CS-512

Breast Tumor Segmentation using Attention Enriched Deep Learning Architecture

---

Authors

Saptarshi Maiti (A20447671)

Rishab Panyam (A20427149)

Abstract:

Incorporating human domain knowledge for breast tumor diagnosis is challenging, since shape, boundary, curvature, intensity, or other common medical priors vary significantly across patients and cannot be employed. This work proposes a new approach for integrating visual saliency into a deep learning model for breast tumor segmentation in ultrasound images. Visual saliency refers to image maps containing regions that are more likely to attract radiologists' visual attention. The proposed approach introduces attention blocks into a U-Net architecture, and learns feature representations that prioritize spatial regions with high saliency levels. The validation results demonstrate increased accuracy for tumor segmentation relative to models without salient attention layers. The salient attention model has potential to enhance accuracy and robustness in processing medical images of other organs, by providing a means to incorporate task- specific knowledge into deep learning architectures.

Overview:

Breast cancer is the second most common cancer in women after skin cancer. Early detection and localization of Breast Cancer can greatly increase the effectiveness of treatment for the cancer. Hence segmentation and localization of breast cancer is a valuable step in improving healthcare accessibility and accuracy. Ultrasound images are used regularly by radiologists to determine the occurrence of a tumor so we use the ultrasound images to segment the tumor present in it. Breast tumor provides a unique problem as shape, boundary, curvature, intensity, or other common medical priors vary significantly across patients and cannot be employed. Hence we plan to use attention blocks into a U-Net architecture, and learn feature representations that prioritize spatial regions with high saliency levels. Saliency maps refers to image maps containing higher values regions that are more likely to contain tumors . We propose this method as an improvement on previous U-Net based architectures for semantic segmentation in medical images.

Dataset:

- Images - the dataset consists of 163 breast ultrasound images.
- Masks - segmentation masks corresponding to the images.
- Saliency - saliency maps for the 163 breast ultrasound images; the maps are obtained based on our approach presented in Xu et al. (2019) A Hybrid Framework for Tumor Saliency Estimation.

Dataset link :
https://drive.google.com/drive/folders/1VQIfST8jjCBm95d7XgvuWlzagf1Aw7Yy?usp=sharing

Proposed Solution:

The proposed solution is based on the well-known U-Net architecture (Ronneberger et al. 2015), which consists of fully convolutional encoder and decoder sub-networks with skip connections. The layers in the encoder employ a cascade of convolutional and max-pooling layers, which reduce the resolution of input images and extract increasingly abstract features. The decoder comprises convolutional and up-sampling layers that provide an expanding path for recovering the spatial resolution of the extracted feature maps to the initial level of the input images. A unique characteristic of the U-Net architecture is the presence of skip connections from the feature maps in the encoder's contracting path to the corresponding layers in the decoder. The features from the respective encoder and decoder's layers are merged via concatenation that allows to recover the spatial accuracy of the objects in images and improves the

resulting segmentation masks. Namely, although the central layer of the network offers high-level features with semantic rich data representation and a large receptive field, it also has low level of spatial context detail due to the down-sampling max-pooling layers along the contracting path, and impacts the localization accuracy around the object boundaries in the predictions. The skip connections provide a means to transmit low-level feature information from the initial high-resolution layers in the encoder to the reconstructing layers in the decoder, thereby restoring the local spatial information in predicted segmentations. Despite the introduction of deeper and more powerful models for image segmentation in recent years, the U-Net architecture has remained popular especially in medical image segmentation, where datasets have small size and large models can overfit on the available sets.

A graphical representation of the proposed model is presented in Figure 1. Besides the main input consisting of BUS images, the network has an auxiliary input consisting of the corresponding salient maps. Attention blocks introduce salient maps with reduced scale in all layers on the contracting path of the encoder in the form of an image pyramid. This enforces the network to focus the attention onto regions in the saliency maps with high intensity values. More specifically, the introduced attention blocks put more weights on areas in the extracted feature maps at each layer that have higher levels of saliency in the salient maps. Thus, the topology of the salient maps influences the learned feature representations.

The images and saliency maps are gray-scale 8-bit data resampled into floating point with normalization. Resized images and saliency maps to 256 × 256 pixels are used as inputs to the model. The number of convolutional filters per layer in the network is (32, 32, 64, 64, 128), which is reduced in comparison to the original U-Net, to account for the relatively small dataset. The output segmentation probability maps have the same spatial dimension as the inputs. The proposed network is trained in an end-to-end fashion; however, the saliency maps are precomputed and used at both training and inference.

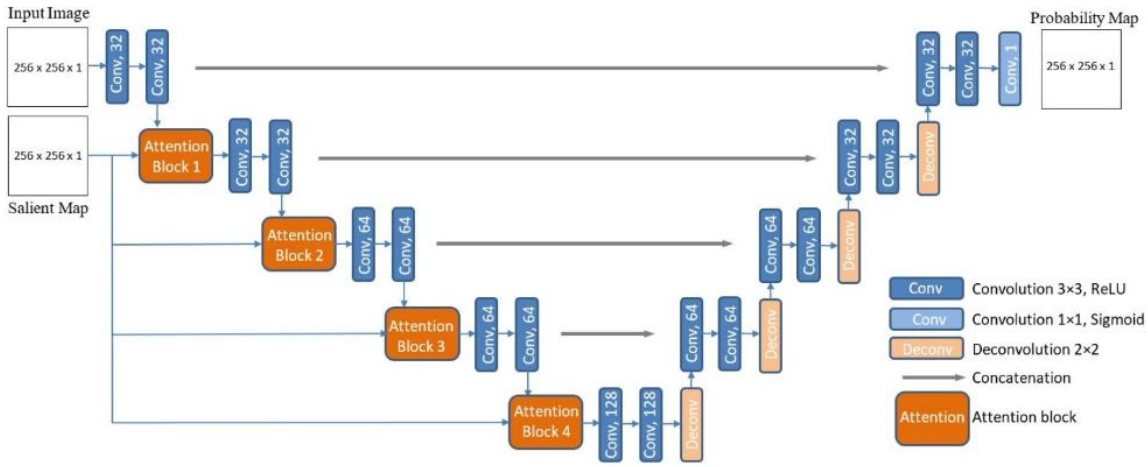


Figure 1: Architecture of the proposed U-Net model with salient attention. The model uses BUS images and saliency maps as inputs, and produces segmentation probability maps as outputs.

Attention Block

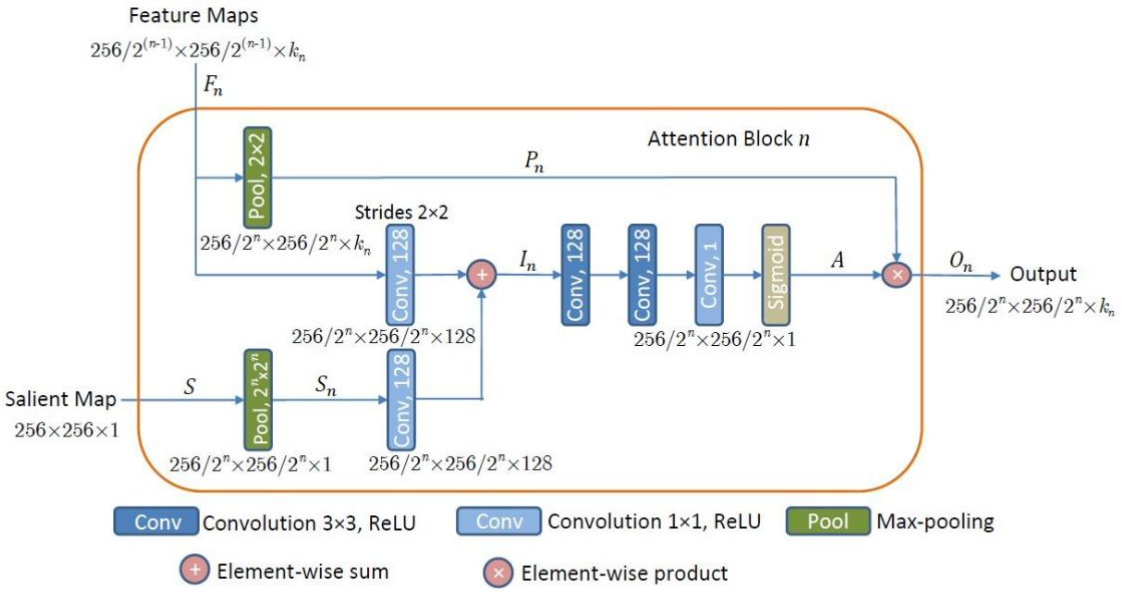


Figure 2: Attention block n, for $n \in \{1, 2, 3, 4\}$. Inputs to the block are feature maps from layer n with spatial dimension $256/2^{(n-1)} \times 256/2^{(n-1)}$ with k_n number of channels, and a salient map, and the output are down-sampled weighted maps with spatial dimension $256/2^n \times 256/2^n$ and k_n number of channels.

The goal of Attention Mechanisms is to recognize discriminative features in the inner activation maps(either through ml techniques or domain knowledge) and use this knowledge for enhanced task-specific data representation and improved model performance. This mechanism leads to suppression of less important features and leads to emphasizing more important features. For eg In image classification , important features lie in salient spatial locations in the image.

Attention blocks introduce salient maps with reduced scale in all layers on the contracting path of the encoder in the form of an image pyramid. This enforces the network to focus the attention onto regions in the saliency maps with high intensity values. More specifically, the introduced attention blocks put more weights on areas in the extracted feature maps at each layer that have higher levels of saliency in the salient maps. Thus, the topology of the salient maps influences the learned feature representations.

In the above block diagram, the input feature maps to the attention blocks denoted by $F = \{f, f, \ldots, f\}$, where each feature map has horizontal and vertical spatial dimensions of $256/2^{(n-1)} \times 256/2^{(n-1)}$ pixels for the block in the layer level $n \in \{1, 2, 3, 4\}$. The symbol k_n is the channel dimension of the feature maps in block *n*, i.e., $k_n \in \{32, 32, 64, 64\}$.

The input Salient Map in the above block diagram is denoted by S and it is down-sampled through a max-pooling layer, resulting in S_n, which matches the spatial dimension of the input feature maps F in Attention Block n. Next, 1 × 1 convolutions followed by rectified linear unit (ReLU) activation functions are used to increase the number of channels of the saliency map S to 128. An element-wise sum block

performs addition of F_n and S_n producing intermediate maps I_n of size $256/2^{(n)} \times 256/2^{(n)} \times 128$. The intermediate maps I_n are further refined through a series of linear $128 \times 3 \times 3$ and $1 \times 1 \times 1$ convolutions, followed by nonlinear ReLU activations. A sigmoid activation function normalizes the values of the activation maps into the [0, 1] range. The produced output is the attention map $A = (\alpha_i)$ with a spatial size of $256/2^{(n)} \times 256/2^{(n)} \times 1$, where the attention coefficients α_i have scalar values for each pixel *i*. Next, soft attention is applied via element-wise multiplication of the attention map A with the max-pooled features P, i.e., $O_n = A * P_n$. The activation maps O with size $256/2^{(n)} \times 256/2^{(n)} \times k$ are the Output of Attention Block *n*, and they are further propagated to the next layer, as depicted in Fig. 1.

Implementation Details:

- Packages :
    - Keras
    - Sklearn
    - Tensorflow
    - Matplotlib

Baseline Model :

- **Unet**

  Skip Connections:

    - Skip connections between encoder and decoder.
    - Upsampled Output is concatenated with cropped input

  Contracting Path

```
conv_layer1 -> conv_layer2 -> max_pooling
```

  Expansive Path

```
conv_2d_transpose -> concatenate -> conv_layer1 -> conv_layer2
```

  Fully Convolutional:

    - 3 × 3 convolution produces segmented output.
    - The number of filters in the final convolution layer = number of classes.

SA Unet Model:

Apart from the Unet Layer used as above, we have used attention blocks to focus on important areas of the image using Saliency maps. The attention blocks were created by first convolving the saliency maps. We find the feature maps obtained by the UNet architecture and add them with the convolved Saliency maps. We then convolve them by 2 layers of 3 by 3 of 128 size filters . We then expand this tensor into 32 times (i.e. repeat the operations). We finally end with an activation layer with a sigmoid function. Finally we multiply this block with the pool layer to assign weights to high saliency areas.

Testing:

We used four-fold cross-validation, where four folds (75% of images) are used for training, and one fold (25% of images) is used for testing. Validation during training is performed on 15% of the training set of images. All images in the dataset are first resized to a 256 × 256 pixels resolution. Since we focus on understanding the impact of the introduced salient attention on the model performance, we did not apply image augmentation.

Evaluation Metrics:

Loss Function :

- Dice Similarity coefficient:

$$DSC = \frac{2|A_g \cap A_p|}{|A_g| + |A_p|}$$

- Dice Loss Function:

$$\mathcal{L} = 1 - DSC = 1 - \frac{2|A_g \cap A_p|}{|A_g| + |A_p|}$$

Accuracy:

- Jaccard Index:

$$JI = \frac{|A_g \cap A_p|}{|A_g \cup A_p|}$$

- True Positive Rate:

$$TPR = \frac{|A_g \cap A_p|}{|A_g|}$$

- False Positive Rate:

$$FPR = \frac{|A_g \cup A_p - A_g|}{|A_g|}$$

- Global Accuracy:

$$ACC = \frac{|A_g \cap A_p| + |\overline{A_g} - A_g \cup A_p|}{|A_g| + |\overline{A_g}|}$$

- Dice Score:

$$Dice = \frac{2 \times TP}{(TP + FP) + (TP + FN)}$$

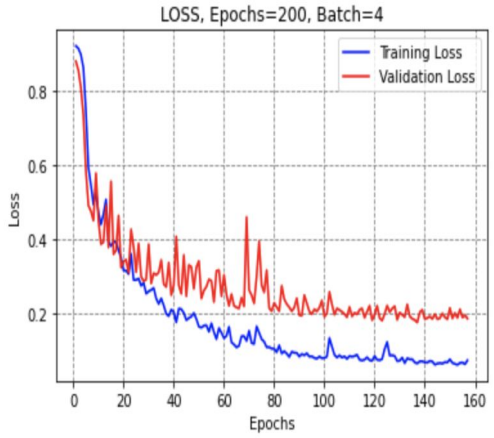
In the above equations, A_9 is the set of pixels that belong to a tumor region in the ground truth segmented images, $\overline{A_9}$ is the set of pixels that belong to the background region without tumors in the ground truth segmented images, and $A\square$ is the corresponding set of pixels that are predicted to belong to a tumor region by the segmentation method.

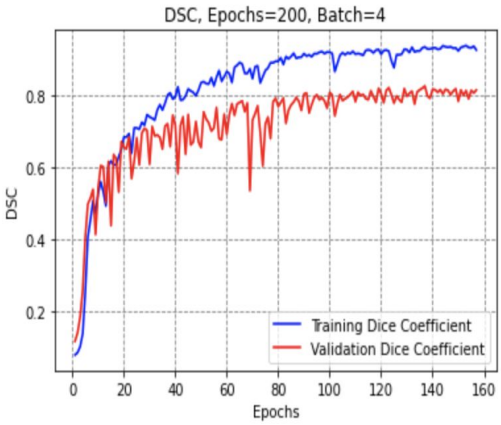
Results:

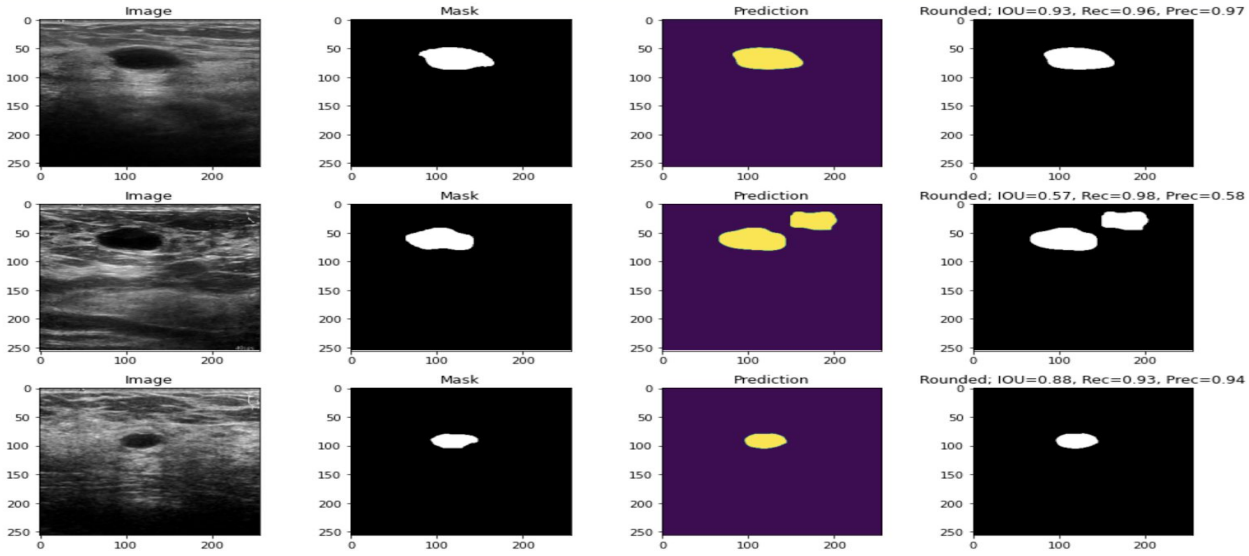
Base Model (Unet) :

```
USING THRESHOLD 0.5

  DSC              0.733
  IOU              0.649
  Recall           0.744
  Precision        0.860
```







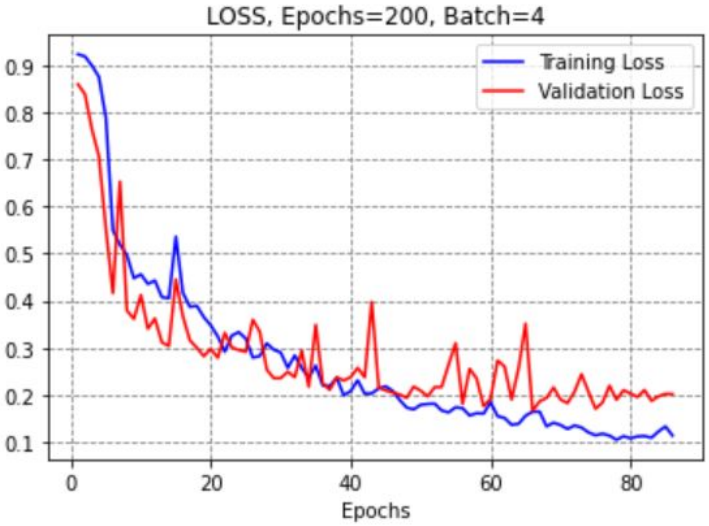
Mean Value of the scores

Epochs Number Mean	Dice Score Mean	IOU Score Mean	Recall (Sensitivity) Mean	Precision Mean	Global Accuracy Mean	AUC-ROC Mean
5	112.75	0.687141	0.590409	0.675406	0.844382	0.97825	0.924145

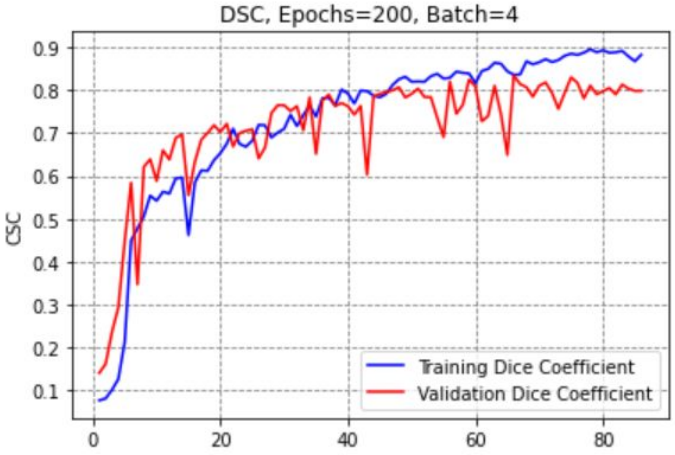
SA Unet Model: --

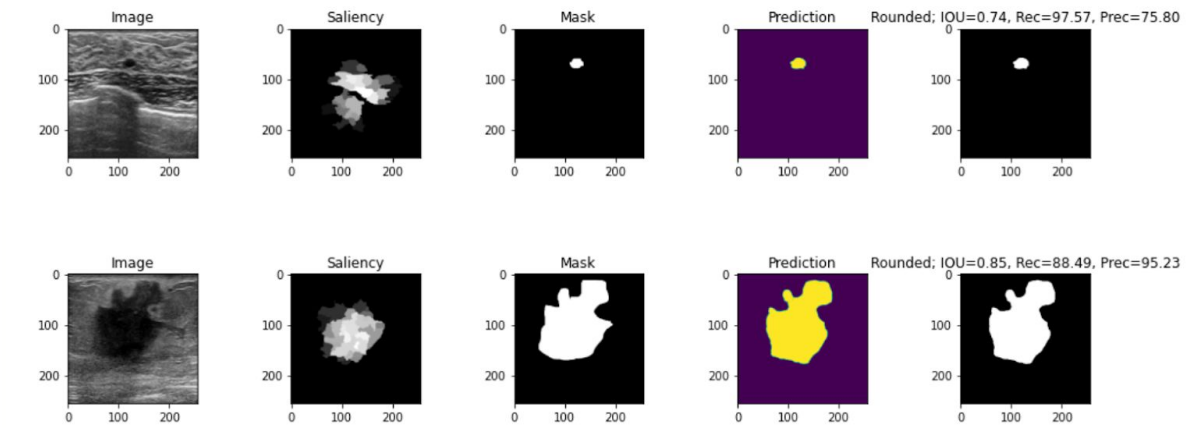
Fold3 Metrics and Graphs:

DSC	0.738
IOU	0.636
Recall	77.520
Precision	78.758



DSC, Epochs=200, Batch=4





Final Mean Scores:

	Dice_score	IOU_Score	Recall	Precision	Global_Accuracy	AOC_score
1	0.692075	0.591017	70.418967	80.310967	0.979881	0.918472

Conclusion:

Model	DSC	IOU	Recall	Precision	Accuracy	AUC-ROC
Unet	0.687	0.59	0.675	0.844	0.97825	0.924145
SA-Unet	0.692	0.591	0.7041	0.8031	0.979	0.918472

We can see that adding attention blocks improves on the model if only marginally. This process can be improved by getting better Saliency maps which have an accurate representation of the probability of breast tumor presence in the image which could further improve on the performance of the UNet model.

References:

- (Main) Attention-Enriched Deep Learning Model for Breast Tumor Segmentation in Ultrasound Images. Authors: Aleksandar Vakanski, Serestina Viriri. Published on 14 August 2020
- (Reference) CIU-Net: Convolutional Networks for Biomedical Image Segmentation. Authors : Olaf Ronneberger, Philipp Fischer, and Thomas Brox
- (Reference) A Benchmark for Breast Ultrasound Image Segmentation (BUSIS) Authors :Min Xian†1 , Yingtao Zhang†2 , H. D. Cheng*2,3, Fei Xu3 , Kuan Huang3 , Boyu Zhang3 , Jianrui Ding4 , Chunping Ning5 , Ying Wang6
- (Reference) https://towardsdatascience.com/unet-line-by-line-explanation-9b191c76baf5