ECG Abnormalities Recognition Using Convolutional Network With Global Skip Connections and Custom Loss Function

Tomas Vicar[1], Jakub Hejc[1,2], Petra Novotna[1], Marina Ronzhina[1], Oto Janousek [1]

[1] Department of Biomedical Engineering, Faculty of Electrical Engineering and Communications, Brno University of Technology, Brno, Czech Republic
[2] Department of Pediatric, Children's Hospital, University Hospital Brno, Brno, Czech Republic

Abstract

The latest trends in clinical care and telemedicine suggest a demand for a reliable automated electrocardiogram (ECG) signal classification methods. In this paper, we present customized deep learning model for ECG classification as a part of the Physionet/CinC Challenge 2020. The method is based on modified ResNet type convolutional neural network and is capable to automatically recognize 24 cardiac abnormalities using 12-lead ECG. We have adopted several preprocessing and learning techniques including custom tailored loss function, dedicated classification layer and Bayesian threshold optimization which have major positive impact on the model performance. At the official phase of the Challenge, our team - BUTTeam - reached a challenge validation score of 0.696, and the full test score of 0.202, placing us 21 out of 40 in the official ranking. This implies that our method performed well on data from the same source (reached first place with validation score), however, it has very poor generalization to data from different sources.

1. Introduction

A large number of automated ECG classification methods have been reported over last decade, most of which have only been evaluated on small or homogeneous datasets. In this paper, we present deep learning model for ECG classification as a part of the Physionet/CinC Challenge 2020 [1, 2]. The Challenge data consists of 12-lead ECGs from wide range of sources and recording platforms including signals of low quality or highly variable length. As another challenging task, the data contains various numbers of reported non-exclusive abnormalities obtained with inhomogeneous annotation methods.

The aim of the model is to classify these signals in a multi-label manner into one or more of 24 given classes. Major improvements of the method were achieved by: (1) customized convolutional neural network architecture with local and global skip connections, (2) data augmentation, (3) custom loss function based on the challenge metric, and (4) class specific threshold optimization.

2. Material and Methods

2.1. Data

Training dataset is composed of 43,101 labeled recordings from 6 different sources [1, 2]. Recordings are sampled with various sampling frequencies (257, 500 or 1000 Hz) and resolution settings. The dataset includes 24 scored pathologies and it consists of data from 6 different databases (see [2] for more details).

2.2. Data preprocessing

Preprocessing pipeline consisted of time-domain resampling with the fixed sampling frequency 125 Hz (linear interpolation combined with decimation and anti-aliasing FIR filter, if needed) for both training phase and inference. Additionally, augmentation tasks such as $\pm 10\%$ stretch along temporal axis (p=0.8) or $\pm 20\%$ amplification along voltage axis (p=0.8) were randomly applied during training to prevent the model from overfitting. Long-term Holter recordings are due to possible memory issues only allowed to pass through the model during the inference. Then, divided into smaller segments, they are fed into the model as a variable-sized batch. Non-redundant binary encoded labels from the entire batch are then joined together into a single vector.

2.3. Model Architecture

In our work, we have adopted a Convolutional Neural Network (CNN) architecture based on a residual neural network (ResNet) [3], which is simple yet proven design for image multi-label classification tasks. Original 2D convolutional filters were replaced by its 1D equivalents. In order to address vanishing gradient problem efficiently,

 ISSN: 2325-887X DOI: 10.22489/CinC.2020.189

skip connections across residual layers were extended by global skip connection [4] providing a direct shortcut to the model input layer (Input Gate in Figure 1). ResNet based feature extractor consists of 6 residual blocks each of which contains 3 convolutional layers ($k = 3$). Since the data can be assigned into c non-exclusive classes, final classification layer is composed of $c = 24$ independent fully connected binary classifiers. This structure allows arbitrary combination of class labels [5] and, based on its principle, can be referred to as a Binary Units Training Technique (BUTT). Implementation details of proposed model architecture are depicted in Figure 1.

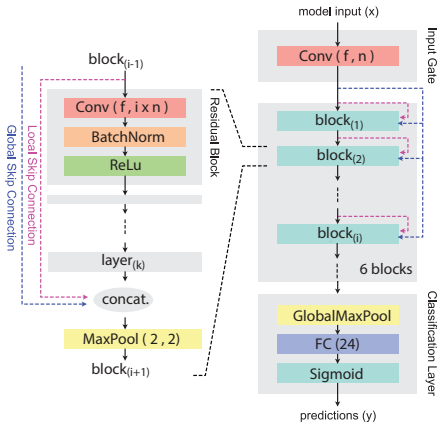

Figure 1: Architecture of proposed model. i – block number of ResNet; k – number of layer in residual block; n – number of filters in first layer

In order to use a mini-batch optimization, collected data samples were padded by zeros to equal length. This procedure may adversely affect extracted features if global or adaptive pooling are applied over temporal axis. To address this problem the zero-padded equivalent is cut out just after leaving ResNet layers. Global Max Pooling is then performed over temporal dimension and pooled tensor is fed into BUTT classifier.

2.4. Loss Function

Cross-entropy is commonly used as a loss function for classification tasks [6]. In the case of class imbalanced data, weighted variant of cross-entropy (WCE) or generalised dice loss function [7] can be introduced instead to provide better stability of a model during an onset phase of training. In multi-label classification problems WCE can be defined as follows:

$$WCE = -\sum_{c} w_c^+ t_c \log(p_c) + w_c^- (1 - t_c) \log(1 - p_c), \tag{1}$$

where c are individual classes, $t_c \in \{0, 1\}$ are binary encoded labels, $p_c \in [0, 1]$ are model output scores, and w_c^+ and w_c^- are weights for positive and negative classes, respectively.

Weights are inversely proportional to its frequency in the training dataset, specifically, $w_c^+ = N/N_c^+$ and $w_c^- = N/N_c^- = N/(N - N_c^+)$. N is the total number of samples, N_c^+ and N_c^- are the numbers of positive and negative samples within given class.

2.5. Custom Loss Function

The official Challenge Metric (CM) is based on modified confusion matrix $\mathbf{A}$, where each entry a_{ij} (confusion) is weighted according to given clinical importance w_{ij}:

$$CM = \sum_{ij} a_{ij} w_{ij}. \tag{2}$$

On this basis, we have derived a custom loss function, which includes the same clinical importance measures. In a multi-label case, modified confusion matrix $\mathbf{A}$ can be expressed by a matrix multiplication:

$$\mathbf{A} = \mathbf{L}^{\mathbf{T}} (\mathbf{R} \oslash \mathbf{N}), \tag{3}$$

where operator $\oslash$ represents point-wise division, $\mathbf{L}$ and $\mathbf{R}$ are $N \times c$ binary matrices, formed by one-hot encoded ground truth labels and thresholded model output scores, respectively. $\mathbf{N}$ is $N \times c$ normalizing matrix factoring in number of unique labels for each sample from $\mathbf{L}$ and $\mathbf{R}$ according to:

$$\mathbf{N} = ((\mathbf{L} | \mathbf{R}) \mathbf{1_{c \times 1}}) \mathbf{1_{1 \times c}}, \tag{4}$$

where $\mathbf{1_{c \times 1}}$ is $c \times 1$ all-ones matrix and symbol $|$ stands for bit-wise binary OR operator. Operator OR as well as binary matrix R can now be replaced with its continuous equivalent using raw output scores p_c as follows:

$$\mathbf{N} = ((\mathbf{L} + \mathbf{R} - \mathbf{L} \odot \mathbf{R})) \mathbf{1_{c \times 1}}) \mathbf{1_{1 \times c}}, \tag{5}$$

where operator $\odot$ represents point-wise multiplication. This makes the metric differentiable, thus it can be used as a loss function.

2.6. Class Specific Threshold Optimization

Mapping a raw score of class membership onto a class label with the default threshold value of 0.5 does not always guarantee the best model performance with respect to a given metric. Thus, proposed model transforms raw output scores using a set of class-specific thresholds τ_c. Each τ_c was estimated via Python implementation of Bayesian Optimization [8] with respect to CM loss function. Optimization was performed on a validation part of the dataset.

2.7. Model Training

Weights and biases of convolutional layers were initialized with Xavier [9] and constant (c=0) initialization, respectively. Model training was performed by Adam optimizer ($\beta_1 = 0.9$; $\beta_2 = 0.999$) [10] with decoupled weight decay regularization ($\lambda = 10^{-5}$) [11], modified learning rate schedule ($\alpha_0 = 10^{-3}$) and mini-batch size of 32. Learning rate schedule consisted of two cycles with decaying learning rate strategy $0.1\alpha_0$ every 30 epochs in each cycle. During the first cycle, the model was trained with the WCE, while in the second cycle custom CM Loss was used to retrain the model.

To bridge the generalization gap, we have adopted a Stochastic Weight Averaging algorithm [12] which captures model weights w_m in the end of every epoch and then sets new model with weights w_{SWA} as a running average of w_m from the last m captured models.

2.8. Ensemble Modeling

Combination of multiple models can decrease variance and may produce a more generalized output [13]. Final class label is thus given by a majority vote of 3 bootstrap aggregated ensembles, each fitted to a 90 % subset of the training data. The number of ensembles has been chosen to meet the computational limits of the Challenge.

3. Results and Discussion

Model performance was internally evaluated with a hold-out method using 10 % randomly sampled subset of the training data. Final evaluation has been performed on the Challenge hidden datasets - which consist of 3 databases. A comparison of a base model with various hyper-parameter setting using the Challenge Metric (CM) and the Dice Coefficient (DC) [14] is listed in Table 1.

Table 1: The results of models with various hyper-parameters and without (w/o) applied customization on the hold-out training subset (n - number of filters in first layer ($i \times n$ for $i-th$ layer) ; k – number of convolutional layers in block; f - filter size).

Model	CM	DC
Small model ($n = 12$; $k = 3$, $f = 7$)	0.665	0.542
Deep model ($n = 12$, $k = 12$, $f = 7$)	0.667	0.544
Wide model ($n = 48$**,** $k = 3$**,** $f = 7$**)**	**0.687**	**0.571**
Smaller filters ($n = 48$, $k = 3$, $f = 3$)	0.679	0.570
Larger filters ($n = 48$, $k = 3$, $f = 11$)	0.677	0.558
w/o augmentation (Wide model)	0.646	0.559
w/o CM Loss (Wide model)	0.636	0.552
w/o global skip con. (Wide model)	0.679	0.563

The best results were reached with the "Wide" model in which the number of convolutional filters (kernel size = 7) were preferred to a model depth. Model widening was performed by simply increasing the number of filters within each convolutional layer. This has led to the largest improvement of the DC by 0.03. Any further increment in the model depth by adding more convolution layers did not improve the model performance as well as either reduction or expansion of filter size.

Table 2: The results of the best performing model on the hold-out training subset and the hidden datasets.

Data	CM	DC
Training subset	0.687	0.571
Validation data	0.696	0.522
Database 1	0.892	0.245
Database 2	0.235	0.259
Database 3	0.104	0.251
Full Test Set	0.202	0.277

Another improvement has been achieved by using the CM Loss function. Efficiency of the CM Loss is greatly dependent on mini-batch size and has been proven to cause an unstable learning (from scratch) of the model. However, when used as a secondary loss function during cycling learning rate schedule, it has led to an increase of DC by 0.02. Data augmentation strategies mentioned previously showed an increment in DC by 0.01. On the other hand, some of preprocessing methods (e.g. baseline drift removal, temporal shifting, noise addition or simulation of electrode switch) have had no effect neither on the model performance nor on generalization capability most likely due to a sufficient diversity of the dataset.

Classification results for individual classes are shown in Figure 2. The best performance (DC > 0.8) has been reached for AF, SNR and PR. DC > 0.6 was obtained in more than 45 % of classes. It should be noted that the model performance is still substandard in some of well distinguishable pathologies such as PVC, PAC, Brady, etc. This may be primarily caused by low occurrence of those classes in the dataset, and also by a presence of annotation inconsistencies (typically confusion of PVC for PAC).

Results of the best performing model on training hold-out and hidden sets are listed in Table 2. The results from ongoing leaderboard are referred as validation results, which data contained hidden subset from training databases. The final full test set contains 3 databases, where Database 1 contains data from the same source as one of the training databases, however, Database 2 and Database 3 contain data from a different source, thus they are more challenging and require a good generalization of the model. As can be seen in Table 2, our model performs well on data from the same source as training data (Validation data and Database 1), where it performs best out of 40 challenge teams, however, it fails on data from different

sources (Database 2 and Database 3). For this reason, our overall results on full test set reached 0.202 of challenge metric, which placed us 21 out of 40 teams in the official ranking.

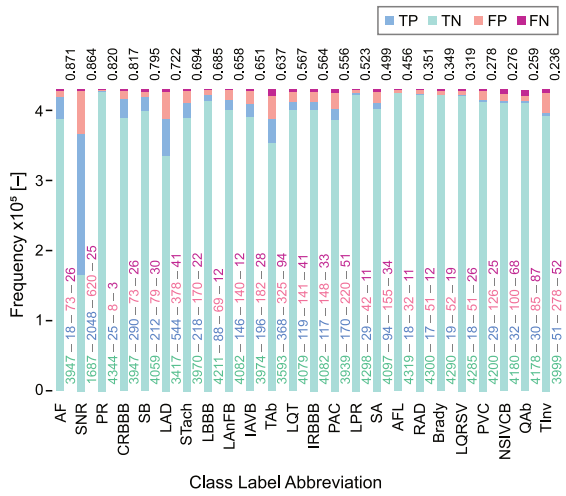

Figure 2: The results for individual classes. Total counts of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) cases are given on the right side of each bar. Individual DC is given at the top of the related bar. Abbreviation for pathologies are listed in [2]. Evaluation is done on hold-out training subset.

4. Conclusions

Our ResNet based CNN architecture with BUTT as dedicated classification layer and custom CM Loss function has met the main aim of the Challenge and is capable to automatically recognize 24 of given pathologies. At the official phase of the Challenge, our team - BUTTeam - reached a challenge validation score of 0.696, and the full test score of 0.202, placing us 21 out of 40 in the official ranking. This implies that our method performed well on data from the same source (reached first place with validation score), however, it has very poor generalization to data from different sources. The code is available at `https://github.com/tomasvicar/BUTTeam`.

References

[1] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 2000;101(23):215–220.

[2] Alday EAP, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, Liu F, Rad AB, Elola A, Seyedi S, Li Q, Sharma A, Clifford GD, Reyna MA. Classification of 12-lead ECGs: the Physionet/Computing in Cardiology Challenge 2020. Physionet. Physiological Measurement 2020;(In Press).

[3] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; 770–778.

[4] Vicar T, Novotna P, Hejc J, Ronzhina M, Smisek R. Sepsis Detection in Sparse Clinical Data Using Long Short-Term Memory Network with Dice Loss. In 2019 Computing in Cardiology (CinC). IEEE, 2019; 1–4.

[5] Zhang ML, Zhou ZH. A Review on Multi-Label Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering 2013;26(8):1819–1837.

[6] Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep Learning, volume 1. MIT press Cambridge, 2016.

[7] Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, 2017; 240–248.

[8] Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2. 2012; 2951–2959.

[9] Glorot X, Bengio Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In JMLR W&CP: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010). 2010; 249–256.

[10] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv preprint arXiv14126980 2014;.

[11] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. arXiv preprint arXiv171105101 2017;.

[12] Izmailov P, Podoprikhin D, Garipov T, Vetrov D, Wilson AG. Averaging Weights Leads to Wider Optima and Better Generalization. arXiv preprint arXiv180305407 2018;.

[13] Hansen LK, Salamon P. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 1990;12(10):993–1001.

[14] Powers DMW. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies 2011; 2(1):37–63.

Address for correspondence:

Tomas Vicar
Department of Biomedical Engineering,
Faculty of Electrical Engineering and Communication,
Brno University of Technology,
Technicka 12, 616 00 Brno, Czech Republic.
vicar@vutbr.cz