Deep Classification of Time-Series Data with Learned Prototype Explanations

Diego Garcia-Olano [*1 2] **Alan H. Gee** [*1 2] **Joydeep Ghosh** [1] **David Paydarfar** [2]

Abstract

The emergence of deep learning networks raises a need for explainable AI so that users and domain experts can be confident applying them to high-risk decisions. In this paper, we leverage data from the latent space induced by deep learning models to learn stereotypical representations or "prototypes" during training to elucidate the algorithmic decision-making process. We study how leveraging prototypes effect classification decisions of two dimensional time-series data in two settings: (1) electrocardiogram (ECG) waveforms to detect clinical bradycardia, a slowing of heart rate, in preterm infants, and (2) audio waveforms to classify spoken digits[1]. We improve upon existing models by optimizing for increased prototype diversity and robustness, visualize how these prototypes in the latent space are used by the model to distinguish classes, and show that prototypes are capable of learning features on two dimensional time-series data to produce explainable insights during classification tasks.

1. Introduction

Despite the recent surge of machine learning, adoption of deep learning models in decision critical domains, such as healthcare, has been slow because of limited transparency and explanations in black-box algorithms. This observation points to the critical need for black-box models to offer interpretable, faithful explanations of their decisions so that practitioners in high-risk domains can trust model outputs and leverage their results.

Prototypes are representative examples learned in-process during model training that describe influential regions in latent representations and can provide insight into the features utilized by the model for classification. In contrast to post-hoc explainability, which trains a secondary model to infer decision reasoning from a primary model by only leveraging inputs and outputs, in-process explainable methods offer faithful explanations of a primary model's decisions (Rudin, 2018).

Explainable methods (Ribeiro et al., 2016; Caruana et al., 2015; Zhou et al., 2015) have largely focused on labeled image and tabular data sets where classes are clearly separable and less so on time-series data in general. Time-series classification on 1-D data with deep neural networks is a rapidly growing field, with almost 9,000 deep learning models (Fawaz et al., 2018; Pons et al., 2017; Faust et al., 2018; Goodfellow et al., 2018), but with limited application to the medical domain (Faust et al., 2018; Yildirim et al., 2018). These deep time-series methods, however, are missing in-process interpretability that explain exactly what the model believes is important. We extend the in-process example-based explainability work of (Li et al., 2017) to consider real-world time-series data and to promote diversity in latent data representation.

On data with unclear class boundaries, in-process methods can misbehave. For example when the model in (Li et al., 2017) is applied to the MNIST dataset, the prototypes easily separate in the latent space because the latent data representation is separable and well-structured (see appendix). However, when class boundaries and features do not form distinguishable clusters, learned prototypes become archetypes (extreme corner cases) that exist near the convex hull of the latent space (Fig. 3). This phenomenon yields prototypes that represent extreme class types and underperform on classifying data in overlapping class regions.

In this work, we provide an explainable method for time-series data while aiming to remedy the formation of archetypes. Our model improves upon the autoencoder framework of Li *et al.*, and introduces a prototype diversity penalty that explicitly accounts for prototype clustering and encourages the model to learn more diverse prototypes that focus on areas of the latent space where class separation is most difficult and least defined. We show the utility of this approach on two dimensional time-series classification in two cases: bradycardia, a slowing of heart rate, events

[*]Equal contribution [1]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA [2]Dell Medical School, The University of Texas at Austin, Austin, TX, USA. Correspondence to: {diegoolano, alangee} <@utexas.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[1] http://www.github.com/diegoolano/icml19-ts-prototypes

from electrocardiogram (ECG) waveforms of preterm infants and spoken digit recognition from audio waveforms. The two-dimensional representation of time-series provides an interpretable method for experts (e.g. clinicians) to understand the evolution of relevant features based on visible phenotypes present in time-series data. To the best of our knowledge this is the first application of prototypes and latent space analysis for health time-series data that could help reveal clinically relevant and explainable phenotypes.

2. Methods

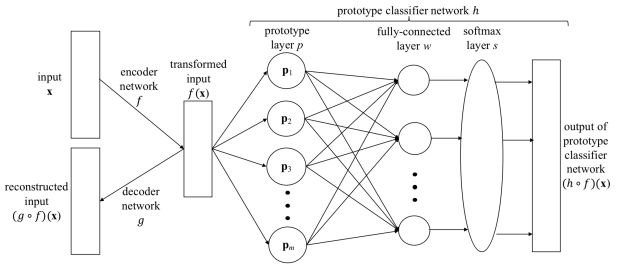

Figure 1. Prototype Architecture from (Li et al., 2017)

2.1. Time-Series Explanation via Prototypes

We adopt the autoencoder-prototype architecture from (Li et al., 2017). Let $\mathcal{X} = (x_i, y_i)_i^n$ be the training set with $x_i \in R^p$ and class labels $y_i \in \{1, ..., K\}$ for each training point $i \in \{1, ..., n\}$. The front-end autoencoder network learns a lower-dimension latent representation of the data with an encoder network, $f : R^p \rightarrow R^q$. The latent space is then projected back to the original dimension using a decoder function, $g : R^q \rightarrow R^p$. The latent representation, $f(x)$ is also passed to a feed-forward prototype network, $h : R^q \rightarrow R^K$, for classification. The prototype network learns m prototype vectors, $p_1, p_2, ..., p_m \in R^q$ using a four-layer fully-connected network over the latent space that learns a probability distribution over the class labels y_i (Fig 1). The learned prototypes can then be decoded using g and examined to infer what the network has learned. The choice of m is determined *a priori*, with larger values allowing for higher throughput and model capacity, but potentially less interpretable prototypes.

We revise the loss function by adding a penalty for learned prototypes that are too close to one another:

$$PDL(p_1,...,p_m) = \frac{1}{log\left(\frac{1}{m}\sum_{j=1}^{m} min_{i>j\in[1,m]} \|p_i - p_j\|_2^2\right)} \quad (1)$$

We calculate the average minimum squared l_2 norm between any two prototypes, p_i, p_j. By applying the inverse log to the prototype distances, we penalize prototypes that are close in distance while making sure the minimum distance between prototypes does not get too large. This prototype diversity loss (PDL) promotes prototype diversity and coverage over the latent space. The updated loss function is:

$$\begin{aligned}\mathcal{L}((f,g,h),X) = &E(h \circ f, X) + \lambda_R\, R(g \circ f, X) \\ &+ \lambda_1\, R_1(p_1, ..., p_m, X) \\ &+ \lambda_2\, R_2(p_1, ..., p_m, X) \\ &+ \lambda_{pd}\, PDL(p_1, ..., p_m)\end{aligned} \quad (2)$$

$$R_1(p_1, ..., p_m, X) = \frac{1}{m}\sum_{j=1}^{m} min_{i\in[1,n]} \|p_j - f(x_i)\|_2^2, \quad (3)$$

$$R_2(p_1, ..., p_m, X) = \frac{1}{n}\sum_{i=1}^{n} min_{j\in[1,m]} \|f(x_i) - p_j\|_2^2. \quad (4)$$

where E is the classification (cross entropy) loss, R is the reconstruction loss of the autoencoder, and R_1 and R_2 are the loss terms that relate the feature vectors to the prototype vectors in latent space (Li et al., 2017).

2.2. Datasets

The neonatal intensive care unit (NICU) dataset is composed of two sources: (1) ECG waveforms from PhysioNet's PICS database (Gee et al., 2017; Goldberger et al., 2000); and (2) ECG waveforms (500 Hz, Intellivue MP450) collected from a preterm infant at Seton Medical Center Austin. Class breakdowns for bradycardia in the ECG signal follow clinical thresholds (Perlman & Volpe, 1985): X_{ECG} = { normal (>100 beats per minute (bpm)): 1039, mild (100-80 bpm): 634, moderate (80-60 bpm): 306, severe (<60 bpm): 132 }. Moderate and severe events were combined into a single class. For full details on pre-processing of the data, please refer to (Gee et al., 2019).

The Free Spoken Digit Dataset (Jackson et al., 2018) consists of 2000 audio clips (8 kHz) of four speakers repeating the digits 0 through 9, 50 times each. Each segment was normalized to zero-mean, unit-variance and clipped for white

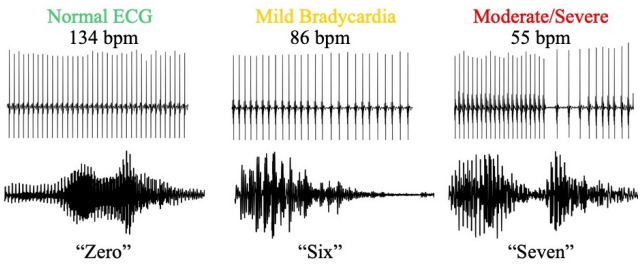

Figure 2. Examples of (1) ECG segments in classification of bradycardia (top), and (2) a speaker (Jackson) saying 0, 6 & 7 (bottom)

space (Fig. 2). This data can be thought of as "spoken MNIST". We perform speaker classification and digit classification within a speaker.

2.3. Prototype Diversity Score

We adopt a version of the group fairness metric presented in (Mehrotra et al., 2018) and refer to it as the prototype diversity score, Ψ:

$$\Psi = \frac{1}{Z}\sum_{i=1}^{t}\sqrt{|\phi_i|} \tag{5}$$

where $\phi_i, i \in \{1, ..., t\}$ is defined for a metric and Z is the normalization constant. For the neighbor diversity metric Ψ_N, ϕ_i is the set of prototypes that have nearest neighbor i and Z is the number of prototypes m. For the class diversity metric Ψ_C, ϕ_i is the set of prototypes that are from class i and Z is the number of classes K. Note, $max(\Psi_D) = 1$.

3. Results

3.1. Classification of ECG with 2-D Prototypes

We observe more diverse prototypes and comparable or better test accuracy with our model 93.1% compared with 92.1% from the baseline model in (Li et al., 2017) (Table 1). Both models perform well on the classification of the normal class, as expected since normal waveforms have near-constant phase. Both models have difficulty separating between the mild and moderate/severe classes, often confusing the classification between the two (see appendix). This behavior is expected since data near these two class boundaries are difficult to discern, even for domain experts, due to events existing in both classes with possible subtle time differences in cardiac firing. Nonetheless, we find that

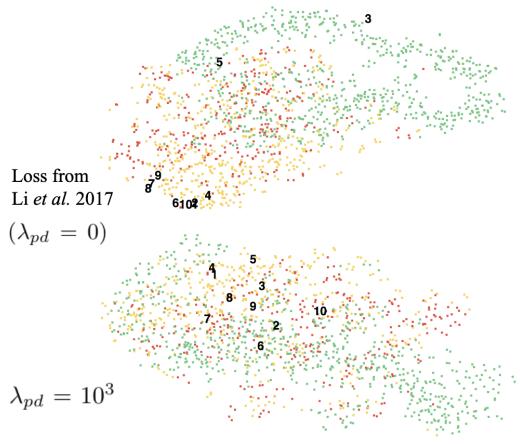

Figure 3. Effect of loss regularization on the latent space and spread of prototypes for the NICU classification task using 10 prototypes with $\lambda_{pd} = 0$ (baseline) and $\lambda_{pd} = 10^3$.

	ECG: Bradycardia		
λ_{pd}	Accu.	Ψ_N	Ψ_C
0	$92.1 \pm 0.1\%$	0.83 ± 0.04	0.78 ± 0.19
500	$92.7 \pm 1.0\ \%$	0.86 ± 0.07	0.89 ± 0.19
$1e3$	$92.4 \pm 1.3\%$	0.87 ± 0.11	0.89 ± 0.19
$2e3$	$\mathbf{93.1 \pm 0.4\%}$	$\mathbf{0.90 \pm 0.04}$	$\mathbf{1.00 \pm 0.00}$

Table 1. Diversity score for neighbors Ψ_N and class Ψ_C. Our model, $\lambda_{pd} > 0$, returns better accuracies and diversity scores (bolded) than the baseline model (row $\lambda_{pd} = 0$). (Model details: *3-class, 10-prototypes, learning rate = 0.002*).

the addition of a prototype diversity loss performs at least, if not better, than the baseline model.

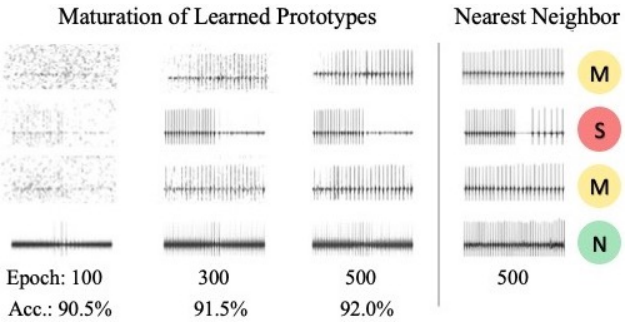

Figure 4. Prototype evolution with in-process explainability over training time. High level features are easily learned in early epochs of training, while more complex features are developed over time. The final nearest neighbors are depicted on the right. The prototypes correspond to a subset of the $\lambda_{pd} = 10^3$ latent space cloud in Figure 3. *Model details: NICU data, 3-class, 10-prototypes*.

Because prototypes are generated during training, we infer features that the algorithm utilized to classify waveforms at different points during training (Fig 4). For example, at epoch 100, we see that some of the prototypes exhibit global morphological features of the normal waveform class after random initialization at epoch 0.

As training progresses, we observe other complex phenotypes emerging: one prototype learns that large gaps in cardiac firings are important for identifying severe cases and another prototype learns the consistent pattern of spikes are important for mild cases. Since the mild class shares mixed features of both normal and positive events, it is not surprising that more prototypes are needed in this class to learn subtleties of the class features (see appendix). Thus, prototypes highlight waveform structures that the algorithm deemed as important when trying to learn the classification of bradycardia. This finding aligns with the idea of clinicians using visible features present in a bradycardia (i.e. the increasing distance between QRS complexes) to decide

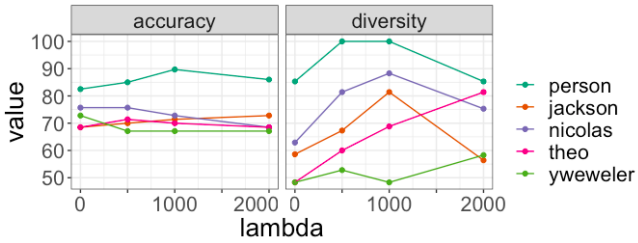

Figure 5. Accuracy and diversity metrics for the NICU and spoken digits experiments. The latter is divided into "person" detection and digit detection within each person in the dataset.

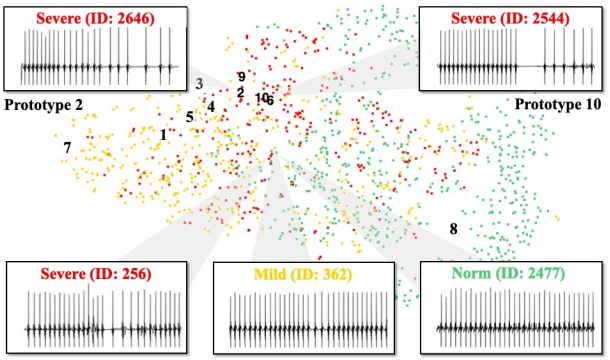

Figure 6. Learned prototypes showcase the diversity of features that are important for understanding ECG morphology while classifying bradycardia events. *(10-prototypes,* $\lambda_{pd} = 10^4$).

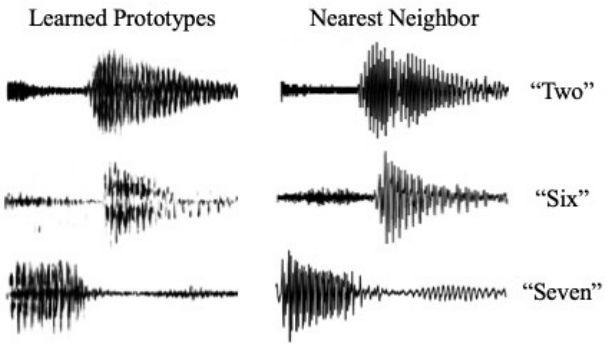

Figure 7. Learned prototypes from audio waveforms of spoken digits by Nicolas from the FSDD *(* $\lambda_{pd} = 500$*).*

whether or not a bradycardia exists in an image.

We compare the latent space of (Li et al., 2017) to the latent space of our model with prototype diversity loss via t-SNE projections, where proximity in 2-D space suggests that the points are "close" in distance in the original latent space. We represent the learned prototypes by mapping each prototype to it's nearest neighbor (Fig 3). We find that by increasing our loss term, PDL, our model increases the local coverage of the prototypes compared with the baseline model (i.e. $\lambda_{pd} = 0$). However, if we regularize our loss term too much (i.e. $\lambda_{pd} > 10^4$), we begin to introduce clustering of prototypes and diversity suffers. Thus with the additional prototype distance penalty, we are able to achieve higher diversity scores and classification accuracies for various hyperparameter settings (Fig 5).

3.2. Case Study with Prototypes: Exploring ECG Morphology and Bradycardia Classification.

We observe that ECG events in a local neighborhood share similar QRS complex morphology, despite having different class labels and cardiac firing periods (Fig. 6). We also observe that the algorithm distinctly separates features within the moderate/severe class that were important in classification (i.e. prototypes 2 and 10 shown in Fig 6). These results suggest that there are physiologic dependencies (i.e. clustering based on cardiac morphology and function) can be learned using our model to investigate physiological phenomena, and possibly applied to other clinical areas, like cardiac ischemia or apnea of prematurity in respiration - both exhibit visible, abnormal waveform behavior.

3.3. Spoken Digits Classification and Analysis

We assess our model on high-frequency audio waveforms of spoken digits (FSDD) as 2-D images for 4 class speaker and 10 digit classification tasks with 4 and 10 prototypes, respectively. The waveform envelope and syllables of these spoken digits are discernible to the eye (see *"six" and "seven"* in Fig 2) and, as such, make good candidates for our image-based explainability model. Experiments show that by varying regularization of the prototype diversity penalty, we observe slightly better or similar accuracies when compared to the baseline model (Fig. 5). With a fine-tuned λ_{pd} we can increase diversity of the prototypes and correspondingly see improved accuracy and data coverage (see appendix). For example, $\lambda_{pd} = 500$ gives a higher diversity score across all tasks, indicating prototypes with more unique nearest neighbors as compared with the baseline model (Fig 5).

Experiments show that increasing the depth of the network and fine-tuning the learning rate lead to both increased accuracy and diversity over all tasks. Similarly, recent data augmentation techniques in medical (Bahadori & Lipton, 2019) and speech recognition (Park et al., 2019) domains could help further improve performance. The purpose of this work, however, is not to obtain the best performance on theses tasks, but rather to show the utility of learned prototypes as faithful explainations of what a model is using to make decisions. We demonstrate some of the learned prototypes in Fig. 7, which shows representations the model finds useful in classifying digits for a given speaker.

Acknowledgement

The authors would like to thank Sinead Williamson and the reviewers for helpful feedback and critical reviews.

References

Bahadori, M. T. and Lipton, Z. C. Temporal-clustering invariance in irregular healthcare time series. *arXiv preprint arXiv:1904.12206*, 2019.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 1721–1730, New York, NY, USA, 2015. ACM. doi: 10.1145/2783258.2788613. URL http://doi.acm.org/10.1145/2783258.2788613.

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., and Acharya, U. R. Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine*, 161:1 – 13, 2018. doi: https://doi.org/10.1016/j.cmpb.2018.04.005.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P. Deep learning for time series classification: a review. *CoRR*, abs/1809.04356, 2018. URL http://arxiv.org/abs/1809.04356.

Gee, A. H., Barbieri, R., Paydarfar, D., and Indic, P. Predicting bradycardia in preterm infants using point process analysis of heart rate. *IEEE Transactions on Biomedical Engineering*, 64(9):2300–2308, 2017. doi: 10.1109/TBME.2016.2632746.

Gee, A. H., García-Olano, D., Ghosh, J., and Paydarfar, D. Explaining deep classification of time-series data with learned prototypes. *CoRR*, abs/1904.08935, 2019. URL http://arxiv.org/abs/1904.08935.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, June 2000. ISSN 0009-7322. doi: 10.1161/01.CIR.101.23.e215. URL http://circ.ahajournals.org/content/101/23/e215.

Goodfellow, S., Goodwin, A., Eytan, D., Greer, R., Mazwi, M., and Laussen, P. Towards understanding ecg rhythm classification using convolutional neural networks and attention mappings. In *Proceedings of Machine Learning for Healthcare*, MLHC '18, pp. 2243–2251, 08 2018. URL http://doi.acm.org/10.1145/3269206.3272027.

Jackson, Z., Souza, C., Flaks, Jason; Pan, Y., Nicolas, H., and Thite, A. Free spoken digit dataset (fsdd). 2018. doi: 10.5281/zenodo.1342401. URL https://github.com/Jakobovski/free-spoken-digit-dataset.

Li, O., Liu, H., Chen, C., and Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. *CoRR*, abs/1710.04806, 2017. URL http://arxiv.org/abs/1710.04806.

Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., and Diaz, F. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pp. 2243–2251, 2018. URL http://doi.acm.org/10.1145/3269206.3272027.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Perlman, J. M. and Volpe, J. J. Episodes of apnea and bradycardia in the preterm newborn: Impact on cerebral circulation. *Pediatrics*, 76(3): 333–338, 1985. URL https://pediatrics.aappublications.org/content/76/3/333.

Pons, J., Nieto, O., Prockup, M., Schmidt, E. M., Ehmann, A. F., and Serra, X. End-to-end learning for music audio tagging at scale. *CoRR*, abs/1711.02520, 2017. URL http://arxiv.org/abs/1711.02520.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL http://arxiv.org/abs/1602.04938.

Rudin, C. Please stop explaining black box models for high stakes decisions. *CoRR*, abs/1811.10154, 11 2018. URL https://arxiv.org/abs/1811.10154.

Yildirim, O., Plawiak, P., Tan, R.-S., and Acharya, U. R. Arrhythmia detection using deep convolutional neural network with long duration ecg signals. *Computers in Biology and Medicine*, 102:411 – 420, 2018.

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015. URL http://arxiv.org/abs/1512.04150.