mTAN: Multi-Time Attention Networks

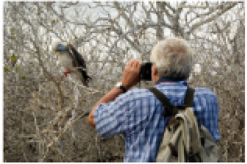
For Irregularly Sampled Time Series

Satya Narayan Shukla, Benjamin Marlin

University of Massachusetts Amherst

Irregularly Sampled Time Series

Time series with non-uniform time intervals between successive measurements



Challenges

- Each time series observed at arbitrary time points
- Different data cases may have different numbers of observations
- Lack of alignment of observation time points across different dimension in multivariate case
- Most machine learning models typically assume fully-observed, fixed-size feature representations

mTAN: Multi-Time Attention Networks

- Continuous-time interpolation-based models
- Continuous-time embedding coupled with Time Attention
- Replace the use of a fixed similarity kernel
- More representational flexibility than previous interpolation-based models
- Time Embedding

$$\phi_h(t)[i] = \begin{cases} \omega_{0h} \cdot t + \alpha_{0h}, & \text{if} \quad i = 0 \\ \sin(\omega_{ih} \cdot t + \alpha_{ih}), & \text{if} \quad 0 < i < d_r \end{cases}$$

Multi-Time Attention

Input Query time point t, keys and values in form of observation time points and values

Output J dimensional embedding at time t

$$\text{mTAN}(t, \mathbf{s})[j] = \sum_{h=1}^{H} \sum_{d=1}^{D} \hat{x}_{hd}(t, \mathbf{s}) \cdot U_{hdj}$$

$$\hat{x}_{hd}(t, \mathbf{s}) = \sum_{i=1}^{L_d} \kappa_h(t, t_{id})\, x_{id}$$

$$\kappa_h(t, t_{id}) = \frac{\exp\left(\phi_h(t) W V^T \phi_h(t_{id})^T / \sqrt{d_k}\right)}{\sum_{i'=1}^{L_d} \exp\left(\phi_h(t) w v^T \phi_h(t_{i'd})^T / \sqrt{d_k}\right)}$$

Learning the time embeddings provides **mTAN** with flexibility to learn complex temporal kernel functions $\kappa_h(t, t')$

Encoder-Decoder Framework

- Discretized mTAN or **mTAND**, produce output representation at a given set of reference time points $\mathbf{r} = [r_1, ..., r_T]$

Generative Process

$$\begin{aligned} \mathbf{z}_k &\sim p(\mathbf{z}_k) \\ \mathbf{h}_{RNN}^{dec} &= \text{RNN}^{dec}(\mathbf{z}) \\ \mathbf{h}_{TAN}^{dec} &= \text{mTAND}^{dec}(\mathbf{t}, \mathbf{h}_{RNN}^{dec}) \\ x_{id} &\sim p(\mathbf{x}_{id} | f^{dec}(\mathbf{h}_{i,TAN}^{dec})[d]) \end{aligned}$$

Inference Network

$$\begin{aligned} \mathbf{h}_{TAN}^{enc} &= \text{mTAND}^{enc}(\mathbf{r}, \mathbf{s}) \\ \mathbf{h}_{RNN}^{enc} &= \text{RNN}^{enc}(\mathbf{h}_{TAN}^{enc}) \\ \mathbf{z}_k &\sim q(\mathbf{z}_k | \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2) \\ \boldsymbol{\mu}_k &= f_{\mu}^{enc}(\mathbf{h}_{k,RNN}^{enc}) \\ \boldsymbol{\sigma}_k^2 &= \exp(f_{\sigma}^{enc}(\mathbf{h}_{k,RNN}^{enc})) \end{aligned}$$

Learning

- Maximize a normalized variational lower bound on the log marginal likelihood based on ELBO

Unsupervised Learning

$$\mathcal{L}_{\mathsf{NVAE}}(\theta, \gamma) = \sum_{n=1}^{N} \frac{1}{\sum_d L_{dn}} \Big(\mathbb{E}_{q_\gamma(\mathsf{z}|\mathsf{r},\mathsf{s}_n)}[\log p_\theta(\mathsf{x}_n|\mathsf{z}, \mathsf{t}_n)] - D_{\mathrm{KL}}(q_\gamma(\mathsf{z}|\mathsf{r}, \mathsf{s}_n)||p(\mathsf{z})) \Big)$$

$$D_{\mathrm{KL}}(q_\gamma(\mathsf{z}|\mathsf{r}, \mathsf{s}_n)||p(\mathsf{z})) = \sum_{i=1}^{T} D_{\mathrm{KL}}(q_\gamma(\mathsf{z}_i|\mathsf{r}, \mathsf{s}_n)||p(\mathsf{z}_i))$$

$$\log p_\theta(\mathsf{x}_n|\mathsf{z}, \mathsf{t}_n) = \sum_{d=1}^{D} \sum_{j=1}^{L_{dn}} \log p_\theta(x_{jdn}|\mathsf{z}, t_{jdn})$$

Supervised Learning

$$\mathcal{L}_{\mathsf{sup}}(\theta, \gamma, \delta) = \mathcal{L}_{\mathsf{NVAE}}(\theta, \gamma) + \lambda \mathbb{E}_{q_\gamma(\mathsf{z}|\mathsf{r},\mathsf{s}_n)} \log p_\delta(y_n|\mathsf{z})$$

$$y^* = \operatorname*{arg\,max}_{y \in \mathcal{Y}} \ \mathbb{E}_{q_\gamma(\mathsf{z}|\mathsf{r},\mathsf{s})}[\log p_\delta(y|\mathsf{z})]$$

Experiments

- **mTAN** performs better than state-of-the-art models
- $1 \sim 2$ orders of magnitude faster training

Table 1: PhysioNet: Interpolation

Model	MSE ($\times 10^{-3}$)
RNN-Impute	3.243 ± 0.275
RNN-Δ_t	3.520 ± 0.276
RNN-Decay	3.215 ± 0.276
RNN GRU-D	3.384 ± 0.274
RNN-VAE	5.390 ± 0.249
ODE-RNN	2.361 ± 0.086
L-ODE (RNN)	3.907 ± 0.252
L-ODE (ODE)	2.118 ± 0.271
mTAND-Full	$\mathbf{0.424 \pm 0.018}$

Table 2: PhysioNet: Classification

Model	AUC Score	time
RNN-Impute	0.764 ± 0.016	0.5
RNN-Δ_t	0.787 ± 0.014	0.5
RNN-Decay	0.807 ± 0.003	0.7
RNN GRU-D	0.818 ± 0.008	0.7
RNN-VAE	0.515 ± 0.040	2.0
ODE-RNN	0.833 ± 0.009	16.5
L-ODE-RNN	0.781 ± 0.018	6.7
L-ODE-ODE	0.829 ± 0.004	22.0
IP-Nets	0.819 ± 0.006	1.3
mTAND-Enc	$\mathbf{0.854 \pm 0.001}$	0.08
mTAND-Full	$\mathbf{0.858 \pm 0.004}$	0.19