The Advantage of Layer Freezing for Fine-Tune Deep Learning Algorithms in ECG Quality Assessment

Álvaro Huerta[1], Arturo Martinez-Rodrigo[1], José J Rieta[2], Raúl Alcaraz[1]

[1] Research Group in Electronic, Biomedical and Telecomm. Eng., Univ. of Castilla-La Mancha, Spain
[2] BioMIT.org, Electronic Engineering Department, Universitat Politecnica de Valencia, Spain

Abstract

Wearable ECG recording devices are becoming popular for long-term screening and monitoring of cardiovascular diseases. However, because of captured signals are broadly disturbed by artifacts, their quality assessment is essential to avoid ECG-based automated misinterpretation. For that purpose, pre-trained convolutional neural networks have recently reported promising performance. During the fine-tuning stage, some layers of these networks can be frozen to preserve their initial knowledge, thus allowing them to learn more generic features. Hence, the present work aims to analyze how freezing of different layers affects the performance of a well-known CNN scheme in ECG quality assessment. Precisely, several versions of a pre-trained 2-D AlexNet architecture, fed with continuous Wavelet transformed scalograms of ECG intervals, were derived by progressively freezing 7 out of their 8 layers with learnable parameters. They were then validated with two separate databases, which contained almost 70,000 5 second-length ECG intervals. After 10 validation iterations, no relevant differences in classification were noticed among the models, but an increasing trend in accuracy was noticed as the number of frozen layers increased. These results therefore recommend freezing all layers, except the last one, during the fine-tuning of AlexNet for ECG quality assessment to achieve a good trade-off between accuracy and computational resources.

1. Introduction

A common test to diagnose many cardiac diseases is the rest ECG, but it is insufficient for those cardiac disorders characterized by an intermittent and random nature, then continuous monitoring being needed [1]. Novel wearable devices, which allow us to continuously record ECG signals for several days, weeks and even months, are recently emerging [1]. Their contribution to the improvement of the diagnosis of heart rhythm alterations, such as paroxysmal atrial fibrillation (AF), in which initial episodes are mostly asymptomatic and only last for a few seconds or minutes, is expected to be huge in the short term [2].

However, although the diagnostic possibilities of these devices are unquestionable, the fact that the ECG signal is acquired during daily routine of the patients reduces its general quality [3]. Thus, to avoid misinterpretation of the recording, it is essential to face the problem of the automated identification of strong environmental disturbances associated with wearable or portable systems, such as motion artifacts, impulse noise, or electrode contact noise. Moreover, it is well-known that the ECG signal is also affected by other common noises, like powerline interference, muscle contractions, and baseline drifts. As a consequence, the quality of long-term ECG recordings must be evaluated as a previous step to be processed and diagnosed.

In recent years, automated quality assessment of long-term ECG recordings has experimented a great development, starting from simple systems based on decision rules to complex artificial intelligence systems. The first methods that obtained promising results were constructed by extracting fiducial points and morphological features from the ECG recording and combining them via common machine learning techniques, including support vector machines or k-nearest neighbors classifiers [4]. However, recent deep learning algorithms have overcome the generalization ability of those methods to discern between high- and low-quality ECG intervals [5]. Moreover, these techniques are also able to directly deal with the raw ECG signal, without requiring additional stages of preprocessing or feature extraction and selection [5].

Due to the large amount of data required for appropriate training of deep learning algorithms, there exists the possibility of using transfer learning, i.e., of fine-tuning the knowledge acquired by previously pre-trained convolutional neural networks (CNNs) on a new task [6]. In this process, it is possible to fine-tune only some layers, freezing the rest ones and then preserving their knowledge. The main goal of the present work is hence to analyze how freezing of a variable number of layers affects performance of a well-known CNN in ECG quality assessment.

Although several pre-trained CNN architectures are today available for transfer learning, which have been commonly trained to discern among more than 1,000 different kinds of images [7], the AlexNet architecture was chosen because it has been widely used in the context of ECG [5].

2. Methods

2.1. Algorithm for ECG quality assessment

A previously proposed deep learning approach to classify different ECG segments into the categories of high-quality and low-quality was used [5]. Because AlexNet architecture has been designed to receive 2-D images as input [8], the ECG signals were segmented into 5 second-length intervals, which were transformed into an image by making use of a Continuous Wavelet Transform (CWT) [9]. This kind of ECG-based image has been widely used as input for many deep learning algorithms in a broad variety of scenarios [9], since it is able to successfully enhance features of non-stationary signals, such as the ECG. The parameters used in this transformation were a Morlet function as mother wavelet and a number of wavelet scales determined by 48 voices per octave. More details can be found in [5].

2.2. Original AlexNet architecture

AlexNet was initially proposed to distinguish among more than 1,000 different image classes [8], and it has become popular thanks to the ImageNet Challenge [10]. Its architecture is composed of eight layers with learning ability, where five are convolutional layers and the other three fully connected layers [8]. The first convolutional layer has a spatial dimension of $11 \times 11 \times 3$ with 96 kernels. The second convolutional layer reduces the input of the first one and uses filters of size $5 \times 5 \times 3$ with 256 kernels. The third convolutional layer is composed of 384 kernels with a size of $3 \times 3 \times 256$. The fourth convolutional layer holds 384 kernels of dimensions $3 \times 3 \times 192$, whereas the last convolution layer applies 256 kernels of size $3 \times 3 \times 192$. Each one of the fully connected layers contains 4,096 neurons, and the output of the last one is connected to a 1,000-way softmax classifier. The number of learnable parameters in these layers, which will be frozen during the fine-tuning process, are detailed in Table 1. The AlexNet architecture is completed by several activation layers and functions. Precisely, linear activation functions are placed next to all the learnable layers to reduce the spatial proportions of the feature map, and three pooling layers are also inserted after the first, second and fifth convolutional layers. Two dropout regularization functions with the task of preventing over-fitting are included before and after the first fully connected layer. The full description of the AlexNet model can be found in [8]. Figure 1 shows a simplied scheme of the original architecture, in which learnable layers are highlighted in blue.

2.3. Experimental setup

Two databases were used for the experimental setup in this study, i.e., a proprietary database (PDB) and the public PhysioNet/CinC Challenge 2017 database (PC2017DB) [11]. Both contain ECG recordings with very different morphologies, which were registered under diverse conditions. On the one hand, PDB was employed for the fine-tuning process of the different conduced experiments. In this dataset, ECGs were acquired with a textile wearable Holter, at a sampling rate of 250 Hz and 12 bits of resolution. The labeling of the high- and low-quality excerpts was carried out by two experts, who reported as noisy signals those in which the R-peaks cannot be clearly distinguished. In total, this dataset contained 20,000 5 second-length ECG intervals, equally distributed between the two groups. It should be noted that high-quality class was composed both of normal sinus rhythm (NSR) and AF episodes. On the other hand, the freely available PC2017DB was utilized to test the fine-tuned algorithms. It was obtained using a portable device that recorded the ECG signals through two fingers with a sampling frequency of 300 Hz and a resolution of 16 bits. In this case, quality annotations are freely available and are divided into four different groups such as AF, NSR, other rhythms, and noisy signals. After the segmentation into 5 second-length intervals, this set was composed of 48,607 5- second-length ECG segments, highly unbalanced between the two groups by belonging 47,439 to high-quality class and 1,168 to low-quality class.

In the fine-tuning process of a pre-trained CNN, the last layer has always to be adapted to the new task. In this study, the last layer of AlexNet, which is able to classify up to 1,000 different classes, was modified to only discern between high- and low-quality ECG excerpts. In addition to this change, eight different experiments can be performed by progressively freezing 7 out of their 8 layers with learnable parameters. Thus, none of the layers were firstly frozen and therefore all learnable parameters from them acquired new knowledge. Next, only the learnable parameters belonging to the first convolutional layer were frozen, thus preserving their initial knowledge. Later, this procedure was repeated by freezing the two first layers, the three first layers and so on until the seven first layers were frozen. Clearly, the first experiments required more time and computational resources than the last ones, because higher number of learnable parameters had to be updated during the fine-tuning process.

Table 1. Learnable parameters in the convolutional (conv.) and fully connected (F.C.) layers of AlexNet architecture.

Layer	Conv. 1	Conv. 2	Conv. 3	Conv. 4	Conv. 5	F.C. 1	F.C. 2	F.C. 3	**Total**
Learnables	34,944	307,456	885,120	663,936	442,624	37,752,832	16,781,312	4,097,000	60,965,224

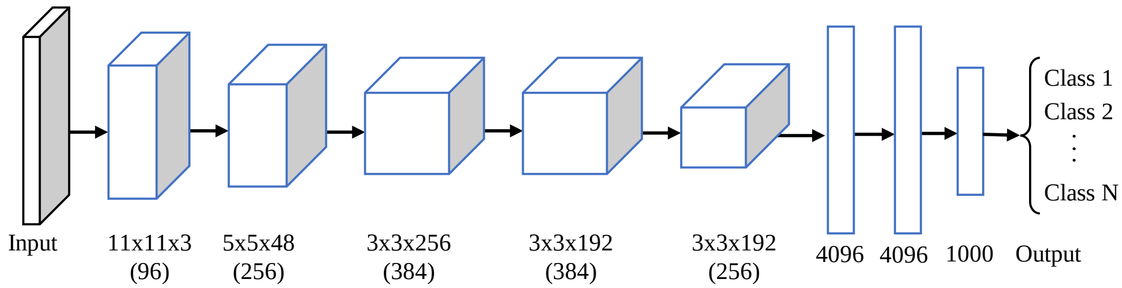

Figure 1. Simplified representation of AlexNet architecture. Layers with learnable parameters are represented in blue.

2.4. Performance analysis

The resulting eight CNN models were trained with the PDB database using a batch size of 32 and 10 epochs and later validated on the PC2017DB. Although this last set is the highly imbalance with 98% of the segments belong to the high-quality class and the 2% remaining belong to the low-quality class, this aspect is not relevant in the validation phase. Classification outcomes were evaluated in terms of the classical statistics of sensitivity (Se), specificity (Sp), and accuracy (Acc). Se was considered as the rate of rightly identified high-quality ECG segments, whereas Sp was the number of low-quality segments properly detected. Finally, Acc was the total number of all ECG intervals correctly detected. To obtain a robust validation, should be noted that each one of the eight experiments were repeated 10 times and classification results from the testing database were averaged. In next section, these results will be expressed in mean $\pm$ standard deviation for the 10 iterations.

3. Results

Table 2 presents the classification results obtained for the 8 fine-tuned and validated AlexNet-based models. As can be seen, comparable values of Acc, Se and Sp were reported by all of them. Indeed, a McNemar test did not provided statistically significant differences between the results of any pair of models. Moreover, it is of note that, even though the experiment were repeated 10 times, the standard deviation values ranged only between 1 and 3% for all cases. Also, Sp and Se were well-balanced, despite the high difference between ECG segments belonging to each class in the PC2017DB.

Table 2. Classification results obtained by the eight AlexNet-based models trained by progressively freezing the layers with learning ability.

Frozen layers	Se	Sp	Acc
0 (Original)	0.881±0.009	0.838±0.09	0.875±0.009
1	0.883±0.014	0.841±0.024	0.873±0.013
1 to 2	0.885±0.013	0.834±0.030	0.878±0.011
1 to 3	0.878±0.010	0.850±0.017	0.878±0.009
1 to 4	0.877±0.019	0.849±0.036	0.876±0.017
1 to 5	0.884±0.008	0.837±0.018	0.883±0.007
1 to 6	0.882±0.008	0.845±0.017	0.881±0.008
1 to 7	0.887±0.009	0.837±0.018	0.885±0.009

4. Discussion

The obtained results suggest that freezing of all the layers, except the last one, during the fine-tuning of AlexNet for ECG quality assessment does not significantly improve classification performance. Nonetheless, this approach is still interesting because training time and computation resources could be optimized by minimizing the number of parameters that should be fine-tuned. This finding is pioneering in the field of ECG quality assessment since, to the best of our knowledge, there are no previous studies about the impact of layer freezing on deep learning algorithms for that purpose. However, it might not be directly extrapolated to other algorithms based on CNN schemes different from AlexNet. In contrast to the sequential layer structure of AlexNet, other networks can present parallel branches, such as GoogleNet, and freezing of learnable parameters could have a different effect [12].

Anyway, the results obtained in the present work are in line with those obtained by other previous studies in contexts far from ECG quality assessment. For instance, Xiao et. al. proposed a new approach to intelligently calculate the number of layers to freeze in the case of natural image classification [13]. The algorithm was tested on different CNN schemes, including VGG, ResNets, and DenseNets, providing in all the cases an acceleration in the training, while similar classification performance was obtained. Likewise, Chudzik et. al. [14], assessed the impact of freezing layers in the detection of microaneurysm from fundus photographies via a previously proposed and pre-trained CNN-based algorithm. As before, the trainable parameters were reduced by increasing the number of frozen layers, whereas classification performance was approximately maintained.

5. Conclusions

Freezing of layers in AlexNet-based algorithms for quality assessment of long-term ECG recordings involves interesting advantages. Thus, although classification performance between high- and low-quality ECG excerpts is not significantly improved, training time and computation resources could be optimized. Future studies will address the possibility of using this approach in other CNN architectures, as well as in other ECG-based applications.

Acknowledgments

This research has received financial support from public grants PID2021-00X128525-IV0 and PID2021-123804OB-I00 of the Spanish Government 10.13039/501100011033 jointly with the European Regional Development Fund (EU), SBPLY/17/180501/000411 and SBPLY/21/180501/000186 from Junta de Comunidades de Castilla-La Mancha, and AICO/2021/286 from Generalitat Valenciana, as well as from the company Daiichi Sankyo S.A.U.

References

[1] Sana F, Isselbacher EM, Singh JP, Heist EK, Pathik B, Armoundas AA. Wearable devices for ambulatory cardiac monitoring: JACC state-of-the-art review. Journal of the American College of Cardiology 2020;75(13):1582–1592.

[2] Stachon P, Ahrens I, Faber T, Bode C, Zirlik A. Asymptomatic atrial fibrillation and risk of stroke. Panminerva Med Dec 2015;57(4):211–5.

[3] Smital L, Haider CR, Vitek M, et. al. Real-time quality assessment of long-term ECG signals recorded by wearables in free-living conditions. IEEE Trans Biomed Eng 2020; 67(10):2721–2734.

[4] Satija U, Ramkumar B, Manikandan MS. An automated ECG signal quality assessment method for unsupervised diagnostic systems. Biocybern Biomed Eng 2018;38:54–70.

[5] Huerta Herraiz Á, Martínez-Rodrigo A, Bertomeu-González V, Quesada A, Rieta JJ, Alcaraz R. A deep learning approach for featureless robust quality assessment of intermittent atrial fibrillation recordings from portable and wearable devices. Entropy 2020;22(7):733.

[6] Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. Journal of Big data 2016;3(1):1–40.

[7] Krishna ST, Kalluri HK. Deep learning and transfer learning approaches for image classification. International Journal of Recent Technology and Engineering IJRTE 2019; 7(5S4):427–432.

[8] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems. 2012; 1097–1105.

[9] Hong S, Zhou Y, Shang J, Xiao C, Sun J. Opportunities and challenges of deep learning methods for ECG data: A systematic review. Comput Biol Med 2020;122:103801.

[10] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision IJCV 2015;115(3):211–252.

[11] Clifford GD, Liu C, Moody B, Lehman LWH, Silva I, Li Q, Johnson AE, Mark RG. AF classification from a short single lead ECG recording: The PhysioNet/Computing in Cardiology Challenge 2017. Comput Cardiol 2017;44.

[12] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2015; 1–9.

[13] Xiao X, Mudiyanselage TB, Ji C, Hu J, Pan Y. Fast deep learning training through intelligently freezing layers. In International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). 2019; 1225–1232.

[14] Chudzik P, Majumdar S, Caliva F, Al-Diri B, Hunter A. Microaneurysm detection using deep learning and interleaved freezing. In Medical Imaging 2018: Image Processing, volume 10574. 2018; 379–387.

Address for correspondence:

Álvaro Huerta Herráiz
ITAV, Campus Universitario s/n, 16071, Cuenca, Spain
Phone: +34-969-179-100
e-mail: alvaro.huerta@uclm.es