NTNU

Norwegian University of Science and Technology

Shape and Time Distortion Loss for Training Deep Time Series Forecasting Models

Le Guen, Vincent
Thome, Nicolas
2019

A paper presentation for course TDT99
Modern Machine Learning for Time Series Analysis

Eivind Strøm

Shape and Time Distortion Loss for Training Deep Time Series Forecasting Models

Vincent Le Guen [1,2]
vincent.le-guen@edf.fr

Nicolas Thome [2]
nicolas.thome@cnam.fr

(1) EDF R&D
6 quai Watier, 78401 Chatou, France

(2) CEDRIC, Conservatoire National des Arts et Métiers
292 rue Saint-Martin, 75003 Paris, France

Abstract

This paper addresses the problem of time series forecasting for non-stationary signals and multiple future steps prediction. To handle this challenging task, we introduce DILATE (DIstortion Loss including shApe and TimE), a new objective function for training deep neural networks. DILATE aims at accurately predicting sudden changes, and explicitly incorporates two terms supporting precise shape and temporal change detection. We introduce a differentiable loss function suitable for training deep neural nets, and provide a custom back-prop implementation for speeding up optimization. We also introduce a variant of DILATE, which provides a smooth generalization of temporally-constrained Dynamic Time Warping (DTW). Experiments carried out on various non-stationary datasets reveal the very good behaviour of DILATE compared to models trained with the standard Mean Squared Error (MSE) loss function, and also to DTW and variants. DILATE is also agnostic to the choice of the model, and we highlight its benefit for training fully connected networks as well as specialized recurrent architectures, showing its capacity to improve over state-of-the-art trajectory forecasting approaches.

1 Introduction

Time series forecasting [6] consists in analyzing the dynamics and correlations between historical data for predicting future behavior. In one-step prediction problems [39, 30], future prediction reduces to a single scalar value. This is in sharp contrast with multi-step time series prediction [49, 2, 48], which consists in predicting a complete trajectory of future data at a rather long temporal extent. Multi-step forecasting thus requires to accurately describe time series evolution.

This work focuses on multi-step forecasting problems for non-stationary signals, *i.e.* when future data cannot only be inferred from the past periodicity, and when abrupt changes of regime can occur. This includes important and diverse application fields, *e.g.* regulating electricity consumption [63, 36], predicting sharp discontinuities in renewable energy production [23] or in traffic flow [35, 34], electrocardiogram (ECG) analysis [9], stock markets prediction [14], *etc.*

Deep learning is an appealing solution for this multi-step and non-stationary prediction problem, due to the ability of deep neural networks to model complex nonlinear time dependencies. Many approaches have recently been proposed, mostly relying on the design of specific one-step ahead

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Overview

1. **Motivation**
2. **Problem description**
   - Introduction
   - Examined paper
3. **Related work**
4. **Methodology**
   - General method
   - Loss function
   - Training networks
5. **Experimental results**
   - Setup
   - Evaluation of loss function
   - Comparison to state-of-the-art
6. **Conclusion**
7. **References**

Motivation

- Addresses the problem of time series forecasting for non-stationary, non-linear signals and multiple future steps prediction.

- Multi-step forecasting from non-stationary signals tend to **require more than just past periodicity** and includes the need to handle **abrupt regime shifts**.

- Important in the fields of:
  - Regulating electricity consumption [36, 63]
  - Predicting sharp discontinuities in renewable energy production [23]
  - Traffic flow [34, 35]
  - Electrocardiogram (ECG) analysis [9]
  - Stock markets prediction [14]
  - … *etc*

- **Little research has been published** on loss functions for shape and time localization.

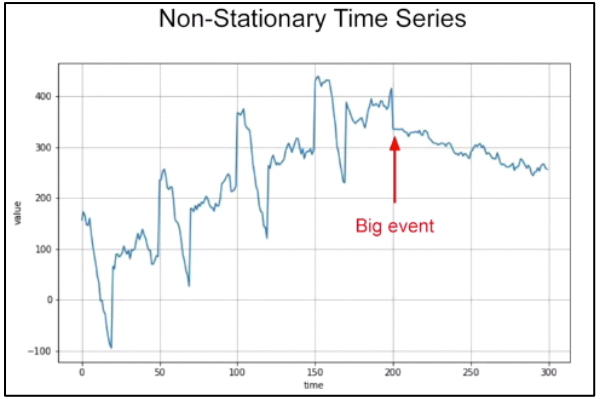


Figure source: https://morioh.com/p/c9290f02e74f

Problem description – introduction

- Time series forecasting include two main categories:
  - Time series with **linear and stationary dynamics** (or non-stationary made stationary by differencing)
  - Time series with **non-linear and non-stationary dynamics**

- The first category is mainly handled by traditional **State Space Models (SSMs)** [17] like the linear auto-regressive model, ARIMA [6], Exponential Smoothing [27] and is a well researched topic.

- For the second category where linear SSMs are inadequate, **deep learning has become a popular tool to forecast non-stationary and multivariate time series.**
- Research in this area has focused on:
  - Leveraging attention mechanisms [30, 39, 50, 12]
  - Tensor factorizations [60, 58, 46]
  - Combining Deep Learning and State Space Models (SSMs) to model uncertainty [45, 44, 40, 56]

Problem description – examined paper

- The examined paper focuses on **deterministic multi-step forecasting:**

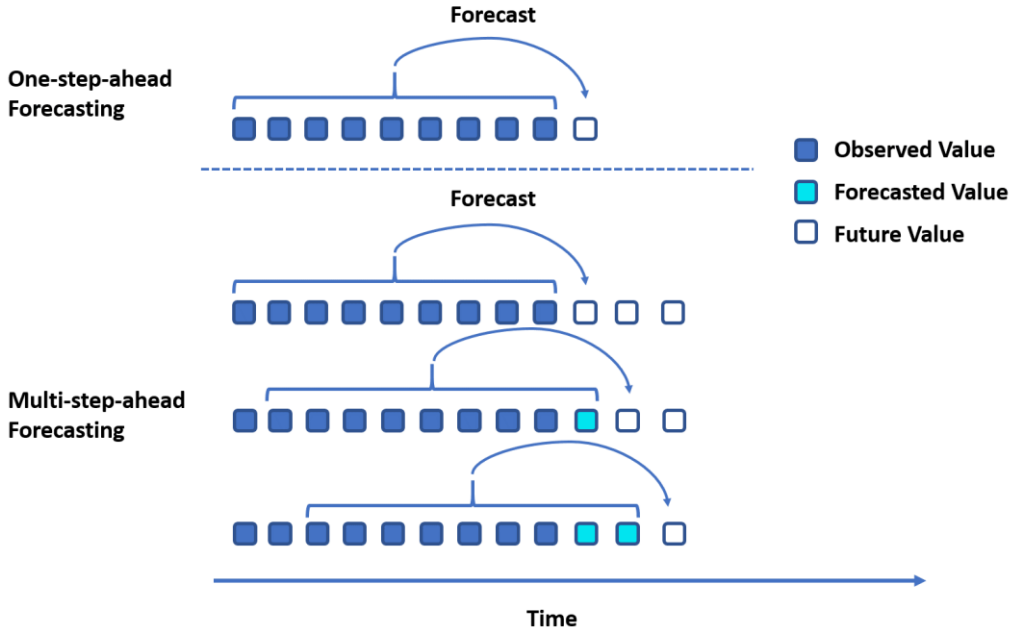
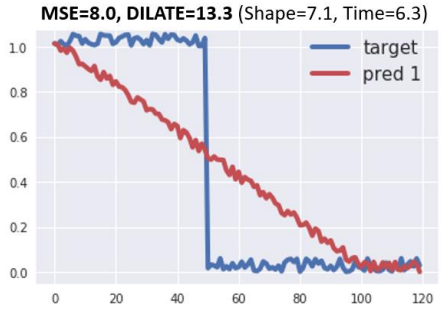
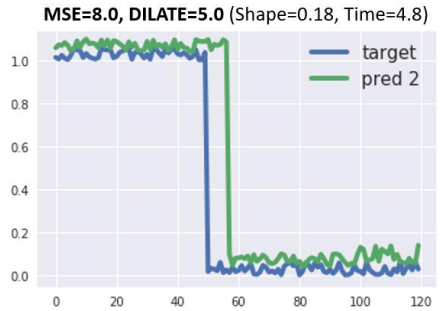


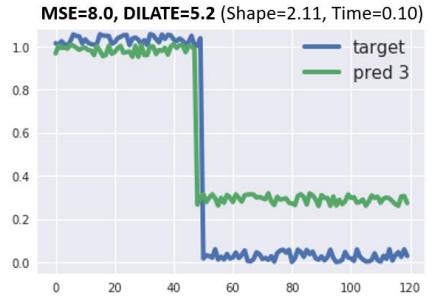
Figure source: https://github.com/sassoftware/python-dlpy/issues/101

Problem description – examined paper

- Deep learning approaches to tackle multi-step forecasting for non-stationary time series mostly rely on **recursively applied single-step architectures** or on **direct multi-step models** such as **Seq2Seq** [55].
- Most of these proposed models rely on Mean Squared Error (**MSE**) as the loss function for training.
- This paper argues that **MSE is inadequate in the context of non-stationary time series**, as it **does not capture** the distinct regime shifts **(shape)** nor temporal aspects **(time).** This is illustrated with the following figures, all having equal MSE=8:







- To rectify this, the authors introduces a new objective function for training: **DILATE**

Related work

- **Research suggests direct multi-step models are to be preferred.**
  - Seq2Seq RNN models which has been used for machine translation [44, 31, 60, 57, 19]
  - WaveNET for audio generation [53]
  - CNNs with dilation for time series forecasting [5]
- **These models, however, rely mainly on MSE and its variants as loss functions.**

- Some work exist on incorporating shape and temporal localization as a **metric.**
  - Shape is considered by Dynamic Time Warping (DTW) [43]
  - Time can be considered by Change Point Detection [8, 33], or Hausdorff distance [22, 51]
- These however are **non-differentiable and therefore useless** for training deep neural nets.

- Recently, **some attempts have been made** to include these aspects to the loss function.
  - A smooth approximation of DTW enables focus on the shape loss [13, 37, 1], **however, ignores temporal localization.**
  - A differentiable timing error loss function based on DTW has been proposed [41], however only applicable for binary time series.
- Therefore, **the paper specifically focus on both shape and temporal localization.**

Methodology – general method

- Define a set of N input time series:

$$\mathcal{A} = \{\mathbf{x}_i\}_{i \in \{1:N\}}$$

- For each input sample of length *n*, predict a k-step ahead forecast by a neural network forecasting model:

$$\mathbf{x}_i = (\mathbf{x}_i^1, ..., \mathbf{x}_i^n) \in \mathbb{R}^{p \times n} \longrightarrow \hat{\mathbf{y}}_i = (\hat{\mathbf{y}}_i^1, ..., \hat{\mathbf{y}}_i^k) \in \mathbb{R}^{d \times k}$$

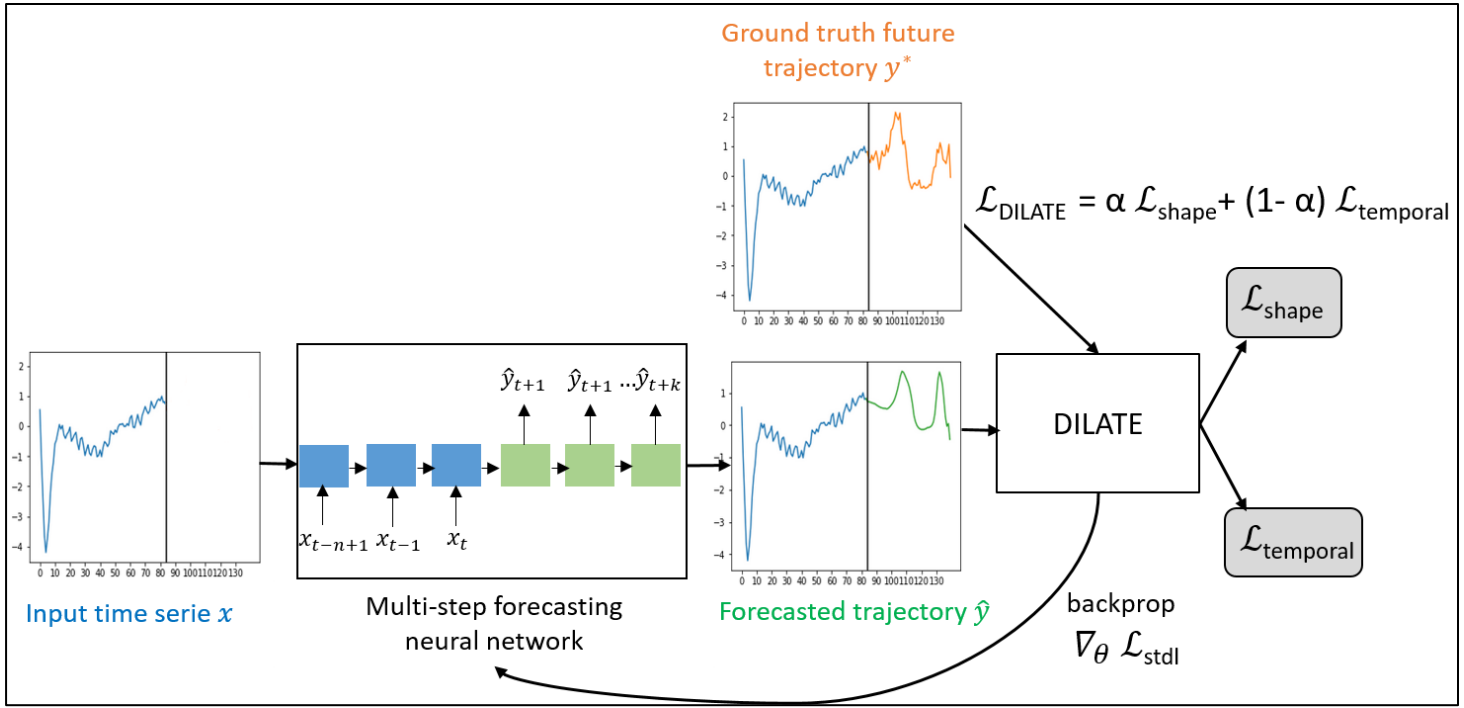
- Compare predicted forecast with actual future trajectory by the DILATE objective function:

$$\overset{*}{\mathbf{y}}_i = (\overset{*}{\mathbf{y}}_i{}^1, ..., \overset{*}{\mathbf{y}}_i{}^k)$$

$$\mathcal{L}_{DILATE}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = \alpha \, \mathcal{L}_{shape}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) + (1 - \alpha) \, \mathcal{L}_{temporal}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i)$$

Methodology – general method

- The described method is summarised by the following picture:



Methodology – loss function

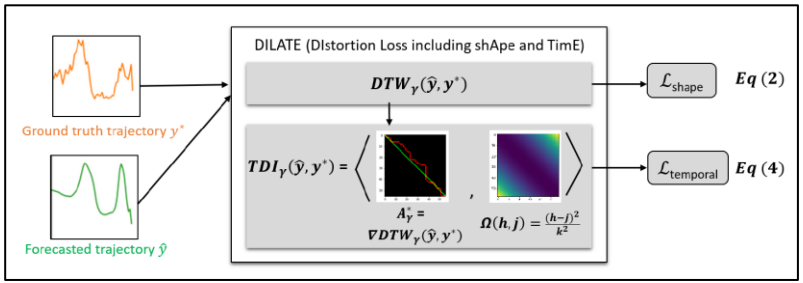
- A modified DTW function made differentiable to account for shape loss:

$$\mathcal{L}_{shape}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = DTW_\gamma(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := -\gamma \log \left(\sum_{\mathbf{A} \in \mathcal{A}_{k,k}} \exp \left(-\frac{\left\langle \mathbf{A}, \mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle}{\gamma} \right) \right)$$

- Smoothed temporal loss based on Time Distortion Index (TDI) for temporal misalignment, derived from the DTW function.

$$\mathcal{L}_{temporal}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := \left\langle \mathbf{A}^*_\gamma, \mathbf{\Omega} \right\rangle = \frac{1}{Z} \sum_{\mathbf{A} \in \mathcal{A}_{k,k}} \langle \mathbf{A}, \mathbf{\Omega} \rangle \exp^{-\frac{\left\langle \mathbf{A}, \mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle}{\gamma}}$$

- **Key takeaway: Modified DTW and TDI functions made differentiable to account for shape and temporal loss**.



Methodology – training networks

- Direct computation of the previous equations is intractable due to the cardinal of $\mathcal{A}_{k,k}$ growing exponentially with *k*. Therefore, **a custom implementation is devised.**

- Custom shape loss is computation:
  - Forward pass: A **Bellman Dynamic Programming** approach [13]
  - Backward pass: **Recursion**, implemented in Pytorch [13]

- Custom temporal loss computation:
  - Forward pass: **Bellman Dynamic Programming** (same as above)
  - Backward pass: **Custom Dynamic Programming** implementation in Pytorch for computing the Hessian of the DTW function [37].

- The final time complexity turns out to be O(k^2).
  - (k being number of steps in a k-step ahead prediction)

Experimental results - setup

- **Experiments done on 3 non-stationary time series data sets from different domains.**
  - Synthetic (k = 20) consisting of 500 generated time series
  - ECG5000 (k = 56) from UCR Time Series Classification Archive [10]
  - Traffic (k = 24) from California Department of Transportation

- **Performance is evaluated in three different ways.**
  - Evaluation of DILATE against networks trained on MSE and the smooth DTW function used in previous research (which only accounts for shape) [13, 37].
  - Evaluation by comparing forecast results using external metrics: Ramp score (shape) and Hausdorff distance (time).
  - Evaluation against state-of-the-art models trained on MSE.

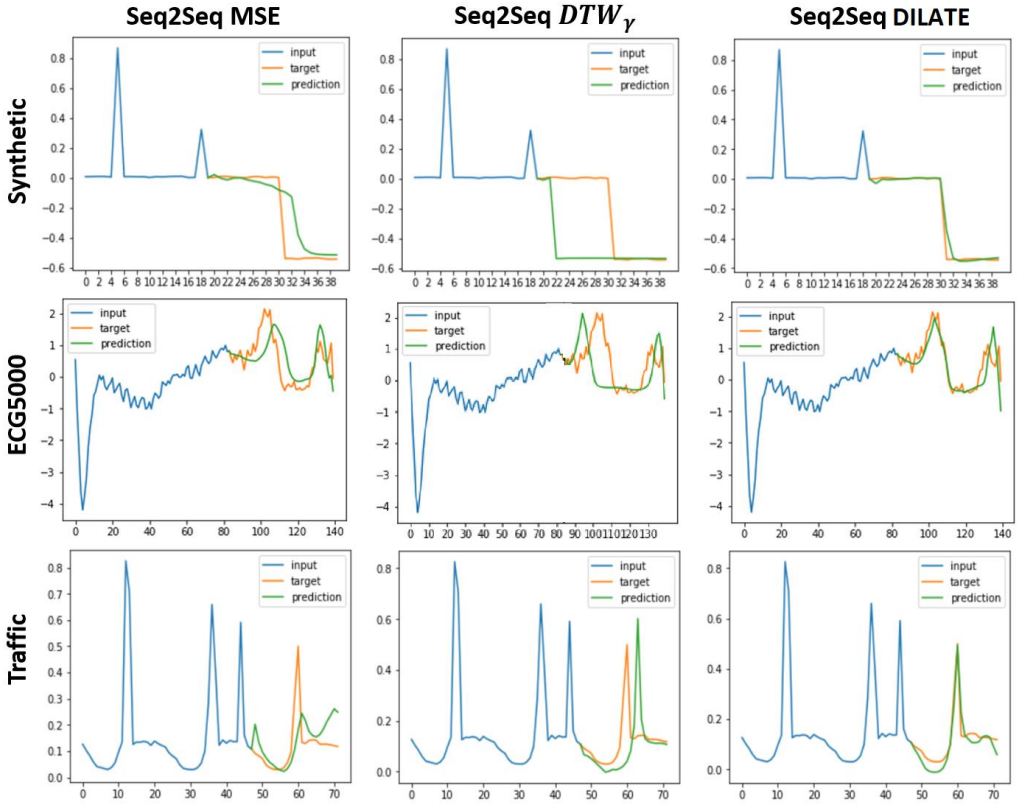
Experimental results – evaluation of loss function

- Evaluation against other loss functions. **DILTAE outperforms in most cases.**

		Fully connected network (MLP)			Recurrent neural network (Seq2Seq)		
Dataset	Eval	MSE	DTW_γ [13]	DILATE (ours)	MSE	DTW_γ [13]	DILATE (ours)
Synth	MSE	**1.65 ± 0.14**	4.82 ± 0.40	**1.67± 0.184**	**1.10 ± 0.17**	2.31 ± 0.45	**1.21 ± 0.13**
	DTW	38.6 ± 1.28	**27.3 ± 1.37**	32.1 ± 5.33	**24.6 ± 1.20**	**22.7 ± 3.55**	**23.1 ± 2.44**
	TDI	15.3 ± 1.39	26.9 ± 4.16	**13.8 ± 0.712**	17.2 ± 1.22	20.0 ± 3.72	**14.8 ± 1.29**
ECG	MSE	**31.5 ± 1.39**	70.9 ± 37.2	37.2 ± 3.59	**21.2 ± 2.24**	75.1 ± 6.30	30.3 ± 4.10
	DTW	19.5 ± 0.159	18.4 ± 0.749	**17.7 ± 0.427**	17.8 ± 1.62	17.1 ± 0.650	**16.1 ± 0.156**
	TDI	**7.58 ± 0.192**	38.9 ± 8.76	**7.21 ± 0.886**	8.27 ± 1.03)	27.2 ± 11.1	**6.59 ± 0.786**
Traffic	MSE	**0.620 ± 0.010**	2.52 ± 0.230	1.93 ± 0.080	**0.890 ± 0.11**	2.22 ± 0.26	**1.00 ± 0.260**
	DTW	24.6 ± 0.180	**23.4 ± 5.40**	**23.1 ± 0.41**	24.6 ± 1.85	**22.6 ± 1.34**	**23.0 ± 1.62**
	TDI	**16.8 ± 0.799**	27.4 ± 5.01	**16.7 ± 0.508**	**15.4 ± 2.25**	22.3 ± 3.66	**14.4± 1.58**

Table 1: Forecasting results evaluated with MSE (×100), DTW (×100) and TDI (×10) metrics, averaged over 10 runs (mean ± standard deviation). For each experiment, best method(s) (Student t-test) in bold.

Experimental results – evaluation of loss function



Experimental results – comparison to state of art

- Comparison to state-of-the-art models trained by MSE.
  - LSTNet recursive prediction for multi-step [30]
  - Tensor-Train RNN trained for multi-step [60, 61]

- **Seq2Seq trained by DILATE outperforms the more complex models on shape and time metrics.** TT-RNN outperforms on MSE.

Eval loss		LSTNet-rec [30]	TT-RNN [60, 61]	Seq2Seq DILATE
Euclidian	MSE (x100)	1.74 ± 0.11	$\mathbf{0.837 \pm 0.106}$	1.00 ± 0.260
Shape	DTW (x100)	42.0 ± 2.2	25.9 ± 1.99	$\mathbf{23.0 \pm 1.62}$
	Ramp (x10)	9.00 ± 0.577	6.71 ± 0.546	$\mathbf{5.93 \pm 0.235}$
Time	TDI (x10)	25.7 ± 4.75	17.8 ± 1.73	$\mathbf{14.4 \pm 1.58}$
	Hausdorff	$\mathbf{2.34 \pm 1.41}$	$\mathbf{2.19 \pm 0.125}$	$\mathbf{2.13 \pm 0.514}$

Table 4: Comparison with state-of-the-art forecasting architectures trained with MSE on Traffic, averaged over 10 runs (mean $\pm$ standard deviation).

Conclusion

- The authors introduce **DILATE**, a **new differentiable loss function** that explicitly consider **shape and time localization** of non-stationary time series with sudden changes.

- DILATE is comparable to MSE and **outperforms on shape and time metrics**. DILATE also **compares favourably to state-of-the-art models**.

- Possibilities for future work:
  - **Extending the ideas to probabilistic forecasting,** for example in Bayesian deep learning [21]**.**
  - Adapt the training scheme for semi-supervised or weakly supervised [16, 42] contexts for full trajectory forecasting using only categorical labels.

- Bonus: Code is available on Github! https://github.com/vincent-leguen/DILATE

References

[1] Abubakar Abid and James Zou. Learning a warping distance from unlabeled time series using sequence autoencoders. In Advances in Neural Information Processing Systems (NeurIPS), pages 10547–10555, 2018.

[5] Anastasia Borovykh, Sander Bohte, and Cornelis W Oosterlee. Conditional time series forecasting with convolutional neural networks. arXiv preprint arXiv:1703.04691, 2017.

[6] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. Time series analysis: forecasting and control. John Wiley & Sons, 2015.

[8] Wei-Cheng Chang, Chun-Liang Li, Yiming Yang, and Barnabás Póczos. Kernel change-point detection with auxiliary deep generative models. In International Conference on Learning Representations (ICLR), 2019.

[9] Sucheta Chauhan and Lovekesh Vig. Anomaly detection in ECG time signals via deep long short-term memory networks. In International Conference on Data Science and Advanced Analytics (DSAA), pages 1–7. IEEE, 2015.

[10] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The UCR time series classification archive. 2015.

[12] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In Advances in Neural Information Processing Systems (NIPS), pages 3504–3512, 2016.

[13] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In International Conference on Machine Learning (ICML), pages 894–903, 2017.

[14] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In International Joint Conference on Artificial Intelligence (IJCAI), 2015.

[16] Thibaut Durand, Nicolas Thome, and Matthieu Cord. Exploiting negative evidence for deep latent structured models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2):337–351, 2018.

[17] James Durbin and Siem Jan Koopman. Time series analysis by state space methods. Oxford university press, 2012.

[19] Ian Fox, Lynn Ang, Mamta Jaiswal, Rodica Pop-Busui, and Jenna Wiens. Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories. In ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1387–1395. ACM, 2018.

[21] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In International Conference on Machine Learning (ICML), pages 1050–1059, 2016.

[22] Damien Garreau, Sylvain Arlot, et al. Consistent change-point detection with kernels. Electronic Journal of Statistics, 12(2):4440–4486, 2018.

[23] Amir Ghaderi, Borhan M Sanandaji, and Faezeh Ghaderi. Deep forecast: Deep learning-based spatio-temporal forecasting. In ICML Time Series Workshop, 2017.

[27] Rob Hyndman, Anne B Koehler, J Keith Ord, and Ralph D Snyder. Forecasting with exponential smoothing: the state space approach. Springer Science & Business Media, 2008.

[30] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and shortterm temporal patterns with deep neural networks. In ACM SIGIR Conference on Research & Development in Information Retrieval, pages 95–104. ACM, 2018.

[31] Nikolay Laptev, Jason Yosinski, Li Erran Li, and Slawek Smyl. Time-series extreme event forecasting with neural networks at Uber. In International Conference on Machine Learning (ICML), number 34, pages 1–5, 2017.

[33] Shuang Li, Yao Xie, Hanjun Dai, and Le Song. M-statistic for kernel change-point detection. In Advances in Neural Information Processing Systems (NIPS), pages 3366–3374, 2015.

[34] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In International Conference on Learning Representations (ICLR), 2018.

References

[35] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: a deep learning approach. IEEE Transactions on Intelligent Transportation Systems, 16(2):865–873, 2015.

[36] Shamsul Masum, Ying Liu, and John Chiverton. Multi-step time series forecasting of electric load using machine learning models. In International Conference on Artificial Intelligence and Soft Computing, pages 148–159. Springer, 2018.

[37] Arthur Mensch and Mathieu Blondel. Differentiable dynamic programming for structured prediction and attention. International Conference on Machine Learning (ICML), 2018.

[39] Yao Qin, Dongjin Song, Haifeng Cheng, Wei Cheng, Guofei Jiang, and Garrison W Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. In International Joint Conference on Artificial Intelligence (IJCAI), pages 2627–2633. AAAI Press, 2017.

[40] Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. In Advances in Neural Information Processing Systems (NeurIPS), pages 7785–7794, 2018.

[41] François Rivest and Richard Kohar. A new timing error cost function for binary time series prediction. IEEE transactions on neural networks and learning systems, 2019.

[42] Thomas Robert, Nicolas Thome, and Matthieu Cord. Hybridnet: Classification and reconstruction cooperation for semi-supervised learning. In European Conference on Computer Vision (ECCV), pages 153–169, 2018.

[43] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. Readings in speech recognition, 159:224, 1990.

[44] David Salinas, Valentin Flunkert, and Jan Gasthaus. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. In International Conference on Machine Learning (ICML), 2017.

[45] Matthias W Seeger, David Salinas, and Valentin Flunkert. Bayesian intermittent demand forecasting for large inventories. In Advances in Neural Information Processing Systems (NIPS), pages 4646–4654, 2016.

[46] Rajat Sen, Yu Hsiang-Fu, and Dhillon Inderjit. Think globally, act locally: a deep neural network approach to high dimensional time series forecasting. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

[50] Yunzhe Tao, Lin Ma, Weizhong Zhang, Jian Liu, Wei Liu, and Qiang Du. Hierarchical attentionbased recurrent highway networks for time series prediction. arXiv preprint arXiv:1806.00685, 2018.

[51] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Supervised kernel change point detection with partial annotations. In International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3147–3151. IEEE, 2019.

[53] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

[55] Arun Venkatraman, Martial Hebert, and J Andrew Bagnell. Improving multi-step prediction of learned time series models. In Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[56] Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. Deep factors for forecasting. In International Conference on Machine Learning (ICML), pages 6607–6617, 2019.

[57] Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A multihorizon quantile recurrent forecaster. NIPS Time Series Workshop, 2017.

[58] Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. Temporal regularized matrix factorization for high-dimensional time series prediction. In Advances in neural information processing systems (NIPS), pages 847–855, 2016.

[60] Rose Yu, Stephan Zheng, Anima Anandkumar, and Yisong Yue. Long-term forecasting using tensor-train RNNs. arXiv preprint arXiv:1711.00073, 2017.

[63] Jian Zheng, Cencen Xu, Ziang Zhang, and Xiaohua Li. Electric load forecasting in smart grids using long-short-term-memory based recurrent neural network. In 51st Annual Conference on Information Sciences and Systems (CISS), pages 1–6. IEEE, 2017.