First Steps Towards Self-Supervised Pretraining of the 12-Lead ECG

Daniel Gedon, Antônio H. Ribeiro, Niklas Wahlström, Thomas B. Schön

Uppsala University, Sweden

Abstract

Self-supervised learning is a paradigm that extracts general features which describe the input space by artificially generating labels from the input without the need for explicit annotations. The learned features can then be used by transfer learning to boost the performance on a downstream task. Such methods have recently produced state of the art results in natural language processing and computer vision. Here, we propose a self-supervised learning method for 12-lead electrocardiograms (ECGs). For pretraining the model we design a task to mask out subsegements of all channels of the input signals and try to predict the actual values. As the model architecture, we use a U-ResNet containing an encoder-decoder structure. We test our method by self-supervised pretraining on the CODE dataset and then transfer the learnt features by finetuning on the PTB-XL and CPSC benchmarks to evaluate the effect of our method in the classification of 12-leads ECGs. The method does provide modest improvements in performance when compared to not using pretraining. In future work we will make use of these ideas in smaller dataset, where we believe it can lead to larger performance gains.

1. Introduction

Supervised learning has been used for some of the most successful applications of deep learning. However, it requires a large amount of labeled training data to be successful [1] which is restricting its applicability especially when data is scarce and annotating data is expensive, which is true for ECGs. Self-supervised learning, on the other hand, use unlabeled data, such as images or raw text, to *generate a supervision signal from the data itself* and then learns the most relevant features for the given task without the need for explicit labels [2]. It opens up for interesting new possibilities, since the feature representations learned from the data can then be used with transfer learning to learn a downstream task on a new dataset with a smaller amount of labeled data. Improved performance can be obtained with this method compared with training from scratch.

Self-supervised learning has been successfully applied to a number of different fields, such as natural language processing [3], computer vision [4], speech recognition and for audio-visual data. This learning paradigm has been particularly successful in areas where a large corpus of unlabeled data is easily available to improve the performance on downstream tasks with a small labeled dataset.

The ECG is a highly utilized diagnostic tool, where accurate labeling of abnormalities for training deep learning algorithms requires expensive hours of specialized doctors. Here, we are motivated by the availability of large quantities of unlabeled ECG data and the recent progress of automated analysis of ECGs to present a method of applying self-supervised methods to ECGs. For this, we: 1) define a self-supervised learning task and pretraining procedure which can learn generalisable features of ECG data and 2) develop and show that a ResNet based architecture can successfully be used in combination with our learning task. We illustrate our development by first pretraining on the CODE training dataset [5] and, then, we use transfer learning with the ECG benchmarks: PTB-XL [6] and the dataset made available during the 2018 China Physiological Signal Challenge, the CPSC dataset [7].

2. Self-Supervised Learning Task and Model Architecture

Our self-supervised learning task is inspired by the completion task of BERT [3], which is a popular model developed in the context of natural language processing. The model architecture takes inspiration from the context encoder [8]. The sequential and bi-directional nature of the problem is the reason for us to choose a BERT-like self-supervision task. The convolution based architecture is motivated by the correlations between nearby samples in the ECG.

Pretraining task description: Our self-supervised pretraining task requires the model to reconstruct the input from a corrupted version of it. Specifically, we modify the input to the network by masking out the true values for random subsequences of N samples from the ECGs across all channels. The masked input is fed to the model which has to predict the true values for the masked subsequences. In total we mask out P percent of the input samples in this way. Similar to [3], 90% of the masked input subse-

 ISSN: 2325-887X DOI: 10.22489/CinC.2021.162

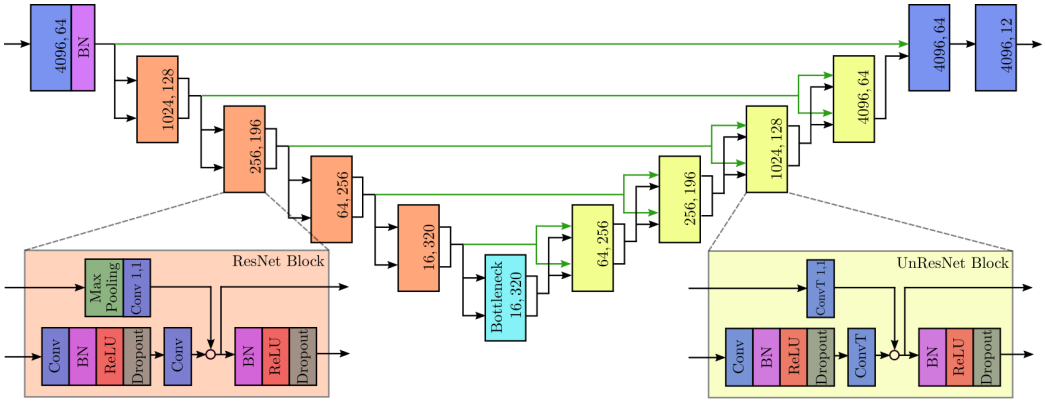

Figure 1: **U-ResNet Architecture:** Output dimension of each block are indicated. The input and output to the architecture are the ECG traces with a length of 4096 samples and 12 channels. In the decoder the input channels are concatenated.

quences are replaced by zeros and the remaining 10% are left unchanged. We use an MSE reconstruction loss.

An important design parameter is the number of subsequent samples N in a subsequence that are masked out. If the value of N is too small, then the task might reduce to a simple interpolation problem due to the significant temporal correlation in the ECG signal. If on the other hand, N is too large the patch might become too hard to reconstruct, rendering the task ineffective for self-supervision.

Model architecture: We use a convolutional ResNet architecture adapted from image classification [9] to unidimensional signals as described in [5]. Due to the convolutional model, a fixed input size is required. An extension for variable input size for this model is possible [10]. For pretraining we choose an encoder-decoder architecture based on [8] where the decoder mirrors the encoder using transposed convolutions. The bottleneck layer is a channel-wise fully connected layer. To improve the reconstruction ability of the network, we append the architecture with U-Net based skip-connections [11]. Details are shown in Figure 1.

Training on downstream task: Once a model is trained on the pretraining task, we transfer learn the pretrained model to a new labeled dataset as downstream classification task. Specifically, we use only the encoder, exclude the bottleneck layer, remove the U-Net skip connections and add a linear classifier.

3. Experiments

In this section we present the results of our experiments. We show that our pretraining task and model architecture works and generalises to new datasets. The results for the highly competitive ECG classification performance indicates that our approach is an interesting and promising path forward compared to randomly initialized models.

We use three datasets for our experiments. (1) CODE obtained by the Telehealth Network of Minas Gerais [12] which contains 2,322,513 ECGs from 1,676,384 patients out of 811 counties in Minas Gerais, Brazil with a length between 7 and 10 seconds. CODE also includes a test dataset with 827 ECGs. (2) CPSC 2018 [7] with 6,877 ECGs from 11 hospitals in China with lengths varying between 6 and 60 seconds including labels for 8 different anomalies. (3) PTB-XL [6] with 21,837 ECGs from 18,885 patients from 51 different sites with a length of 10 seconds including labels for 71 different anomalies. The anomalies are not mutually exclusive. Hence, for the classification task we use a sigmoid output layer together with a binary cross-entropy loss.

For self-supervised pretraining we use the CODE training data [5] and for the downstream task training we classify the anomalies in CPSC and PTB-XL. For CPSC we use a 70%-30% training/validation split (using the validation split also for reporting the final results), and for PTB-XL we take the predefined 80%-10%-10% training/validation/test split. We apply a high-pass filter to the input ECGs: an elliptic filter with a cutoff frequency at 0.8Hz and an attenuation of 40dB, applied in the forward and reverse direction to obtain zero-phase distortion. The filter removes biases and low frequency trends. In sequence, we resample all ECGs to a sampling frequency of 400 Hz and use zero padding if necessary to obtain a signal of a fixed length corresponding to 4096 samples. ECGs with longer traces are split over multiple batches and the individual classification predictions are combined by averaging the logits.

3.1. Pretraining

The pretraining reconstruction task requires two parameters, the number of subsequently masked samples N and the percentage of masked samples P in total. For the first, we choose $N = 64 \mathrel{\widehat{=}} 0.16$ sec which cover approximately one ECG segment, i.e. a characteristic subpart of the ECG. Therefore, the model is required to reconstruct complete segments and to learn the characteristics of ECG segments which can then be utilized in downstream tasks. For the second parameter, we choose $P = 30\%$ in order to make the problem hard enough, but still feasible. We train the

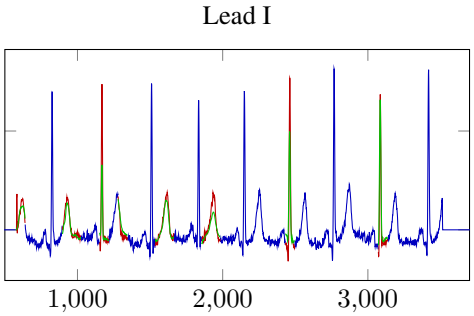

Figure 2: **ECG Reconstruction:** Examples for lead I of one CODE test ECG. Inputs to the model are blue where the gaps ($N = 64$, $P = 30\%$) are filled with zeros. The ground truth which should be reconstructed is red and the model output is green.

Table 1: Average MSE loss for the reconstruction on different datasets when removing blocks from the U-ResNet.

blocks	CODE train	CODE test	CPSC	PTB-XL
all	1.20	1.80	17.40	22.01
-1	2.54	3.06	19.63	19.66
-2	4.58	5.90	22.46	18.12
-3	19.72	19.70	32.12	28.92

model for 10 epochs on a single GPU. As an example, Figure 2 illustrates that the model can reconstruct complete segments and even partially cover the peaks.

To quantify how well the model generalises, we analyse the average MSE reconstruction loss of the obtained model for all datasets. Note that for the CODE training dataset, which we use for pretraining we report results on the first 100,000 ECGs only. The results are listed in Table 1. In the first row we see that the model generalizse reasonably well to the CODE test set, but not to the other test sets. Visual inspections of the ECGs show that while CODE is a rather homogeneous dataset, CPSC and PTB-XL have many outliers accounting for this difference in loss.

We additionally inspect the importance of depth for the reconstruction. We remove up to three ResNet blocks in the encoder/decoder, replace the bottleneck layer with an identity and try to reconstruct without re-training. Due to the U-ResNet skip connections, the model could be able to reconstruct the ECG only from the first blocks, which limits the usefulness of deeper ResNet blocks. We observe that the loss increases sharply when removing layers. We therefore conclude that the model learns useful representations in deeper layers, confirming our choice of architecture.

3.2. ECG Classification

The main results of the paper are summarized in Table 2. We compare multiple performance metrics on the classification of the CPSC and the PTB-XL dataset with and without

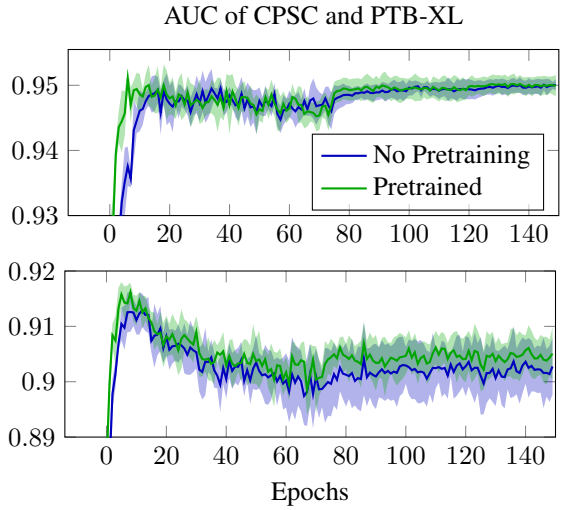

Figure 3: **AUC evolution:** AUC on the validation set during training for CPSC (top) and PTB-XL (bottom). Mean and standard deviation over 6 models are shown.

Table 2: Classification performance metrics with and without pretraining. 'PT' indicates with pretraining. For F_β and G_β we use $\beta = 2$.

(a) Metrics for CPSC

Model	F_β	G_β	AUC
Ours	$.775 \pm .004$	$.533 \pm .016$	$.953 \pm .001$
Ours + PT	$\mathbf{.780 \pm .013}$	$\mathbf{.538 \pm .019}$	$\mathbf{.954 \pm .001}$

(b) Metrics for PTB-XL

Model	F_{max}	AUC
ResNet1d [13]	$.767 \pm .008$	$.919 \pm .008$
Ours	$.667 \pm .037$	$.917 \pm .004$
Ours + PT	$.638 \pm .034$	$\mathbf{.919 \pm .003}$

pretraining. The results in the table correspond to the model with the largest area under the ROC curve (AUC). Plots for the evolution of the AUC on the validation set over the training epochs are given in Figure 3. The improvements we obtained are modest, but we believe the procedure has the potential to open interesting new possibilities in the low data regime.

3.3. Diversity of Predictions

For many real world applications, an ensemble of models outperforms standalone models in terms of accuracy [14] while it provides uncertainty bounds. It is theoretically proven that a high performing ensemble model requires the base models to be accurate and diverse in terms of their errors on the input [15]. Diversity in deep neural networks is achieved by initialising the network randomly and shuffling the training examples. For the pretrained model we always initialize at the same point. We analyse the diversity of the trained models initialized with our pretraining method

measured by the pairwise disagreement [14] and compare to randomly initialised models. We observe, that the models from our method on CPSC are only slightly less diverse than purely randomly initialised models while achieving higher AUC values. For PTB-XL we even notice, that our method produces models with higher AUC and greater diversity, see Figure 4. The results indicate that our method is well suited for ensemble based models.

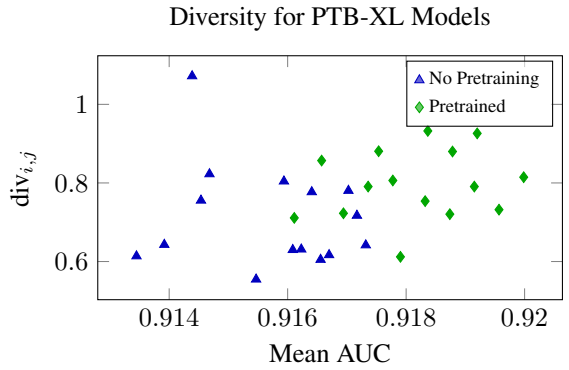

Figure 4: Diversity plot of models for PTB-XL. Pairwise diversity over the mean between the AUC of two models.

4. Conclusion and Future Work

In this paper, we introduce a simple completion based pretraining task for ECGs based on ResNet models. We illustrate the use of the method in the 12-lead ECG classification of the benchmarks: CPSC and PTB-XL. Our performance improvement is limited, nonetheless, we still believe this line of research to be interesting and relevant. This work should be seen as initial exploratory step that can yield interesting future work. Continuing in this direction can lead to more general ECG models that can be used across many datasets to solve tasks for which data is scarce.

Acknowledgments

This research was supported by the Brazilian research agency CAPES, the *Wallenberg AI, Autonomous Systems and Software Program (WASP)* funded by Knut and Alice Wallenberg Foundation, the *Kjell och Märta Beijer Foundation* and the project *AI4Research* at Uppsala University.

References

[1] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 2013;1798–1828.

[2] Zhang R. Image Synthesis for Self-Supervised Visual Representation Learning. Ph.D. thesis, University of California at Berkeley, 2018.

[3] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota, 2019; 4171–4186.

[4] Hjelm RD, Fedorov A, Lavoie-Marchildon S, Grewal K, Bachman P, Trischler A, Bengio Y. Learning deep representations by mutual information estimation and maximization. In International Conference for Learning Representations (ICLR). 2019; .

[5] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MPS, Andersson CR, Macfarlane PW, Meira Jr. W, Schön TB, Ribeiro ALP. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications 2020;1760.

[6] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data 2020;.

[7] Liu F, Liu C, Zhao L, Zhang X, Wu X, Xu X, Liu Y, Ma C, Wei S, He Z, Li J, Yin Kwee EN. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. Journal of Medical Imaging and Health Informatics 2018;1368–1373.

[8] Pathak D, Krähenbühl P, Donahue J, Darrell T, Efros AA. Context Encoders: Feature Learning by Inpainting. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; 2536–2544.

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

[10] Ribeiro AH, Gedon D, Teixeira DM, Ribeiro MH, Ribeiro ALP, Schön TB, Meira Jr. W. Automatic 12-lead ECG classification using a convolutional network ensemble. In Computing in Cardiology. 2020; .

[11] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI. 2015; 234–241.

[12] Alkmim MB, Figueira RM, Marcolino MS, Cardoso CS, Pena de Abreu M, Cunha LR, da Cunha DF, Antunes AP, de A Resende AG, Resende ES, Ribeiro ALP. Improving patient access to specialized health care: the telehealth network of Minas Gerais, Brazil. Bulletin of the World Health Organization 2012;373 – 378.

[13] Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl, 2020.

[14] Zhou ZH. Ensemble methods: foundations and algorithms. CRC press, 2012.

[15] Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning. In Advances in Neural Information Processing Systems. 1995; 231–238.

Address for correspondence:

Daniel Gedon

Department of Information Technology, Uppsala University, Lägerhyddsvägen 2 SE-752 37, Sweden

daniel.gedon@it.uu.se