_{j,i} = [f(\mathcal{S}), s_j]^T W_i$

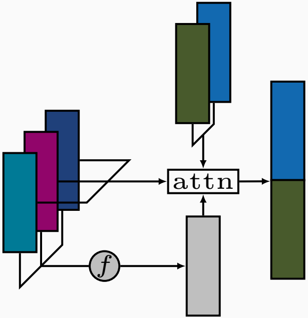
Set Attention Mechanism



Keys: $K_{j,i} = [f(\mathcal{S}), s_j]^T W_i$

Queries: $Q \in \mathbb{R}^{m \times d}$

Set Attention Mechanism

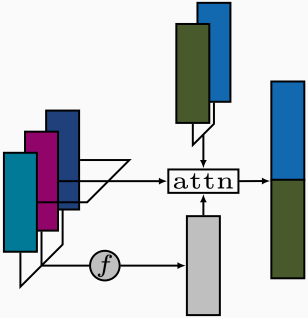


Keys: $K_{j,i} = [f(\mathcal{S}), s_j]^T W_i$

Queries: $Q \in \mathbb{R}^{m \times d}$

Preattentions: $e_{j,i} = \frac{K_{j,i} \cdot Q_i}{\sqrt{d}}$

Set Attention Mechanism



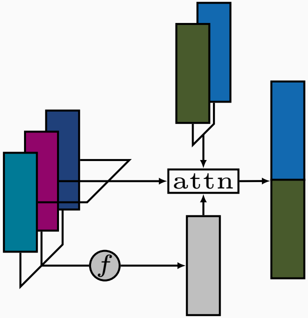
Keys: $K_{j,i} = [f(\mathcal{S}), s_j]^T W_i$

Queries: $Q \in \mathbb{R}^{m \times d}$

Preattentions: $e_{j,i} = \dfrac{K_{j,i} \cdot Q_i}{\sqrt{d}}$

Attentions: $a_{j,i} = \dfrac{\exp(e_{j,i})}{\sum_j \exp(e_{j,i})}$

Set Attention Mechanism



Keys: $K_{j,i} = [f(\mathcal{S}), s_j]^T W_i$

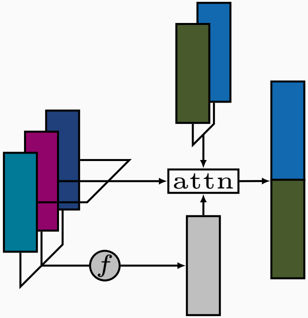
Queries: $Q \in \mathbb{R}^{m \times d}$

Preattentions: $e_{j,i} = \frac{K_{j,i} \cdot Q_i}{\sqrt{d}}$

Attentions: $a_{j,i} = \frac{\exp(e_{j,i})}{\sum_j \exp(e_{j,i})}$

Values: $V_i = \sum_j a_{j,i} h_\theta(s_j)$

Set Attention Mechanism



$$\text{Keys:} \quad K_{j,i} = [f(\mathcal{S}), s_j]^T W_i$$

$$\text{Queries:} \quad Q \in \mathbb{R}^{m \times d}$$

$$\text{Preattentions:} \quad e_{j,i} = \frac{K_{j,i} \cdot Q_i}{\sqrt{d}}$$

$$\text{Attentions:} \quad a_{j,i} = \frac{\exp(e_{j,i})}{\sum_j \exp(e_{j,i})}$$

$$\text{Values:} \quad V_i = \sum_j a_{j,i} h_\theta(s_j)$$

$$\mathcal{L}(\theta, \psi) = \mathbb{E}_{(\mathcal{S},y)\in\mathcal{D}} \left[\ell \left(y; g_\psi \left(\sum_{s_j \in \mathcal{S}} a(\mathcal{S}, s_j) h_\theta(s_j) \right) \right) \right]$$

Experimental setup

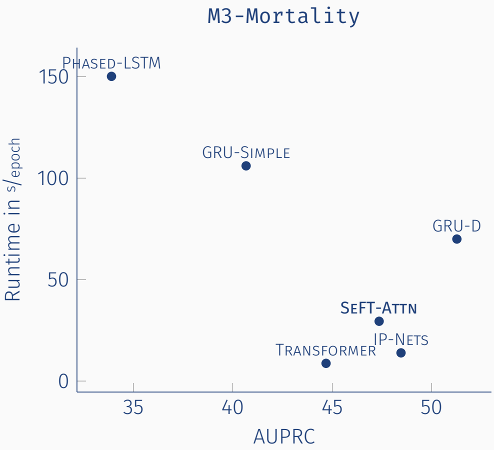
Datasets

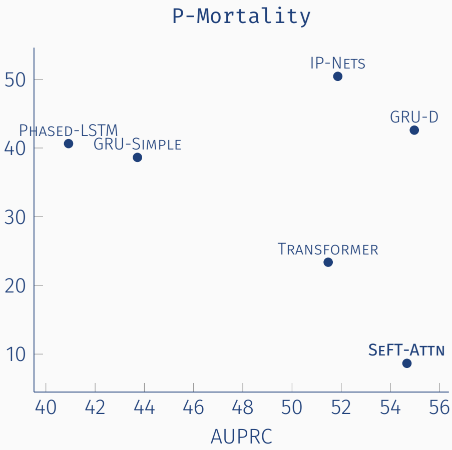
- Two mortality prediction tasks - MIMIC-III (`M3-Mortality`) and Physionet 2012 (`P-Mortality`)
- Sepsis early recognition task - Physionet 2019 Challenge

Comparison partners

- PHASED-LSTM – *Neil et al., NeurIPS 2017*
- TRANSFORMER – *Vaswani et al., NeurIPS 2017*
- GRU-SIMPLE & GRU-D – *Che et al., Scientific reports 2018*
- IP-NETS – *Shukla & Marlin, ICLR 2019*

Results - Performance vs. Runtime





Results - Sepsis Early Prediction

Model	B-Accuracy	AUPRC	U_{norm}	s/epoch
GRU-D	51.15	5.82	0.021 21	190.41
GRU-Simple	50.69	6.97	0.013 09	92.90
IP-Nets	**78.02**	*37.60*	**0.513 27**	232.92
Phased-LSTM	50.09	6.40	0.001 59	110.49
Transformer	*77.84*	**55.30**	*0.499 74*	*71.70*
SeFT-Attn	74.50	8.78	0.341 20	**62.91**

Results - Sepsis Early Prediction

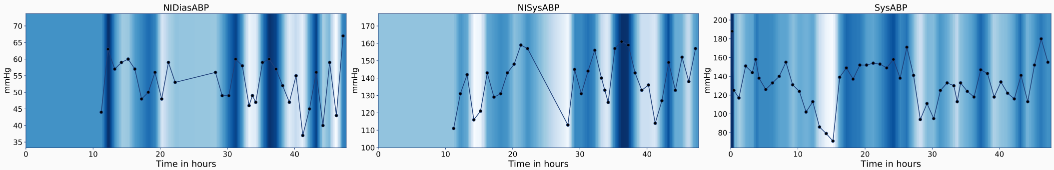
Model	B-Accuracy	AUPRC	U_{norm}	s/epoch
GRU-D	51.15	5.82	0.021 21	190.41
GRU-SIMPLE	50.69	6.97	0.013 09	92.90
IP-NETS	**78.02**	*37.60*	**0.513 27**	232.92
PHASED-LSTM	50.09	6.40	0.001 59	110.49
TRANSFORMER	*77.84*	**55.30**	*0.499 74*	*71.70*
SEFT-Attn	74.50	8.78	0.341 20	**62.91**

Possible Leakage of Future Information

IP-NETS Through unmasked interpolation

TRANSFORMER Through layer normalization

Results - Interpretability

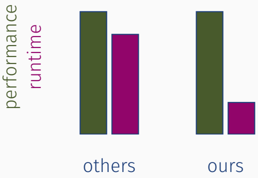


Uniquely allows a **per-observation** quantification of importance

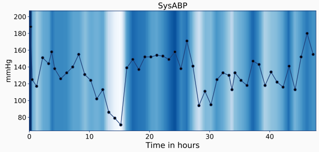
Summary

$$\left\{ \begin{array}{l} (t_3, z_3, m_1), \\ (t_5, z_5, m_1), \ldots \\ (t_1, z_1, m_2), \\ (t_4, z_4, m_2), \ldots \\ (t_2, z_2, m_3), \ldots \\ (t_{11}, z_{11}, m_3) \end{array} \right\}$$

New approach for irregularly-sampled time series



Competitive performance with lower runtime



Per observation contributions

For further information please check out our paper.