FOANet: A Focus of Attention Network with Application to Myocardium Segmentation

Zhou ZHAO, Elodie Puybareau, Nicolas Boutry, Thierry Géraud
EPITA Research and Development Laboratory (LRDE), Le Kremlin-Bicêtre, France
Email: zhou.zhao@lrde.epita.fr
Cardiovascular diseases (CVDs) are the leading cause of death globally. According to the World Health Organization (WHO).

Medical imaging:
- Computed tomography (CT)
- Magnetic resonance imaging (MRI)
- Positron emission tomography (PET)
- Single photon emission computed tomography (SPECT)
- Ultrasound (US)

Diagnosis and Treatments for CVDs.
The difficulties of cardiac segmentation on the MR Images are shown as follow:

- **Poor contrast**
  - between myocardium and surrounding structures

- **Brightness**
  - in left ventricular/right ventricular cavities due to blood flow

- **Non-homogeneous partial volume**
  - limited CMR resolution (1.5T, 3.0T...) along the long-axis

- **Noise**
  - motion artifacts and heart dynamics

- **Shape and intensity variability**
  - different patients and pathologies
Methodology

Global overview of the proposed method

The network framework consists of two different fully convolutional networks (FCN), one for positioning and the other for segmentation.

**Localization**

**Segmentation**
Architecture of our networks

1) Part 1 belongs to Network1;
2) Part 2 belongs to Network2.
Hybrid Loss

To increase the boundary quality, the hybrid loss (only used in segmentation network) is defined:

\[ l = l_{cce} + l_{ssim} + l_{dc} \]

*\( cce \): Categorical Cross Entropy;
*\( ssim \): Structural Similarity;
*\( dc \): Dice Coefficient.

\[ l_{cce} \text{ is defined:} \]

\[ l_{cce} = - \sum_{i}^{C} \sum_{a}^{H} \sum_{b}^{M} y^i_{(a,b)} \ln y^i_{*(a,b)} \]

*C*: the numbers of class for each image;
*\( H, M \): the height and width of image;
*\( y^i_{(a,b)} \in \{0,1\} \): the ground truth one-hot label of class \( i \) in the position (a, b);
*\( y^i_{*(a,b)} \): the predicted probability of class \( i \).
SSIM loss can assess image quality, and it can be used to capture the structural information. $l_{ssim}$ is defined:

$$
\begin{align*}
    l_{ssim} &= 1 - \frac{(2\mu_S \mu_G + C_1)(2\sigma_{SG} + C_2)}{(\mu_S^2 + \mu_G^2 + C_1)(\sigma_S^2 + \sigma_G^2 + C_2)} \\
S: \text{predicted probability map;} \\
G: \text{ground truth mask;} \\
\mu, \sigma: \text{the mean and standard deviations of } S \text{ and } G \text{ respectively;} \\
\sigma_{SG}: \text{the covariance of } S \text{ and } G; \\
C_1 = 0.01^2 \text{ and } C_2 = 0.03^2 \text{ are used to avoid dividing by zero.}
\end{align*}
$$

DC loss is used to measure the similarity of two sets. $l_{dc}$ is defined:

$$
\text{dice}_i = \frac{2 \sum_{n}^{N_i} y_n^i y_{*n}^i + \epsilon}{\sum_{n}^{N_i} (y_n^i + y_{*n}^i) + \epsilon}
$$

$$
l_{dc} = 1 - \sum_{i}^{c} \frac{\text{dice}_i}{N_i + \epsilon}
$$

$N_i$: the numbers of class $i$; 
$\epsilon$: smooth factor.
To decrease the effects of similar tissues, Focus of Attention (FOA) is defined:

$$I_{FOA} = \sum_a^{H} \sum_b^{M} I_{(a,b)} \omega_{FOA}$$

$I_{(a,b)}$ denotes image intensity at location $(a,b)$. $\omega_{FOA}$ is a Gaussian weighted function defined by:

$$\omega_{FOA} = \alpha \exp\left(\frac{-(a^* - a)^2 - (b^* - b)^2}{\delta^2}\right)$$

$(a^*, b^*)$: the object center; 
$\alpha$: normalization constant; 
$\delta$: scale parameter.

Cropping after locating original image

FOA

$\omega_{FOA}$
Experiments and Results

Dataset description

1) LVQuan19
- 56 patients processed SAX MR sequences
- 20 temporal frames correspond to a whole cardiac cycle
- In-plane resolution: 0.6836 ~1.5625 mm/pixel
- Image sizes: 256 × 256 or 512 × 512 pixels

2) MM-WHS2017
- 20 MRI and 20 CT image
- In-plane resolution: 0.78 ~1.2 mm/pixel
- Slice spacings: 0.899 ~1.60 mm/pixel
- Average image sizes: 324 × 325 × 171 pixels
Pre-processing in the localization network

1) Gauss normalization

For the (2D+ t) image $I$ corresponding to a given patient, we compute $I$:

$$I = \frac{I - \mu}{\sigma}$$

$\mu$, $\sigma$: is the mean and standard deviation of $I$ ($\sigma$ is assumed not to be equal to zero).

2) Temporal-like image

Note: For 2D model, although it can have a much larger field of view, is not able to fully explore inter-slice information.
Pre-processing in the segmentation network

1) Data augmentation;
   ---using rotations and flips or not
2) Resizing;
   ---with a fixed inter-pixel spacing
3) Focus of Attention (FOA);
4) Gauss normalization;
5) Temporal-like image.

Post-processing in the segmentation network

1) **Keeping** the greatest connected component;
   ---for the segmented (2D+t) image
2) **Get back** the initial inter-pixel spacing;
   ---computing the inverse interpolation on the x and y axes

Post-processing

Adding a zero-valued border to get back initial image shape
Implementation and Experimental Setup

(1) Keras/TensorFlow using a NVidia Quadro P6000 GPU;

(2) $l_{cce} \rightarrow$ localization network, hybrid loss $l \rightarrow$ segmentation network;

(3) Adam optimizer;
   ---batchsize=1, $\beta_1=0.9$, $\beta_2=0.999$, epsilon=0.001, lr = 0.002

(4) Epoch=10;

(5) Dividing all classes into the same class given in the ground truth(GT) $\rightarrow$ localization network.
Evaluation Methods

(1) Dice Coefficient

(2) Boundary of Dice Coefficient (BDC)
Ablation study includes three parts (architecture, loss and FOA).

<table>
<thead>
<tr>
<th>Ablation</th>
<th>Configurations</th>
<th>DC</th>
<th>95HD</th>
<th>BDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>a: B. + $\ell_{CCE}$</td>
<td>0.842</td>
<td>3.186</td>
<td>0.269</td>
</tr>
<tr>
<td></td>
<td>b: B. + L. + $\ell_{CCE}$</td>
<td>0.867</td>
<td>2.209</td>
<td>0.281</td>
</tr>
<tr>
<td></td>
<td>c: BLP + $\ell_{CCE}$</td>
<td>0.877</td>
<td>2.019</td>
<td>0.303</td>
</tr>
<tr>
<td>Loss</td>
<td>d: BLP + $\ell_{SSIM}$</td>
<td>0.873</td>
<td>2.094</td>
<td>0.297</td>
</tr>
<tr>
<td></td>
<td>e: BLP + $\ell_{DC}$</td>
<td>0.871</td>
<td>2.193</td>
<td>0.295</td>
</tr>
<tr>
<td>FOA (our)</td>
<td>i: BLP + FOA + $\ell_{CSD}$</td>
<td>0.879</td>
<td>1.826</td>
<td>0.306</td>
</tr>
<tr>
<td>UNet</td>
<td>-</td>
<td>0.862</td>
<td>3.976</td>
<td>0.291</td>
</tr>
</tbody>
</table>

“B.” means “baseline” (Net.1); “L.” means “localization”; “P2.” means “Part 2” (Net.2); “BLP” means “baseline + localization + Part2”.

Note: $\ell_{CSD} = \ell_{CCE} + \ell_{SSIM} + \ell_{DC}$

![Image of ablation study results](image)
One patient
Red: Ground truth
Green: Prediction
Yellow: Interaction
Testing our method on **MM-WHS2017** for segmenting the myocardium.

<table>
<thead>
<tr>
<th>Method</th>
<th>DC(train)</th>
<th>DC(test)</th>
<th>Computation time</th>
<th>Data augmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Method</td>
<td>0.851</td>
<td>?</td>
<td>&lt;2s</td>
<td>No</td>
</tr>
<tr>
<td>Champion</td>
<td>0.796</td>
<td>0.781</td>
<td>&lt;2min</td>
<td>No</td>
</tr>
<tr>
<td>The second place</td>
<td>0.752</td>
<td>0.778</td>
<td>-</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: **Red**: best; Champion : the first place in MM-WHS2017 Challenge
(1) **Boundary-aware network**.
---FOANet consists of one localization network and one segmentation network.

(2) **Hybrid loss**.
---combines CCE, SSIM and DC to guide the training process on three levels: pixel level, patch-level and map-level.

(3) **Focus of Attention (FOA)**.
---decreases the effect of surrounding similar tissues.

(4) **Temporal-like method**.
---let the FOANet take advantage of the temporal information by stacking 3 successive 2D frames.

**In the future work**

(1) **Studying the impact of the hybrid loss**
---by weighting differently the segmentation loss.
(2) **Adding constraints on shapes in the network**.
Thank You!