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

Vincent Le Guen / Nicolas Thome

EDF R&D / CEDRIC, Conservatoire National des Arts et Métiers

NIPS 2019

2020.07.02

최영제

0 SIL Index

1 Introduction

Abstract

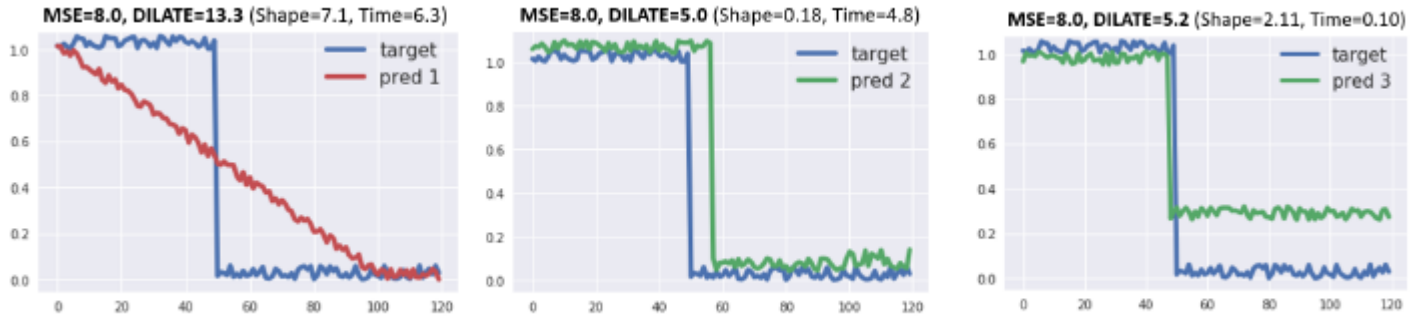
- 본 논문은 Time series forecasting task 중 non-stationary signals data, multiple future steps prediction을 다룸

- Time series forecasting에서 정확한 예측은 무엇보다도 중요하며 이는 크게 두가지 측면에서 바라볼 수 있음

  1) Ground truth를 정확하게 예측함 = 예측 시점(t)별 residual(true value - prediction)이 작음 = 실제 값과 예측 값의 shape이 동일함
  2) Change point 시점을 정확하게 예측함 = time lagging이 발생하지 않음

- 일반적으로 neural net을 설계할 때 regression problem은 MSE, MAE를 최소화 하는 것을 목표로 삼으며 time series forecasting도 동일함

  → 그러나 MSE나 MAE 및 기타 MSE 파생 loss function은 위의 두가지를 만족시킬 수 없음



(a) Non informative prediction (b) Correct shape, time delay (c) Correct time, inaccurate shape

1 Introduction

Abstract

- 본 논문은 Time series forecasting task 중 non-stationary signals data, multiple future steps prediction을 다룸

- Time series forecasting에서 정확한 예측은 무엇보다도 중요하며 이는 크게 두가지 측면에서 바라볼 수 있음

  1) Ground truth를 정확하게 예측함 = 예측 시점(t)별 residual(true value - prediction)이 작음 = 실제 값과 예측 값의 shape이 동일함
  2) Change point 시점을 정확하게 예측함 = time lagging이 발생하지 않음

- 일반적으로 neural net을 설계할 때 regression problem은 MSE, MAE를 최소화 하는 것을 목표로 삼으며 time series forecasting도 동일함

  → 그러나 MSE나 MAE 및 기타 MSE 파생 loss function은 위의 두가지를 만족시킬 수 없음

- 따라서 논문의 저자들은 DILATE(DIstortion Loss including shApe and TimE) Loss를 제안함

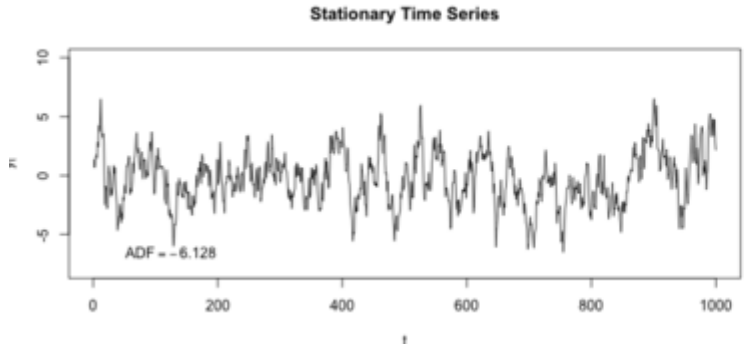
  → DILATE aims at accurately predicting sudden changes, and explicitly incorporates two terms supporting precise shape and temporal change detection

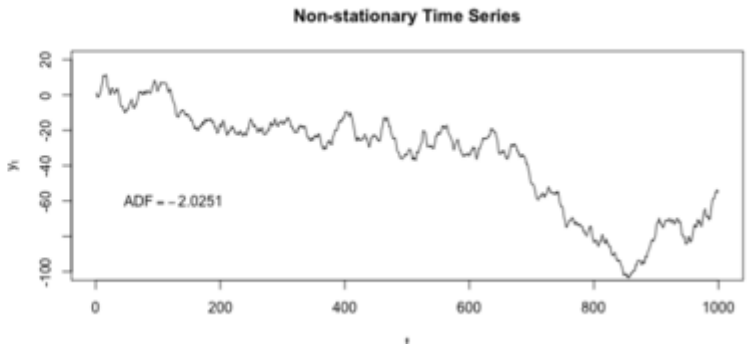
$$\mathcal{L}_{DILATE}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_i^*) = \alpha\, \mathcal{L}_{shape}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_i^*) + (1-\alpha)\, \mathcal{L}_{temporal}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_i^*) \qquad (1)$$

2 Related work

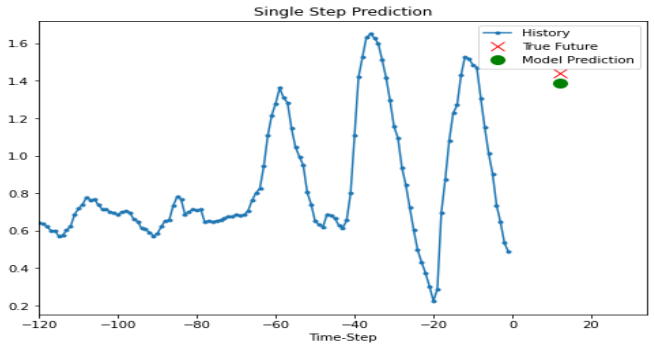
Background

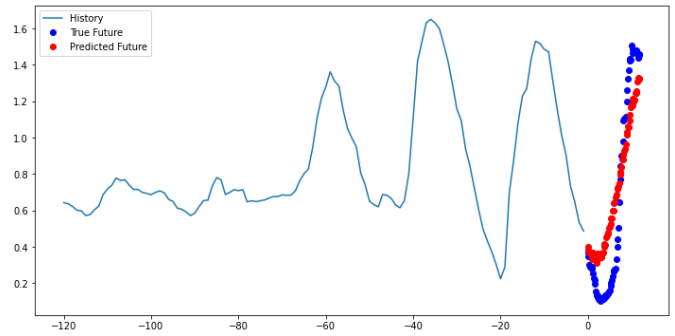
- Stationary process란 확률론에서 확률변수 간의 확률 분포가 시간에 상관없이 일정한 확률 과정을 말하며 non-stationary는 반대의 경우
  → stock market prediction이 대표적인 non-stationary이며 real data는 대부분 non-stationary에 해당





- multiple future steps prediction이란 t-n ~ t시점의 데이터를 학습하여 t+1~t+k 시점의 값을 예측하는 것을 말함
  → one-step prediction을 여러 번 진행하거나 multi-step model(ex.Seq2seq) 사용

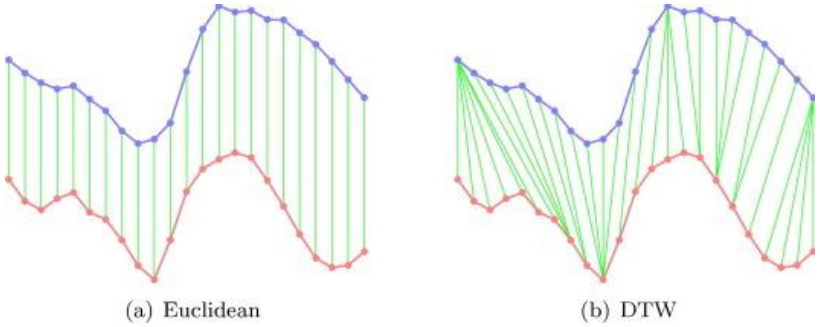


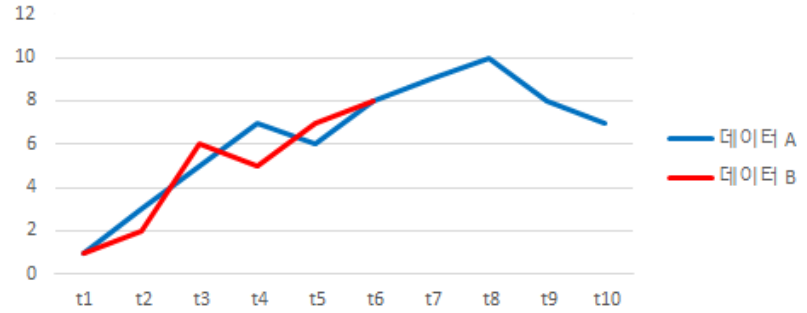


2 Related work

DTW(Dynamic Time Warping)

- Dynamic time warping이란 두개의 다른 속도의 시간 축의 파장의 유사성을 측정 및 매칭하는 알고리즘





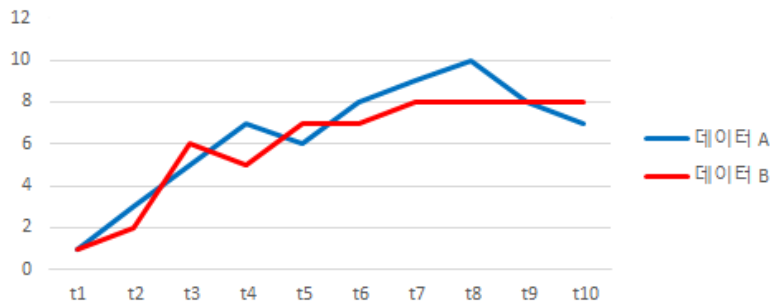
- 예를 들어 A - [1, 3, 5, 7, 6, 8, 9, 10, 8, 7], B - [1, 2, 6, 5, 7, 8]로 두 시계열 간 길이가 다른 경우 DTW를 이용하여 유사도를 구할 수 있음

  → distance matrix 상에서 좌 상단부터 우 하단까지 최소 distance elements를 이어서 matching을 판단

  → distance matrix의 각 축에서 음의 방향으로는 이동하지 못함

	Index	0	1	2	3	4	5	6	7	8	9
Index	Data	1	3	5	7	6	8	9	10	8	7
0	1	0	2	4	6	5	7	8	9	7	6
1	2	1	1	3	5	4	6	7	8	6	5
2	6	5	3	1	1	0	2	3	4	2	1
3	5	4	2	0	2	1	3	4	5	3	2
4	7	6	4	2	0	1	1	2	3	1	0
5	8	7	5	3	1	2	0	1	2	0	1

	Index	0	1	2	3	4	5	6	7	8	9
Index	Data	1	3	5	7	6	8	9	10	8	7
0	1	**0**	2	4	6	5	7	8	9	7	6
1	2	1	**1**	3	5	4	6	7	8	6	5
2	6	5	3	**1**	1	0	2	3	4	2	1
3	5	4	2	**0**	2	1	3	4	5	3	2
4	7	6	4	2	**0**	**1**	1	2	3	1	0
5	8	7	5	3	1	2	**0**	**1**	**2**	**0**	**1**



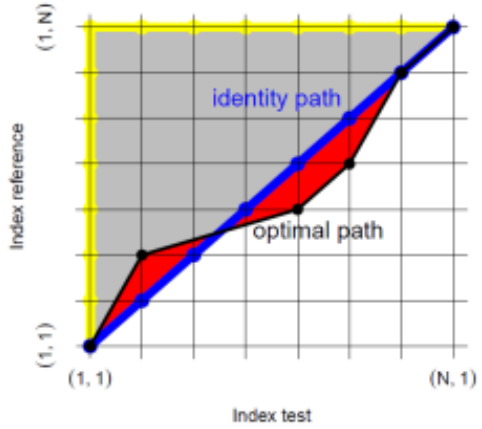
Related work

TDI(Time Distortion Index)

- Time distortion index란 DTW를 통해서 찾은 optimal path의 왜곡도를 표현하는 지표

  → TDI is in the interval[0,1]. 0 corresponds with the null temporal distortion, 1 with the maximum temporal distortion

  → TDI would be calculated as the quotient between the red area and the grey one



- 즉 DILATE loss는 DTW를 이용하여 predicted line의 shape을 학습하고, TDI를 이용하여 time delay를 보정하는 loss function

  → hyper-parameter alpha를 이용하여 반영 비율 조정

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

3 Proposed Approach

Overview

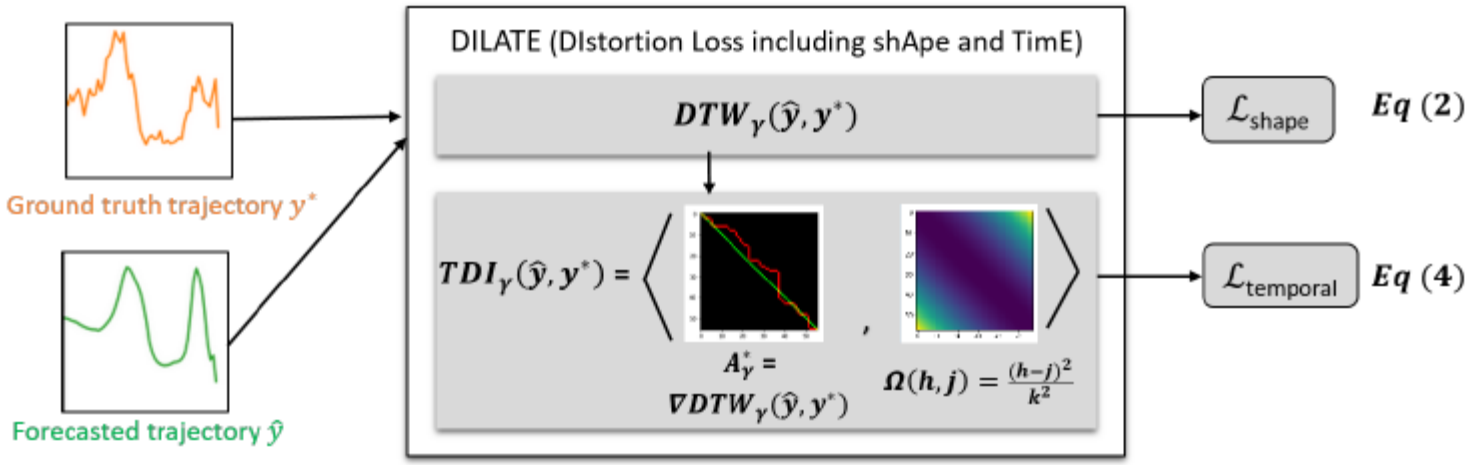


Figure 3: DILATE loss computation for separating the shape and temporal errors.

3 SIL Proposed Approach

Notations

- a set of N input time series $\mathcal{A} = \{\mathbf{x}_i\}_{i \in \{1:N\}}$
- each input example of length n $\mathbf{x}_i = (\mathbf{x}_i^1, ..., \mathbf{x}_i^n) \in \mathbb{R}^{p \times n}$
- predicted values the future k-step ahead trajectory $\hat{\mathbf{y}}_i \in \mathbb{R}^{d \times k}$
- ground truth $\overset{*}{\mathbf{y}}_i \in \mathbb{R}^{d \times k}$
- warping path as binary matrix $\mathbf{A} \subset \{0,1\}^{k \times k}$

  → set of all valid warping paths connecting the endpoints (1, 1) to (k, k) $\mathcal{A}_{k,k}$
- pairwise cost matrix $\boldsymbol{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := [\delta(\hat{\mathbf{y}}_i^h, \overset{*}{\mathbf{y}}_i^{\mathrm{j}})]_{h,j}$

  → where δ is a given dissimilarity between $\hat{\mathbf{y}}_i^h$ and $\overset{*}{\mathbf{y}}_i^{\mathrm{j}}$, likes the euclidean distance.

	Index	0	1	2	3	4
Index	Data	1	3	5	7	6
0	1	0	2	4	6	5
1	2	1	1	3	5	4
2	6	5	3	1	1	0
3	5	4	2	0	2	1
4	7	6	4	2	0	1
5	8	7	5	3	1	2

3 Proposed Approach

Shape loss function

- 앞서 말했듯, DTW를 사용하여 shape loss를 구하며 DTW와 optimal path의 notation은 아래와 같음

$$DTW(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = \min_{\mathbf{A} \in \mathcal{A}_{k,k}} \left\langle \mathbf{A}, \boldsymbol{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle \qquad \mathbf{A}^* = \arg\min_{A \in \mathcal{A}_{k,k}} \left\langle \mathbf{A}, \boldsymbol{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle$$

- 또한 DTW는 미분이 불가한 함수기 때문에 이를 smooth min operator로 변환(=Soft-DTW, ICML 2017)
  → 이동 가능한 모든 path의 distance 합을 최소화시키는 목적함수, gamma = 0이면 DTW와 동일
  → The cost of an alignment is equal to the sum of entries visited along the path

$$\mathcal{L}_{shape}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = DTW_\gamma(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := -\gamma \log \left(\sum_{\mathbf{A} \in \mathcal{A}_{k,k}} \exp \left(-\frac{\left\langle \mathbf{A}, \boldsymbol{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle}{\gamma} \right) \right) \qquad (2)$$

- Ground truth k – [1,4,7,10,13], predicted value k – [2,5,8,11,14] 일 때 MSE는 1, soft-DTW(gamma = 1)는 4.2931의 값을 갖음(DTW = 5)
  → predicted value k2 – [3,3,7,10,13]인 경우 MSE는 1로써 동일하지만 soft-DTW(gamma=1)은 4.6728로 더 큰 값을 갖음

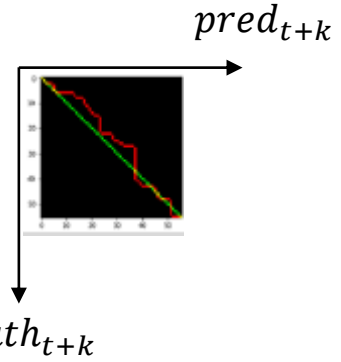
```
array([[  1.,  16.,  49., 100., 169.],
       [  4.,   1.,  16.,  49., 100.],
       [ 25.,   4.,   1.,  16.,  49.],
       [ 64.,  25.,   4.,   1.,  16.],
       [121.,  64.,  25.,   4.,   1.]])
```

```
array([[  4.,   4.,  36.,  81., 144.],
       [  1.,   1.,   9.,  36.,  81.],
       [ 16.,  16.,   0.,   9.,  36.],
       [ 49.,  49.,   9.,   0.,   9.],
       [100., 100.,  36.,   9.,   0.]])
```

3 SIL Proposed Approach

Temporal loss function

- Loss function의 두번째 term인 temporal loss는 실제 값과 예측 값의 시간적 왜곡에 penalty를 주는 것이 목표



- DTW의 distance matrix에서 identity path에서 벗어났다는 것은 개별 t시점의 정확한 예측에서 벗어났다는 말과 동일
  → 두 시계열이 예측 값과 실제 값이기 때문, 오른쪽 그림은 예측 값의 time lagging이 발생한 경우

- 따라서 identity path와 optimal path를 이용하여 time distortion이 발생하면 penalty를 가함

- Ω matrix를 optimal path에 내적함으로써 penalty를 부여하며 Ω matrix는 $h \neq j$인 elements에 값을 갖고 있음
  → prior knowledge can also be incorporated in the Ω matrix structure, e.g. to penalize more heavily late than early predictions (and vice versa)

$$TDI(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = \langle \mathbf{A}^*, \boldsymbol{\Omega} \rangle = \left\langle \underset{\mathbf{A} \in \mathcal{A}_{k,k}}{\arg\min} \left\langle \mathbf{A}, \boldsymbol{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle, \boldsymbol{\Omega} \right\rangle \qquad e.g.\ \boldsymbol{\Omega}(h,j) = \frac{1}{k^2}(h-j)^2$$

3 SIL Proposed Approach

Temporal loss function

- 그러나 마찬가지로 TDI는 미분이 불가, 그 이유는 optimal path를 찾는 argmin operator가 존재하기 때문

- 따라서 논문의 저자들은 DTW의 미분값 $\mathbf{A}^* = \nabla_\Delta DTW(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i)$이 argmin operator의 smooth approximation이라는 정의를 이용함

$$\mathbf{A}^*_\gamma = \nabla_\Delta DTW_\gamma(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = 1/Z \sum_{\mathbf{A}\in\mathcal{A}_{k,k}} \mathbf{A} \exp^{-\frac{\left\langle \mathbf{A}, \Delta(\hat{\mathbf{y}}_t, \overset{*}{\mathbf{y}}_t)\right\rangle}{\gamma}} \qquad Z = \sum_{\mathbf{A}\in\mathcal{A}_{k,k}} \exp^{-\frac{\left\langle \mathbf{A}, \Delta(\hat{\mathbf{y}}_t, \overset{*}{\mathbf{y}}_t)\right\rangle}{\gamma}}$$

- 최종적으로 얻어진 smoothed temporal loss는 다음과 같음

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

→ 구현된 코드는 아래와 같음

```
path_dtw = path_soft_dtw.PathDTWBatch.apply
path = path_dtw(D,gamma)
Omega =  soft_dtw.pairwise_distances(torch.range(1,N_output).view(N_output,1)).to(device)
loss_temporal =  torch.sum( path*Omega ) / (N_output*N_output)
```

4 Experiments

Experimental setup – Datasets & network architecture

- Datasets

1) Synthetic : 인공적인 데이터, step function의 형태를 띄며 500/500/500으로 train, valid, test를 나눔, Gaussian noise(σ = 0.01)가 추가됨

2) ECG5000 : UCR Time Series Classification dataset. 길이가 140인 5000개의 electrocardiograms. train 500 test 4500으로 구성.
84 time steps를 input으로 넣어 나머지 56steps를 predict하도록 구성

3) Traffic : 2015-2016년 California 도로 점유 비율(교통량, 0-1) 데이터로 48개월 동안 1시간 단위로 기록.
17,544 length의 univariate time series를 60/20/20으로 분리하였으며 168points로 24points를 예측

- Network architecture

- GRU(1layer of 128 units)로 구성된 seq2seq 모델을 사용
- Max epochs = 1000 with early stopping with the ADAM optimizer
- Smoothing parameter γ of DTW and TDI is set to 10^{-2}
- Balancing parameter α is 0.5 for Synthetic and ECG50000 and 0.8 for Traffic dataset

4 Experiments

DILATE forecasting performances

		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.

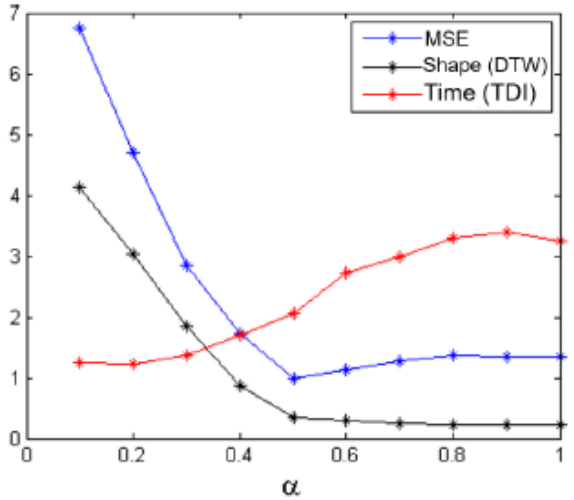


Figure 5(b): Influence of α

4 Experiments

Shape of predicted values

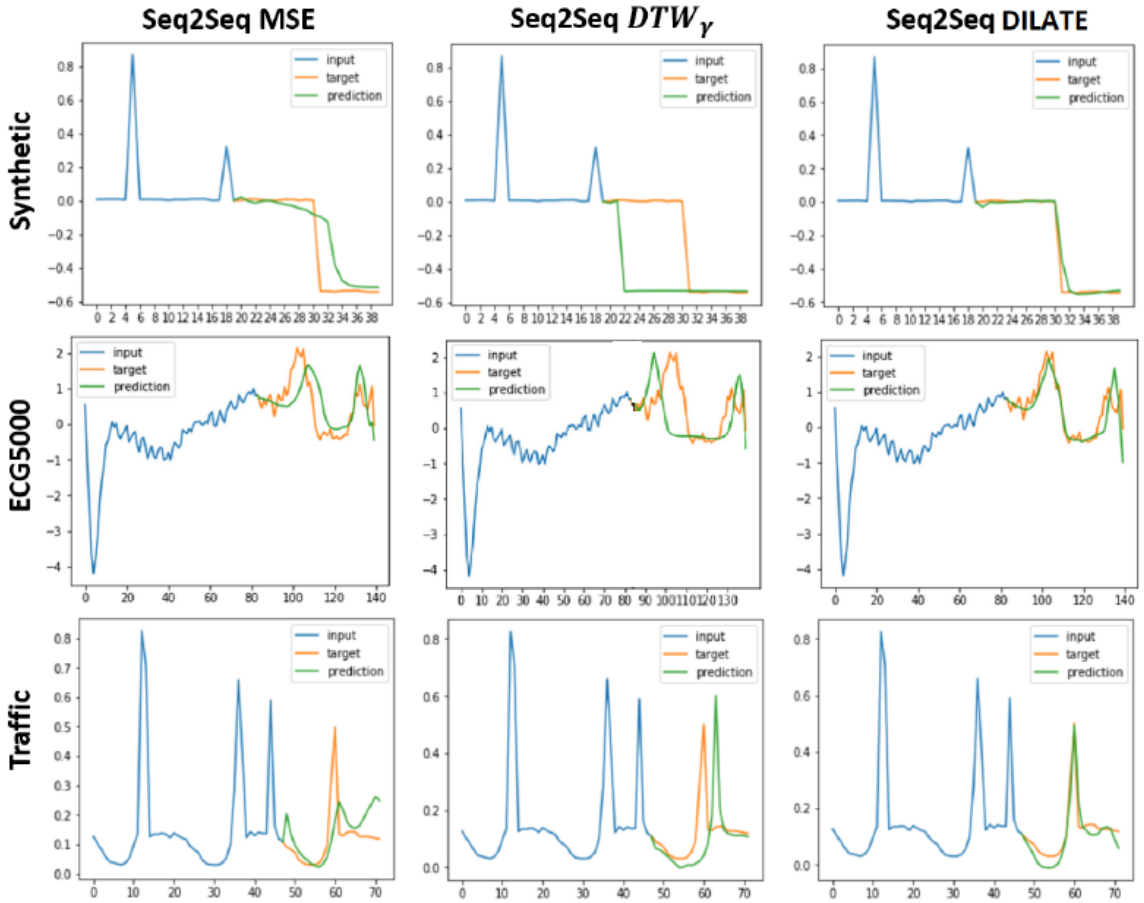


Figure 4: Qualitative forecasting results.

4 Experiments

Comparison to SOTA models

- Seq2seq 모델을 제외한 나머지는 MSE로 training
- LSTNet-rec : LSTNet의 one-step prediction을 k-step 반복
- TT-RNN : Tensor-train RNN, multi-step prediction
- 평가지표
  - Hausdorff : 서로 다른 두 개의 time series의 change point의 차이를 이용, 이를 통해 time의 어긋남을 측정함
  - Ramp score : 두 시계열의 기울기(SD)를 이용하여 shape을 판단하는 지표

$$\text{Hausdorff}(\mathcal{T}^*, \hat{\mathcal{T}}) := \max(\max_{\hat{t}\in\hat{\mathcal{T}}} \min_{t^*\in\mathcal{T}^*} |\hat{t} - t^*|, \max_{t^*\in\mathcal{T}^*} \min_{\hat{t}\in\hat{\mathcal{T}}} |\hat{t} - t^*|)$$

$$ramp\ \ score = \frac{1}{t_{max} - t_{min}} \int_{t_{min}}^{t_{max}} |SD(T(t)) - SD(R(t))| dt$$

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 ± standard deviation).

Reference

- DTW
  - Cuturi, M., & Blondel, M. (2017). Soft-DTW: a differentiable loss function for time-series. arXiv preprint arXiv:1703.01541.

    Ramp score : accuracy: A new ramp and time alignment metric. Solar energy, 150, 408-422.

  - https://medium.com/@Aaron__Kim/dynamic-time-warping-%EB%8F%99%EC%A0%81-%EC%8B%9C%EA%B0%84-%EC%9B%8C%ED%95%91-ac80777f49a

- TDI
  - Gastón, M., Frías, L., Fernández-Peruchena, C., & Mallor, F. (2017, June). The temporal distortion index (TDI). A new procedure to analyze solar radiation forecasts. In AIP Conference Proceedings (Vol. 1850, No. 1, p. 140009). AIP Publishing LLC.

- Ramp score
  - Vallance, L., Charbonnier, B., Paul, N., Dubost, S., & Blanc, P. (2017). Towards a standardized procedure to assess solar forecast

Q&A