Fully convolutional regression network for accurate detection of measurement points

Michal Sofka, Fausto Milletari, Jimmy Jia, and Alex Rothberg

4Catalyzer

Abstract. Accurate automatic detection of measurement points in ultrasound video sequences is challenging due to noise, shadows, anatomical differences, and scan plane variation. This paper proposes to address these challenges by a Fully Convolutional Neural Network (FCN) trained to regress the point locations. The series of convolutional and pooling layers is followed by a collection of upsampling and convolutional layers with feature forwarding from the earlier layers. The final location estimates are produced by computing the center of mass of the regression maps in the last layer. The temporal consistency of the estimates is achieved by a Long Short-Term memory cells which processes several previous frames in order to refine the estimate in the current frame. The results on automatic measurement of left ventricle in parasternal long axis view of the heart show detection errors below 5% of the measurement line which is within inter-observer variability.

1 Introduction

Regression modeling is an approach for describing a relationship between an independent variable and one or more dependent variables. In machine learning, this relationship is described by a function whose parameters are learned from training examples. In deep learning models, this function is a composition of logistic (sigmoid), hyperbolic tangent, or more recently rectified linear functions at each layer of the network. In many applications, the function learns a mapping between input image patches and a continuous prediction variable.

Regression has been used to detect organ [4] or landmark locations in images [2], visually track objects and features [8], and estimate body poses [14,13]. The deep learning approaches have outperformed previous techniques especially when a large annotated training data set is available. The proposed architectures used cascade of regressors [14], refinement localization stages [11,4], and combining cues from multiple landmarks [9] to localize landmarks. In medical images, the requirements on accurate localization are high since the landmarks or measurement points are used to help in diagnosis. When tracking the measurements in video sequences, the points must be accurately detected in each frame while ensuring temporal consistency of the detections.

This paper proposes a Fully Convolutional Network for accurate localization of anatomical measurement points in video sequences. The advantage of the

Fully Convolutional Network is that the responses from multiple windows covering the input image can be computed in a single step. The network is trained end-to-end and outputs the locations of the points. The regressed locations are mapped at the last convolutional layer into a location using a new center-of-mass layer which computes mean position of the predictions. This approach has advantages to regressing heatmaps, since the predictions can have subpixel values and the regression objective can penalize measurement length differences from the ground truth. The temporal consistency of the measurements is improved by Convolutional Long Short-term Memory (CLSTM) cells which process the feature maps from several previous frames and produce updated features for the current frame in order to refine the estimate. The evaluation is fast to process each frame of a video sequence at near frame rate speeds.

2 Related Work

Regression forests were previously trained to predict locations and sizes of anatomical structures [2]. The initial estimates were refined via Hough regression forests [3] or local regressors guided by probabilistic atlas [4]. Automatic X-ray landmark detection in [1] estimated landmark positions via a data-driven non-convex optimization method while considering geometric constraints defined by relative positions.

Recently, deep learning approaches have been shown to effectively train representations that outperform traditional methods [7,10]. Multiple landmark localization in [9] was achieved by combining local appearance each landmark and spatial configuration of all other landmarks. The final combined heatmap of likely landmark location was obtained from appearance and spatial configuration heatmaps computed by convolutional layers. This approach requires to specify a hyperparameter of the heatmap Gaussian at the ground truth locations.

Long short-term memory (LSTM) architectures [5] were proposed to address the difficulties of training Recurrent Neural Networks (RNNs). The regression capability of Long Short-Term Memory (LSTM) networks in the temporal domain can be used to concatenate high-level visual features produced by CNNs with region information [8]. The target coordinates are directly regressed taking advantage of the joint spatio-temporal model. Convolutional LSTMs [15] replace the matrix multiplication by the weight vector with a convolution. As a result, the model captures spatial context.

3 Regressing Point Locations

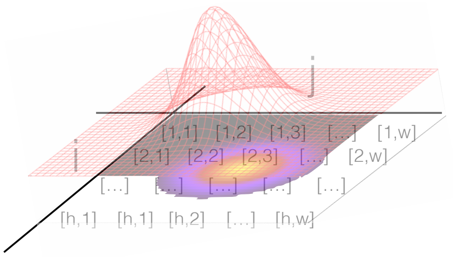
Denote an input image of width w and height h as $I \in \mathcal{R}^{w \times h}$ (independent variable) and the keypoint positions stacked columnwise into $\mathbf{p}$ (dependent variable). The goal of the regression is to learn a function $f(I; \theta) = \mathbf{p}$ parametrized by θ. We approximate f by a convolutional neural network and train the parameters θ using a database of images $\bar{\mathcal{I}} = \bar{I_1}, \ldots, \bar{I_n}$ and their corresponding annotations $\bar{\mathcal{P}} = \{\bar{\mathbf{p}}_1, \ldots, \bar{\mathbf{p}}_n\}$. Typically, a Euclidean loss $L(\mathcal{I}, \mathcal{P}; \theta) = \frac{1}{2N} \sum_{k=1}^{N} ||f(I_k; \theta) - \bar{\mathbf{p}}_k||_2^2$ is employed to train f using each annotated image.

Previously, regression estimates were obtained directly from the last layer of the network, which was fully connected to previous layer. This is a highly nonlinear mapping [13], where the estimate is computed from the fully connected layers after convolutional blocks.

3.1 Fully Convolutional Network with Center of Mass Layer

Instead of fully connected network, we propose to regress keypoint locations using a Fully Convolutional Network (FCN). FCNs have been previously used for image segmentation [6], for regressing heatmaps [9], and object localization [12]. Their advantage is that the estimates can be computed in a single evaluation step. In our architecture, we obtain point coordinate estimates at each image location.

The point coordinate predictions are computed in a new center of mass layer from input at each predicting location $\mathbf{l}_{ij}$ (see Fig. 1).


j
i
[1,1] [1,2] [1,3] [...] [1,w]
[2,1] [2,2] [2,3] [...] [2,w]
[...] [...] [...] [...] [...]
[h,1] [h,1] [h,2] [...] [h,w]

$$\hat{\mathbf{p}} = \frac{1}{w \times h} \sum_{i=1}^{h} \sum_{j=1}^{w} \mathbf{l}_{ij} \tag{1}$$

Fig. 1. Center of Mass layer computes the estimate as a center of mass computed from the regressed location estimates at each location.

Center of mass layer makes it possible to design a loss function with a penalty on the error of the measurement line length. Our penalty is defined as an absolute value of the difference between estimated and ground truth lengths relative to the ground truth length. This penalty is combined with the Euclidean loss discussed above. The model is trained with an Adam optimizer with learning rate set as 0.0002 and converges within 100 epochs. The best model is selected based on the lowest error of the point location estimates.

3.2 Convolutional Long Short-term Memory for Temporal Consistency

Recurrent neural networks (RNN) can learn sequential context dependencies by accepting input x_t and updating a hidden vector h_t at every time step t. The RNN network can be composed of Long-short Term Memory (LSTM) units, each controlled by a gating mechanism with three types of updates, $i_t, f_t, o_t \in R^n$ that range between 0 and 1. The value i_t controls the update of each memory cell, f_t controls the forgetting of each memory cell, and o_t controls the influence of the memory state on the hidden vector. In Convolutional LSTMs (CLSTMs), the input weights and hidden vector weights are convolved instead of multiplied to model spatial constraints. The function introduces a non-linearity which we chose as *tanh*. Denoting the convolutional operator as $*$, the values at the gates are computed as follows:

$$\text{forget gate:} \quad f_t = \text{sigm}(W_f * [h_{t-1}, x_t] + b_f) \tag{2}$$

$$\text{input gate:} \quad i_t = \text{sigm}(W_i * [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\text{output gate:} \quad o_t = \text{sigm}(W_o * [h_{t-1}, x_t] + b_o) \tag{4}$$

$$\tag{5}$$

The parameters of the weights W and biases b are learned from training sequences. In addition to the gate values, each CLSTM unit computes state candidate values

$$g_t = \tanh(W_g * [h_{t-1}, x_t] + b_g),$$

where $g_t \in R^n$ ranges between -1 and 1 and influences memory contents. The memory cell is updated by

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

which additively modifies each memory cell. The update process results in the gradients being distributed during backpropagation. The symbol $\odot$ denotes the Hadamard product. Finally, the hidden state is updated as:

$$h_t = o_t \odot \tanh(c_t).$$

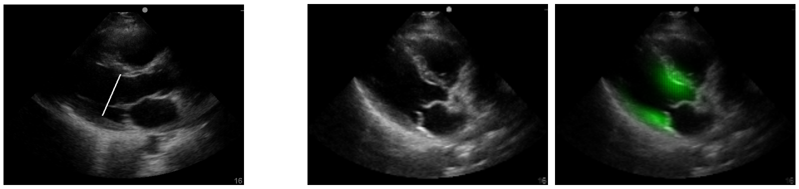
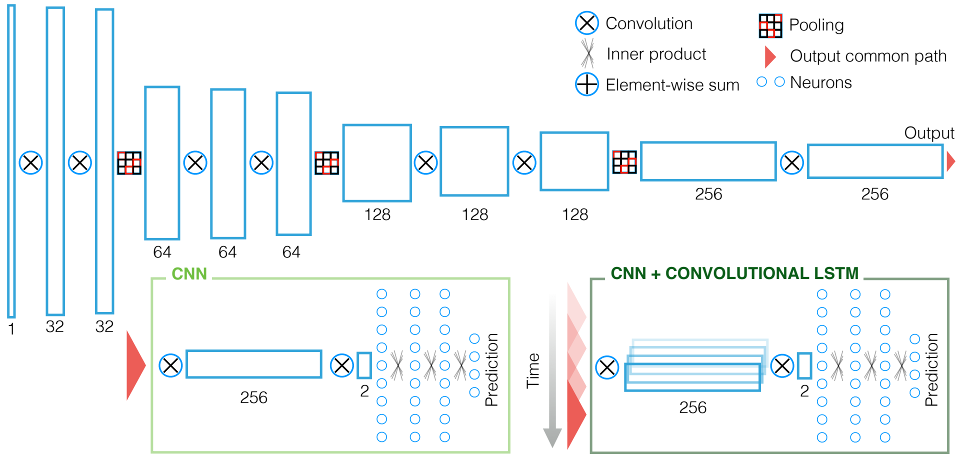


Fig. 2. (Left) Frame from an ultrasound sequence of the PLAx view of the left ventricle and overlaid measurement line. There is an ambiguity in the annotation points that can slide along the interface between myocardial wall and cavity and along the interface between wall and pericardium as reflected by aggregated prediction maps of the FCN regression model (Right).

In sequential processing of image sequences, the inputs into the LSTM consist of the feature maps computed from a convolutional neural network. In this work, we propose to use two architectures to compute the feature maps. The first architecture is a neural network with convolutional and pooling layers. After sequential processing the feature maps in CLSTM, the output is fed into fully connected layers to compute the point location estimate (Fig. 3). In the second architecture, the CLSTM inputs is the final layer of a convolutional path of the Fully Convolutional Network (FCN). The point location estimates are computed from the CLSTM output processed through the transposed convolutional part of the FCN network (Fig. 4). Similarly to [7,10], the feature maps are forwarded using connections from the previous layers.

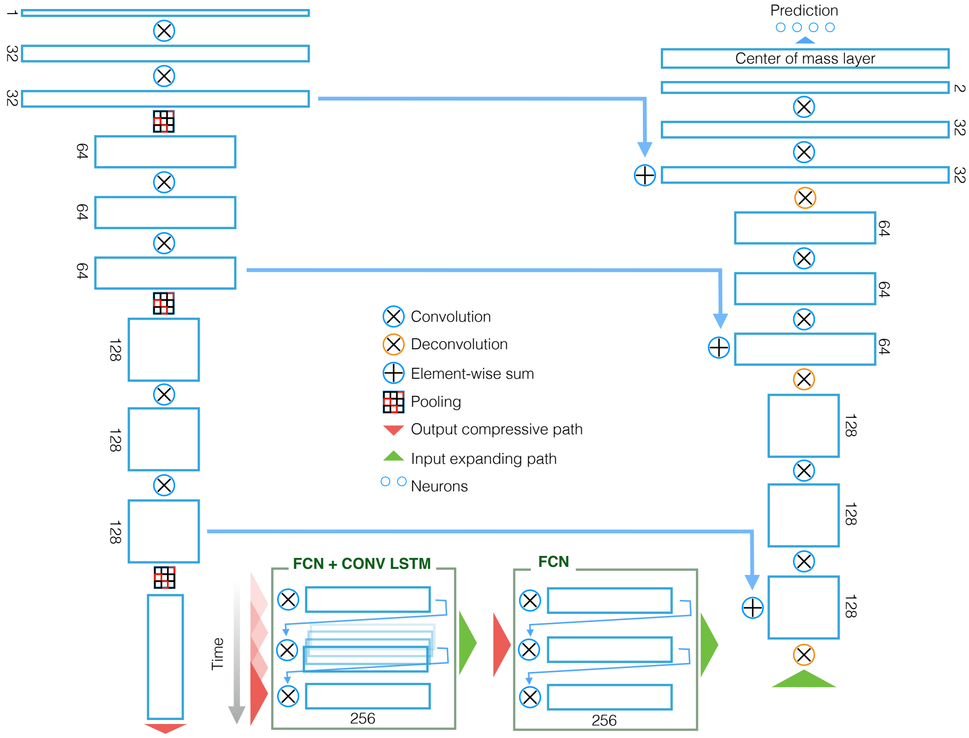


(a) (b)

Fig. 3. (a) Convolutional Neural Network (CNN) architecture to regress the keypoint locations. (b) CNN with feature maps processed by a Convolutional LSTM to model temporal constraints. CLSTM processes 256 feature maps and its output is used to compute the point location estimate.

4 Results

We evaluated the proposed network architectures on a dataset of ultrasound videos showing parasternal long axis (PLAx) view of the heart (Fig. 2). The data was acquired in several clinics and hospitals with four ultrasound systems: Siemens Acuson Aspen 7.0 and X300, Phillips iE33, and Sonosite M-Turbo. A total of 4981 annotated video frames were used for training and 628 for validation (model selection). The testing data set had 90501 frames of which 2048 were annotated. Our datasets are substantially larger than data sets often used in the medical literature. Two experienced sonographers annotated the frames by manually placing two measurement line calipers (keypoints) perpendicular to the left ventricle (LV) long axis, and measured at the level of the mitral valve leaflet tips. Calipers were positioned on the interface between myocardial wall and cavity and the interface between wall and pericardium. Average locations across annotators were used for training.



(a) (b)

Fig. 4. (a) FCN with feature maps processed by a Convolutional LSTM to model temporal constraints. (b) Fully Convolutional Network (FCN) architecture to regress the keypoint locations. CLSTM processes 256 feature maps and its output is used to compute the point location estimate. In both cases, center of mass layer is used to compute the final estimate.

We computed the following error measures to compare the different architectures. Define the ground truth measurement line based on two keypoints as $\bar{l}_k = ||\bar{\mathbf{r}}_k - \bar{\mathbf{s}}_k||$, $\bar{\mathbf{p}}_k = (\bar{\mathbf{r}}_k, \bar{\mathbf{s}}_k)^\top$. Estimated measurement line $\hat{l}_k$ is defined similarly using points detected at $\hat{\mathbf{r}}_k$ and $\hat{\mathbf{s}}_k$. The length error is defined as $el_k = |\bar{l}_k - \hat{l}_k|/\bar{l}_k$. The temporal error is defined as $et_k = \big| ||\hat{\mathbf{r}}_k - \hat{\mathbf{r}}_{k-1}|| + ||\hat{\mathbf{s}}_k - \hat{\mathbf{s}}_{k-1}|| \big| / 2\hat{l}_k$. We experimented with various frame sequence lengths and report results on sequences of 3 frames for the CNN + CLSTM model and 6 frames for the FCN + CSLSTM model. The results summarized in Tab. 1 show 50th, 75th, and 95th percentiles to present distribution of errors and evaluate difficult cases more directly.

Overall, the temporal modeling with CLSTM improves results to frame-wise processing. The final detection accuracy at the 50th percentile is 4.87% of the measurement length which is within average inter-observer error of 4.98%.

5 Conclusion

This paper proposed to detect measurement keypoint locations by computing their regression estimates with a Fully Convolutional Network (FCN). The esti-

error	length × 100%			temporal × 100%	
network	50th	75th	95th	75th	95th
CNN	5.67	10.02	21.10	3.52	8.73
CNN+CLSTM	4.89	8.68	17.51	2.92	7.13
FCN	5.00	9.36	19.32	3.36	8.28
FCN+CLSTM	4.87	8.86	18.27	2.67	6.70

Table 1. Average length and temporal errors computed on the testing data set. The errors are computed relative to the length of the measurement line for different percentiles (50th, 75th, 95th). Convolutional LSTM improves the accuracy and temporal stability. FCN + CLSTM model performs best overall.

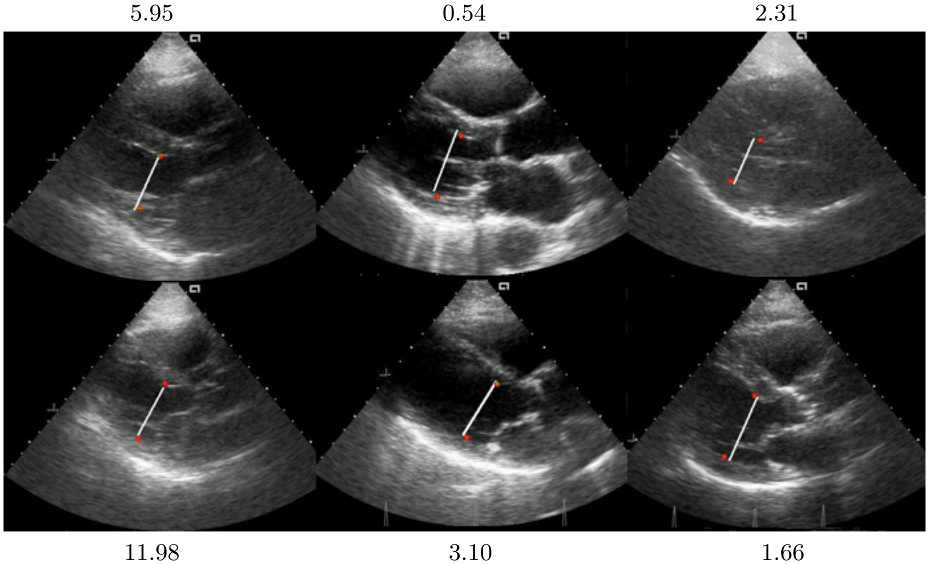


Fig. 5. Several examples of the detection results showing the measurement line (white) and ground truth annotations (red). Errors are shown as a percentage of the ground truth line length. Accurate measurements are obtained despite the shape and appearance variability and despite the ambiguity of the point annotations that can slide along the myocardial wall.

mates at each pixel location are mapped into the predicted location with a new center-of-mass (CoM) layer. The CoM layer makes it possible to define penalty loss on the measurement line. Spatial context is modeled with Convolutional Long-Short Term Memory (CLSTM) cells.

The results showed errors below 5% of the left ventricle measurement which is within inter-observer variability. The automated measurement was computed in the Parasternal Long Axis (PLAx) view of the heart which has not been previously proposed in the literature. The measurement is an important indicator of the left ventricular function and can be used to compute ejection fraction.

Our current work focuses on exploiting variance of the regressed predictions for regularization and on estimating additional measurements.

References

1. Chen, C., Xie, W., Franke, J., Grutzner, P., Nolte, L.P., Zheng, G.: Automatic X-ray landmark detection and shape segmentation via data-driven joint estimation of image displacements. Medical image analysis 18(3), 487–499 (2014)
2. Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., Siddiqui, K.: Regression forests for efficient anatomy detection and localization in computed tomography scans. Medical image analysis 17(8), 1293–1303 (2013)
3. Donner, R., Menze, B.H., Bischof, H., Langs, G.: Global localization of 3d anatomical structures by pre-filtered hough forests and discrete optimization. Medical image analysis 17(8), 1304–1314 (2013)
4. Gauriau, R., Cuingnet, R., Lesage, D., Bloch, I.: Multi-organ localization with cascaded global-to-local regression and shape prior. Medical image analysis 23(1), 70–83 (2015)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
6. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
7. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), 2016 Fourth International Conference on. pp. 565–571 (2016)
8. Ning, G., Zhang, Z., Huang, C., He, Z., Ren, X., Wang, H.: Spatially supervised recurrent convolutional neural networks for visual object tracking. arXiv preprint arXiv:1607.05781 (2016)
9. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using cnns. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 230–238. Springer (2016)
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241 (2015)
11. Szegedy, C., Toshev, A., Erhan, D.: Deep neural networks for object detection. In: Advances in Neural Information Processing Systems. pp. 2553–2561 (2013)
12. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 648–656 (2015)
13. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in neural information processing systems. pp. 1799–1807 (2014)
14. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1653–1660 (2014)
15. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems. pp. 802–810 (2015)