Final Report: Automated Semantic Segmentation of Volumetric Cardiovascular Features and Disease Assessment

Tony Lindsey[1,3], Xiao Lu[1] and Mojtaba Tefagh[2]

[1]Department of Biomedical Informatics, Stanford University, Palo Alto, CA
[2]Department of Electrical Engineering, Stanford University, Palo Alto, CA
[3]NASA Ames Research Center, Mountain View, CA

Abstract

Cardiac magnetic resonance imaging provides high spatial resolution, enabling improved extraction of important functional and morphological features for cardiovascular disease staging. Segmentation of the heart in cardiac cine sequencing is clinically used for cardiac function assessment. We present a method that curtails the expense and observer bias of manual cardiac evaluation by combining semantic segmentation and disease classification into a fully automatic processing pipeline. The initial processing element consists of a robust dilated convolutional neural network architecture for voxel-wise segmentation of the myocardium and ventricular cavities. The resulting comprehensive volumetric feature matrix captures diagnostic clinical procedure data and is used by the final processing element to model a cardiac pathology classifier. Our approach evaluated anonymized cardiac images from a training dataset of 100 patients (4 pathology groups, 1 healthy group, 20 patients per group) examined at the University Hospital of Dijon. We achieved top average Dice index scores of 0.910, 0.913, 0.851 for structure segmentation of the left ventricle (LV), right ventricle (RV) and myocardium respectively. A 5-ary pathology classification accuracy of 85% was recorded on an independent test set using a trained model. The performance results demonstrate potential for advanced machine learning methods to deliver accurate, efficient and reproducible cardiac pathological assessment.

1. Motivation

Cardiac genome expression resulting in molecular, cellular and interstitial changes is generally accepted as a determinant in the progressive course of heart failure [1]. Clinical manifestations of ventricular remodeling include changes in heart size, mass, geometry, function and regional wall motion. Magnetic resonance imaging (MRI) has successfully been used as a noninvasive cardiac health evaluation tool for timely adverse event monitoring. Cardiovascular MRI is a tomographic, nonionizing technique clinically used for assessing expansion of infarcted segments, late wall thinning of infarcted regions, LV volumes, distortion of LV shape and compensatory hypertrophy of non-infarcted myocardium tissue [2]. In today's modern clinical setting the tremendous benefits of comprehensive quantitative measurements remain largely untapped due to assessment costs, interpretation variability, and lack of reproducible medical scenarios. Therefore, accurate automatic approaches that differentiate multiple coronary event features with computer-aided diagnosis are desirable assets for a large spectrum of cardiovascular diseases [3]. Computers are becoming increasingly capable of supplementing and enhancing medical imaging, morphologic tissue information, and molecular classification with diagnostic, prognostic, and theragnostic predictions. Thus, rich opportunity exists for computer-aided interpretation and multi-modality integration to provide new insights into myocardial disease mechanism, severity and prognosis [4].

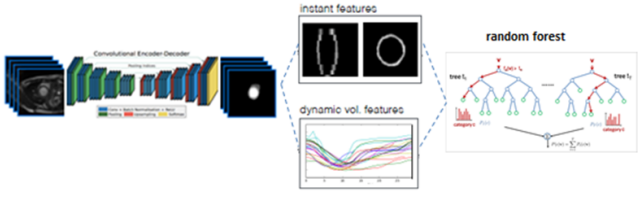

Figure 1: Proposed pipeline - Voxel-wise segmentation quantifies volumetric cardiac features which are fed into a random forest ensemble classifier for pathology assessment.

In this project report, we present an approach for automatic classification of cardiac pathologies associated with

ventricular remodeling. Deep learning algorithms enabled cardiac volume dynamics quantification and subsequent training of classifiers that predict coronary disease. In particular, based on multi-structure segmentation for each slice increment of the cardiac MRI, we extracted domain-specific features that were motivated by a cardiologist's workflow, to then train an ensemble classifier for disease prediction see Fig. 1. Classification of five categories (normal case, heart failure with infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, abnormal right ventricle) for 20 patients per group was performed at each cine-MRI time step see Fig. 2. We evaluated our cardiac health interpretation methods as part of an international automated cardiac diagnosis challenge (ACDC) [5].

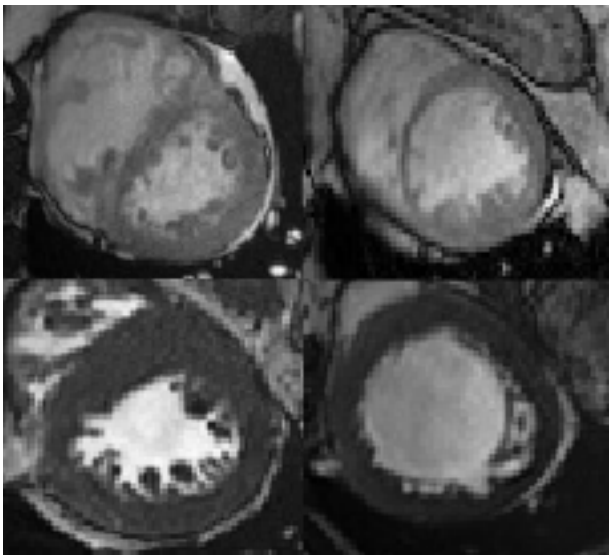

Figure 2: Representative cardiac images for various pathologies, from top left to bottom right: abnormal right ventricle, myocardial infarction, hypertrophic cardiomyopathy, dilated cardiomyopathy.

2. Method

Pre-trained convolutional neural networks use activations available before the last fully connected layer as the feature vector for a machine learning classifier. We trained several such models, obtained from the ImageNet object recognition database, to establish a baseline for 5-ary classification accuracy of our cardiac image dataset. Motivated by less than stellar test set accuracy results, we explored alternative learning methods. Semantic segmentation is a dense prediction problem that is structurally different from image classification. We employed a dilated convolutional network module that systematically aggregates multiscale contextual information without resolution loss or coverage. The architecture consists of a rectangular prism of convolutional layers, without pooling or subsampling, that supports exponential expansion of the receptive field. In the one-dimensional case, we define

$$y[i] = \sum_{k=1}^{K} x[i + rk]w[k]$$

where, x is a 1D input, y an output signal and w a filter of size K.The rate parameter r corresponds to the dilation factor. The dilated convolution operator can reuse the weights from the filters that were trained on downsampled feature maps by sampling the unreduced feature maps with an appropriate rate.

We developed a state-of-art deep learning system for semantic image segmentation using tensorflow to express arbitrary computation as a graph of dataflows. The dilated autoencoder architecture described above was implemented as a fully convolutional variant of ResNet-101. The trainable parameters in the autoencoder were optimized using a loss function based on the Dice similarity coefficient [6]. This partly corrects for class imbalance in the voxel labels. A soft Dice loss was used,

$$Dice_c = \frac{sum_i^N R_c(i)A_c(i)}{sum_i^N R_c(i) + sum_i^N A_c(i)}$$

where R_c is the binary reference image for class c, A_c is the probability map for class c, N is the number of voxels, and $Dice_c$ is the Dice coefficient for class c. This coefficient was computed for all eight classes (BG_{ED}, LV_{ED}, RV_{ED}, MYO_{ED}, BG_{ES}, LV_{ES}, RV_{ES}, MYO_{ES}), i.e. background, left & right ventricles, myocardium mass for both end diastolic and systolic portions of the cardiac cycle. The volumes were averaged to ensure joint optimization for all classes.

Stochastic gradient descent with momentum (0.9), weight decay (5^{-4}) and L2-regularization on the autoencoder parameters were used as initial training parameters. Pairs of ED and ES images were processed between 4,000 and 6,000 iterations for model generation. In each iteration, the network was optimized with mini-batch containing 10 images and default learning rate of 2.5^{-4}. Cardiac features for each patient were subsequently classified with a random forest ensemble learning algorithm.

2.1. Preprocessing

Cardiac segmentation masks, one provided for each image, were converted from 3D RGB vectors to a 1D label. Images were normalized by subtracting the minimum pixel intensity from each channel and dividing by the mean pixel intensity to represent pixels in the range 0 to 1. A filtering method located and removed from our dataset defective images with non-interpretable pixel color distributions. The Pearson correlation coefficient measure was applied to our

volumetric feature matrix constructed from voxel segmentation class mappings. Any pair-wise correlation exceeding a conservative cutoff threshold of 0.9 was removed from the dataset. Finally, we z-score normalized our cardiac feature data by determining the probability of a score occurring within our normal distribution with respect to the mean.

2.2. Data Augmentation

We augmented the number of images in real-time to improve network localization capability and reduce overfitting. During each epoch, a random augmentation of images that preserve collinearity and distance ratios was performed. We implemented random padding with zeros, zoom, rolling and rotation. The training dataset was further augmented by adding new contrast limited adaptive histogram equalization filtered images randomly selected from each pathology category. There wasn't a noticeable improvement in segmentation performance using these synthetic techniques. One possible explanation is that significantly larger amounts of augmentation are required for improvement. However, it should be noted that the gap between training error % and test error % steadily decreased as the amount of augmentation was increased, thus indicating a reduction in overfitting.

2.3. Feature Selection

The volumetric feature matrix generated by semantic segmentation contains 20 records for each of 5 classes and 16 quantitative covariates including weight, height, ventricular volumes, calculated ejection fractions, myocardium mass and permuted ratios for both end diastolic and systolic time points of the cardiac cycle. Feature selection determined a subset of maximally varying predictors for classifier model construction. Random forests ranked feature importance based on amount of mean Gini impurity decrease see Fig 3.

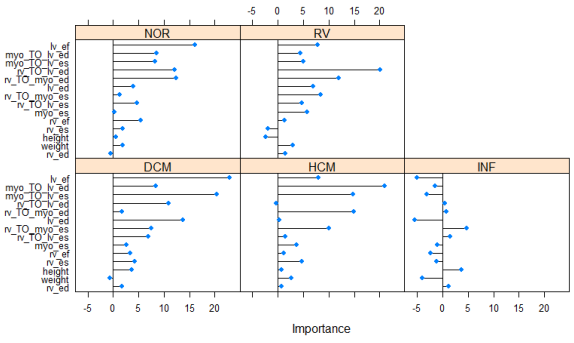

Figure 3: Most important features ranked by Gini impurity decrease.

3. Experiments and Results

A deep learning GPU training system (DIGITS) with pre-trained convolutional neural networks for image classification enabled rapid prototype training, real-time performance monitoring and visualizations. Our dataset contained 934 parasternal short axis cine-MRI image slices partitioned into 50% training, 25% validation and 25% test subsets. The GoogLeNet architecture using the stochastic gradient descent (SGD) optimizer and transfer learning, which retains initial model weights and extracts image features via a final network layer, performed the best and was chosen as our baseline classifier (see Fig 4).

Rapid Prototyping Results - Transfer Learning						
5-ary Classifier	Solver	Learning Rate	Policy	Validation Accuracy%	Test Set 95% CI	Median Specificity
AlexNet	RMSProp	1e-4	Exponential Decay	98.96	(0.319, 0.461)	84.71
AlexNet	SGD	1e-3	(17,33) Step Down	98.44	(0.384, 0.529)	84.88
AlexNet	Adam	1e-4	(25,43) Step Down	98.96	(0.329, 0.472)	85.63
AlexNet	AdaGrad	1e-4	(17,33) Step Down	92.71	(0.271, 0.408)	82.08
AlexNet	NAG	1e-3	Exponential Decay	90.10	(0.334, 0.477)	83.43
GoogLeNet	Adam	1e-4	(33,66) Step Down	94.27	(0.379, 0.524)	86.06
GoogLeNet	SGD	1e-3	(33,66) Step Down	95.31	(0.502, 0.646)	91.67
GoogLeNet	AdaGrad	1e-3	(33,66) Step Down	92.71	(0.310, 0.451)	83.14
GoogLeNet	NAG	1e-3	(33,66) Step Down	91.67	(0.379, 0.524)	85.63
GoogLeNet	RMSProp	1e-3	(50,85) Step Down	95.31	(0.415, 0.560)	86.42

Figure 4: Hyperparameter optimization of the ACDC dataset using pretrained models with transfer learning.

We targeted comparison of this result with a pipeline consisting of semantic segmentation followed by classical machine learning of volumetric features. The initial processing model was trained on a mini-batch of images and corresponding ground truth masks with the softmax classifier at the top. During training, the masks were downsampled to match the size of the output from the network. Bilinear upsampling was applied during inference to acquire equivalent output and input dimension sizes. The final segmentation mask was computed using argmax over the logits. The SGD optimizer with momentum was used for model training to predict a 3D voxel-wise label map $T_v \in \{0, 1, 2, 3\}$. The labels were tallied to quantify background tissues, left ventricle endocardium, epicardium and right endocardium volumes respectively (see Table 1). Ventricular ejection fraction and all volumetric ratio permutations, for both end diastolic and systolic time points, were computed from these basic quantities and averaged over all minibatches.

	RV	MYO	LV
ED	0.935	0.862	0.932
ES	0.890	0.840	0.887
Average	0.913	0.851	0.910

Table 1: Dice scores for the 3D voxel-wise label map prediction using the semantic segmentation SGD optimizer.

The resulting feature matrix contained 10 averaged car-

diac volumes, 2 ejection fractions plus height and weight for each of the 100 patients. Optimal features were selected via random forest importance ranking and a pruned decision tree (see Fig 5). The later was used for interpretability

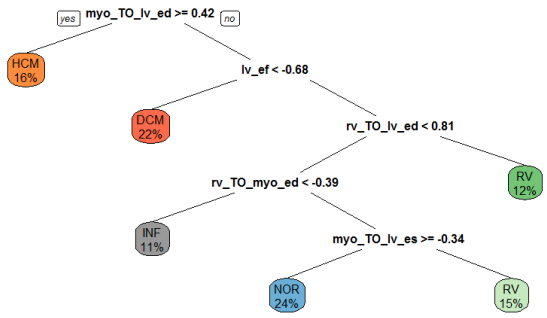

Figure 5: Pruned decision tree for optimal feature selection.

and to compare volumetric thresholds with clinical guidelines. Our dataset was partitioned into train (65%), holdout (15%) and test (20%). A random forest bagging model was trained with the top 7 feature selected attributes using 5-fold cross validation, and hyperparameters tuned on a holdout set from the same distribution. The best performance accuracy of 85% was recorded by this model when applied to the random test set partition (see Fig 6 and Table 2).

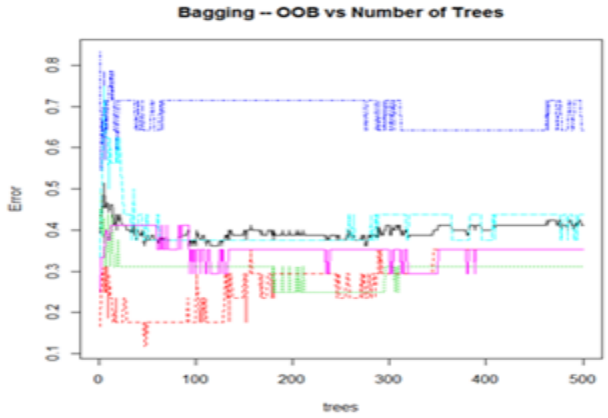

Figure 6: Computed out-of-bag error performance metric over 500 trees. Blue line is prediction accuracy.

Accuracy	Sensitivity	PPV	Specificity	NPV
85.0%	86.6%	87.7%	96.1%	96.4%

Table 2: Random forest classifier model test set performance metrics summary table.

3.1. Error Analysis

We performed error analysis during each of the major processing elements within our fully automated pipeline. The semantic segmentation element was analyzed by tuning weight decay, batch size, learning rate and momentum strength. The difference in Dice coefficient similarity error was computed over separate runs for each hyperparameter. Learning rate and optimizer type, Adam outperformed SGD with momentum, had the largest impact in reducing error according to the analysis. The choice of selected features substantially affected misclassification error during the second processing pipeline element. After data preprocessing and feature selection, the difference in classification accuracy varied by as much as 10% depending on the model features chosen.

4. Conclusion

In this project report we presented a fully automatic processing pipeline for pathology classification on cardiac cine-MRI. The system achieved higher accuracy on 5-ary classification compared to a baseline robust pretrained CNN architecture with transfer learning. Our image pipeline includes semantic segmentation with dilated convolutions, volumetric feature extraction and random forest model classification. We trained, hyperparameter tuned and summarized test set partition performance metrics for a 5-ary classifier (see Table 2). Rigorous ablative analysis revealed that learning rate, optimizer type and feature selection were the greatest contributors to overall improved pipeline processing element performance. Data augmentation, weight decay and momentum strength variance failed to influence semantic segmentation performance by an appreciable amount according to our experiments.

References

[1] J N Cohn, R Ferrari, and N Sharpe. Cardiac remodeling concepts and clinical implications: A consensus paper from an international forum on cardiac remodeling. *JACC*, 35:569–582, 2000.

[2] Maythem Saeed, Tu Anh Van, Roland Krug, Steven W Hetts, and Mark W Wilson. Cardiac mr imaging: Current status and future direction. *Cardiovascular Diagnosis and Therapy*, pages 290–310, 2015.

[3] K Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *CMIG*, 31 (4):198–211, 2008.

[4] Scott Ritter and Kenneth B Margulies. Emerging tools for computer-aided diagnosis and prognostication. *Journal of Clinical Trials*, 4(2):e120, 2014.

[5] Jacques Dubloit. Automated cardiac diagnosis challenge. *https://www.creatis.insalyon.fr/Challenge/acdc*, 2017.

[6] F Milletari, N Navab, and S Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *3D Vision*, pages 565–571, 2016.