Automated Quality Assessment of Cardiac MR Images Using Convolutional Neural Networks

Le Zhang[1](✉), Ali Gooya[1], Bo Dong[1], Rui Hua[1], Steffen E. Petersen[2],
Pau Medrano-Gracia[3], and Alejandro F. Frangi[1]

[1] Centre for Computational Imaging and Simulation Technologies
in Biomedicine (CISTIB), Department of Electronic and Electrical Engineering,
University of Sheffield, Sheffield, UK
le.zhang@sheffield.ac.uk

[2] William Harvey Research Institute, Queen Mary University of London, London, UK

[3] Anatomy with Medical Imaging, University of Auckland, Auckland, New Zealand

Abstract. Image quality assessment (IQA) is crucial in large-scale population imaging so that high-throughput image analysis can extract meaningful imaging biomarkers at scale. Specifically, in this paper, we address a seemingly basic yet unmet need: the automatic detection of missing (apical and basal) slices in Cardiac Magnetic Resonance Imaging (CMRI) scans, which is currently performed by tedious visual assessment. We cast the problem as classification tasks, where the bottom and top slices are tested for the presence of typical basal and apical patterns. Inspired by the success of deep learning methods, we train Convolutional Neural Networks (CNN) to construct a set of discriminative features. We evaluated our approach on a subset of the UK Biobank datasets. Precision and Recall figures for detecting missing apical slice (MAS) (81.61 % and 88.73 %) and missing basal slice (MBS) (74.10 % and 88.75 %) are superior to other state-of-the-art deep learning architectures. Cross-dataset experiments show the generalization ability of our approach.

1 Introduction

Cardiac Magnetic Resonance Imaging (CMRI) can not only reflect anatomic information of the heart but also provide physiological information associated with cardiovascular diseases. Although low image quality can be minimized by careful design of the imaging acquisition protocols, it cannot be fully avoided; particularly in large-scale imaging studies, where data is acquired at different imaging sites, across subjects with a diverse constitution and at a big pace [5].

On the other hand, few objective guidelines exist, clinical or otherwise, that establish what constitutes, in general, a good image and, in particular, a good CMRI study [6]. To ensure that the quality of data collected in such imaging studies is maintained, Image Quality Assessment (IQA) is crucial. Surprisingly, IQA is still usually carried out by visual inspection of the images which can be exhaustive, costly, subjective, error prone, and time consuming [1]. Thus, Automatic IQA (AIQA) methods are required to detect deviations from the desired

S.A. Tsaftaris et al. (Eds.): SASHIMI 2016, LNCS 9968, pp. 138–145, 2016.
DOI: 10.1007/978-3-319-46630-9_14

quality, intervene to correct problems in data collection as soon as possible, and discard low-quality images, whose analysis would otherwise impair any aggregated statistics over the cohort. Additionally, *a priori* and objective knowledge on image quality of a given dataset (and possibly the type of artifact affecting it) could assist in choosing the most appropriate image analysis method to be used. This paves the way to “quality-aware image analysis” [16].

In multimedia, AIQA is a mature research field and usually concerned with detecting specific image distortions [15,17]. Unfortunately, most of these methods cannot be directly translated to medical imaging due to different properties in image statistics and the more complex nature of image artifacts [9]. Thus, AIQA remains as a relatively unexplored research area in medical imaging. It is acknowledged that lack of basal and/or apical slices is probably the most common problem affecting image quality in CMRI and has a major impact on the accuracy of quantitative parameters of cardiac performance [7]. In this paper, we mainly focus on short axis (SA) cine MRI. More specifically, we aim to identify missing apical slice (MAS) or missing basal slice (MBS). To address this problem, we are motivated by the success of deep learning techniques and, in particular, Convolutional Neural Network (CNN) [2,4]. They can achieve effective generalization properties, when applied to complex classification problems such identifying missing SA slices.

To the best of our knowledge, this is the first paper tackling the problem of detecting the missing slices in CMRI. Apart from introducing a new application for the CNN’s, and addressing a pressing need, we propose an effective strategy for their training. In practice, the lack of sufficient number of CMR data sets with MBS/MAS deficiencies imposes a severe class imbalance problem. To alleviate this issue, only the bottom and top SA slices are examined to ensure the full coverage of the heart. This allows us to use the middle slices as non BS/AS training samples. We present results for various depth of the networks, and identify the optimal number of the layers. We also compare our framework with an array of other deep learning methods such as Deep Boltzman Machines (DBM) and Stack Auto Encoders (SAE), and show its better performance. In the next section, we briefly introduce the architecture of our networks and provide the specification of our data sets. We then present our classification results and conclude the paper in the final section.

2 Methodology

2.1 Convolutional Neural Network for Feature Learning

As mentioned, we are interested in detecting missing apical and basal slices in CMRI data sets. To this end, for each cardiac subject, the top and bottom SA slices in the scan are classified using two CNNs, each particularly trained for detecting missing slices in basal or apical positions. Each CNN is composed of alternating convolutional and sampling layers, and one fully-connected output layer. Figure 1 shows the configuration of CNNs with total number of 5 layers

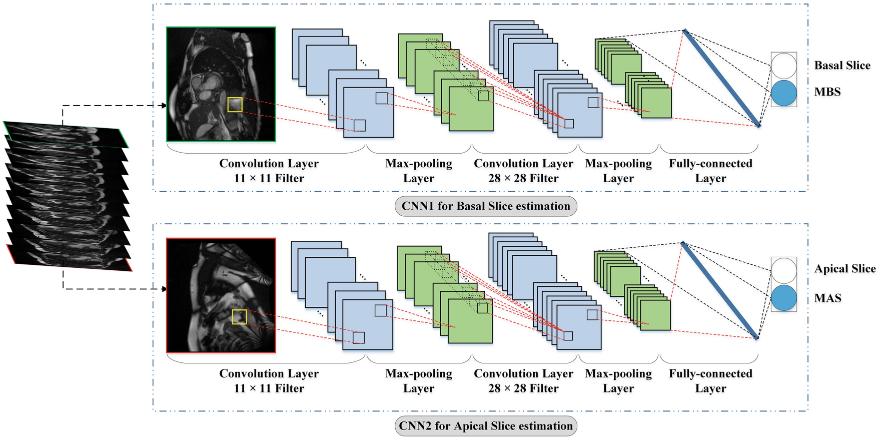

Fig. 1. Overview of our proposed deep learning model for cardiac MRI quality assessment. The CNNs are composed of 5 layers: four multi perceptron convolutional layers plus one fully-connected layer. The bottom and top SA slices are examined individually.

(showing overall the best classification performance). Here, we briefly review the various components in the proposed CNNs with a further detail.

Convolutional Feature Layers: Convolutional layers implement kernels that are used to detect discriminative features from input images [11]. During the training, these kernels are optimized to compute some salient features (such as edges, corners, etc.) that are relevant for discrimination of the observed categorical variables. We define $\mathbf{X}_i^{l-1}$ and $\mathbf{X}_i^l$ as input and output ith feature map of the lth layer. Let $m \times n$ and $k \times k$ be the size of input maps and the convolution kernel for layer l. With this setting of parameters, we can get N output maps with the size $(m-k+1) \times (n-k+1)$. The output of a convolutional layer l is given by

$$\mathbf{X}_j^l = f\left(\sum_{i \in M_j} \mathbf{X}_i^{l-1} * \mathbf{k}_{ij}^l + b_j^l\right), \tag{1}$$

where $\mathbf{k}_{ij}^l$ denotes the convolution kernel linking the ith input to the jth output map; b_j^l is the bias vector for the jth output-feature-map of lth layer; f is the activating function $1/(1+e^{-x})$, and M_j is the input feature map in the former layer.

Sampling Layers: These layers are designed to reduce the number of kernel parameters, minimize the computational complexity, and make the features robust to zoom, shift and rotation. The output of convolution layers are divided into sub-regions having the size of $w \times h$ pixels. Then, each output pixel of a sampling layer is defined as the maximum value in the corresponding input sub-region. These operations can be formulated using the following relationship

$$\mathbf{X}_j^l = f\left(\beta_j^l down\left(\mathbf{X}_j^{l-1}\right) + b_j^l\right), \tag{2}$$

where $down(\cdot)$ symbolizes the down sampling function; j, l, β and b denote the feature map index, the layer number, the weighting coefficients, and the bias vector, respectively.

Softmax classifiers: To predict the final labels, the CNN detected low-dimensional features are used to train softmax classifiers. Given the feature vector $\mathbf{x}^{(i)}$, we computed the posterior probabilities for $k = 1, 2, ..., K$ classes using

$$p(y^{(i)} = k|\mathbf{x}^{(i)}) = \frac{e^{\boldsymbol{\theta}_j^T \mathbf{x}^{(i)}}}{\sum_{l=1}^{K} e^{\boldsymbol{\theta}_l^T \mathbf{x}^{(i)}}}, \tag{3}$$

where $\boldsymbol{\theta}$ denotes the parameters of the softmax classifier, obtained from the pre-trained CNN network. The neural network was trained over 3 days for 100 epochs with a fixed learning rate 0.01. In the framwork, Rectified Linear Unit (ReLU) [8] was used as a activation function, and back-propagation technique [14] was used for adjusting weights of connections in the network. To test a single image with size 100 × 100, it only took approximate 0.2 s.

3 Results

3.1 Pre-processing and Data Description

To minimize the influence from the background region, a global mask covering the heart and its vicinity was employed prior to training. We define three classes of qualities in this paper: MAS, MBS, and no missing slices (normal). The last label is obtained by logical combination of the results from the MAS and MBS classifiers. The criterion used to determine a correct basal slice position is to verify if the left ventricular outflow tract (LVOT) is observable at the end-systolic phase [7].

Table 1. The average precision and recall rates of each type of missing slices using different deep learning models.

	Precision rate			Recall rate		
	MAS	MBS	Normal	MAS	MBS	Normal
SAE	79.08 %	68.63 %	78.54 %	88.48 %	88.72 %	88.15 %
DBM	66.67 %	70.09 %	71.47 %	88.38 %	88.71 %	88.32 %
3-CNNs	80.77 %	70.92 %	78.43 %	88.52 %	88.75 %	87.85 %
5-CNNs	**81.61 %**	**74.10 %**	**79.42 %**	**88.73 %**	**88.75 %**	**88.01 %**
7-CNNs	82.19 %	69.43 %	75.06 %	88.62 %	88.76 %	87.01 %

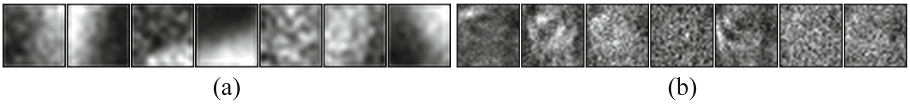

Fig. 2. The learned convolution kernels on basal and mid-slices of the first (a), and the second (b) layers of the trained CNN.

We apply our framework to 100 *UK Biobank (UKB)* cardiac MRI pilot data sets. These data sets are obtained by 1.5T MR scanners [12,13] and show overall good quality and no missing slices. Therefore, to generate synthetic deficiencies in the data, we manually removed basal slices from 50 subjects and apical slices from another 50 subjects. For each kind of the considered defect, we randomly selected 80 % of generated data sets as training sets and the left the rest as the testing sets. In order to evaluate our proposed framework's performance, we use the *Precision Rate* $= TP/(TP + FP)$, and the *Recall Rate* $= TP/(TP + FN)$, where TP, FP, and FN denote the number of true positive, false positive, and false negative samples, respectively.

3.2 Evaluation and Comparison to Other Deep Learning Models

We systematically compared our proposed CNNs framework with different types of CNNs architectures and traditional deep learning methods. Table 1 lists the results for different CNNs architectures and other state-of-the-art deep learning methods. As seen, the CNNs with a total number of 5 layers shows the best precision rate and recall rates.

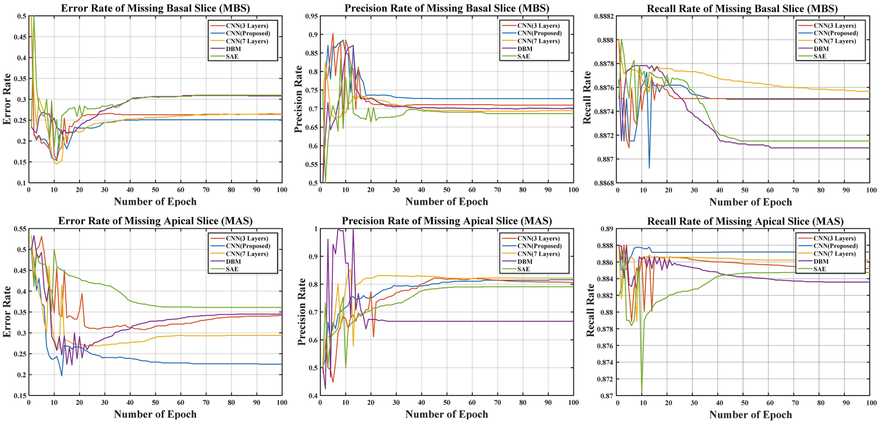

Fig. 3. The distributions of the error, precision, and recall rates over 100 training epochs, showing a superior performance of the CNNs with 5 layers.

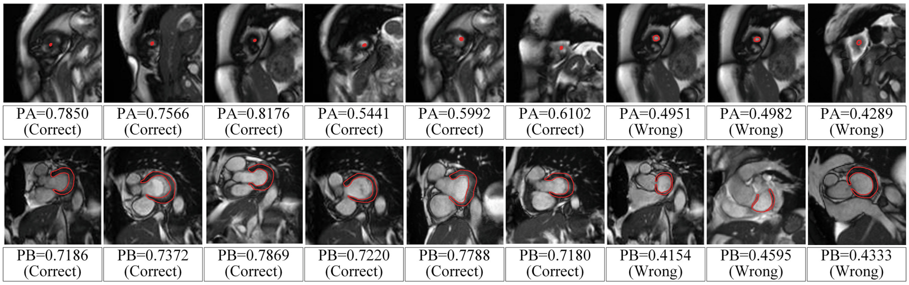

Fig. 4. Sample test slices and their probability values of being apical (top row) or basal slice (bottom row) are shown. 'PA' means the Probability value of being Apical slice; 'PB' means the Probability value of being Basal slice. The 'correct' and 'wrong' subscripts indicates the classification results.

We also visually examined the learned convolution kernels, and found only a few kernels present structure related appearances. Figure 2 shows the kernels learned for classifying missing basal slices. It is not surprising that some of these kernels show noisy, rather than strong structural and interpretable patterns. This is because our features are trained to be discriminative. In fact, to obtain user interpretable features, generative models such as those outlined in [10] is usually considered.

Furthermore, to demonstrate the convergence behaviour of the compared methods, in Fig. 3 we show the distributions of the error, precision, and recall rates over 100 training epochs. It can be seen the CNNs with 5 layers outperforms other CNN architectures and learning models.

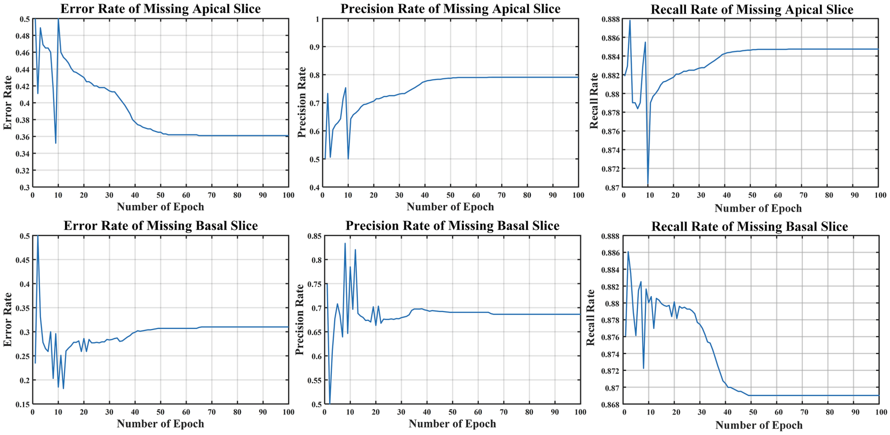

Fig. 5. The error, precision, and recall rates in cross dataset test.

In Fig. 4, a few apical (top row) and basal (bottom row) slices in the test datasets along with their corresponding posterior probability values are shown. We can observe that our framework correctly classifies a few challenging basal slices, but also fails in a few other cases. Furthermore, the basal slices with existing LVOT's indicate higher probability values of being correctly classified. This shows that the training has been successful in capturing the LVOT as a prominent feature in the correctly positioned basal slices.

We also designed a validation experiment with a second collection of CMR data sets to show the generalization ability of our method. To this end, we trained the proposed model using the UK Biobank datasets and tested it using the data sets available from Data Science Bowl Cardiac Challenge data sets [3]. This experiment was repeated for 100 training epochs and the values for error, precision and recall rates are shown in Fig. 5. These results show that our trained convolutional neural network achieves a good generalization efficacy.

4 Conclusion

In this paper, we tackled the problem of identifying the missing apical and basal slices in large imaging databases. We illustrated the concept by applying the method to CMRI studies from the UK Biobank pilot datasets. We designed slice classifiers and learned a set of discriminative features directly by training Convolutional Neural Networks. Casting this problem as a slice classification task, we were able to alleviate the class imbalance issue and effectively train the CNNs using the available data. Different numbers of network layers were examined and compared to other deep learning models (such as Stacked Auto-Encoder and Deep Boltzmann Machines). We showed that a CNN model with 5 layers outperforms the other models. We also validated our model by training the 5-CNNs using UKB pilot datasets and applying them to CMR data sets from Data Science Bowl Cardiac Challenge. The proposed model shows a high consistency with human perception and becomes superior compared to the state-of-the-art methods, showing its high potential. In this paper, the kernel sizes in the convolutional layers of the network were selected somehow arbitrarily. However, in principle these parameters can be optimized by performing exhaustive cross validation experiments. In future, we will further refine the current structure of our model by tuning such parameters.

References

1. Attili, A.K., Schuster, A., Nagel, E., Reiber, J.H., van der Geest, R.J.: Quantification in cardiac MRI: advances in image acquisition and processing. Int. J. Cardiovasc. Imaging **26**(1), 27–40 (2010)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
3. Bowl, K.: Data science bowl cardiac challenge data. https://www.kaggle.com/c/second-annual-data-science-bowl/data, Accessed 17 Mar 2016

4. Chen, X., Xu, Y., Yan, S., Wong, D.W.K., Wong, T.Y., Liu, J.: Automatic feature learning for glaucoma detection based on deep learning. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 669–677. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24574-4_80
5. Ferreira, P.F., Gatehouse, P.D., Mohiaddin, R.H., Firmin, D.N.: Cardiovascular magnetic resonance artefacts. J. Cardiovasc. Magn. Reson. **15**(1), 41 (2013)
6. van der Graaf, A., Bhagirath, P., Ghoerbien, S., Götte, M.: Cardiac magnetic resonance imaging: artefacts for clinicians. Neth. Heart J. **22**(12), 542–549 (2014)
7. Klinke, V., Muzzarelli, S., Lauriers, N., Locca, D., Vincenti, G., Monney, P., Lu, C., Nothnagel, D., Pilz, G., Lombardi, M., et al.: Quality assessment of cardiovascular magnetic resonance in the setting of the european CMR registry: description and validation of standardized criteria. J. Cardiovasc. Magn. Reson. **15**, 55 (2013)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
9. Krupa, K., Bekiesińska-Figatowska, M.: Artifacts in magnetic resonance imaging. Pol. J. Radiol. **80**, 93 (2015)
10. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 609–616. ACM (2009)
11. Oskoei, M.A., Hu, H.: A Survey on Edge Detection Methods. University of Essex, UK (2010)
12. Petersen, S.E., Matthews, P.M., Bamberg, F., Bluemke, D.A., Francis, J.M., Friedrich, M.G., Leeson, P., Nagel, E., Plein, S., Rademakers, F.E., et al.: Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of uk biobank-rationale, challenges and approaches. J. Cardiovasc. Magn. Reson. **15**(1), 46 (2013)
13. Petersen, S.E., Matthews, P.M., Francis, J.M., Robson, M.D., Zemrak, F., Boubertakh, R., Young, A.A., Hudson, S., Weale, P., Garratt, S., et al.: UK biobanks cardiovascular magnetic resonance protocol. J. Cardiovasc. Magn. Reson. **18**(1), 1 (2016)
14. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Cogn. Model. **5**(3), 1 (1988)
15. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: a natural scene statistics approach in the DCT domain. IEEE Trans. Image Process. **21**(8), 3339–3352 (2012)
16. Wang, Z., Wu, G., Sheikh, H.R., Simoncelli, E.P., Yang, E.H., Bovik, A.C.: Quality-aware images. IEEE Trans. Image Process. **15**(6), 1680–1689 (2006)
17. Xue, W., Mou, X., Zhang, L., Bovik, A.C., Feng, X.: Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. IEEE Trans. Image Process. **23**(11), 4850–4862 (2014)