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

Manjot Singh Yiyu Wang
{manjot.singh, yiyu.wang}@epfl.ch

Abstract

In this paper, some results from DILATE(Distortion Loss including shape and time) (2) are reproduced and additional studies on hyperparameter explorations and other Neural Network models are done as a part of NeurIPS Reproducibility Challenge. The comparison between DILATE and Euclidean Loss (MSE) has been shown when training Deep Neural Networks on non-stationary time series forecasting problems.

1 Introduction

Time series forecasting for non-stationary signals and multi-future-step predictions has been a challenging task. Deep Neural Networks such as RNNs, LSTMs and GRUs have been intensively studied in related domains because of their abilities to model complex non-linear time dependencies. The authors of (2) introduced DILATE(Distortion Loss including shape and time) for training Deep Neural Networks in the context of multi-step and non-stationary time series forecasting. DILATE is a differentiable loss function which penalizes two errors, shape and temporal localisation errors of change detection (2). Mostly, mean squared error (MSE) or its variants are used as a loss function for training Deep Neural Nets. In (2), the authors have showed that in some cases, MSE or its variants doesn't capture a sharp drop or is not a better fit for regulation purposes whereas DILATE reflects on the sharp changes of regime although with a slight delay or with a slight inaccurate amplitude(refer to Figure1:b,c in (2)).

The source code provided by the authors was used and it includes the implementation of DILATE loss function and seq2seq model. We tried to make contact with the authors for supplementary material and data preprocessing guidelines but didn't get a response back. We followed the directions as mentioned in (2) for data preprocessing and also evaluated the performance of DILATE on additional dataset section 3.1. In this paper, the main results from (2) were reproduced to evaluate the performance of DILATE. We propose Convolutional-LSTM model to make certain that DILATE can be used with other Neural Network architectures specifically designed for multi-step and non-stationary forecasting. In addition to the above, several tests were done to study the impact of α and γ on the overall performance of the model and results are shown in 3.4.

2 DILATE

A very brief introduction to the working of DILATE is given in this section. The framework proposed for multi-step time series forecasting is depicted in Figure2 in (2). For each input example of length *n*, the forecasting model predicts the output $\hat{y}$. The DILATE objective function (2) is composed of two terms balanced by the hyperparameter $\alpha \in [0, 1]$ which compares the prediction $\hat{y}_i$ with the ground truth value y_i^*:

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

Shape term The shape loss function is based on Dynamic Time Warping (5) . DTW is non-differentiable but the smooth minimum operator γ proposed in (1) defines the shape term $\mathcal{L}_{shape}$ as:

$$\mathcal{L}_{shape}(\hat{y}_i, y_i^*) = DTW_r(\hat{y}_i, y_i^*) := -\gamma \log(\sum_{A \in \mathcal{A}_{k,k}} \exp(-\frac{\langle \mathbf{A}, \mathbf{\Delta}(\hat{y}_i, y_i^*) \rangle}{\gamma})) \tag{2}$$

Temporal Term The temporal term penalizes the temporal irregularities between $\hat{y}_i$ and y_i^* (2). The loss function is inspired from computing the Time Distortion Index (TDI) for temporal misalignment estimation (4). The loss function is given as:

$$\mathcal{L}_{temporal}(\hat{y}_i, y_i^*) := \langle \mathbf{A}_\gamma^*, \mathbf{\Omega} \rangle \frac{1}{Z} \sum_{A \in \mathcal{A}_{k,k}} \langle \mathbf{A}, \mathbf{\Omega} \rangle \exp(-\frac{\langle \mathbf{A}, \mathbf{\Delta}(\hat{y}_i, y_i^*) \rangle}{\gamma}) \tag{3}$$

For more details on the above algorithm, please refer to (2), (1), (4).

3 Experiments

The results were reproduced from section 4.2 in (2) on 3 non-stationary time series datasets from different domains as implemented in (2). We did an evaluation on one more additional dataset to analyse the relevance of DILATE against Euclidean loss (MSE). The multi-step evaluation consists in forecasting the future trajectory on k future time steps (2).

3.1 Experimental Setup

Synthetic and ECG5000 Data preprocessing on these 2 datasets is same as in (2).

Traffic(k=24) dataset corresponds to the road occupancy rates (between 0 and 1) from the California Department of Transportation (48 months from 2015-2016) measured every 1h (2). For this, there were no data pre-processing instructions, this allowed us to work on univariate series of length 17544 with train/test on 75/25 weeks of data and predictions on 24 future points given the past 168 points (past week).

Wafer (k=62) comes from the UCR Time Series classification Archive (6) with sequence length 152. It consists of the collection of inline process control measurements (1000/6000: train/test) recorded from various sensors during the processing of silicon wafers for semiconductor fabrication.

Network Architectures and Training For multi-step forecasting, we used the Seq2Seq model with Gated Recurrent Units (GRU) as provided by the authors of (2). In addition to Seq2Seq, an implementation of CNN-LSTM model (section 3.3) is carried out so as to compare the performance of DILATE and MSE on different Neural Network architectures. The models are trained in Pytorch for a maximum of 1000 epochs with the ADAM optimizer.

3.2 Evaluating DILATE Forecasting

The similar procedure is used as mentioned in Section4.2 in (2) to evaluate the forecasting performance of DILATE against Euclidean Loss (MSE) and DTW_γ. The results were averaged over 5 runs as opposed to 10 runs in (2). The same experiment was repeated with CNN-LSTM Neural Network as well. The results are evaluated using three metrics: MSE, shape(DTW) and TDI(temporal) (2). Overall results are presented in Table1. For ECG dataset when evaluated on DTW, our results don't quite match with that of the authors in (2) and are highlighted in bold in Table1. Also, the results differ significantly for traffic dataset(seq2seq) and this could be due to the difference in ways of processing the data.

We display a few qualitative examples for Synthetic and Wafer dataset in Figure 1 and 2 when implemented in Se2Seq and CNN-LSTM. As in paper (2), our results for Synthetic dataset matches with that of the authors where DILATE is better than MSE in predicting sharp changes in regime. DTW_γ leads to very sharp predictions in shape, but with a large temporal misalignment. In the case of Wafer dataset, the performance of MSE is better than DILATE when evaluated on MSE.

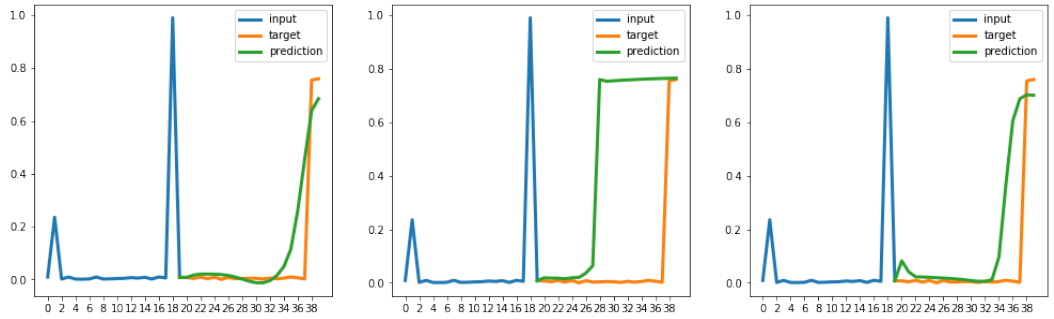
Table 1: **Forecasting results evaluated with MSE (x100), DTW (x100) and TDI (x10) metrics, averaged over 5runs.**

		Convolutional-LSTM Network			**Recurrent Neural Network**		
Dataset	Eval	MSE	DTW_γ	DILATE	MSE	DTW_γ	DILATE
Synth	MSE	1.64 ± 0.03	5.06 ± 0.44	1.7 ± 0.07	1.07 ± 0.08	1.66 ± 0.08	1.3 ± 0.05
	DTW	32.67 ± 0.24	28.18 ± 1.69	28.66 ± 0.42	22.78 ± 1.5	14.59 ± 0.81	19.24 ± 1.39
	TDI	12.42 ± 0.47	30.44 ± 2.27	14.06 ± 0.12	19.15 ± 1.78	20.06 ± 0.47	14.79 ± 1.56
ECG	MSE	17.70 ± 0.19	67.18 ± 15.66	34.22 ± 4.54	20.38 ± 0.32	72.29 ± 1.77	35.98 ± 5.19
	DTW	$\mathbf{167.07 \pm 1.27}$	$\mathbf{169.95 \pm 4.16}$	$\mathbf{168.29 \pm 4.76}$	$\mathbf{177.62 \pm 5.24}$	$\mathbf{311.99 \pm 2.67}$	$\mathbf{180.7 \pm 12.78}$
	TDI	6.29 ± 0.226	23.17 ± 8.47	7.65 ± 0.52	8.04 ± 0.28	77.68 ± 18.58	9.27 ± 2.1
Traffic	MSE	0.103 ± 0.004	0.13 ± 0.009	0.10 ± 0.009	0.1 ± 0.004	0.13 ± 0	0.1 ± 0
	DTW	11.66 ± 0.10	10.21 ± 0.41	11.34 ± 0.39	12.01 ± 0.32	12.04 ± 0.26	11.11 ± 0.15
	TDI	0.77 ± 0.18	0.87 ± 0.02	1.23 ± 0.25	0.71 ± 0.01	0.70 ± 0.03	1.32 ± 0.10
Wafer	MSE	4.25 ± 0.24	105.17 ± 8.3	75.07 ± 15.48	2.88 ± 0.54	165.82 ± 1.53	43.63 ± 5.24
	DTW	94.59 ± 3.36	29.35 ± 0.59	52.81 ± 5.25	74.37 ± 10.72	38.74 ± 0.6	90.22 ± 14.37
	TDI	11.52 ± 0.14	67.91 ± 9.26	31.19 ± 9.76	20.83 ± 1.18	97.87 ± 1.14	22.13 ± 1.47

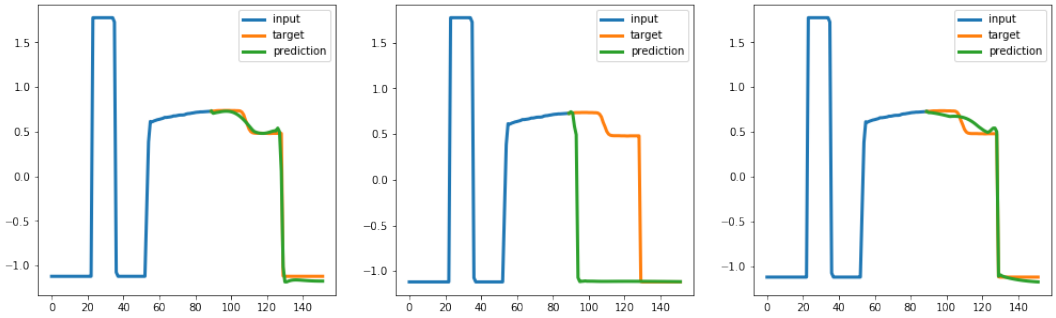
MSE Comparison: DILATE is better than MSE when evaluated on shape(DTW) in 5/8 experiments. On evaluation on time(TDI), MSE performs much better than DILATE in 7/8 experiments. When evaluated on MSE, DILATE is equivalent to MSE except on Wafer dataset.

DTW_γComparison: On evaluation with shape(DTW), DILATE performs equivalently to DTW_γ (2 reductions, 1 significant improvement and 5 similar performances). When evaluated on time, DILATE outperforms DTW_γ in 6/8 experiments except for Traffic dataset. For MSE evaluation, DILATE is significantly better than DTW_γ in all experiments.

In conclusion, DILATE performs better than DTW_γ. The performance of DILATE is better than MSE but in some cases, MSE is comparable to DILATE and when evaluated on MSE, MSE outperforms DILATE.

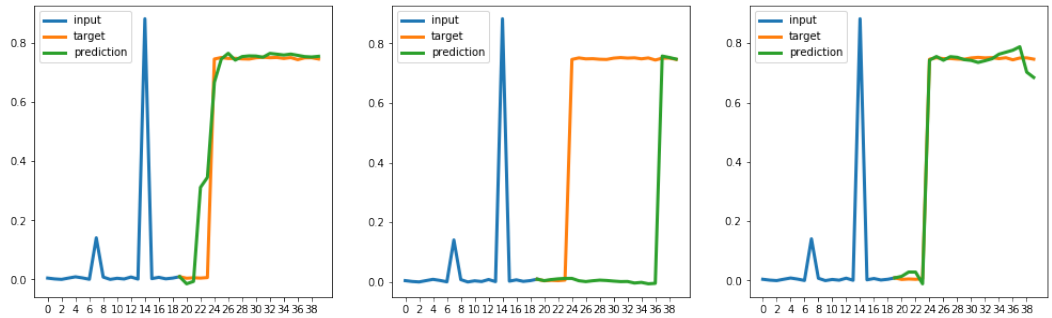


(a) Synthetic data

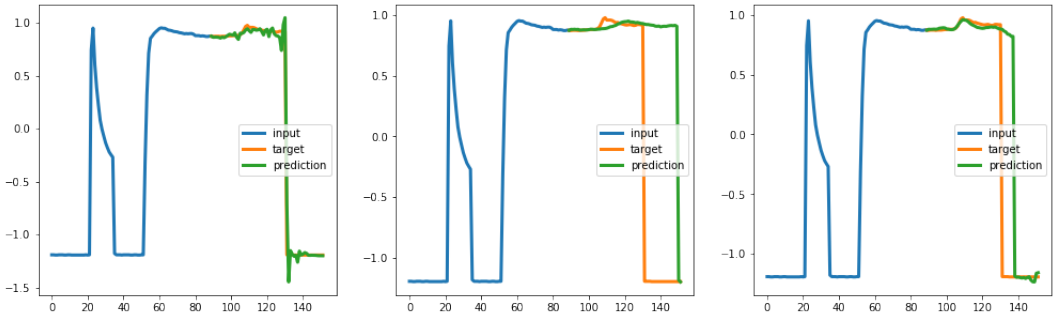


(b) Wafer data

Figure 1: Forecasting results for Synthetic and Wafer data based on Seq2Seq model. **Leftmost:** Seq2Seq with MSE loss, **Middle:** Seq2Seq with DTWγ loss, **Rightmost:** Seq2Seq with DILATE loss.



(a) Synthetic data



(b) Wafer data

Figure 2: Forecasting results for Synthetic and Wafer data based on CNN_LSTM model. **Leftmost:** CNN_LSTM with MSE loss, **Middle:** CNN_LSTM with DTWγ loss, **Rightmost:** CNN_LSTM with DILATE loss.

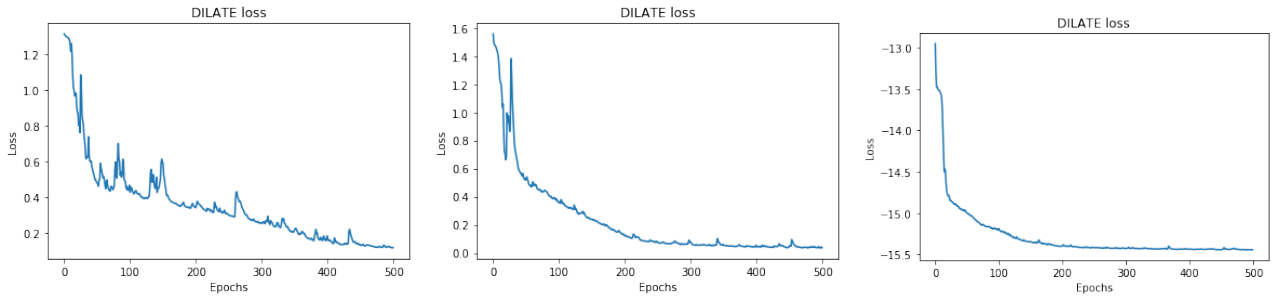
3.3 Comparison with CNN-LSTM

We compared DILATE performance on Seq2Seq model and a basic CNN-LSTM. CNN-LSTM model is developed with the goal to perform 1-D convolution to capture local patterns in the multi-step time series and then LSTM for capturing the complex nonlinear time dependencies. The Seq2Seq DILATE performs similarly to CNN-LSTM when evaluated on all the three metrics and four datasets used in this study. This highlights the relevance of DILATE loss function, which reaches better performances with simpler architectures. If DILATE and MSE perform similar when used in complex Neural Network architectures (3), then MSE seems to be a better choice because of computational reasons.

3.4 DILATE Analysis

Impact of γ: We carried out a thorough analysis of γ on the overall performance of the model using Synthetic dataset. We chose from 6 log-spaced values between 10^{-3} and 10. We observed that for low values of γ, DILATE training loss converges to an acceptable minimum value but as seen from the Figure 3, it has a possibility to get stuck in a very bad local minima. On the other hand, for high values of γ, loss converges smoothly to a reasonable solution. As γ decreases, DTW_γ achieves low loss values which verifies that DILATE is more compliant to optimisation by gradient-descent methods. The convergence of training loss for 3 different values of gamma is shown in the Figure 3. The same analysis can carried out on different datasets for more validity.

Impact of α: We run experiments on 10 different values of α from 0 to 1 and observe its impact on the performance of DILATE. We carry out this analysis on synthetic and ECG5000 dataset. When $\alpha = 0$, temporal loss is minimized without any shape constraint. Both MSE and shape errors explode in this case illustrating the fact from (2). We reproduced the similar results as in (2) for Synthetic dataset. For ECG5000 dataset, we observed the similar performance between $\mathcal{L}_{temporal}$ and $\mathcal{L}_{shape}$ but MSE in this case seems to be independent of alpha as seen from Figure 4. This suggests that MSE loss does not depend on the value of α. For more insights on this, similar analysis can be carried out on different datasets.



(a) γ = 0.001 (b) γ = 0.01 (c) γ = 1

Figure 3: DILATE training loss on different γ values.

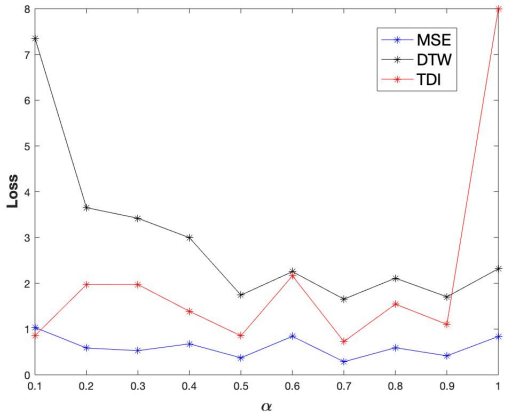


Figure 4: α on ECG data

Custom backward implementation speedup: In order to verify the speedup, we apply the Pytorch build in function *TORCH.AUTOGRAD.PROFILER* to check the runtime of the back propagation of DILATE and MSE loss. According to section 3.2 in paper(2), the time complexity of DILATE is $\mathcal{O}(k^2)$ and the time complexity of MSE algorithm is known i.e. $\mathcal{O}(k)$. We tested different input length and different models, then estimated the speedup by computing $\frac{runtime^2_{mse}}{runtime_{dilate}}$. We verified that the custom backward implementation in (2) is very effective, but DILATE is still computationally expensive than MSE.

4 Conclusion

We reproduced some results which are similar to that of the paper (2). We did additional experiments using different datasets and studied the effect of two hyperparamters α and γ on the model performance. The results were easy to reproduce given one has to run the tests many times for quantitative reasons. We could not reproduce the Hausdorff distance and Ramp score as there was no access to the supplementary material. DILATE works better than MSE but from the above experiments, in some cases, MSE is as good as DILATE. For more understanding, one can compare the performance of DILATE and MSE loss on transformer like models(3) with respect to better predictions and computational complexity.

References

[1] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International Conference on Machine Learning*.

[2] Vincent Le Guen and Nicolas Thome. Shape and time distortion loss for training deep time series forecasting models. In *NIPS Neural Information Processing Systems*.

[3] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *NIPS Neural Information Processing Systems*.

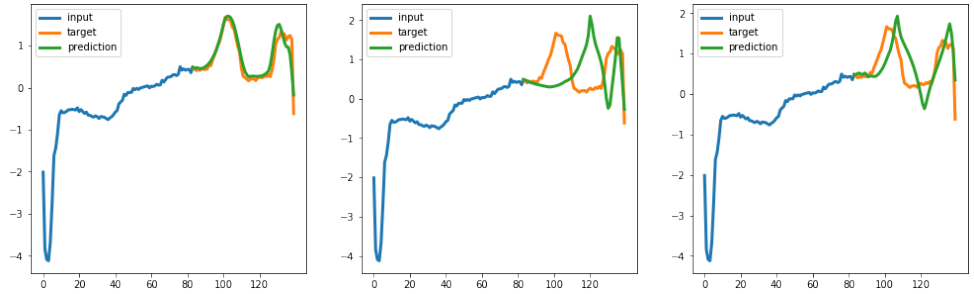
[4] Laura Frías Paredes, Fermín Mallor, Teresa Leon, and Martín Gaston Romeo. Introducing the temporal distortion index to perform a bidimensional analysis of renewable energy forecast. *Elsevier*.

[5] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pages 43 – 49. IEEE, 1990.

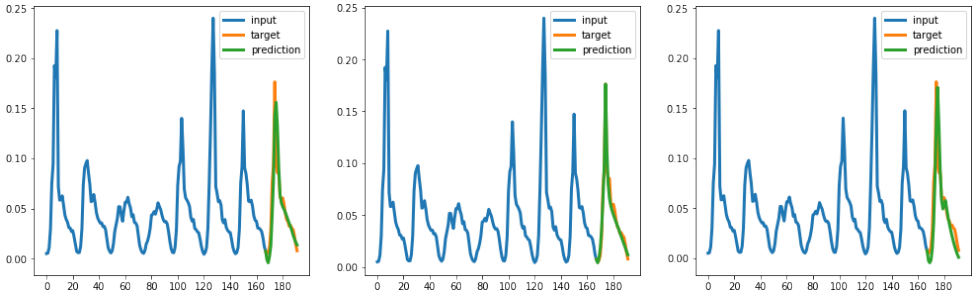
[6] Chen Yanping, Keogh Eamonn, Hu Bing, Begum Nurjahan, Bagnall Anthony, Mueen Abdullah, and Gustavo Batista. Ucr time series classification archive.

5 Appendix

The following are some more figures from the reproduction work.

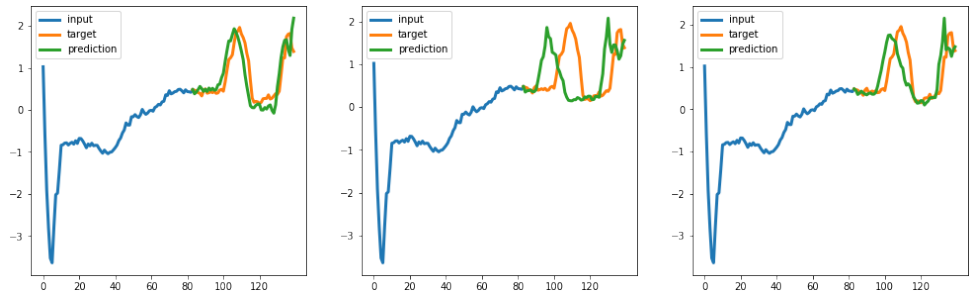


(a) ECG data

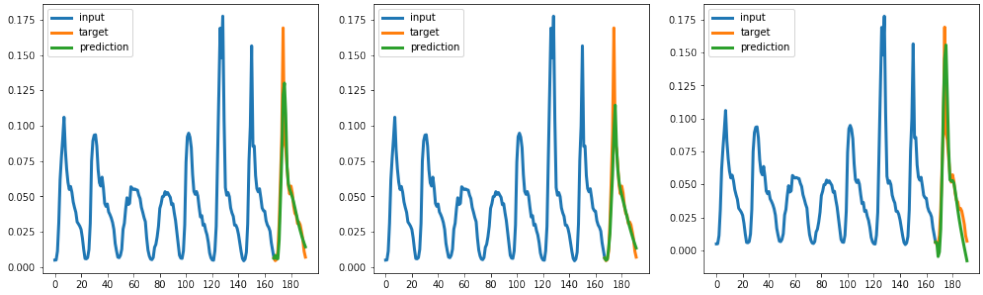


(b) Traffic data

Figure 5: Forecasting results for ECG and Traffic data based on Seq2Seq model. **Leftmost:** Seq2Seq with MSE loss, **Middle:** Seq2Seq with DTWγ loss, **Rightmost:** Seq2Seq with DILATE loss.

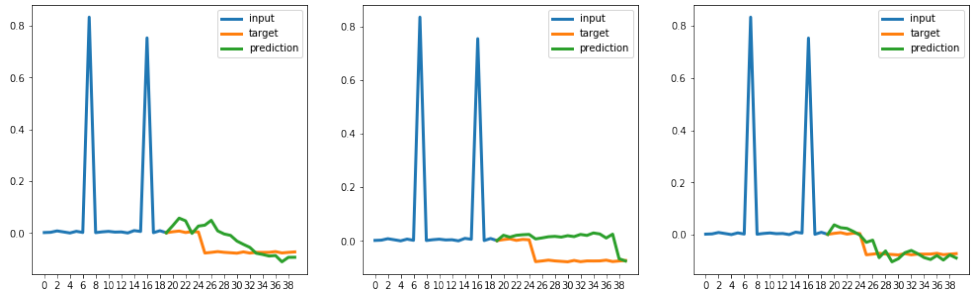


(a) ECG data

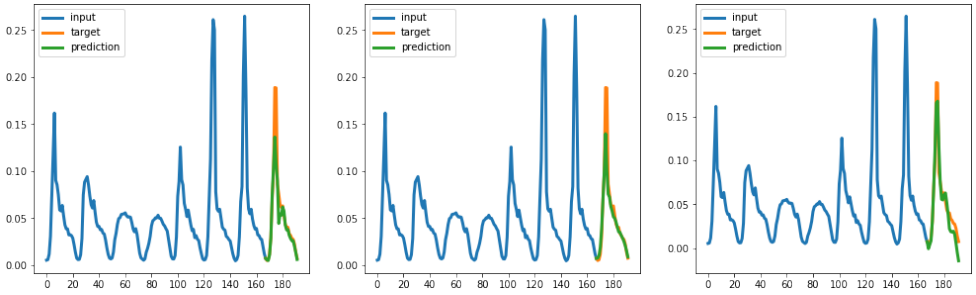


(b) Traffic data

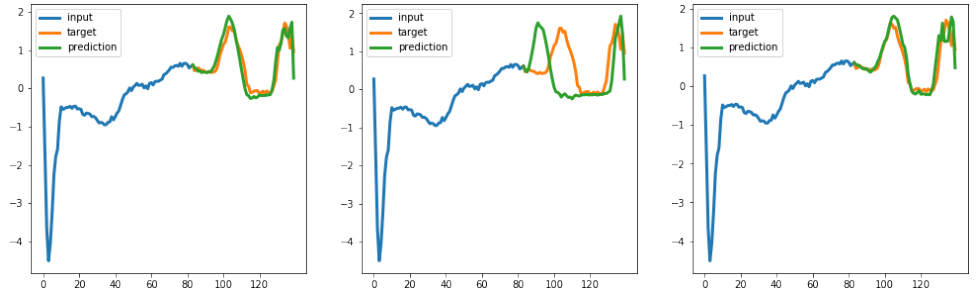
Figure 6: Forecasting results for ECG and Traffic data based on CNN_LSTM model. **Leftmost:** CNN_LSTM with MSE loss, **Middle:** CNN_LSTM with DTWγ loss, **Rightmost:** CNN_LSTM with DILATE loss.



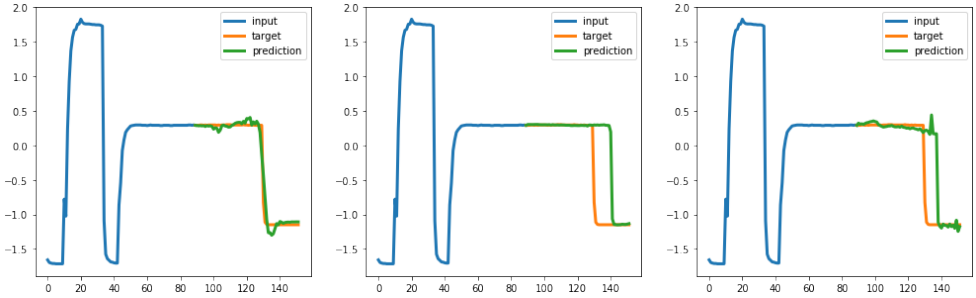
(a) Synthetic data



(b) Traffic data



(c) ECG data



(d) Wafer data

Figure 7: Forecasting results for 4 datasets based on Fully connected network(MLP) model. **Leftmost:** MLP with MSE loss, **Middle:** MLP with DTWγ loss, **Rightmost:** MLP with DILATE loss.