Mini-Project (ML for Time Series) - MVA 2023/2024

Ben KABONGO ben.kabongo_buzangu@ens-paris-saclay.fr
Martin BROSSET martin.brosset@student-cs.fr

December 18, 2023

1 Introduction and contributions

La prévision des séries temporelles constitue une problématique centrale dans le domaine du Deep Learning, où la maîtrise des signaux non stationnaires et la prédiction précise à plusieurs étapes temporelles futures restent des défis de taille.
Traditionnellement, les modèles prédictifs s'appuient sur la loss function de l'Erreur Quadratique Moyenne (MSE), qui, malgré sa popularité, montre des limites dans la capture des dynamiques complexes et des transitions abruptes propres aux séries temporelles. La MSE tend à lisser les prédictions, ce qui peut nuire à la détection des pics ou des changements soudains.
Dans ce contexte, la DTW offre une alternative attrayante, capable de mesurer avec précision la similarité entre deux séquences temporelles, en tenant compte des distorsions possibles dans le temps. Cependant, sa non-différentiabilité constitue un obstacle majeur à son intégration dans les réseaux neuronaux, qui nécessitent des fonctions de perte permettant la rétropropagation pour l'ajustement des poids.
De plus la DTW est calculée de telle manière à ne pas considérer les faibles décalages temporels ce qui représente une autre limite de son utilisation pour prédire des séries temporelles.

Ainsi le modèle DILATE, (DIstortion Loss including shApe and TimE), proposé par Le Guen et Thomes, apporte une contribution significative. En introduisant une fonction de perte différentiable qui intègre à la fois la forme et la temporalité des séries temporelles, DILATE pallie les insuffisances de la MSE et de la DTW classique.
Cette nouvelle fonction offre une granularité fine dans la prévision des changements soudains, en alignant les prédictions avec les événements critiques des séries, là où la MSE pourrait les diluer. Pour ce faire, DILATE généralise le DTW en une forme lisse et temporellement contrainte, permettant son utilisation efficace dans l'optimisation des réseaux de neurones profonds.

Pour étudier cet article, Ben a cherché à trouver des datasets de séries temporelles aux propriétés différentes pour voir si le modèle étudié performait particulièrement bien ou non dans certains cas, et a adapté une partie du code original. Martin a lancé le modèles sur les différents datasets en optimisant les paramètres et extrayant les résultats.
Nous avons réutilisé une partie du code (60%) des auteurs afin de coder les implémentations forward et backward des loss de manière astucieuse et sans quoi les calculs prennaient beaucoup trop de temps (on en discute à la fin de la section méthode). On a lancé les modèles avec un taux d'apprentissage adaptatif, choix sans lequel le modèle proposé convergeait rarement et testé l'influence du paramètre de pondération entre les deux composantes de la loss dilate.

2 Method

Soit un ensemble d'apprentissage $\{\mathbf{x}_i\}_{i=1,...,N}$, tel que $\mathbf{x}_i = (x_i^1, ..., x_i^n) \in \mathbb{R}^{d \times n}$ est une série temporelle de taille n et de dimension d.
On souhaite prédire pour chaque série temporelle la suite de la série sur k pas de temps. On note donc $\hat{\mathbf{y}}_i = (\hat{y}_i^1, ..., \hat{y}_i^k)$ la prédiction pour une série temporelle $\mathbf{x}_i$ et $\mathbf{y}_i^* = (y_i^{*1}, ..., y_i^{*k})$ la suite réelle de la série temporelle.

Etant donné un couple $(\hat{\mathbf{y}}_i, \mathbf{y}_i^*)$, la loss DILATE se décompose en deux termes distincts : un terme de forme (shape) et un terme temporel (temporal).

Terme de forme. Le terme de forme, noté L_{shape}, est basé sur la Dynamic Time Warping (DTW), mais contrairement à la DTW qui est non-différentiable, L_{shape} est rendu différentiable grâce à l'introduction d'un opérateur de minimum lisse $\min_\gamma$:

$$\min_\gamma(a_1, \dots, a_n) = -\gamma \log \left(\sum_{i=1}^{n} \exp\left(-\frac{a_i}{\gamma}\right) \right), \tag{1}$$

où $\gamma > 0$ est un paramètre qui contrôle le degré de lissage.
Comme on peut exprimer la DTW comme un minimum :

$$\text{DTW}(\hat{y}_i, y_i^*) = \min_{A \in \mathcal{A}_{k,k}} \langle A, \Delta(\hat{y}_i, y_i^*) \rangle$$

avec $\mathcal{A}_{k,k}$ représentant l'ensemble de tous les chemins de warping valides reliant les points de départ et d'arrivée des séries temporelles, et Δ est la matrice des coûts pair à pair qui calcule la dissimilarité, comme la distance euclidienne, entre les points des trajectoires prédites $\hat{y}_i$ et ceux des trajectoires réeles y_i^*.
On peut donc l'implémenter son opérateur lisse, le terme de forme différentiable est alors exprimé comme :

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

L'optimisation de cette fonction permet au modèle d'effectuer des prédictions dont la forme sera similaire mais sans réellement pénaliser les faibles distortions temporelles.

Terme temporel. Le terme temporel $L_{temporal}$, vise à pénaliser les distorsions temporelles entre les séries prédites et les séries réelles.
Il est inspiré de l'Indice de Distorsion Temporelle (TDI) pour l'estimation du désalignement temporel et utilise une matrice Ω pour quantifier les écarts temporels.
La matrice Ω est définie de manière à pénaliser plus fortement les associations entre points qui sont éloignés temporellement :

$$\Omega(h, j) = \frac{1}{k^2}(h - j)^2, \tag{2}$$

ce qui renforce la pénalisation des grands écarts temporels dans la série.
La fonction de perte temporelle lissée est donnée par :

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

Où Z est la fonction de partition, qui sert à normaliser la contribution de chaque chemin de warping et est définie comme :

$$Z = \sum_{A \in \mathcal{A}_{k,k}} \exp\left(-\frac{\langle A, \Delta(\hat{y}_i, y_i^*)\rangle}{\gamma}\right). \tag{3}$$

Dilate. En combinant ces deux termes, on obtient la fonction de perte DILATE :

$$\begin{aligned} L_{dilate} &= \alpha L_{shape} + (1-\alpha) L_{temporal} \\ &:= -\gamma \log\left(\sum_{A \in \mathcal{A}_{k,k}} \exp\left(-\frac{\langle A, \alpha\Delta(\hat{y}_i, y_i^*) + (1-\alpha)\Omega\rangle}{\gamma}\right)\right) \end{aligned}$$

L'intérêt de cette fonction de loss par rapport aux travaux qui ont précédés l'article qui implémentent une DTW différentiable est que l'ajout de la pénalisation de distortion temporelle. Son ajout permet de mieux appréhender les moments où des changements brusques surviennent.

Les auteurs présentent des méthodes astucieuses que nous avons réutilisées pour le calcul des passes avant et arrière des pertes de forme (L_{shape}) et temporelle ($L_{temporal}$) dans l'implémentation de leur fonction de perte DILATE. En effet le calcul de $\mathcal{A}_{k,k}$ pose un problème car est très long à calculer, car croissant exponentiellement avec k.

Forward et backward du terme de forme L_{shape} :

Passe Avant (Forward Pass). Pour résoudre ce problème d'optimisation, la méthode de programmation dynamique de Bellman est employée. Elle permet de décomposer le problème complexe en sous-problèmes plus simples, en calculant les coûts de distorsion pour chaque paire de points de manière itérative. Cette approche trouve le chemin optimal de distorsion minimale dans $\mathcal{A}_{k,k}$ sans avoir à évaluer tous les chemins possibles, ce qui serait prohibitif en termes de calcul

Passe Arrière (Backward Pass). Lors du passage arrière, l'implémentation personnalisée dans Pytorch prend avantage des résultats intermédiaires obtenus pendant le passage avant (c'est-à-dire les informations de $\mathcal{A}_{k,k}$ et les coûts de distorsion calculés).
En réutilisant ces données, la méthode évite de recalculer le chemin optimal et les coûts associés, ce qui serait nécessaire avec l'auto-différenciation standard. Cela rend le passage arrière beaucoup plus efficace en termes de calcul

Forward et backward du terme de temporel $L_{temporal}$:

Passage Avant (Forward Pass). Pour le calcul du gradient de la soft DTW, on utilise la même implémentation que le passage arrière pour L_{shape}.

Passage Arrière (Backward Pass). Calcul du Hessien de la soft DTW avec une méthode de programmation dynamique, rendant le processus plus rapide que l'auto-différenciation standard.

Ces implémentations astucieuses permettent de passer en une complexité polynomiale (en $O(k^2)$)

3 Data

Les différents datasets que nous utilisons dans nos expérimentations sont les suivants :

Données synthétiques. Une partie de nos expérimentations portent des jeux de données synthétiques générés comme dans le papier que nous avons étudié. Les données synthétiques sont

utilisées pour évaluer les différentes loss sur les changements soudains dans les séries temporelles. Les signaux comprennent donc deux pics placés aléatoirement pour chaque signal afin d'incorporer les changements brusques.

ECG5000. Les données originales du dataset ECG5000 ont été obtenues à partir de la base de données BIDMC Congestive Heart Failure Database (CHFDB) de PhysioNet. Les données ont été prétraitées en deux temps : chaque battement de cœur a d'abord été extrait, puis chaque battement de cœur a été interpolé pour avoir la même longueur de 140 pas de temps. On retrouve 5 classes de battements cardiaques différents dans le dataset : normal (1), et d'autres classements anormaux repartis en quatre classes différentes. (Voir Figure 2)

Classes	1 (normal)	2	3	4	5
Ensemble d'apprentissage	291	177	10	19	2
Ensemble de test	2626	1590	86	175	22
Dataset	2917	1767	96	194	24

Table 1: Dataset ECG5000

Traffic. Le dataset Traffic comprend les taux d'occupations des routes, normalisé entre 0 et 1, du département des transports de Californie, sur 48 mois à partir de 2015-2016. Nous sélectionnons 1000 séries temporelles d'une longueur de 192, dont 500 pour l'apprentissage lors de nos expérimentations et le reste pour l'évaluation. (Voir Figure 3)

Insect Sound. Le dataset Insect Sound a été généré par le groupe d'entomologie informatique de l'UCR. Il s'agit d'enregistrements audios des sons émis par différentes espèces de mouches, des segments de 10 ms échantillonnés à 6000 Hz. Chaque série temporelle a une longueur de 600. Le dataset comprend 10 espèces d'insectes. Nos expériences ne portent que sur une seule : l'*aedes female*. On retrouve un dataset réduit à 1000 signaux en apprentissage et 9000 signaux en test. Pour la plupart des signaux, le début et la fin ressemble à du bruit blanc. Dans nos expérimentations, nous intéressons à une fenêtre centrale de chaque signal. Ce dataset a été choisi pour sa nature complètement non stochastique. (Voir Figure 4)

4 Results

Résultats quantitatifs. Nous avons testé la prévision à plusieurs étapes en utilisant un modèle Seq2Seq spécialisé, constitué d'une couche de 128 unités GRU (Gated Recurrent Units). On entraîne les modèle jusqu'à convergence.
On obtient le tableau comparatif des résultats suivants qui représentent les moyennes $\pm$ les écarts types calculés sur 5 runs dans la table 2.

Choix de l'hyper-paramètre α. Cet hyper-paramètre est crutial pour l'optimisation du modèle. Il représente un compromis entre l'optimisation de la prédiction de la forme (mesurée par la DTW) et celle de la justesse de prédiction du moment où un changement brusque apparaît (mesurée par la TDI). Visuellement, on peut bien observer l'impact de ce paramètre sur le dataset synthétique 5. Un α élevé entraîne globalement une plus faible DTW et donc une justesse dans la forme prédite, un α faible permet de mieux capter les moments où le signal connait un changement soudain.
Cependant cette dépendance entre α et la valeur de la DTW et celle de la TDI est plus complexe, on observe empiriquement qu'au delà d'un seuil, propre à chaque dataset, augmenter sa valeur

peut dégrader la DTW et la diminuer peut dégrader la TDI.
On peut d'ailleurs l'observer dans le tableau des résultats, sur le dataset ECG500, DILATE avec un $\alpha = 0.5$ permet d'obtenir une DTW plus faible que la Soft DTW (équivalente à DILATE avec $\alpha = 1$.)
Pour les autres datasets, on a choisi α de manière empirique en testant différentes valeurs et en choisissant celle qui permet d'obtenir un compromis satisfaisant, notons cependant que ce choix dépendra de ce qu'on cherche à prédire. Les valeurs de α sélectionées sont 0.5 pour ECG5000 et Synth et Traffic et 0.7 pour Insect.

Résultats qualitatifs et réflexion sur la pertinence des loss par type de données à prédire. On présente quelques prédictions qualitatives dans la figure 6.

On observe que DILATE permet d'obtenir de meilleures performances en terme de DTW et TDI (les mesures qu'elle est censée optimiser) par rapport à la loss MSE sur les datasets ECG5000 et synthétique. Ces datasets présentent des formes caractéristiques que le terme de forme arrive à bien saisir, alors que même si la MSE a des performance assez proches de DILATE, elle a tendance à lisser les prédictions et perdre une partie de l'information.

Sur les dataset Traffic et Insect, leur performances sont similaires, voire la loss MSE performe mieux en terme de MSE.
Une explication est que la DTW est particulièrement adaptée pour des séries temporelles où l'alignement temporel peut varier, permettant une certaine flexibilité dans la correspondance des motifs.
Avec un dataset présentant des motifs très réguliers sans décalages ou étirements temporels (comme Traffic et Insect), la MSE pourrait être plus appropriée car elle assume une correspondance un-à-un stricte entre les points de temps.
Même si l'ajout de la composante temporelle permet de réduire cette erreur, on observe empiriquement qu'elle persiste.

En comparant à la loss SOFT DTW, DILATE performe globalement mieux dans l'ensemble de nos expérimentations. Elle permet au modèle de capter à la fois la forme de la série (faible DTW) mais également le temporalité (faible TDI), ce que la SOFT DTW ne permet pas de faire.
On note une exception, dans le dataset Synth, la loss SOFT DTW permet d'obtenir une meilleure DTW, comme discuté plus tôt.

Quelques conclusions suite aux expérimentations. On peut noter que l'entraînement du modèle avec la loss DILATE est plus long que celui avec la loss SOFT DTW lui-même plus long que celui avec la loss MSE. Ceci est dû à la complexité du calcul dans les passes avant et arrière.
On note également que pour que le modèle converge vers de paramètre optimaux/intéressant (i.e. pas de prédictions constantes par exemple), la loss DILATE est très sensible aux hyperparamètres comme le taux d'apprentissage, la taille des batch et leur calibration peut s'avérer laborieuse.

Appendix

Datasets

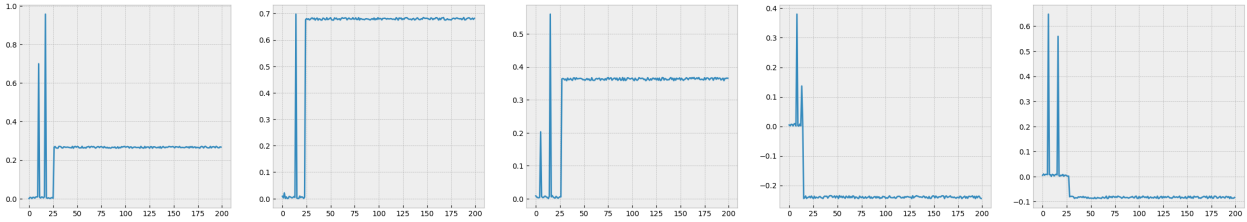

Figure 1: Données synthétiques

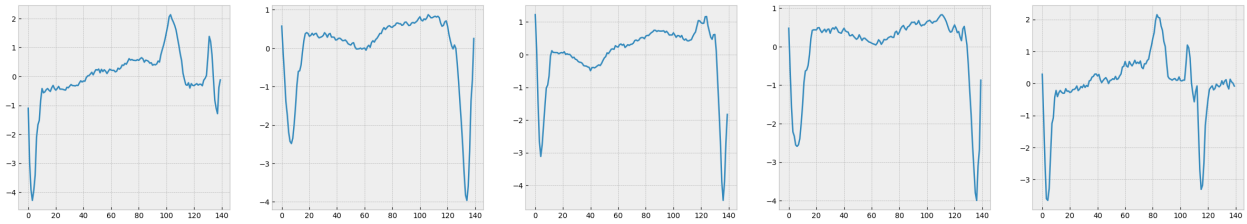

Figure 2: Dataset ECG 5000

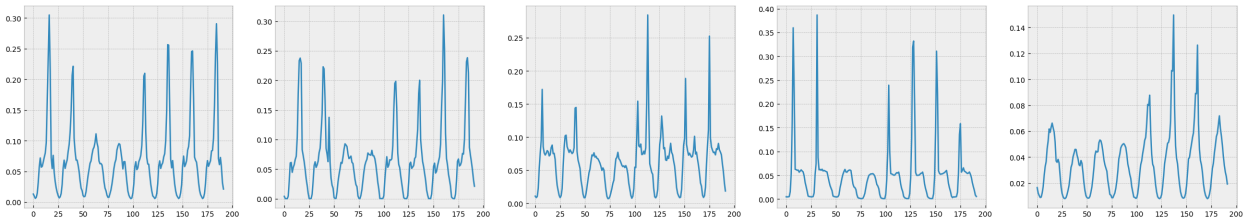

Figure 3: Dataset Traffic

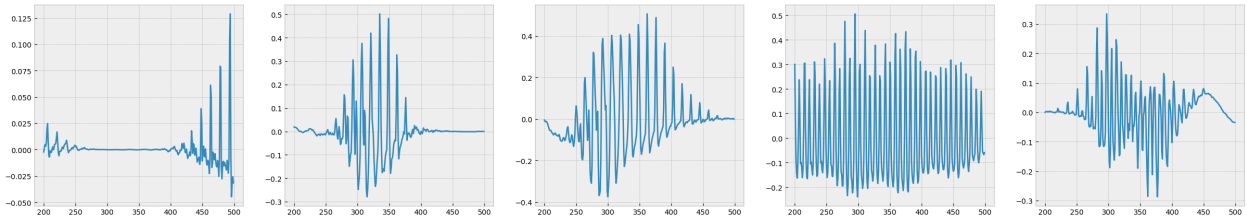

Figure 4: Dataset Insect Sound - Espèce *aedes female*

Résultats

	DILATE	MSE	SOFT DTW
ECG5000			
MSE	0.457 ± 0.026	0.403 ± 0.030	1.169 ± 0.328
DTW	1.850 ± 0.032	2.114 ± 0.052	1.860 ± 0.143
TDI	0.458 ± 0.026	0.556 ± 0.041	1.981 ± 0.750
Traffic			
MSE	0.567 ± 0.030	0.363 ± 0.004	1.085 ± 0.132
DTW	1.831 ± 0.005	1.836 ± 0.010	1.836 ± 0.080
TDI	0.408 ± 0.023	0.402 ± 0.003	2.415 ± 0.467
Synth			
MSE	0.01689 ± 0.001	0.0128 ± 0.001	0.02117 ± 0.00367
DTW	0.2083 ± 0.0402	0.2633 ± 0.0128	0.1517 ± 0.0022
TDI	1.5688 ± 0.2115	1.7504 ± 0.1241	1.8313 ± 0.0603
Insect			
MSE	1.187 ± 0.058	0.817 ± 0.00049	1.185 ± 0.077
DTW	3.361 ± 0.052	3.568 ± 0.083	3.213 ± 0.0997
TDI	3.259 ± 0.411	4.778 ± 0.07	6.645 ± 0.265

Table 2: Tableau comparatif des résultats de prédiction par loss et par dataset

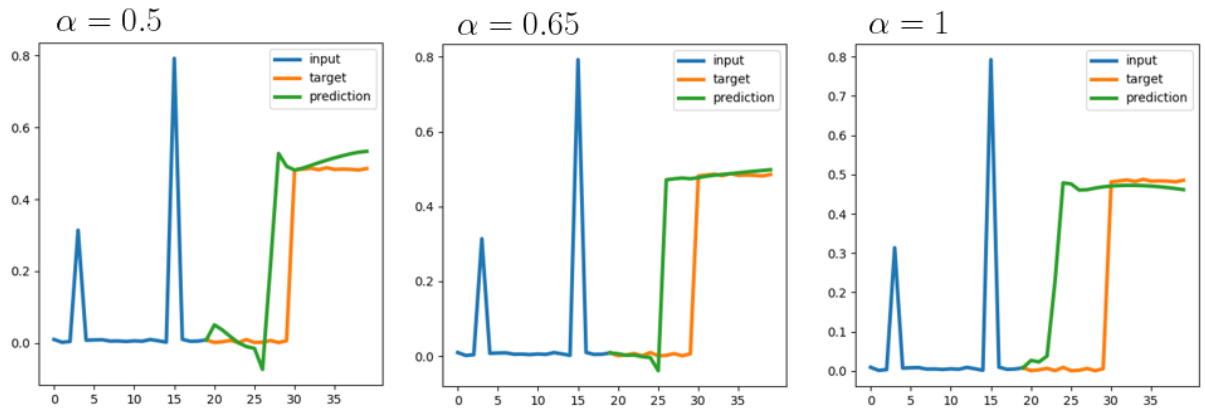

Figure 5: Influence du paramètre α sur une série temporelle du dataset synthétique

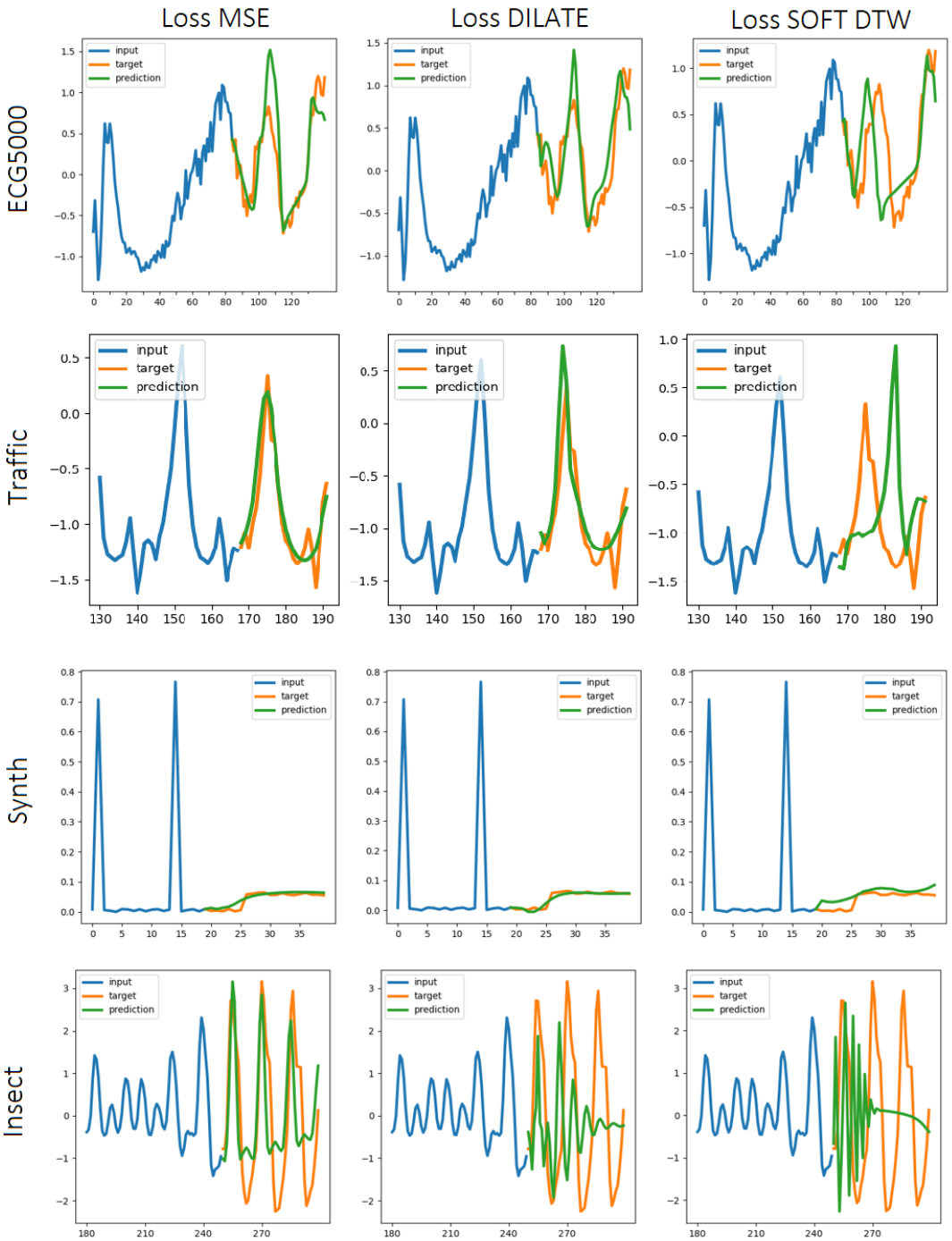

Figure 6: Résultats qualitatifs de prédiction par loss et par dataset