Multitask and Transfer Learning for Cardiac Abnormality Detections in Heart Sounds

João L. Costa[1,3], Paula Couto[2,4], Rui Rodrigues[2,4]

[1] Department of Mathematics, Iscte-IUL, Portugal
[2] Department of Mathematics, Caparica, FCT NOVA, Portugal
[3] Centro de Análise Matemática, Geometria e Sistemas Dinâmicos, IST-ULisboa, Portugal
[4] Center for Mathematics and Applications (CMA), FCT NOVA, Caparica, Portugal

Abstract

We present a deep learning model for the automatic detection of murmurs and other cardiac abnormalities from the analysis of digital recordings of cardiac auscultations. This approach was developed in the context of the George B. Moody PhysioNet Challenge 2022.

More precisely, we consider multi-objective neural networks, with several Transformer blocks at their core, trained to perform 3 distinct tasks simultaneously: murmur detection, outcome classification and audio signal segmentation. We also perform pre-training with the 2016's Challenge data.

We entered the challenge under the team name matLisboa. Our (best) results on the hidden validation dataset (public Challenge leaderboard) were:
Murmur score (weighted accuracy): 0.754.
Outcomes score (cost): 9512.

1. Introduction

The detection of murmurs and other cardiac abnormalities via cardiac auscultation provides critical insights into heart malfunctioning. The design of algorithms for the automatic detection of such pathologies can provide valuable information in situations where there is difficulty in accessing health services. This is in essence the goal of the George B. Moody PhysioNet Challenge 2022 [1].

More specifically, this year's challenge proposes two distinct but related classification tasks of patient data [2] (that, in particular, include audio recordings of cardiac auscultation and meta data regarding sex, age and other individual features):

1. Murmur classification*: classify the data for each individual patient in terms of the existence of murmurs, by associating a label "Present", "Unknown" or "Absent";*

2. Outcome classification*: obtain, for each patient, a* clinical outcome classifier *in terms of "Abnormal" or "Normal".*

The performance of the submitted algorithms is then evaluated in each task separately by its own scoring metric that aggregates the algorithmic predictions of all patients: in the case of murmur classification *the final goal is to maximize a given* weighted accuracy metric *(s_{murmur}) and in the second task the goal is to minimize a* total outcome coast *($c_{outcome}$) that models the algorithm's efficiency/safety.*

Our approach is to try to solve both problems at once by using multi-objective deep neural networks, that receive as input (a pre-processed version) of a recording (of a cardiac auscultation) and output two distinct probability distributions, one for each classification task. In fact, in order to reduce overfitting, we also train our networks to perform a third (unrequested) task, which corresponds to the segmentation of each recording, in terms of S1, Systolic, S2 and Diastolic waves. This final task works as a regularization mechanism, that complements the use of dropout.

The core of our network's architecture is composed of several Transformer blocks [3]. As is well known, this type of architecture was first developed, with remarkable success, to tackle problems in natural language processing (NLP) and its main defining feature is a learnable self-attention mechanism that provides valuable contextual and positional information about the relation between individual data segments/tokens (for instance "words") within the overall structure of a given complete data input (for instance a "sentence"). In the meantime Transformers havebeen applied in multiple contexts from image processing to reinforcement learning.

In our framework, after applying the Short Time Fourier Transform and a convolution network block, each digital recording of a cardiac auscultation becomes a sequence of vectors, with fixed dimension, that, in analogy, we may consider as token embedding. By an appropriate supervised learning procedure (that we will describe below) we expect the Transformers to learn and encode relevant

relations between the individual time-steps that can then be further processed to yield the desired classifications. Stretching the NLP analogy a bit further, we can think of the Murmur and Outcome classification tasks as a kind of Sentiment Analysis.

Following a suggestion by one of the referees, we also tried to input meta data (age, sex, etc) information into our networks in order to understand its influence in the learning procedure [1].

Finally, to reduce overfitting, the biggest challenge in dealing with these large network models, we also tested pre-training with the 2016's Challenge data [4] and introduce data augmentation mechanisms.

2. Methods

2.1. Preprocessing

Motivated by both the need to compress the size of the original data, and by the fact that heart murmurs "are more significant when the flow is more turbulent" [2] and turbulence has characteristic profiles in frequency space [5], we start by applying the Short Time Fourier Transform to the audio signals. This requires the choice of two hyper-parameters: frame length*, that we set to* 32×10^{-3} *seconds, and* frame step*, fixed at* 16×10^{-3} *seconds; these values correspond to the best compromise between quality of results and processing time that we were able to find experimentally.*

Part of the provided training data was segmented into S1, Systolic, S2 and Diastolic waves, a relevant information that we used as labels for the training of the segmentation block of our networks (see Fig. 1). As, in general, only a partial segmentation of each signal was provided, we designed the loss function to only be sensitive to the segmented parts of the training input signals.

Our networks process each audio signal individually, while the goal of the challenge is to provide classifications at the level of each patient – for which, typically, there exists multiple audio recordings. For the Murmur classification task this creates no major problems, since the training data clearly identifies the corresponding label for each recording. However, for the Outcomes classification task, the training labels were provided at the level of each individual patient and, for the lack of a better idea, we decided to assign the same label to all of the patient's recordings.

We have also designed a simple data augmentation mechanism: instead of considering the complete audio signals as inputs – recall that these signals have, in general quite different sizes to start with – we took as inputs sequential portions of the signals with a randomly chosen size and randomly chosen initial starting time. This allowed to increase the size of the data considerably, which, as is well known, is of paramount importance to deal with overfitting.

[1] Note for instance that, in the training data, all "Adolescent" patients with audible murmurs were labeled "Abnormal", in the outcome category.

2.2. Model's architecture

We designed our neural networks to simultaneously output a probability distribution over the labels "Present", "Unknown" or "Absent", concerning the presence of murmurs, and also a probability of "Abnormal" and "Normal", as clinical outcome; this information is later aggregated to produce classifications at the level of each patient. Our models also produce a third kind of output corresponding to the auxiliary task, introduced by the authors to reduce overfitting, that attempts to obtain the segmentation of the original audio signals into "S1", "Systolic", "S2", "Diastolic" and "Unclassified" (see Fig. 1).

To perform this multitask learning our network starts with a common block that later splits into three branches, one for each task (see Fig. 2). In our approach this requires a considerably big network with approximately 654,000 trainable parameters.

Common block:

As the name indicates this component is the same for all tasks and, moreover, it is by far the largest component of our network containing 567,000 parameters. It starts with 3 convolution blocks with 32, 64 and 128 filters, respectively; each of these blocks is interleaved with batch normalization, maxpooling and spatial dropout layers. Then we use a positional embedding layer that adds learnable positional information to each vector. In the final step, of this common component, this information is used by 3 consecutive Transformer blocks, with 8 heads each.

Segmentation block:

Is essentially a linear layer with the same weights for each time-step and a five dimensional output vector (one for each segmentation label).

Output/Murmur classification blocks:

These blocks are essentially similar, the main difference being the output dimensions: 3d for the Murmur classification and 1d for the Outcomes. They are composed by another Transformer block, with 8 heads, followed by a feedforward network (with dropout) applied solely to the first time-step of the sequence outputted by the last Transfomer.

We have also considered other variations of the described model: either by including metadata information (regarding sex, age, etc) as a delayed input added in the final feedforward stages of the Output/Murmur classification blocks; or by feeding intermediate information of the

Murmur classification block into the Outcomes classification block; or both.

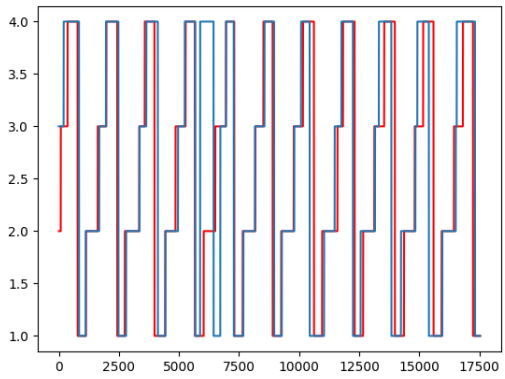

Figure 1. Segmentation task: the red line corresponds to the labels and the blue line to the model's prediction.

2.3. Training procedure

We perform a pre-train using the 2016's Challenge data. Afterwards, we train the models with this year's data by setting the initial weights of the common block to be the pre-trained weights; the weights of the task-specific blocks are trained from scratch using the 2022 data.

During training, the loss functions used in the Murmur/Outcome tasks are adapted/simplified versions of the Challenge scoring metrics (s_{murmur} and $c_{outcome}$) suitably weighted with cross-entropy losses. For the segmentation task we use a squared error loss that only considers the segmented components of the recordings in the training set.

2.4. Post-processing – the classification of each patient's data

As said before, our networks work at the level of individual recordings. To obtain classifications at the level of each patient, we collect the outputted probability distributions associated to all the recordings of a given patient's data and assign labels according to the following simple rules:

- Murmur classification*: If at least one of the murmur outputs predicts maximum probability for "Present", we assign that label to that patient's data; else, if at least one of the outputs predicts maximum probability for "Unknown", that's the label assigned; otherwise we assign the label "Absent" to the patient's data.*
- Output classification*: If at least one of the outcome outputs has (a probability) value below 0.5 we assign the label "Abnormal" to that patient; otherwise, we assign the label "Normal".*

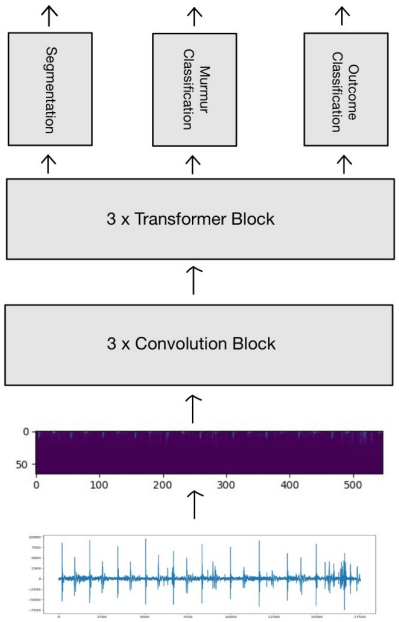

Figure 2. A schematic version of our network's architecture.

Table 1. Results on the public training dataset (10-fold cross-validation).

Model	Murmur	Outcomes
transf learn & pretraining	0.682	12947
transf learn & no pretraining	0.674	12962

3. Results

Our model's results on the 10-fold cross-validation on the public training dataset are presented in Table 1. For comparison, we present results with and without pre-training.

Moreover, our (best) results on the hidden validation dataset (public Challenge leaderboard) were:
Murmur score (weighted accuracy): 0.754.
Outcomes score (cost): 9512.

4. Discussion

In this paper we present some results concerning the use of multitask and transfer learning techniques for the

training of large neural networks (with Transformer core blocks) to deal with the automatic classification of audio recordings of cardiac auscultations. These results show that this is a promising approach, a conclusion that we base of the following facts: our networks systematically outperform the safe strategy *of classifying each patient's data as "Present" and "Abnormal"; we obtained "good" scores in both tasks of the official phase, specially in the Murmur classification task (see Sec. 3).*

There is a significant variance in the results of our model with different training runs: concerning murmur classification, our first and second official submissions contain exactly the same code and pre-trained weights, but their weighted accuracy on the hidden validation dataset were 0.754 and 0.716.

The performance of our networks in the Outcomes task is weaker than in the Murmur classification. We believe that this unfortunate situation as its roots in the fact that for each patient's data, in the training set, we have labeled all its audio recordings in the same way. To see why this might be problematic, consider, for instance, that this forces our networks, that work at the level of individual recordings, to learn to assign a label of "Abnormal" to recording where no abnormality is detectable.

Our multitask approach, in particular, allows our networks to learn relations between the different tasks, which are clearly correlated: for instance, a murmur classification of "Present" strongly suggests an outcome classification of "Abnormal", a relation that, unfortunately, our networks tend to take to literally. In this context it is also intriguing to inquire if the use of other (meta) data, e.g. the patient's sex and age, might help the networks create more interconnections between tasks. With this in mind, in some experimental runs, we have inputed meta-data into our networks, but due to the instability in training it is very hard to access if this has any particular practical value.

Moreover, multitasking works as a regularization technique against overfitting and for that reason we have decided to train the networks to perform a third task. Nonetheless, overfitting still plagues our models, even after using transfer learning by pre-training part of our network with the 2016's Challenge data.

Even though we believe that our approach is promising, it is clear that new ideas are needed, as well as more work to fine tune our models, in order to deal with overfitting, make the training procedure more stable and obtain better and more reliable results.

Acknowledgments

This work is funded by national funds through the FCT - Fundação para a Ciência e Tecnologia, I.P., under the scope of the projects UIDB/04459/2020 and UIDP/04459/202 (CAMGSD, IST-ID), and UIDB/00297/2020 (Center for Mathematics and Applications).

References

References

[1] Reyna, Matthew A. and Kiarashi, Yashar and Elola, Andoni and Oliveira, Jorge and Renna, Francesco and Gu, Annie and Perez Alday, Erick A. and Sadr, Nadi and Sharma, Ashish and Mattos, Sandra and Coimbra, Miguel T. and Sameni, Reza and Rad, Ali Bahrami and Clifford, Gari D., Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022, 2022.08.11.22278688, 2022, 10.1101/2022.08.11.22278688, Cold Spring Harbor Laboratory Press, medRxiv

[2] Oliveira, J., Renna, F., Costa, P. D., Nogueira, M., Oliveira, C., Ferreira, C., … Coimbra, M. T, "The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification," IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2524 - 2535, Jun, 2022.*

[3] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin, Attention Is All You Need, CoRR, abs/1706.03762,2017, http://arxiv.org/abs/1706.03762.

[4] Liu, C., et al Physiol. Meas.*, vol. 37, no. 12, pp. 2181, 2016.*

[5] Pope, S.B., "Turbulent Flows," Cambridge University Press*, 2000.*

Address for correspondence:

João L. Costa
Iscte – Instituto Universitário de Lisboa,
Avenida das Forças Armadas, 1649-026 Lisboa, Portugal
jlca@iscte-iul.pt