Uncertainty-DTW for Time Series and Sequences

Lei.Wang@data61.csiro.au*,1,2 Piotr.Koniusz@data61.csiro.au*,2,1
*equal contribution [1]Australian National University [2]Data61/CSIRO

Motivation and key ideas

- The role of soft-DTW is to evaluate the (relaxed) DTW distance between a pair of sequences $\boldsymbol{\Psi} \equiv [\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_\tau] \in \mathbb{R}^{d' \times \tau}$, $\boldsymbol{\Psi}' \equiv [\boldsymbol{\psi}'_1, ..., \boldsymbol{\psi}'_{\tau'}] \in \mathbb{R}^{d' \times \tau'}$ of lengths τ and τ', respectively. Under its transportation plan $\mathcal{A}_{\tau,\tau'}$, each path $\boldsymbol{\Pi} \in \mathcal{A}_{\tau,\tau'}$ is evaluated to ascertain the path distance, and the smallest distance is 'selected' by the soft minimum: $d^2_{\text{DTW}}(\boldsymbol{\Psi}, \boldsymbol{\Psi}') = \text{SoftMin}_\gamma \left([\langle \boldsymbol{\Pi}, \boldsymbol{D}(\boldsymbol{\Psi}, \boldsymbol{\Psi}') \rangle]_{\boldsymbol{\Pi} \in \mathcal{A}_{\tau,\tau'}} \right)$, where $\text{SoftMin}_\gamma(\boldsymbol{\alpha}) = -\gamma \log \sum_i \exp(-\alpha_i/\gamma)$ is the soft minimum, $\gamma \geq 0$ controls its relaxation.
- However, the path distance $\langle \boldsymbol{\Pi}, \boldsymbol{D}(\boldsymbol{\Psi}, \boldsymbol{\Psi}') \rangle$ of path $\boldsymbol{\Pi}$ ignores the observation uncertainty of frame-wise feature representations by simply relying on the Euclidean distances stored in $\boldsymbol{D}$.
- We model each path distance by the product of likelihoods of Normal distributions.

Our (soft) uncertainty-DTW takes the following generalized form:

$$\begin{cases} d^2_{\text{uDTW}}(\boldsymbol{D}, \boldsymbol{\Sigma}^\dagger) = \text{SoftMin}_\gamma \Big(\underbrace{[\langle \boldsymbol{\Pi}, \boldsymbol{D} \odot \boldsymbol{\Sigma}^\dagger \rangle]_{\boldsymbol{\Pi} \in \mathcal{A}_{\tau,\tau'}}}_{\boldsymbol{w}} \Big) & (1) \\ \Omega(\boldsymbol{\Sigma}) = \text{SoftMinSel}_\gamma \Big(\boldsymbol{w}, [\langle \boldsymbol{\Pi}, \log \boldsymbol{\Sigma} \rangle]_{\boldsymbol{\Pi} \in \mathcal{A}_{\tau,\tau'}} \Big), & (2) \\ \text{where } \boldsymbol{D} \equiv \boldsymbol{D}(\boldsymbol{\Psi}, \boldsymbol{\Psi}'),\ \boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}(\boldsymbol{\Psi}, \boldsymbol{\Psi}') \text{ and } \boldsymbol{\Sigma}^\dagger = \text{inv}(\boldsymbol{\Sigma}), \end{cases}$$

where $\odot$ is the Hadamard product, $\boldsymbol{\Sigma}^\dagger(\boldsymbol{\Psi}, \boldsymbol{\Psi}')$ is the element-wise inverse of matrix $\boldsymbol{\Sigma} \in \mathbb{R}_+^{\tau \times \tau'} \equiv [\sigma^2(\boldsymbol{\psi}_m, \boldsymbol{\psi}'_n)]_{(m,n) \in \mathcal{I}_\tau \times \mathcal{I}_{\tau'}}$ which contains pair-wise variances between all possible pairings of frame-wise feature representations from sequences $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}'$.

  - Eq. (1) yields the uncertainty-weighted time warping distance $d^2_{\text{uDTW}}(\boldsymbol{D}, \boldsymbol{\Sigma}^\dagger)$ between sequences $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}'$ because $\boldsymbol{D}$ and $\boldsymbol{\Sigma}^\dagger$ are both functions of $(\boldsymbol{\Psi}, \boldsymbol{\Psi}')$.
  - Eq. (2) provides the regularization penalty $\Omega(\boldsymbol{\Sigma})$ for sequences $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}'$ (as $\boldsymbol{\Sigma}$ is a function of $(\boldsymbol{\Psi}, \boldsymbol{\Psi}')$) which is the aggregation of log-variances along the path with the smallest distance, *i.e.*, path matrix ($(\boldsymbol{\Pi}_{i^*} \in \{0,1\}^{\tau \times \tau'})$: $i^* = \arg\min_k w_k$) if $\gamma = 0$, and vector $\boldsymbol{w}$ contains path-aggregated distances for all possible paths of the plan $\mathcal{A}_{\tau,\tau'}$.

- Below is an example of a generic similarity learning loss:

$$\arg\min_{\mathcal{P}} \sum_n \ell \left(d^2_{\text{uDTW}}(\boldsymbol{D}(\boldsymbol{\Psi}_n, \boldsymbol{\Psi}'_n), \boldsymbol{\Sigma}^\dagger(\boldsymbol{\Psi}_n, \boldsymbol{\Psi}'_n)),\ \delta_n \right) + \beta \Omega(\boldsymbol{\Sigma}(\boldsymbol{\Psi}_n, \boldsymbol{\Psi}'_n)), \text{or} \quad (3)$$

$$\arg\min_{\mathcal{P}, \boldsymbol{\Sigma} > 0} \sum_n \ell \left(d^2_{\text{uDTW}}(\boldsymbol{D}(\boldsymbol{\Psi}_n, \boldsymbol{\Psi}'_n), \boldsymbol{\Sigma}^\dagger),\ \delta_n \right) + \beta \Omega(\boldsymbol{\Sigma}), \quad (4)$$

where $\boldsymbol{\Psi}_n = f(\mathbf{X}_n; \mathcal{P})$ and $\boldsymbol{\Psi}'_n = f(\mathbf{X}'_n; \mathcal{P})$ are obtained from some backbone encoder $f(\cdot; \mathcal{P})$ with parameters $\mathcal{P}$ and $(\mathbf{X}_n, \mathbf{X}'_n) \in \mathcal{X}$ is a sequence pair to compare with the similarity label $\delta_n \in \{0, 1\}$ (where $\delta_n = 0$ if $y_n = y'_n$ and $\delta_n = 1$ otherwise), (y_n, y'_n) is a pair of class labels for $(\boldsymbol{\Psi}_n, \boldsymbol{\Psi}'_n)$, and $\beta \geq 0$ controls the penalty for high matching uncertainty.

- Minimizing Eq. (4) w.r.t. $(\mathcal{P}, \boldsymbol{\Sigma})$ assumes that $\boldsymbol{\Sigma} \in \mathbb{R}_+^{\tau \times \tau'}$ is a free variable to minimize over. However, as sequence pairs vary in length, *i.e.*, $\tau \neq \tau'$, optimizing one global $\boldsymbol{\Sigma}$ is impossible (its size changes). Thus, we minimize loss functions with the distance/penalty in Eq. (3) and (4) where $\boldsymbol{\Sigma}$ is parametrized by $(\boldsymbol{\Psi}_n, \boldsymbol{\Psi}'_n)$:

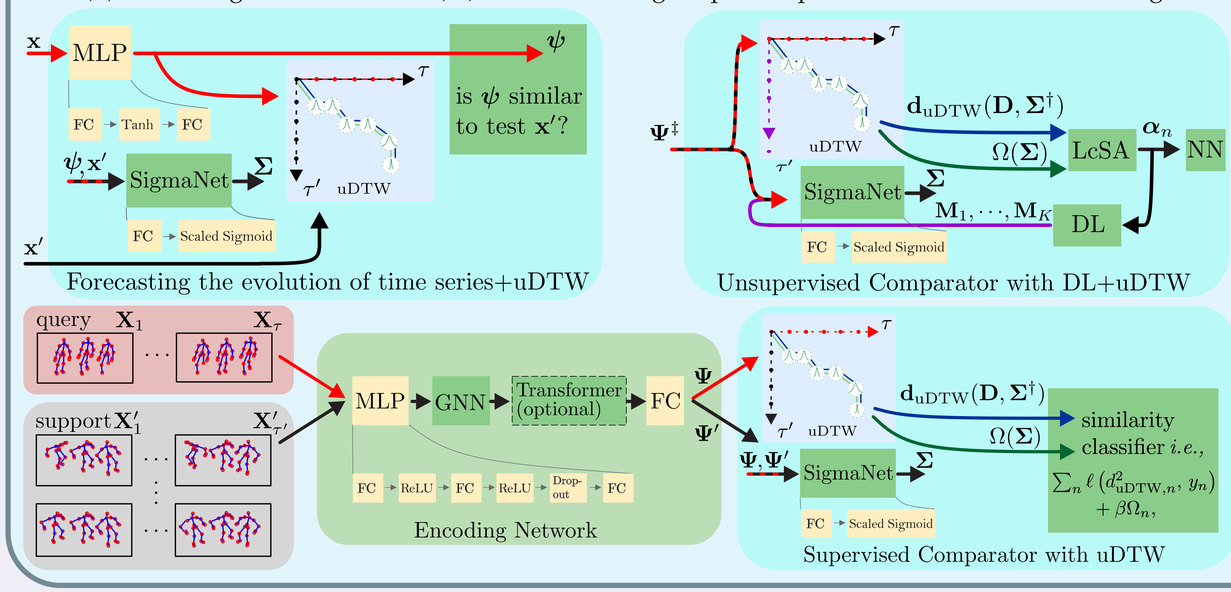
$$d^2_{\text{uDTW}\bullet}(\boldsymbol{\Psi}, \boldsymbol{\Psi}') \equiv d^2_{\text{uDTW}}(\boldsymbol{D}(\boldsymbol{\Psi}, \boldsymbol{\Psi}'), \boldsymbol{\Sigma}^\dagger(\boldsymbol{\Psi}, \boldsymbol{\Psi}')), \quad (5)$$

$$\Omega_\bullet(\boldsymbol{\Psi}, \boldsymbol{\Psi}') \equiv \Omega(\boldsymbol{\Sigma}(\boldsymbol{\Psi}, \boldsymbol{\Psi}')). \quad (6)$$

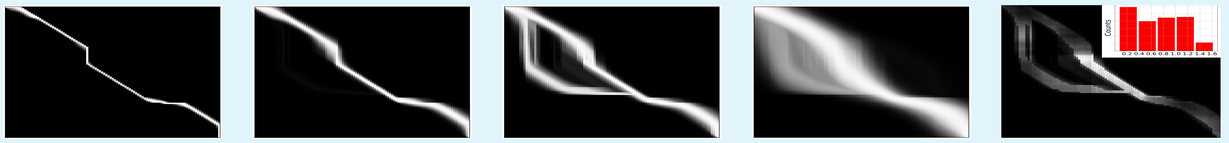
- We devise a small MLP (SigmaNet) for $\sigma(\cdot; \mathcal{P}_\sigma)$ or $\sigma(\cdot, \cdot; \mathcal{P}_\sigma)$ to obtain $\boldsymbol{\Sigma}$ (uses additive variance for individual frames $\boldsymbol{\psi}_m$ and $\boldsymbol{\psi}'_n$) or $\boldsymbol{\Sigma}'$ (uses a jointly generated variance for $(\boldsymbol{\psi}_m, \boldsymbol{\psi}'_n)$).

The pipeline: further details

We define specific loss functions for several problems such as (i) forecasting the evolution of time series, (ii) clustering time series or (iii) even matching sequence pairs in few-shot action recognition.



Results



(a) sDTW$_{\gamma=0.01}$ (b) sDTW$_{\gamma=0.1}$ (c) uDTW$_{\gamma=0.01}$ (d) uDTW$_{\gamma=0.1}$ (e) uncertainty

We power-normalized pixels of plots (by 0.1) to see darker paths better. With higher γ that controls softness, in (b) & (d) more paths become 'active'. In (c), uDTW has two possible routes *vs*. sDTW (a) due to uncertainty modeling. In (e), we visualise uncertainty $\boldsymbol{\Sigma}$. We binarize plot (c) and multiply it by the $\boldsymbol{\Sigma}$ to display uncertainty values on the path (white pixels = high uncertainty).

Few-shot action recognition results on NTU-60 & NTU-120:

#classes	10	20	30	40	50
Supervised					
MatchNets	46.1	48.6	53.3	56.3	58.8
ProtoNet	47.2	51.1	54.3	58.9	63.0
TAP	54.2	57.3	61.7	64.7	68.3
Euclidean	38.5	42.2	45.1	48.3	50.9
sDTW	53.7	56.2	60.0	63.9	67.8
sDTW div.	54.0	57.3	62.1	65.7	69.0
uDTW	**56.9**	**61.2**	**64.8**	**68.3**	**72.4**
Unsupervised					
Euclidean	20.9	23.7	26.3	30.0	33.1
sDTW	35.6	45.2	53.3	56.7	61.7
sDTW div.	36.0	46.1	54.0	57.2	62.0
uDTW	**37.0**	**48.3**	**55.3**	**58.0**	**63.3**

#classes	20	40	60	80	100
Supervised					
MatchNets	20.5	23.4	25.1	28.7	30.0
ProtoNet	21.7	24.0	25.9	29.2	32.1
TAP	31.2	37.7	40.9	44.5	47.3
Euclidean	18.7	21.3	24.9	27.5	30.0
sDTW	30.3	37.2	39.7	44.0	46.8
sDTW div.	30.8	38.1	40.0	44.7	47.3
uDTW	**32.2**	**39.0**	**41.2**	**45.3**	**49.0**
Unsupervised					
Euclidean	13.5	16.3	20.0	24.9	26.2
sDTW	20.1	25.3	32.0	36.9	40.9
sDTW div.	20.8	26.0	33.2	37.5	42.3
uDTW	**22.7**	**28.3**	**35.9**	**39.4**	**44.0**