Semi-Supervised Learning for Network-Based Cardiac MR Image Segmentation

Wenjia Bai[1], Ozan Oktay[1], Matthew Sinclair[2], Hideaki Suzuki[3], Martin Rajchl[1], Giacomo Tarroni[1], Ben Glocker[1], Andrew King[2], Paul M. Matthews[3], and Daniel Rueckert[1]

[1] Biomedical Image Analysis Group, Department of Computing, Imperial College London, UK
[2] Division of Imaging Sciences and Biomedical Engineering, King's College London, UK
[3] Division of Brain Sciences, Department of Medicine, Imperial College London, London, UK

Abstract. Training a fully convolutional network for pixel-wise (or voxel-wise) image segmentation normally requires a large number of training images with corresponding ground truth label maps. However, it is a challenge to obtain such a large training set in the medical imaging domain, where expert annotations are time-consuming and difficult to obtain. In this paper, we propose a semi-supervised learning approach, in which a segmentation network is trained from both labelled and unlabelled data. The network parameters and the segmentations for the unlabelled data are alternately updated. We evaluate the method for short-axis cardiac MR image segmentation and it has demonstrated a high performance, outperforming a baseline supervised method. The mean Dice overlap metric is 0.92 for the left ventricular cavity, 0.85 for the myocardium and 0.89 for the right ventricular cavity. It also outperforms a state-of-the-art multi-atlas segmentation method by a large margin and the speed is substantially faster.

1 Introduction

Recent development in deep learning, especially the proposal of fully convolutional networks (FCN) [8], has greatly advanced the state-of-the-art in semantic image segmentation. The fully convolutional network has the advantage of offering end-to-end training and it has achieved a high accuracy for natural image segmentation [5]. Typically, such a network consists of millions of parameters and learning these parameters requires a large training set, which is formed of pairs of training images and corresponding pixel-wise label maps. In the medical imaging domain, however, it is a major challenge to obtain such a large training set due to several reasons. First, it is not easy to recruit experts who can reliably annotate medical images. Second, accurate pixel-wise annotation is time-consuming and tedious. Third, there are various modalities and imaging protocols, a training set generated for one study may not be easily transferable to another study.

To address this challenge, we propose a semi-supervised learning approach for network-based medical image segmentation, in which the segmentation network is trained from both labelled and unlabelled data, so that the need for a large training set is alleviated. The method is applied to the task of cardiac MR image segmentation, which is a crucial step for quantifying ventricular volumes and assessing cardiac function. Experimental results demonstrate that the proposed method effectively improves the segmentation accuracy, compared to a baseline method that only utilises the labelled data. It achieves a high performance for left and right ventricular segmentations. In addition, it outperforms a state-of-the-art multi-atlas segmentation method in both accuracy and speed.

1.1 Related Works

Many works have proposed using FCN for medical image segmentation [3,4,12]. For cardiac MR image segmentation, FCN and other network architectures have also been explored [1,9,11,16,17]. Most of these approaches learn image features from fine to coarse scales using convolutions and poolings and then combine multi-scale features to predict a pixel- or voxel-wise label map. These networks are normally trained in a fully supervised manner. The contribution of this work is that we have developed a semi-supervised way to train the network so that not only labelled images, but also unlabelled images can be utilised for training.

In the domain of computer vision, several works have proposed weakly-supervised learning, where labelled training data is augmented by data with image-level annotations, bounding boxes or scribbles [7, 10]. Our work explores semi-supervised learning with unlabelled images and evaluates its effect within a medical imaging scenario, or more specifically, cardiac MR image segmentation.

2 Methods

2.1 Semi-Supervised Learning

Let x denote an image and y denote its pixel-wise label map. A training set S consists of pairs of images and label maps, $S = \{X, Y\}$, where $X = \{x_i | i = 1, 2, \ldots, N\}$, $Y = \{y_i | i = 1, 2, \ldots, N\}$ and i denotes the image index. Suppose we have two sets, a labelled set $S_L = \{X_L, Y_L\}$ and an unlabelled set $S_U = \{X_U, Y_U\}$. The label maps Y_L are known and they normally come from manual segmentations by experts on images X_L, whereas the label maps Y_U are unknown. We build a network parameterised by Θ for image segmentation, i.e. to predict label map y from image x.

In the supervised setting, estimating the segmentation model is formulated as an optimisation problem for the following loss function,

$$\min_{\Theta} L(\Theta) = -\sum_{i \in L} \sum_{j} \log P(y_{i,j} | x_i, \Theta), \tag{1}$$

where j denotes the pixel index, $P(y_{i,j}|x_i, \Theta)$ is the softmax probability provided by the network at pixel j for image i and $L(\Theta)$ is the cross-entropy loss function.

This loss function is defined on labelled set S_L and it is usually optimised w.r.t. Θ by stochastic gradient descent (SGD).

In the semi-supervised setting, we introduce the unlabelled set S_U to the optimisation problem,

$$\min_{\Theta, Y_U} L(\Theta, Y_U) = -\sum_{i \in L} \sum_{j} \log P(y_{i,j}|x_i, \Theta) - \lambda \sum_{i \in U} \sum_{j} \log P(y_{i,j}|x_i, \Theta). \quad (2)$$

where the second term on the right is the cross-entropy for the unlabelled set and λ is a weight for this term. The loss function needs to be optimised against both the network parameters Θ and the unknown label maps Y_U. We solve this problem by alternately updating Θ and Y_U:

1. With $\hat{\Theta}$ fixed, estimate Y_U. Only the second term in the loss function Eq.(2) needs to be optimised. This step performs segmentation for the unlabelled images based on the current network.
2. With $\hat{Y}_U$ fixed, estimate Θ. This step updates the network parameters by training on both Y_L and estimated segmentations $\hat{Y}_U$.

The initial values of Θ are obtained by training the network only on the labelled maps Y_L for a number of epochs. Step 1 is performed by computing the softmax probability from the network and deploying a conditional random field (CRF) [6] to estimate a refined segmentation from the probability map. Step 2 is performed by using SGD to optimise the cross-entropy loss function, similar to supervised learning. We iteratively alternate between the two steps, in the hope that after each iteration, the network parameters are improved due to the updated segmentations and vice versa.

2.2 Conditional Random Field (CRF)

During the iterative approach, a CRF is used to refine the segmentation for the unlabelled data. The CRF optimises the following energy function [6],

$$E(y) = \sum_{j} \theta_j(y_j) + \sum_{j,k} \theta_{j,k}(y_j, y_k), \quad (3)$$

where the first term $\theta_j(y_j) = -\log P(y_j)$ is a unary potential which encourages the output to be loyal to the softmax probability, the second term $\theta_{j,k}(y_j, y_k)$ is a pairwise potential between labels on pixel j and pixel k,

$$\mu(y_j, y_k)\Big[w_1 \exp(-\frac{||p_j - p_k||^2}{2\sigma_\alpha^2} - \frac{||x_j - x_k||^2}{2\sigma_\beta^2}) + w_2 \exp(-\frac{||p_j - p_k||^2}{2\sigma_\gamma^2})\Big], \quad (4)$$

where $\mu(y_j, y_k) = 1$ if $y_j \neq y_k$ and 0 otherwise. This term penalises pixels with similar positions p and intensities x but with different labels y. The CRF can improve the localisation property of the network and refine the segmentation as shown in [6].

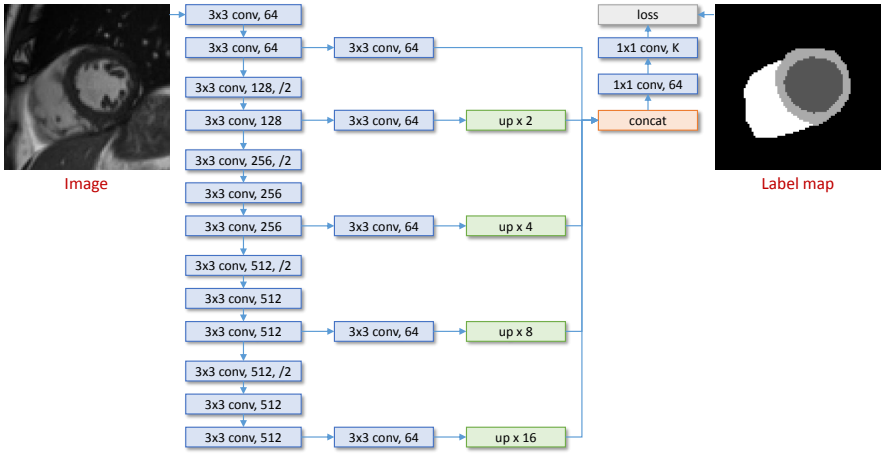


Fig. 1: The network consists of convolutional (conv), upsampling (up), concatenation (concat) and loss (loss) layers. A typical layer name "3×3 conv, 128, /2" means convolutional with 3×3 kernel, 128 output features and a stride of 2.

2.3 Network Architecture

We use a fully convolutional network architecture, as illustrated in Figure 1. It is adapted from the VGG-16 net [14] and similar to the DeepLab architecture used in [5]. A major difference is that DeepLab predicts label maps downsampled by a factor of 8, whereas our network predicts label maps in the original resolution. In the network, each convolutional layer is followed by batch normalisation and ReLu, except the last one, which is followed by the softmax function. After every two or three convolutional layers, a stride of 2 is used in the convolution to downsample the feature map so as to learn features at a more global scale. Feature maps learnt at different scales are upsampled using bilinear interpolation back to the original resolution, then combined using a concatenation layer. Finally, convolutional layers of a kernel size 1×1 are used to generate the softmax probability maps of K classes for pixel-wise segmentation. It has to be noted our main focus in this work is to investigate the idea of semi-supervised learning and other network architectures may also be used here for segmentation.

3 Experiments and Results

3.1 Data and Pre-Processing

Experiments were performed using short-axis cardiac MR images from the UK Biobank study, with the typical image resolution 1.8×1.8×10.0 mm^3. Due to the large spacing between short-axis slices and the possible inter-slice shift caused by respiratory motion, we use a 2D convolutional network and segment each slice separately, similar to how a human would annotate the image. A clinical expert

manually segmented the left-ventricular (LV) cavity, the myocardium and the right-ventricular (RV) cavity for 100 subjects at end-diastolic (ED) and end-systolic (ES) frames. Manual segmentation took about 20 minutes per subject, with each subject containing 10 to 12 slices at two time frames.

Out of 100, 20 subjects were used as testing set. The 20 testing subjects were also manually segmented twice by the same expert for evaluating intra-observer variability of human. The other 80 subjects and some unlabelled subjects were used for training. We experimented with two training settings: (1) 20 training subjects as labelled data (422 slices), 60 unlabelled subjects (1208 slices); (2) 80 training subjects as labelled data (1630 slices), 240 unlabelled subjects (4790 slices). For pre-processing, all training images were cropped to the same size of 224×224, intensity normalised to the range of $[0, 1]$ and randomly shuffled before being fed to the network. Intensity inhomogeneity correction was not performed.

3.2 Parameters

For SGD, a mini-batch size of 20 and a learning rate of 0.001 were used. Data augmentation was performed on-the-fly, including random translation, rotation, scaling and intensity rescaling. The parameters for CRF were $w_1 = 1$, $w_2 = 2$, $\sigma_\alpha = 0.5$, $\sigma_\beta = 1$, $\sigma_\gamma = 1$. These values were chosen by evaluating the segmentation performance on a small validation set.

To initialise semi-supervised learning, the network was trained on the labelled data in a supervised way for 500 epochs until the change of loss function was minimal. This network with CRF refinement is regarded as the baseline method for comparison. For semi-supervised learning, we performed alternate optimisation for 3 iterations, with 100 epochs for each iteration. We found the performance improvement after 3 iterations became negligible. We tested two values, 0.5 and 1.0, for the weight λ in the unlabelled data cross-entropy term, and found $\lambda = 1.0$ performed slightly better so adopted this value.

The method was implemented using Python and Theano [15]. In terms of computation time, it took about 10 hours to train the network for 100 epochs on a Nvidia Tesla K80 GPU, when 20 labelled data and 60 unlabelled data were used. It took about 35 hours to train for 100 epochs, when 80 labelled data and 240 unlabelled data were used. When the trained network was deployed, it took about 6 seconds to segment all the images slices for one subject at ED and ES.

3.3 Evaluation of Segmentation Performance

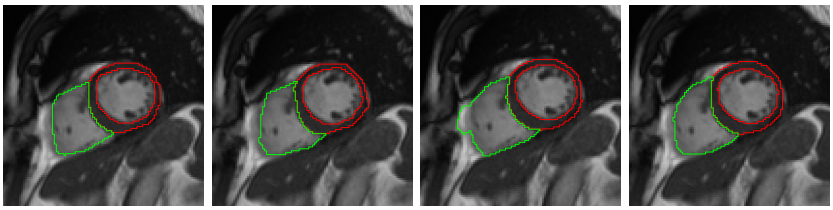
The segmentation performance was evaluated by computing the Dice overlap metric between automated segmentation and expert manual segmentation for three structures: LV cavity, LV myocardium and RV cavity. The average Dice metric of ED and ES time frames is reported.

First, we evaluate the impact of semi-supervised learning. Table 1 compares the segmentation performance between a baseline supervised learning method and the proposed semi-supervised learning method. It shows that if the same number of labelled data is used, semi-supervised learning generally improves

Table 1: Comparison of supervised and semi-supervised learning for varying number of labelled data in terms of the Dice metric.

	#labelled	#unlabelled	LV	Myo	RV
supervised	20	-	0.900	0.808	0.855
semi-super.	20	60	**0.903**	**0.822**	**0.865**
supervised	80	-	0.917	0.841	0.888
semi-super.	80	240	**0.920**	**0.848**	0.888

performance. Adding 60 unlabelled data to 20 labelled data increases the myocardium Dice from 0.808 to 0.822 ($p < 0.001$ for paired t-test) and the RV Dice from 0.855 to 0.865 ($p < 0.001$). When there are more labelled data, however, the increase becomes less prominent. Adding 240 unlabelled data to 80 labelled data only increases the myocardium Dice by 0.007 ($p < 0.001$) and there is no increase for the RV Dice. This is probably because the network can already be trained to perform well when large training data is available and thus the improvement introduced by semi-supervised learning becomes marginal.



(a) Manual 1 (b) Manual 2 (c) Multi-atlas (d) Proposed

Fig. 2: Comparison of manual segmentations of the same observer two times, multi-atlas segmentation and the proposed method. LV cavity and myocardium are delineated using red contours. RV cavity is delineated using green contours.

We also compare the performance of the proposed method to a publicly available multi-atlas based segmentation method[4] [2] and to the intra-observer variability of the human expert. For multi-atlas segmentation, we use the same 80 labelled data as atlases, using non-rigid image registration to warp the atlases [13] and cross-correlation as the similarity metric for patch-based label fusion. Figure 2 compares the manual segmentations of the same observer two times, the multi-atlas segmentation and the proposed method. Both the multi-atlas method and the proposed method achieves a good job for segmenting the LV cavity and myocardium. However, the multi-atlas method segments wrongly at the RV, probably because the weak contrast of the RV myocardium leads to less optimal target to atlas registration.

[4] `https://github.com/baiwenjia/CIMAS`

Table 2: Comparison of the proposed method to multi-atlas segmentation and intra-observer variability, in terms of Dice metric and computation time at testing stage. For intra-observer variability, manual segmentation time is reported.

	#labelled	#unlabelled	LV	Myo	RV	Time
multi-atlas	80	-	0.896	0.828	0.840	$\sim$ 5 hr
proposed	80	240	**0.920**	**0.848**	**0.888**	$\sim$ 6 sec
intra-observer	-	-	0.940	0.860	0.893	$\sim$ 20 min

Table 2 reports the Dice metric and shows that the proposed method outperforms the multi-atlas method in terms of the Dice metric for all the three structures ($p < 0.001$). Compared to intra-observer variability, the proposed method is about 1 or 2% lower in the LV or myocardium Dice but the RV Dice is close to the human performance. Table 2 also compares the computation time at testing stage and the time for manual segmentation. The multi-atlas method takes about 5 hours to segment one subject, when 80 atlases are used. The main computation cost is on the non-rigid image registration for multiple atlases. On the contrary, the proposed method only takes 6 seconds at testing stage.

4 Conclusion and Discussion

In this paper, we propose a novel, semi-supervised and network-based method for cardiac MR image segmentation. The main contribution is that we propose a semi-supervised way to train the network to address a major challenge with medical image segmentation, the limited number of training data. We have shown that the introduction of unlabelled data leads to an improvement in segmentation performance, especially when the size of the existing training set is small. We have also shown that the method outperforms a state-of-the-art multi-atlas segmentation method. Once the network is trained, it only takes a few seconds to segment one subject. Therefore, it can efficiently analyse large-scale cardiac MR image sets, such as the UK Biobank dataset, which will eventually consist of 100,000 subjects.

For future work, we are interested in improving the quality of automated segmentations for unlabelled data. A drawback with the current approach is if an error or bias (over- or under-segmentation) occurs in the initial segmentation of the unlabelled data, the error will be learnt by the network during the following iterations. This negative effect is currently alleviated by refining the segmentation using CRF and by assuming that the majority of the automated segmentations are correct so the average gradient that the network learns is still roughly correct. We are interested in exploring using level-sets to refine the segmentation as in [9] and correcting the segmentation with minimal manual intervention. Another interesting direction is to incorporate segmentation uncertainty estimation into semi-supervised learning.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 18545. This work is supported by EPSRC programme Grant (EP/P001009/1). H.S. is supported by a Research Fellowship from the Uehara Memorial Foundation. P.M.M. gratefully acknowledges support from the Imperial College Healthcare Trust Biomedical Research Centre, the EPSRC Centre for Mathematics in Precision Healthcare and the MRC.

References

1. Avendi, M., et al.: A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. Medical Image Analysis 30, 108–119 (2016)
2. Bai, W., et al.: A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: Application to cardiac MR images. IEEE Transactions on Medical Imaging 32(7), 1302–1315 (2013)
3. Brosch, T., et al.: Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE Transactions on Medical Imaging 35(5), 1229–1239 (2016)
4. Chen, H., et al.: DCAN: Deep contour-aware networks for accurate gland segmentation. In: CVPR. pp. 2487–2496 (2016)
5. Chen, L., et al.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv:1606.00915 (2016)
6. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with gaussian edge potentials. In: NIPS. pp. 1–9 (2011)
7. Lin, D., et al.: ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation. In: CVPR. pp. 3159–3167 (2016)
8. Long, J., et al.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
9. Ngo, T., et al.: Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. Medical Image Analysis 35, 159–171 (2017)
10. Papandreou, G., et al.: Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: ICCV. pp. 1742–1750 (2015)
11. Poudel, R., et al.: Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. arXiv:1608.03974 (2016)
12. Ronneberger, O., et al.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
13. Rueckert, D., et al.: Nonrigid registration using free-form deformations: application to breast MR images. IEEE Transactions on Medical Imaging 18(8), 712–21 (1999)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. pp. 1–14 (2015)
15. Theano Development Team: Theano: A Python framework for fast computation of mathematical expressions. arXiv:1605.02688 (2016)
16. Tran, P.: A fully convolutional neural network for cardiac segmentation in short-axis MRI. arXiv:1604.00494 (2016)
17. Yang, H., et al.: Deep fusion net for multi-atlas segmentation: Application to cardiac MR Images. In: MICCAI. pp. 521–528 (2016)