COMP6248 REPRODUCIBILITY CHALLENGE 2019/20 RE: SHAPE AND TIME DISTORTION LOSS FOR TRAINING DEEP TIME SERIES FORECASTING MODELS

Wen Rui Michael Mak, Virodh D Sok One & Jun Jet Wong
School of Electronics and Computer Science
University of Southampton

{wrmm1y14,vdso1g16,jjw5g16}@ecs.soton.ac.uk

ABSTRACT

DIstortion Loss including shApe and TimE (DILATE) was proposed to improve the training of Deep Neural Network on non-stationary time series forecasting models and this report attempts to reproduce the results using the same architecture with an in-depth analysis into the implementation as well as discussion of the findings observed throughout this reproducibility challenge.

1 INTRODUCTION

Recurrent Neural Network (RNN) is widely used in forecasting models because such neural network is able to remember the past through feedback loops and make predictions by storing past information efficiently through distributed hidden states and updating the hidden states in complicated ways to tackle non-linear dynamics. The variant of RNN utilised in the report by Guen & Thome (2019) was Gated Recurrent Unit which will save computation time due to its lower complexity. The implementation of DILATE is then introduced to improve on the the timing and shape of the predictions because using Mean Square Error (MSE) as the go-to loss function often resulted in inaccurate prediction due to the inability to capture sudden disturbances.

The experiment results were reproduced using the code provided by the author without modifying the network architecture. 3 types of dataset were used in this reproducibility challenge to test the capability of DILATE as compared to MSE where the first dataset is the synthetic dataset created by the author, the second is the ECG5000 dataset and the last dataset is the traffic dataset. Ultimately, the results were evaluated and analysed such as the impact of the hyperparameters to the performance.

2 THE IMPLEMENTATION OF DILATE

The distortion loss function comprised of 2 losses which are the shape loss and the temporal loss. Both the shape and temporal loss function compares the prediction with the ground truth and both terms are added together where the portion of each loss is determined by the hyperparameter α (Guen & Thome (2019)):

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

where $\mathcal{L}_{shape}$ is actually a Dynamic Time Wrapping loss function which has been relaxed by the smooth minimum operator with a relaxation factor γ to make it differentiable.

$$\mathcal{L}_{shape}(\hat{y_i}, y_i^*) = DTW_\gamma(\hat{y_i}, y_i^*) = -\gamma log(\sum_{\mathbf{A} \in \mathcal{A}_{k,k}} exp(-\frac{< \mathbf{A}, \Delta(\hat{y_i}, y_i^*) >}{\gamma})) \tag{2}$$

The shape loss deals with the difference in shape between the ground truth and the prediction. On the other hand, inspired by the Time Distortion Index, the $\mathcal{L}_{temporal}$ is introduced into the calculation

of DILATE to penalise temporal distortion between the prediction and the ground truth to reduce misalignment. The temporal loss function is given as follows:

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

where $Z = \sum_{\mathbf{A} \in \mathcal{A}_{k,k}} exp(-\frac{<\mathbf{A}, \Delta(\hat{y}_i, y_i^*)>}{\gamma})$, is for the smooth approximation for the temporal loss function to make it differentiable. The author also used an efficient implementation of forward and backward pass that was mentioned in Cuturi & Blondel (2017.) and in Mensch & Blondel (2018). It is claimed that the backpropagation has achieved a significant speed-up compared to standard auto-differentiation.

3 Experimental Methodology

Experiments from Guen & Thome (2019) was reproduced by building a seq2seq model with GRU network with the same architecture which is 1 layer of 128 units. Same ADAM optimiser is used and all the hyperparameters that was used in the code provided were unchanged.

The first dataset used was the **synthetic dataset** created by the author that uses an input signal composed of 2 peaks. The setup for training such as number of epoch, learning rate, smoothing parameter γ and etc., remained the same as described in Guen & Thome (2019). Second dataset used was **ECG5000** and the data preprocessing is the same as the author's. However, the **traffic dataset** used by the author couldn't be found so another dataset from the California Department of Transportation is used instead. For simplicity's sake, the data preprocessing used for this dataset is the same as the one used on synthetic dataset. A summary of the dataset setup is shown in the following table:

Dataset	Synthetic	ECG5000	Traffic
α	0.5	0.5	0.8
Time-series Input	20	84	50
k future predictions	20	56	30

Table 1: Dataset setup

4 Analysis and Discussion

The Seq2Seq model was trained using 3 different types of loss function which are MSE, soft-DTW, and DILATE. The performance of the trained models were then evaluated using MSE, DTW (shape), and TDI (temporal) to reproduce the evaluation result presented by the author. The reproduced results were compared with the author's results (shown in Table 2). As there are page restrictions, this report was unable to replicate and analyse the author's work completely. Judging from other reproducibility reports, a complete analysis required at least 8 full pages. Only vital analysis relating to DILATE were included to test and validate the author's original ideas.

The model which was trained using synthetic dataset matches the author's published results. Based on Table 2, the reproduced results for synthetic data using all loss functions were within a reasonable range of the author's results. Referring to Figure 1, DILATE's time series predictions best fits the target in terms of shape and temporal alignment. The MSE model reacted poorly to the sharp decline in value. Additionally, it was noticed that the MSE model tends to produce a smoother or rounded edge when there is a sharp decline or increase. On the other hand, the soft-DTW model was able to accurately predict a sharp decline but a shift in temporal alignment was observed. The results were as expected since no changes were made to the source code to reproduce this result.

In the early attempts to recreate the evaluation results for ECG5000, the author's methodology was ultilised for training. Large disparities were observed for all loss functions, especially for DTW evaluation. By estimation, the reproduced results are at least 1 order of magnitude higher than the

Dataset	Eval	**Seq2Seq RNN (using GRU)**		
		MSE	DTW_γ	DILATE
Synth	MSE	1.1 ± 0.10 (1.1 ± 0.17)	1.81 ± 0.25 (2.31 ± 0.45)	1.32 ± 0.32 (1.21 ± 0.13)
	DTW	23.8 ± 2.25 (24.6 ± 1.20)	15.0 ± 1.19 (22.7 ± 3.55)	20.1 ± 0.70 (23.1 ± 2.44)
	TDI	18.1 ± 1.11 (17.2 ± 1.22)	20.9 ± 1.69 (20 ± 3.72)	15.7 ± 1.8 (14.8 ± 1.29)
ECG	MSE	4.1 ± 0.42 (21.2 ± 2.24)	3.68 ± 0.21 (75.1 ± 6.30)	3.7 ± 0.55 (30.3 ± 4.10)
	DTW	71.9 ± 0.11 (17.8 ± 1.62)	71.8 ± 2.50 (17.1 ± 0.65)	72.3 ± 2.00 (16.1 ± 0.16)
	TDI	28.3 ± 0.92 (8.27 ± 1.03)	22.2 ± 5.1 (27.2 ± 11.1)	20.7 ± 0.80 (6.59 ± 0.79)
Traffic	MSE	5.1 ± 0.37 (0.89 ± 0.11)	8.34 ± 0.67 (2.22 ± 0.26)	6.95 ± 0.52 (1.00 ± 0.26)
	DTW	98.9 ± 7.40 (24.6 ± 1.85)	80.8 ± 7.38 (22.6 ± 1.34)	81.81 ± 11.67 (23.0 ± 1.62)
	TDI	20.0 ± 5.20 (15.4 ± 2.25)	52.27 ± 4.67 (22.3 ± 3.66)	11.85 ± 2.17 (14.4 ± 1.58)

Table 2: Evaluation results reproduced using MSE (x100), DTW (x100), and TDI (x10) metrics (mean $\pm$ standard deviation) averaged over 10 runs with comparison to the results in the original report (the values in bracket)

results published by the author. Referring to Singh & Wang (2019.), the authors suggested that the data should be scaled and normalised to obtain better results. For the second attempt, the data was scaled using a minimum (0.0) and maximum (1.0) standardisation technique. The observed result was much better in comparison to the early attempt. According to Table 2, the reproduced evaluation results still did not reasonably match the published results. Singh & Wang (2019.) reported similar observations for ECG5000 dataset. With that said, some trained model was able to predict time series data with reasonable accuracy. Referring to Figure 1, predictions from the models trained using MSE and DILATE can predict the sudden changes in time series data accurately with respect to its shape, temporal alignment and amplitude. The model trained using soft-DTW was barely able to predict time series data.

A different set of traffic data was used in this analysis because the original dataset used by the author was no longer available. Creating a performance analysis between the reproduced and published results under the current circumstances would not yield a meaningful comparison. Regardless, the results are shown in Table 2. Through observation, the reproduced results were several times worse for all loss functions and evaluation metrics. Still, Figure 1 proved that DILATE made the best predictions in terms of shape and timing (lower DTW and TDI).

The effect of the hyperparameter α is explored to understand how it affects the evaluation metrics. Similar to the author's analysis on DILATE, the temporal error does increase when α is increasing whereas the DTW decreases since the DILATE function focuses more on the shape loss. The plots in Figure 2 had been re-scaled to show the impact of α as the numerical value reproduced is not the same as the original graph.

5 Conclusion

In conclusion, this report has examined the effectiveness of DILATE for time series data predictions. DILATE was clearly superior at making timely predictions to a sharp change in shape when compared to MSE and soft-DTW models. Although this report fails to reproduce the key evaluation metrics published by the author, the predictions made were within a reasonable range of accuracy from a visual perspective. This report finds that both temporal and shape are invaluable factors that have contributed to the superior performance of DILATE.

References

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

Vincent Le Guen and Nicolas Thome. Shape and time distortion loss for training deep time series forecasting models. *NeurIPS 2019*, 2019.

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

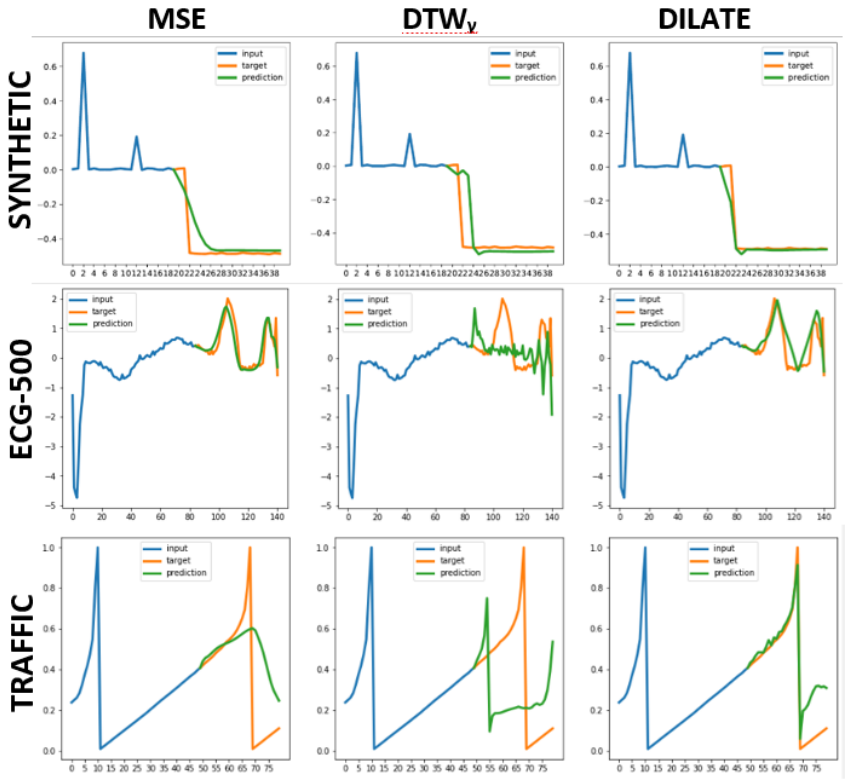

Figure 1: Visualisation of the reproduced results for the seq2seq model trained on different datasets with MSE loss function, soft-dtw loss function, and DILATE.

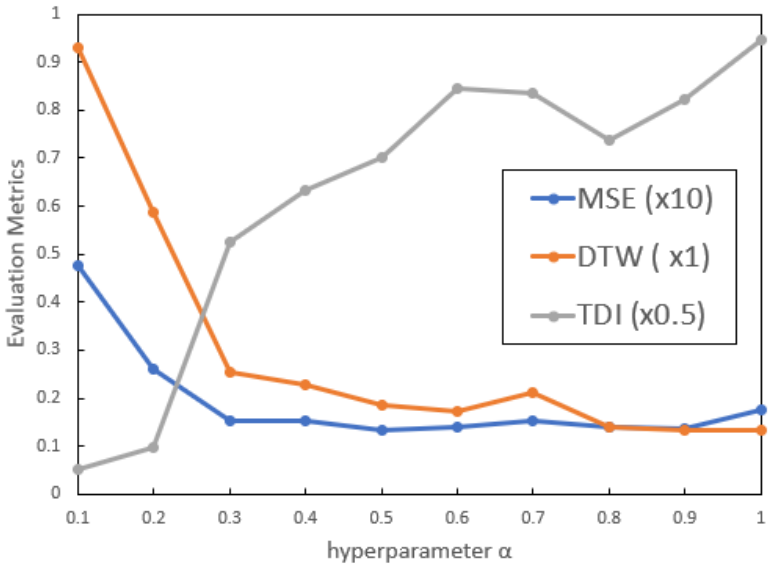

Figure 2: Impact of α

Manjot Singh and Yiyu Wang. Re: Shape and time distortion loss for training deep time series forecasting models. *International Conference on Machine Learning (ICML)*, pp. 2, 2019.