BAYESIAN OPTIMIZATION OF 2D ECHOCARDIOGRAPHY SEGMENTATION

*Tung Tran, Joshua V. Stough**

Bucknell University
Computer Science
Lewisburg, PA

Xiaoyan Zhang, Christopher M. Haggerty

Geisinger
Translational Data Science and Informatics
Danville, PA

ABSTRACT

Bayesian Optimization (BO) is a well-studied hyperparameter tuning technique that is more efficient than grid search for high-cost, high-parameter machine learning problems. Echocardiography is a ubiquitous modality for evaluating heart structure and function in cardiology. In this work, we use BO to optimize the architectural and training-related hyperparameters of a previously published deep fully convolutional neural network model for multi-structure segmentation in echocardiography. In a fair comparison, the resulting model outperforms this recent state of the art on the annotated CAMUS dataset in both apical two- and four-chamber echo views. We report mean Dice overlaps of 0.95, 0.96, and 0.93 on left ventricular endocardium, epicardium, and left atrium respectively. We also observe significant improvement in derived clinical indices, including smaller median absolute errors for left ventricular end-diastolic volume (4.9ml vs. 6.7), end-systolic volume (3.1ml vs. 5.2), and ejection fraction (2.6% vs. 3.7); and much tighter limits of agreement, which were already within inter-rater variability for non-contrast echo. While these results demonstrate the benefits of BO for echocardiography segmentation over even a recent state-of-the-art framework, they must still be validated against large-scale independent clinical data.

***Index Terms*—** Echocardiography, Segmentation, Bayesian Optimization

1. INTRODUCTION

Echocardiography is the most frequently used non-invasive imaging modality for the quantification of heart structure and function [1]. However, such quantification requires precise annotations of the key cardiac structures, including the left ventricle endocardium (LV_{endo}), epicardium (LV_{epi}), and left atrium (LA). Manual annotation is high-cost and prone to inter-rater variability, which has motivated the development of automated segmentation methods, most recently using convolutional neural networks [2].

*Corresponding author: joshua.stough@bucknell.edu

The U-net architecture [3] is well-known for its capacity to learn abstract features and complete a wide range of medical image segmentation tasks. Within the domain of echocardiography segmentation, LeClerc et al. [4] have demonstrated the relative efficacy of U-net models on their large annotated CAMUS dataset.

Among others who have followed this work [5], Stough et al. [6] developed a U-net variant and claimed state-of-the-art results on CAMUS. However, the hyperparameters used in this framework were found heuristically and through small-scale grid search, given the high cost of training such a deep fully convolutional neural network. In this work we leverage Bayesian Optimization (BO) and distributed computing to efficiently search the hyperparameter space, resulting in significant improvements in multi-structure echocardiography segmentation.

Bayesian optimization provides both theoretical guarantees in optimizing expensive black-box functions [7], and has been proven to work well in practice [8]. BO is especially valuable when the observed objective is gradient-free, expensive to evaluate, and have fewer than 20 parameters to be optimized [9]. BO utilizes Gaussian Processes to gain insights into the objective function's probability distribution, and uses an acquisition function on top of this distribution to determine the next set of hyperparameters to be tested.

2. METHODS

The model developed in [6] is an encoder-decoder style network incorporating additive skip connections and group normalization (Fig. 1). We optimize on a variety of architectural and training-related hyperparameters, as shown in Table 1.

We then observe a noisy objective function $f(\mathbf{x})$ that evaluates the performance of the model on a given candidate, or set of hyperparameters $\mathbf{x} \in H$: $y = f(\mathbf{x}) + \epsilon$, where y is the observed value, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, in which σ is given. The known performances on initial candidates will be fed into BO to find candidates $\mathbf{x}$ that maximize $f(\mathbf{x}) : \text{argmax}_x\ f(\mathbf{x})$.

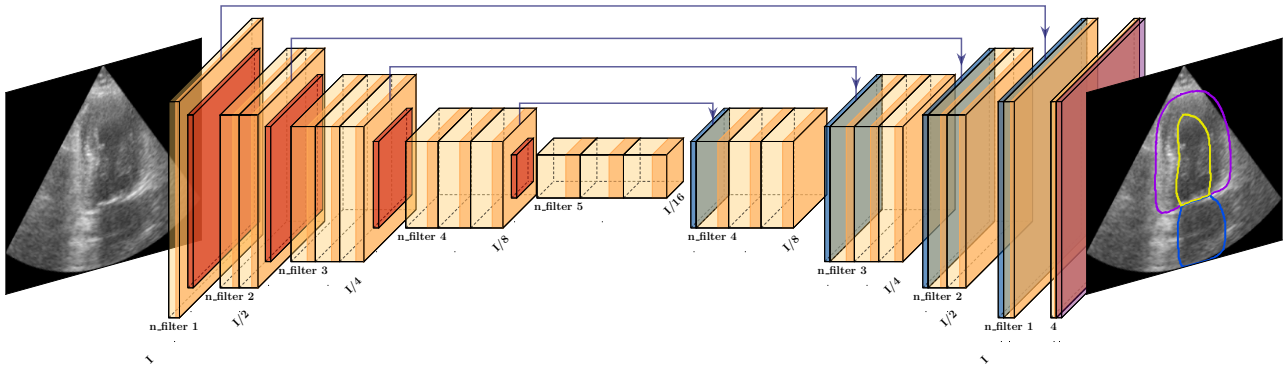

Fig. 1: CNN architecture for multi-structure segmentation of echocardiography [6]. Each block incorporates chained convolution, group normalization, and non-linear activation. During BO optimization, architectural hyperparameters include the number of output filters (n_filter), and whether to use group or batch normalization.

	Hyperparameter	Lower bound	Upper bound
Architectural	n_filters #1	16	32
	n_filters #2	57	128
	n_filters #3	153	256
	n_filters #4	281	512
	n_filters #5	537	1024
	group vs batchnorm	0	1
	number of groups	2	24
Training-related	log learning rate	-9	2
	log weight decay	-9	-2
	batch size	2	10

Table 1: Table showing hyperparameter space to be optimized and their description.

2.1. Objective Functions

We score the goodness of candidate $\mathbf{x}$ through mean test loss in a 5-fold cross validation setting with N total images. Let E_j be test fold j, $h^j_{\mathbf{x}}$ be the model trained using $\mathbf{x}$ with fold j, I be an image in E_j, and L be the Cross Entropy loss from the resulting segmentation $h^j_{\mathbf{x}}(I)$. Since we're maximizing the objective function, we invert this test loss. We call this objective Mean Validation loss (MV), which is observed through:

$$y = o_{MV}(\mathbf{x}) = 1 - \frac{1}{N}\sum_{j=1}^{5}\sum_{I\in E_j} L[h^j_{\mathbf{x}}(I)]$$

2.2. Bayesian Optimization

Implemented with BoTorch [10], we utilize heteroskedastic Gaussian Processes (GP), which wrap another GP to model changing objective noise. Beside the objective, we also also model a GPU constraint with a fixed noise GP. All GPs used the default Matérn 5/2 kernel. Based on these GPs, the acquisition function Noisy Expected Improvement (NEI) [11] is applied on both objective and constraint in scoring candidates.

In this setting, we assume that the observed objective value is corrupted by white noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is set to be the variance of validation loss. Suppose we use Expected Improvement (EI) on the observed objective, we need to know f^*, or known best true objective, which is not apparent. This is due to our assumption that only the noisy observed value y is known. To deal with this drawback, Letham et al. [11] sample multiple random "imaginary" instances of $[f(\mathbf{x}_1), \cdots, f(\mathbf{x}_n)] \mid D_f \sim \mathcal{N}(\mu_f, \Sigma_f)$, where $D_f = \{\mathbf{x}_i, y_i, \sigma_i^2\}_{i=1}^n$ are known values and uncertainty estimates of the objective. Each instance $[f(\mathbf{x}_1), \cdots, f(\mathbf{x}_n)]$ is then used to fit a predetermined k number of different GP models $\mathcal{M}_1, \cdots, \mathcal{M}_k$, which offer insights into different stochastic scenarios. Additionally, a constraint is added to keep GPU usage in check, thus ensuring the worker machines are capable of handling the objective evaluation process. Similar to the objective, multiple "imaginary" instances of constraint are sampled and fitted to multiple GPs $\mathcal{M}'_1, \cdots, \mathcal{M}'_k$. Each output on a potential candidate of the objective GPs $f(\mathbf{x})$ and constraint GPs $c(\mathbf{x})$ is combined as follows:

$$Weighted\ Objective\ of\ \mathbf{x} = f(\mathbf{x})\Big(1 - \frac{1}{1+e^{-c(\mathbf{x})}}\Big)$$

We then calculate EI for the weighted objective of each pair $[\mathcal{M}_i, \mathcal{M}'_i]$. After that, we average all EI values of a particular $\mathbf{x}$ to get the final NEI value.

3. EXPERIMENTAL RESULTS

The CAMUS dataset consists of 450 patients, two (AP2/AP4) views per patient, and two annotated (ED/ES) phases per view, totalling 1800 echocardiographic frames and corresponding label masks (LV_{endo}, LV_{epi}, LA, background). Additional information for each patient includes age, sex, and reported ED/ES LV_{endo} volumes and ejection fraction (EF), along with the observed image quality for each view.

We initially leave out ∼30% (N = 136) of patients for final evaluation. The remaining 70% (N = 314) are then partitioned for 5-fold cross validation with training, validation, and test splits. As in [4], all splits are stratified on both patient EF range ($\leq 45\%$, $\geq 55\%$, else) and reported image quality.

	Label / Score	D_{mean}	ED $D_{mean} \pm \sigma$	ES $D_{mean} \pm \sigma$
AP2	LV_{endo}	**0.950**(0.921)	**0.960**(0.937) ±**0.015**(0.034)	**0.941**(0.905) ±**0.033**(0.059)
	LV_{epi}	**0.966**(0.950)	**0.968**(0.953) ±**0.012**(0.024)	**0.964**(0.947) ±**0.014**(0.028)
	LA	**0.934**(0.879)	**0.929**(0.857) ± **0.038**(0.131)	**0.939**(0.901) ±**0.028**(0.081)
AP4	LV_{endo}	**0.954**(0.935)	**0.962**(0.946) ±**0.016**(0.023)	**0.945**(0.924) ±**0.027**(0.039)
	LV_{epi}	**0.969**(0.958)	**0.971**(0.961) ±**0.010**(0.015)	**0.967**(0.955) ±**0.011**(0.016)
	LA	**0.935**(0.910)	**0.924**(0.890) ±**0.042**(0.071)	**0.947**(0.931) ±**0.025**(0.032)

Table 2: Dice overlaps for the MV optimal candidate shown against the published candidate [6] (in parentheses) on the evaluation set (136 patients).

In training candidate $\mathbf{x}$, the model weights are saved according to performance on the validation splits, while the BO objective is computed against the associated test splits. As in [6], on-the-fly data augmentation is used, including intensity windowing, slight rotation about the transducer point, and additive Gaussian noise. Training is continued to convergence using a standard scheduler that reduces learning rate on a plateau in validation loss.

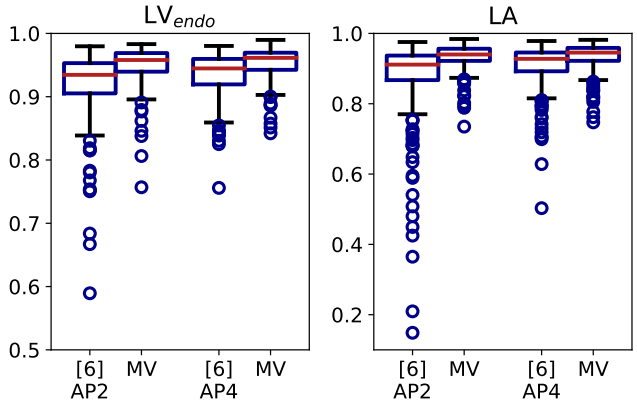

Fig. 2: Box plots of LV_{endo} and LA Dice performance. The MV optimum shows significant improvement over [6] in median and outlier performance on both structures.

We run BO asyncronously in a distributed environment in which each node runs a single GeForce RTX 2080 Ti. We run 100 candidates for each of AP2 and AP4 views, resulting in two best candidates. In segmenting the 30% evaluation set with a particular optimal candidate, we accumulate the outputs from all five folds to obtain an ensemble result.

Table 2 directly compares the MV optimal candidate to the previously published candidate [6] through Dice. The MV optimal candidate shows improved agreement with manual annotation for all views, structures, and phases. Figure 2 provides additional context, showing greatly improved outlier performance of the MV optimal candidate, particularly for the relatively small and thus difficult left atrium. Figure 3 shows representative median and poor segmentation results.

We further derive the LV volume measurements and EF using the Simpsons modified biplane method and the corresponding AP2 and AP4 views for each patient. Compared to

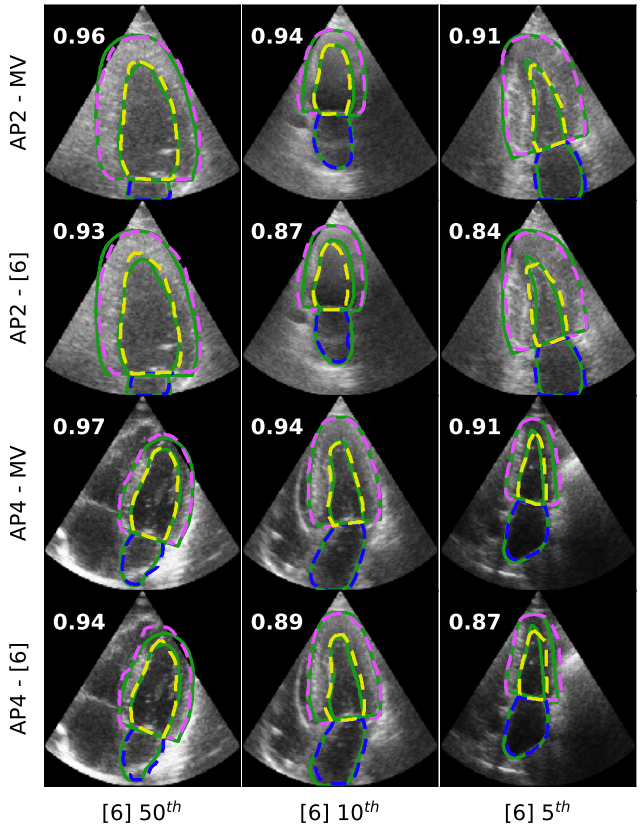

Fig. 3: Segmentation performance of MV and published [6] candidates on 50^{th}, 10^{th} and 5^{th} LV_{endo} Dice score percentile for AP2 view of [6]. Green contour denotes manual segmentation.

the previously published candidate [6], the MV optimal candidate obtains significantly smaller biases and narrower limits of agreement with reported clinical indices. We report median absolute errors of 4.9ml (vs. 6.7) for ED volume, 3.1ml (vs. 5.2) for ES volume, and 2.6% (vs. 3.7%) for EF (Fig. 4).

4. CONCLUSION

In this work, we have utilized Bayesian Optimization to significantly improve upon recent state-of-the-art multi-structure segmentation in echocardiography. In a fair comparison, the optimal candidate boasts tighter limits of agreement and

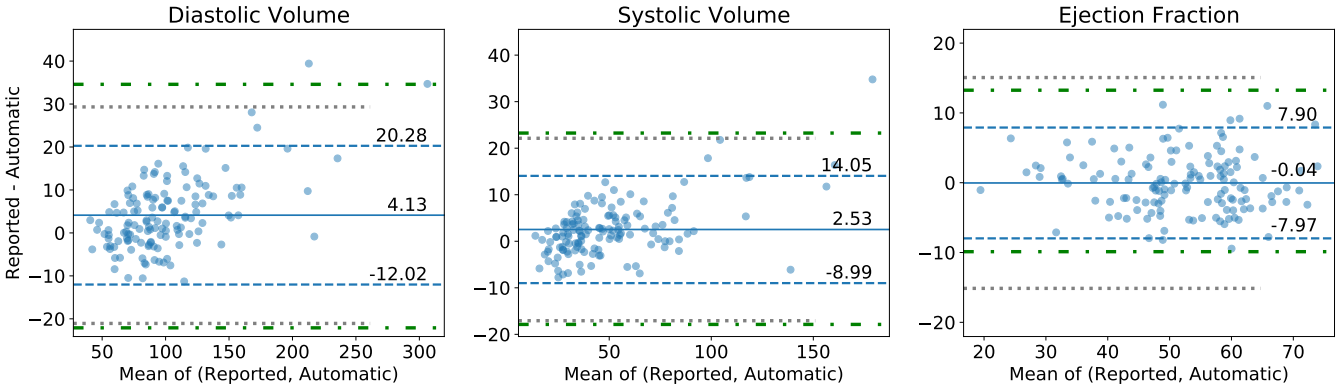

Fig. 4: Bland-Altman plots comparing the optimal MV candidate against manual annotations for LV_{endo} volumes and ejection fraction on the evaluation set, in blue. Additional limits of agreement are shown for both the published candidate [6] on the same data (green, dashdotted) and previously reported inter-observer variability for 2D echocardiography [12] (gray, dotted). Compared to [6], we report bias$\pm 1.96\sigma$ of $4.13\text{mL}\pm 16.15$ (vs $6.25\text{mL}\pm 28.34$) for ED volume, $2.53\text{mL}\pm 11.52$ (vs $2.69\text{mL}\pm 20.56$) for ES volume, and $-0.04\%\pm 7.94$ (vs $1.69\%\pm 11.57$) for EF.

vastly improved outlier performance. The potential absence of catastrophic failures makes more feasible limited auditing in future large-scale historical analyses.

Relative to [6], the optimal hyperparameter set is characterized by smaller learning and decay rates along with deeper feature maps and thus more trainable parameters (40M vs 13M). One resulting concern is that there may be overfitting to the relatively consistent and artifact-free CAMUS images, though care was taken to separate training, BO scoring, and evaluation data. We must assess the optimal model's generalizability on independent clinical datasets [13], which may feature larger variability in acquisition settings and image quality, and even burned-in view and patient information that is more common in the clinic.

5. REFERENCES

[1] Roberto M Lang, Luigi P Badano, Victor Mor-Avi, Jonathan Afilalo, et al., "Recommendations for cardiac chamber quantification by echocardiography in adults...," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

[2] Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, et al., "Fully automated echocardiogram interpretation in clinical practice," *Circulation*, vol. 136, no. 16, pp. 1623–1635, 2018, https://doi.org/10.1161/CIRCULATIONAHA.118.034338.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[4] Sarah Leclerc, Erik Smistad, João Pedrosa, et al., "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE Trans Med Imaging*, 2019, https://doi.org/10.1109/TMI.2019.2900516.

[5] Hongrong Wei, Heng Cao, et al., "Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape," in *MICCAI*, 2020, http://www.digitalimaginggroup.ca/members/Shuo/temporal.pdf.

[6] Joshua V Stough, Sushravya Raghunath, et al., "Left ventricular and atrial segmentation of 2d echocardiography with convolutional neural networks," in *SPIE Medical Imaging 2020: Image Processing*, 2020, https://doi.org/10.1117/12.2547375.

[7] Adam D Bull, "Convergence rates of efficient global optimization algorithms.," *Journal of Machine Learning Research*, vol. 12, no. 10, 2011.

[8] Jasper Snoek, Hugo Larochelle, and Ryan P Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[9] Peter I Frazier, "A tutorial on bayesian optimization," *arXiv*, 2018, https://arxiv.org/abs/1807.02811.

[10] Maximilian Balandat, Brian Karrer, Daniel R Jiang, et al., "Botorch: Programmable bayesian optimization in pytorch," *arXiv preprint arXiv:1910.06403*, 2019.

[11] Benjamin Letham, Brian Karrer, et al., "Constrained bayesian optimization with noisy experiments," *Bayesian Analysis*, vol. 14, no. 2, pp. 495–519, 2019.

[12] Peter W. Wood, Jonathan B. Choy, Navin C. Nanda, and Harald Becher, "Left ventricular ejection fraction and volumes: It depends on the imaging method," *Echocardiography*, vol. 31, no. 1, pp. 87–100, 2014, https://doi.org/10.1111/echo.12331.

[13] David Ouyang, Bryan He, et al., "Video-based ai for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020, https://www.nature.com/articles/s41586-020-2145-8.