Proceedings Paper:
Zhang, L, Gooya, A and Frangi, AF (2017) Semi-supervised assessment of incomplete LV coverage in cardiac MRI using generative adversarial nets. In: Lecture Notes in Computer Science. Second International Workshop on Simulation and Synthesis in Medical Imaging: SASHIMI 2017, 10 Sep 2017, Quebec, Canada. Springer Verlag, pp. 61-68. ISBN: 9783319681269. ISSN: 0302-9743. EISSN: 1611-3349.

https://doi.org/10.1007/978-3-319-68127-6_7

Semi-supervised Assessment of Incomplete LV Coverage in Cardiac MRI Using Generative Adversarial Nets

Le Zhang, Ali Gooya, Alejandro F. Frangi

Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), Department of Electronic and Electrical Engineering
University of Sheffield, United Kingdom

Abstract. Cardiac magnetic resonance (CMR) images play a growing role in diagnostic imaging of cardiovascular diseases. Ensuring full coverage of the Left Ventricle (LV) is a basic criteria of CMR image quality. Complete LV coverage, from base to apex, precedes accurate cardiac volume and functional assessment. Incomplete coverage of the LV is identified through visual inspection, which is time-consuming and usually done retrospectively in large imaging cohorts. In this paper, we propose a novel semi-supervised method to check the coverage of LV from CMR images by using generative adversarial networks (GAN), we call it Semi-Coupled-GANs (SCGANs). To identify missing basal and apical slices in a CMR volume, a two-stage framework is proposed. First, the SCGANs generate adversarial examples and extract high-level features from the CMR images; then these image attributes are used to detect missing basal and apical slices. We constructed extensive experiments to validate the proposed method on UK Biobank with more than 6000 independent volumetric MR scans, which achieved high accuracy and robust results for missing slice detection, comparable with those of state of the art deep learning methods. The proposed method, in principle, can be adapted to other CMR image data for LV coverage assessment.

1 Introduction

Left Ventricular (LV) cardiac anatomy and function are widely used for diagnosis and monitoring disease progression in cardiology and to assess the patient's response to cardiac surgery and interventional procedures. Cardiac ultrasound (US) and cardiac magnetic resonance (CMR) imaging are arguably the most wide-spread techniques for clinical diagnostic imaging of the heart. For population imaging studies, however, CMR remains the modality of choice and provides one-stop-shop access to cardiac anatomy and function non-invasively. The quantification of LV anatomy and function from large population imaging studies or patient cohorts from large clinical trials requires automatic image quality assessment and image analysis tools. A basic criteria for cardiac image quality is LV coverage and detection of missing apical and basal CMR slices [7]. Due to

rapid mechanical motion of the heart, breathing motion, and imperfect triggering, CMR can display incomplete LV coverage, which hampers quantitative LV characterization and diagnostic accuracy [12]. For example, missing basal slices has an important impact on LV volume calculation and several derived LV functional measures like ejection fraction and cardiac output. Even if scout images are acquired to center the LV in the field of view and minimize this problem, incomplete coverage can result at any points throughout the cardiac cycle due to patient breathing and cardiac motion. Automatic quality assessment is important in large-scale population imaging studies, where data is acquired across different imaging sites, from subjects with diverse constitutions, and with strict time constraints on scanner availability [4].

Few guidelines exist, clinical or otherwise, that objectively establish what constitutes a good medical image and a good CMR study [6]. To ensure consistent quantification of CMR data, automatic assessment of complete LV coverage is a first step. LV coverage is still assessed by visual inspection of CMR image sequences, which is subjective, repetitive, error prone, and time consuming [2]. Automatic coverage assessment must intervene and correct data acquisition soon, and/or discard promptly images with incomplete LV coverage whose analysis would otherwise impair any aggregated statistics over the cohort.

In medical imaging it is hard to have access to quality-labelled image databases due to the diversity of image characteristics, and their artifacts, of diverse anatomical locations and image modalities. Therefore, it is essential to devise techniques that do not require manual labelling of visual image quality. Image synthesis models provide a unique opportunity for performing unsupervised learning. These models build a rich prior over natural image statistics that can be leveraged by classifiers to improve predictions on datasets for which few labels exist [11]. Among them, generative adversarial networks (GAN) can synthesize adversarial examples, which increase the loss by a machine learning model [13]. Meanwhile, GAN can perform unsupervised learning by simply ignoring the component of the loss arising from class labels when a label is unavailable for a training image [5].

In this paper, we mainly focus on the analysis of short axis (SA) cine MRI. We aim to identify missing apical slices (MAS) and/or basal slices (MBS) in cardiac MRI volumes. In previous research, Le [14] used convolutional neural network (CNN) constructed on single-slice images and processed them sequentially. But this solution needs large amount of labelled data and lacks the ability to classify examples with perturbations correctly. In this paper, we exploit semi-coupled-GANs (SCGANs), a semi-supervised approach, for incomplete LV coverage detection. To alleviate the lack of sufficient numbers of CMR datasets with MBS or MAS, the proposed SCGANs use two generative models to synthesize adversarial examples. By learning adversarial examples, it improves not only robustness to adversarial examples, but also generalization performance for original examples. This work is the first work we know of to use adversarial examples to improve the robustness of an attribute learning model.

2 Methodology

We present a novel technique of LV coverage assessment for CMRI by using SC-GANs. The motivation behind our proposed method is: In medical image quality assessment problems, we are always faced with a lack of quality-labelled data, especially images with artifacts. Several deep learning models cannot classify the examples with perturbation correctly. Our semi-supervised SCGANs is proposed by using adversarial examples as the outlying observations for discriminative model training. We generate adversarial samples by two generators separately, which confuse the discriminator into mistaking them for genuine images. After that, we obtain the robust attribute classifiers by learning both original data and synthetic data. Our proposed SCGANs represents a strategy to better handle the typical LV coverage assessment problem.

2.1 Generative Adversarial Learning

Recently, GAN [5] was proposed as a novel way for adversarial learning. It consists of a generative model and a discriminative model, both are realized as multilayer perceptrons [9]. The aim of the discriminator is to correctly classify the original examples and adversarial examples. By learning the adversarial examples, the network cannot only becomes robust to adversarial examples, but also generalization improves for unmodified examples. GAN does not need the label information when training the generator and then the discriminator can estimates the probability that a sample came from the original data rather than the generator.

We assume a probability distribution M, which is a black box relative to us. To realize how the black box works, we construct two 'adversarial' models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample from the training data rather than G. Both G and D could be a non-liner mapping function, such as a multi-layer perceptron. Our objective is to learn feature representation to handle a wide range of visual appearances in cardiac MRI and identify images with incomplete LV coverage. We regard adversarial examples as outlying observations regarding other samples in training data. The generative model constantly produce new adversarial samples and the discriminative model classify the positive and negative samples by learning the new produced adversarial samples constantly. Given a particular describable visual attribute - say 'MBS'. An outlier image is expected to be mapped to negative values, which indicates the absence of basal slice. This can happen for two reasons: (1) the image does not belong to the basal slice, (2) the image belongs to the adversarial examples. We consider them all as the outliers.

2.2 Semi-Coupled GANs

Here we introduce our model based on the above discussion. Our model is illustrated in Fig. 1 designed as a semi-coupled-GANs for attribute learning. It

consists of a pair of *Generators*– G_1 and G_2, which share a same discriminator. Each generator synthesizes the adversarial samples Y_1 and Y_2 for positive and negative data, respectively.

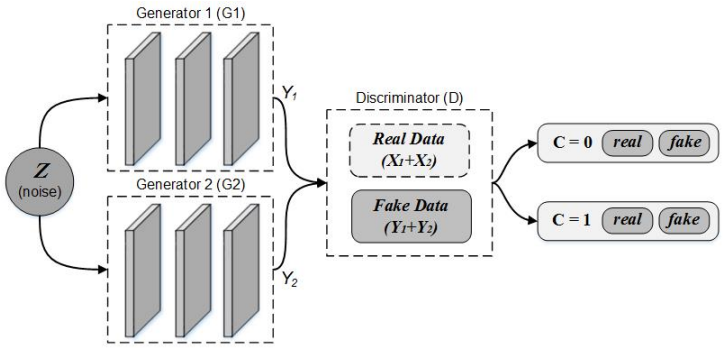

Fig. 1. The Proposed Semi-Coupled-GANs Framework.

Generative Models: We firstly feed the two generators G_1 and G_2 noise data $\boldsymbol{z}$, G_1 and G_2 learn probability distribution from the original positive and negative images respectively, and generate the corresponding adversarial samples. Then, we give the adversarial data to discriminator D. Denote the distributions of $G_1(\boldsymbol{z})$ and $G_2(\boldsymbol{z})$ by p_{G_1} and p_{G_2}. Both G_1 and G_2 are realized as multilayer perceptions:

$$\begin{cases} G_1(\boldsymbol{z}) = G_1^{(m_1)}(G_1^{(m_1-1)}(...G_1^{(2)}(G_1^{(1)}(\boldsymbol{z})))) \\ G_2(\boldsymbol{z}) = G_2^{(m_2)}(G_2^{(m_2-1)}(...G_2^{(2)}(G_2^{(1)}(\boldsymbol{z})))) \end{cases} \tag{1}$$

where $G_1^{(i)}$ and $G_2^{(i)}$ are the ith layers of G_1 and G_2 and m_1 and m_2 are the numbers of layers in G_1 and G_2. In our training process, m_1 and m_2 need not to be the same. In traditional discriminative deep neural network, the feature information is extracted from low-level features in first layers to the high-level features in last layers. While, through multi-layer perceptron operations, our two generator models decode the information with an opposite flow direction from abstract concepts to more material details.

Discriminative Models: Every generated sample has a corresponding class label and the discriminator gives both a probability distribution over dataset and a probability distribution over the class labels. We put both the original samples and the adversarial samples into D for the discriminator training, D output multiple output values between 0 and 1. In this process, if the training samples $\boldsymbol{x}$ is the positive/or real data, the discriminant D ensures the output value is similar with the trained corresponding value, which represents the input data is the positive/or real, while output values close to 0 indicates the input data

is the negative/or fake. The discriminant D equals a classifier with supervision situation, which returns to 1 or 0. Let D be the discriminative model given by:

$$D(\boldsymbol{x}) = D^{(n)}(D^{(n-1)}(...D^{(2)}(D^{(1)}(\boldsymbol{x})))) \tag{2}$$

where $D^{(i)}$ is the ith layer of D and n is the number of layers. The discriminator maps each input image to a probability score which indicates the input is drawn from the positive data or the negative data. In this process, the first layer of the discriminative model extracts low-level features, while the last layer extracts high-level features.

Learning: The Semi-Coupled-GANs framework corresponds to a constrained minimax game given by

$$\begin{aligned}\max_{D} \min_{G_1,G_2} V(G_1, G_2, D) = E_{x\sim p_{x_{data}}}[\log D(\boldsymbol{x} \mid \boldsymbol{y})] + E_{z\sim p_z}[\log(1 - D(G_1(\boldsymbol{z})))] \\ + E_{z\sim p_z}[\log(1 - D(G_2(\boldsymbol{z})))]\end{aligned} \tag{3}$$

There are two terms in (3), each term has an independent generator but share a same discriminator. The two generative models synthesize a pair of adversarial samples for confusing the discriminative models. The discriminator gives both a probability distribution over image data and a probability distribution over the class labels, $D(\boldsymbol{x} \mid \boldsymbol{y})$. Here, there are four kinds of samples for training the discriminator: the positive and negative samples from original images and their corresponding adversarial samples computed by two generators. The inputs discriminative model is data and corresponding labels. Similar to GAN, our SC-GANs can be trained by back propagation with the alternating gradient update steps.

2.3 Quality Estimation

For a given cardiac volume, a dissimilarity score is computed for each representative visual attribute - MAS and MBS. Any visual attributes with a score below an optimal threshold is classified as an artifact. After computing the visual attributes, we could verify the cardiac MRI quality based on the corresponding attributes scores. Let $x_{target} = P_{MAS}(\boldsymbol{X}_{target})$ and $y_{target} = P_{MBS}(\boldsymbol{X}_{target})$ be the outputs of the discriminator. If the quality of target cardiac volume $\boldsymbol{X}_{target}$ is good, the values $P_{MAS}(\boldsymbol{X}_{target})$ and $P_{MBS}(\boldsymbol{X}_{target})$ from the target cardiac volume should be similar with the trained corresponding positive attribute values. We combine the output values so the verification classifier Q can make sense of the data. To address the problem, we use the concatenation of these tuples for both MAS and MBS attribute classifier outputs form the input to the verification classifier Q [8]. Finally, putting both terms together yields the tuples $q(S_{target})$:

$$q(S_{target}) = Q(< p_{MAS}, p_{MBS} >) \tag{4}$$

Training Q requires pairs of positive examples and negative examples. For the classification function, we use SVM with an RBF kernel for $\boldsymbol{X}$, trained using libsvm [3] with the default parameters of $C = 1$ and $\gamma = 1/ndims$, where $ndims$ is the dimensionality of $< p_{MAS}, p_{MBS} >$.

3 Experiment and Related Analysis

Data specifications: In the UK Biobank (UKBB) dataset, we have 3400 subjects, each with 50 time points covering the heart from the base to apex. We use the endocardial contour as the main characteristic to identify the apical, middle and basal slices. For example, we can find the Left Ventricular Outflow Tract (LVOT) in the basal slice. In other slices, LVOT is nonexistant. As for the apical slice, we define it as the LV cavity is still visible at end-systole. Besides the basal slice and apical slice, we can consider the rest slices as the middle slices. To obtain the negative samples, we choose the middle slice as the negative samples for each attribute learning.

Experimental set-up: All experiments used TensorFlow [1] on GPUs. With all 50 time points consideration for each subject, we can obtain 17,0000 and regarded as the ground truth in our experiments. The architecture of the two generators G_1 and G_2 are consisted of several 'deconvolution' layers that transform the noise $\boldsymbol{z}$ and class c into an image [11]. We train the model architecture for generating images at 120×120 spatial resolutions. The discriminator D is a deep convolutional neural network with a Leaky ReLU nonlinearity [10]. In our experiment, 10-fold cross-validation method is used to evaluate the final performance of our attribute classifiers. To evaluate the classification algorithms, we use Accuracy, Precision Rate and Recall Rate defined as: *Accuracy = (TP+TN)/(TP+FP+TN+FN)*, *Precision Rate = TP/(TP + FP)* and *Recall Rate = TP/(TP + FN)*. Where TP, TN, FP, and FN are the numbers of the true positive, true negative, false positive and false negative samples, respectively.

Performance and Discussion: We evaluate the quality of our semi-supervised representation learning algorithms by applying it as a feature extractor on supervised datasets. Table 1 shows the test performance on UK Biobank Dataset with the state-of-art deep learning methods. With supervised deep learning methods, 2D CNN, it achieved accuracies with 77.5% and 74.9%. With adversarial learning approach, traditional GAN, the results are much better with 90.4% and 88.1% accuracies. Compared with the above two methods, our SCGANs achieved performance with significant increase, 92.5% and 89.3% accuracies. This is despite the state of the art models having no ability to discriminate the adversarial samples, whereas our model requires to training the generative model to produce the adversarial examples and can correctly classify both unmodified and adversarial samples. It improves not only robustness to adversarial examples, but also generalization performance for original examples. Meanwhile, our SCGANs also achieved a comparable result with the 3D CNN, which indicates opportunity for future 3D image synthesis models.

Table 1. The accuracy, precision rate and recall rate between the state-of-art deep learning approaches and our method.

	Accuracy		Precision Rate		Recall Rate	
Method	MAS	MBS	MAS	MBS	MAS	MBS
2D CNN	77.5±0.7%	74.9±0.6%	82.6±0.7%	74.9±0.8%	87.7±0.8%	87.8±0.9%
3D CNN	93.1±0.6%	91.8±0.7%	90.1±0.6%	87.3±0.7%	89.9±0.7%	93.3±0.8%
GAN	90.4±0.7%	88.1±0.6%	85.9±0.5%	88.5±0.6%	89.1±0.4%	90.6±0.6%
Our SC-GAN	92.5±0.5%	89.3±0.4%	87.6±0.4%	89.1±0.3%	90.5±0.5%	91.7±0.4%

Our attribute classifiers are trained using nine folds and then evaluated on the remaining fold, cycling through all ten folds. Receiver Operating Characteristic (ROC) curves are obtained by saving the classifier outputs for each test pair in all ten folds and then sliding a threshold over all output values to obtain different false positive/detection rates. In Fig.2, we demonstrate the ROC curve to show that our adversarial training (SCGANs) method can achieve ideal results. These results reinforce that adversarial examples are powerful samples for attribute leaning. In Fig.2 we can see our proposed method can correctly classify a few challenging samples (True Positive) and adversarial samples (False Negative). Experimental results obtained confirm that adversarial training approach makes the model more robust to adversarial examples and generalization performance for original examples. Although the results show that the accuracy of the proposed method is slightly lower but comparable to that of 3D CNN, our SCGAN can reduce the computation cost, which is especially important in population imaging.

4 Conclusion

In this paper, we tackled the problem of defining missing apical and basal slices in large imaging databases. We illustrated the concept by proposing a SCGANs to CMR image studies from the UK Biobank pilot datasets. By training the classifier with the adversarial examples, our model can achieve a significant improvement in attribute representation. A well-trained attribute classifiers are performed on the candidates to corresponding categories. We also validated our model by comparing with traditional deep learning methods and applying them to UK Biobank data sets. 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. Our proposed semi-couple-GANs can also be easily applied and boost the results for other detection and segmentation tasks in medical image analysis.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)

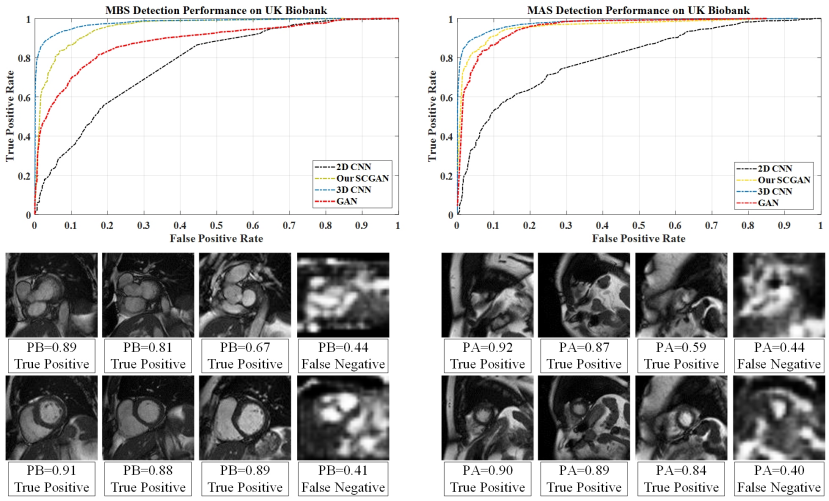

Fig. 2. MAS and MBS detection performance (Top) and sample test slices and their probability values (Bottom). PA means the Probability value of being Apical slice; PB means the Probability value of being Basal slice.

2. 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. The international journal of cardiovascular imaging 26(1), 27–40 (2010)
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27 (2011)
4. Ferreira, P.F., Gatehouse, P.D., Mohiaddin, R.H., Firmin, D.N.: Cardiovascular magnetic resonance artefacts. Journal of Cardiovascular Magnetic Resonance 15(1), 1 (2013)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
6. van der Graaf, A., Bhagirath, P., Ghoerbien, S., Götte, M.: Cardiac magnetic resonance imaging: artefacts for clinicians. Netherlands Heart Journal 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. Journal of Cardiovascular Magnetic Resonance 15(1), 1 (2013)
8. Kumar, N., Berg, A., Belhumeur, P.N., Nayar, S.: Describable visual attributes for face verification and image search. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(10), 1962–1977 (2011)
9. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems. pp. 469–477 (2016)
10. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. ICML. vol. 30 (2013)

11. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. arXiv preprint arXiv:1610.09585 (2016)
12. Pusey, E., Lufkin, R.B., Brown, R., Solomon, M.A., Stark, D.D., Tarr, R., Hanafee, W.: Magnetic resonance imaging artifacts: mechanism and clinical significance. Radiographics 6(5), 891–911 (1986)
13. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
14. Zhang, L., Gooya, A., Dong, B., Hua, R., Petersen, S.E., Medrano-Gracia, P., Frangi, A.F.: Automated quality assessment of cardiac MR images using convolutional neural networks. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 138–145. Springer (2016)