On the Application of Convolutional Neural Networks for 12-lead ECG Multi-label Classification Using Datasets From Multiple Centers

Davide Borra[1], Alice Andalò[1], Stefano Severi[1], Cristiana Corsi[1]

[1]DEI, Campus of Cesena, University of Bologna, Cesena, Italy

Abstract

Cardiac arrhythmia is a group of conditions in which falls changes in the heartbeat. Electrocardiography (ECG) is the most common tool used to identify a pathology in the cardiac electrical conduction system. ECG analysis is usually manually performed by an expert physician. However, manual interpretation is time-consuming and challenging even for cardiologists. Many automatic algorithms relying on handcrafted features and traditional machine learning classifiers were developed to recognize cardiac diseases. However, a large a priori knowledge about ECG signals is exploited. To overcome this main limitation and provide higher performance, recently, deep neural networks were designed and applied for 12-lead ECG classification. In this study, we designed decoding workflows based on three state-of-the-art architectures for time series classification. These were InceptionTime, ResNet and XResNet. Experiments were conducted using the training datasets provided during the PhysioNet/Computing in Cardiology Challenge 2020. The best-performing algorithm was based on InceptionTime, scoring a training 5-fold cross-validation challenge metric of 0.5183±0.0016, while using a low number of parameters (510491 in total). Thus, this algorithm provided the best compromise between performance and complexity.

1. Introduction

Cardiovascular diseases are the main cause of death, responsible for the 31% of the worldwide deaths in 2016 [1] and the electrocardiogram (ECG) is the most common used tool in their diagnoses by non-invasively recording the electrical activity of the heart. Twelve-lead ECG describes the activity from 12 sites, each lead containing features potentially related to a specific arrhythmia. ECG analysis is usually manually performed by an expert physician. However, manual interpretation is time-consuming and challenging [2] and these difficulties promoted the development of automatic ECG interpretation algorithms.

Automatic algorithms relying on handcrafted features (e.g. statistical features, frequency-domain features, time-domain features) and traditional machine learning classifiers were developed to recognize cardiac diseases. However, a large a priori knowledge about characteristics of the signals is exploited and separate feature extraction, selection and classification steps are performed. Lastly, these classic algorithms are limited in performance [3], precluding their usage as a standalone diagnostic tool.

Recently, deep neural networks (DNNs) were designed and applied to electrophysiological signals [4-7]. This provided an end-to-end framework where the most relevant features are automatically learned directly from raw/lightly pre-processed data without separately perform feature extraction and classification. When applied to ECG signals [8-10], this enabled statements that resulted highly difficult to make even for cardiologists [11]. Among DNNs, recently Strodthoff et al. [10] reported outstanding results using deep convolutional neural networks (CNNs), such as InceptionTime [12], ResNet [13] and XResNets [14], on a large public benchmark dataset.

In occasion of the PhysioNet/Computing in Cardiology Challenge 2020 [15], challengers had to build an algorithm to automatically identify the cardiac abnormality / abnormalities among 27 conditions (see Table 1 for the full list of the statements) from 12-lead ECG recordings across 6 different datasets.

To this aim, we participated to the competition as CardioUniBo team and implemented decoding workflows based on InceptionTime, ResNet and XResNets, evaluating their performance on the target decoding task.

2. Methods

In the following sections, the datasets, pre-processing procedure, and decoding algorithms are described. Experiments were GPU-accelerated via a NVIDIA Titan V and PyTorch was used as framework to build and solve the optimization of the CNNs.

2.1. Datasets

Among the 6 datasets provided by the organizers of the PhysioNet/Computing in Cardiology Challenge 2020, we considered signals from the China Physiological Signal Challenge in 2018 sets named CPSC, CPSC-EXTRA (6877+3453 examples) [16], from the Physikalisch Technische Bundesanstalt (PTB) set named PTB-XL (21837 examples) [17] and from the unique set recorded in the Southeastern United States named GEORGIA (10344 examples).

These datasets were selected because resulted the most representative for the target 27 diagnosis and exhibits comparable recording lengths (CPSC and CPSC-EXTRA from 6 to 60 s, while PTB-XL and GEORGIA of 10 s). Each example can be associated with a one or more statements (multi-label classification). The statement distributions of these datasets are reported in Table 1.

Table 1: Statement distributions in the datasets used in the performed experiments.

Statement	CPSC + CPSC-EXTRA	PTB-XL	GEORGIA
1st degree av block	828	797	769
Atrial fibrillation	1374	1514	570
Atrial flutter	54	73	186
Bradycardia	271	0	6
Complete right bundle branch block	113	542	28
Incomplete right bundle branch block	86	1118	407
Left anterior fascicular block	0	1626	180
Left axis deviation	0	5146	940
Left bundle branch block	274	536	231
Low qrs voltages	0	182	374
Nonspecific intraventricular conduction disorder	4	789	203
Pacing rhythm	3	296	0
Premature atrial contraction	689	398	639
Premature ventricular contractions	188	0	0
Prolonged pr interval	0	340	0
Prolonged qt interval	4	118	1391
Qwave abnormal	1	548	464
Right axis deviation	1	343	83
Right bundle branch block	1858	0	542
Sinus arrhythmia	11	772	455
Sinus bradycardia	45	637	1677
Sinus rhythm	922	18092	1752
Sinus tachycardia	303	826	1261
Supraventricular premature beats	53	157	1
T wave abnormal	22	2345	2306
T wave inversion	5	294	812
Ventricular premature beats	8	0	357

2.2. Pre-processing

Twelve-lead ECG signals were collected, and the gain factor was divided for each lead. Then, a zero-phase 2nd order band-pass Butterworth filter was applied between 0.5 and 40 Hz. Lastly, signals were downsampled to 100 Hz to reduce the computational cost and each lead signal was standardized.

The downsampled signals resulting from the 4 datasets were augmented by adopting 2 different procedures. At first, we extracted consecutive 2.5 s chunks from each 12-lead ECG signal without overlap (resulting in an offline augmentation). Therefore, these chunks of 12-lead ECG represented the CNN input with shape of (12,250). This also allowed to keep limited the input time dimension and, thus, keep controlled the number of trainable parameters in the convolutional-to-dense transition. Furthermore, due to the presence of 12-lead ECGs with some leads set to zero in the datasets, during the optimization of the decoders, each lead signal was randomly set to zero with a probability of 0.5 (online augmentation) to give more robustness to the algorithms when handling 12-lead ECGs with zeroed derivations.

Once pre-processed, the datasets were merged together in a multicenter dataset (42511 total examples) with single or multiple statements associated to each 12-lead ECG data. Five-fold stratified cross-validation was performed and, since our team was unable to obtain the scores on the test set during the official phase of the challenge, the evaluation metrics reported in this study correspond to training cross-validation scores.

2.2. Decoding

To solve the objective decoding task, InceptionTime [12], a ResNet-based architecture proposed by Wang et al. [13] specifically for time series classification (here referred as "ResNet" for simplicity) and XResNets [14] with 18, 34, 50, 101 layers, were tested. For each of these architecture designs, the hyper-parameters of the convolutional module were set as in the original papers. A kernel size of 5 was used in the 1-D convolutions in ResNet and XResNets, while kernel sizes of 40, 20 and 10 were used in each Inception block (6 in total) of InceptionTime as in the original paper [12].

On top of the convolutional module a concat-pooling layer was used (concatenation of the output obtained with global average and max poolings) as was done in Strodthoff et al. [10]. In addition, all the re-implemented CNN architectures shared the same dense module, implemented as a first fully-connected layer with 128 units followed by batch normalization, ReLU non-linearity and dropout ($p = 0.5$), and a second fully-connected layer with 27 units (output layer) activated with sigmoid functions. The total number of trainable parameters introduced was 510491, 476187, 697499, 1356955, 1808795, 3645339, respectively for InceptionTime, ResNet and XResNets with 18, 34, 50, 101 layers.

Cross-entropy was used as loss function and Adam as optimizer with a learning rate of 1e-3 and a batch size of 32. While training the architectures with a maximum number of 100 epochs, early stopping was performed (setting the validation set as the 20% of the training set) using the validation loss as stop metric.

Once the trainings ended, for each 12-lead ECG recording the chunk-level probabilities were averaged together for the chunks belonging to the same recording, obtaining recording-level probabilities (from crop-level to recording-level probabilities). Then, these probabilities were binarized into the predicted statement/statements using a threshold of 0.5 for each output neuron. When the prediction was empty (i.e. no output probability exceeded the threshold set to 0.5), the most probable statement was selected.

To evaluate the CNNs, we computed the metric specifically designed for this challenge (here denoted as "challenge metric") [15] and the Area Under the Receiver Operating Characteristics (AUROC).

3. Results

In Table 2 the training cross-validation scores (challenge metric and AUROC) obtained in the experiments with the deep CNNs are reported.

Table 2. Training cross-validation scores (metrics scored adopting a cross-validation scheme on the training sets released for the competition) obtained in the performed experiments.

Architecture	Challenge Metric (mean±std)	AUROC (mean±std)
InceptionTime	0.5183±0.0016	0.9391±0.0017
ResNet	0.5091±0.0071	0.9400±0.0015
XResNet1d (18)	0.5127±0.0051	0.9335±0.0031
XResNet1d (34)	0.5085±0.0051	0.9307±0.0026
XResNet1d (50)	0.5180±0.0042	0.9352±0.0037
XResNet1d (101)	0.5140±0.0047	0.9334±0.0022

All the experiments were conducted on the training datasets released, obtaining an average training cross-validation challenge metric above 0.50 in all cases. The obtained training cross-validated challenge scores were similar across the architectures, scoring a higher average value with InceptionTime.

4. Discussion and Conclusion

In this study, we designed decoding workflows based on three state-of-the-art CNN designs for time series classification and applied them to the multicenter ECG signals provided for the PhysioNet/Computing in Cardiology Challenge 2020.

From our experiments, InceptionTime resulted the architecture with higher training cross-validated challenge metric (on average 0.5183) and with less variability, while ResNet scored higher training cross-validation AUROC (on average 0.9400). Together with ResNet, InceptionTime introduced a lower total number of trainable parameters (510491 in total) compared to the other designs, resulting in a more parsimonious neural network. Indeed, comparable performance to InceptionTime were achieved only with XResNet1d (50) using more than 2x trainable parameters. Therefore, InceptionTime resulted the best compromise between performance and complexity, which could be useful to design models less prone to overfit less-represented conditions (e.g. in particular the condition "Premature ventricular contractions" represented only with 188 examples).

Despite these promising results obtained using CNNs on multicenter ECG signals, these algorithms suffer from limited interpretability of the learned features. Future developments could include designing more explicable architectures by adding interpretable layers [6] (layers that, once trained, allow a direct interpretation of the learned features) and designing even more optimized (in terms of trainable parameters) architectures than InceptionTime.

Acknowledgments

We gratefully acknowledge the support of NVIDIA

Corporation with the donation of the TITAN V used for this research.

References

[1] E. J. Benjamin, S. S. Virani, C. W. Callaway, et al. Heart Disease and Stroke Statistics-2018 Update: A Report From the American Heart Association. Circulation 2018; 137 (12):e67–e492.

[2] S. M. Salerno, P. C. Alguire, H. S. Waxman. Competency in Interpretation of 12-lead Electrocardiograms: A Summary and Appraisal of Published Evidence. Ann Intern Med 2003; 138(9):751-60.

[3] A. P. Shah, S. A. Rubin. Errors in the Computerized Electrocardiogram Interpretation of Cardiac Rhythm. J. Electrocardiol 2007; 40: 385–390.

[4] O. Faust, Y. Hagiwara, T. J. Hong, et al. Deep learning for Healthcare Applications Based on Physiological Signals: A Review. Computer Methods and Programs in Biomedicine 2018; 161:1-13.

[5] D. Borra, S. Fantozzi, E. Magosso. Interpretable and Lightweight Convolutional Neural Network for EEG Decoding: Application to Movement Execution and Imagination. Neural Networks 2020; 129:55-74.

[6] M. Simoes, D. Borra, E. Santamaria-Vázquez, et al. BCIAUT-P300: A Multi-Session and Multi-Subject Benchmark Dataset on Autism for P300-based Brain-Computer-Interfaces. Frontiers in Neuroscience 2020; 14.

[7] S. Hong, Y. Zhou, J. Shang. Opportunities and Challenges of Deep Learning Methods for Electrocardiogram Data: A Systematic Review. Computers in Biology and Medicine 2020; 122: 103801.

[8] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, et al. Cardiologist-level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. Nature Medicine 2019; 25:65-69.

[9] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão. Automatic Diagnosis of the 12-lead ECG Using a Deep Neural Network. Nature Communications 2020. 11(1): 1-9.

[10] N. Strodthoff, P. Wagner, T. Schaeffter. Deep Learning for ECG Analysis: Benchmarks and Insights From PTB-XL. arXiv preprint 2020: 2004.13701.

[11] Z. I. Attia, P. A. Friedman, P. A. Noseworthy, et al. Age and Sex Estimation Using Artificial Intelligence From Standard 12-lead ECGs. Circulation: Arrhythmia and Electrophysiology 2019; 12(9): e007284.

[12] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, et al. InceptionTime: Finding AlexNet for Time Series Classification. arXiv preprint 2019: 1909.04939.

[13] Z. Wang, W. Yan, T. Oates. Time Series Classification From Scratch with Deep Neural Networks: A Strong Baseline. arXiv preprint 2016: 1611.06455.

[14] T. He, Z. Zhang, H. Zhang, et al. Bag of Tricks for Image Classification with Convolutional Neural Networks. arXiv preprint 2018: 1812.01187.

[15] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, Liu F, Rad BA, Elola A, Seyedi S, Li Q, Sharma A, Clifford GD, Reyna MA. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiol. Meas. 2020 (Under Review).

[16] F. Liu, C. Liu, L. Zhao, et al. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. J Med Imaging Health Inform 2018; 8:1368–1373.

[17] P. Wagner, N. Strodthoff, R. Bousseljot, et al. PTB-XL, a Large Publicly Available Electrocardiography Dataset. Nature Scientific Data 2020; 7:154

Address for correspondence:

Davide Borra
DEI, University of Bologna
Via dell'Università 50, 47521, Cesena, Italy
E-mail: davide.borra2@unibo.it