Volumetric Attention for 3D Medical Image Segmentation and Detection

Xudong Wang[1,2], Shizhong Han[1], Yunqiang Chen[1], Dashan Gao[1], Nuno Vasconcelos[2]
[1]12 Sigma Technologies, [2]University of California San Diego

Introduction

A volumetric attention (VA) module for 3D medical image segmentation and detection is proposed. VA attention is inspired by recent advances in video processing, enables 2.5D networks to leverage context information along the z direction, and allows the use of pretrained 2D detection models when training data is limited, as is often the case for medical applications. Its integration in the Mask R-CNN is shown to enable state-of-the-art performance on the Liver Tumor Seg- mentation (LiTS) Challenge, outperforming the previous challenge winner by 3.9 points and achieving top performance on the LiTS leader board at the time of paper submission. Detection experiments on the DeepLesion dataset also show that the addition of VA to existing object detectors enables a 69.1 sensitivity at 0.5 false positive per image, outperforming the best published results by 6.6 points.
Keywords:
Volumetric Attention · 3D images · LiTS · DeepLesion

Motivation

	Pre-training Model	High Spatial Resolution	Context Information
3D Network	✗	✗	✓
2D Network	✓	✓	✗
????	✓	✓	✓

Compared with 2.5D network which is limited by the lack of contextual information and 3D network which lacks pre-training model and need to use low spatial resolution images due to GPU memory limitation, our proposed method can leverage contextual information in the z direction, pre-trained 2D CNN weights for transfer learning and maintain high spatial resolution.

Volumetric Attention Module

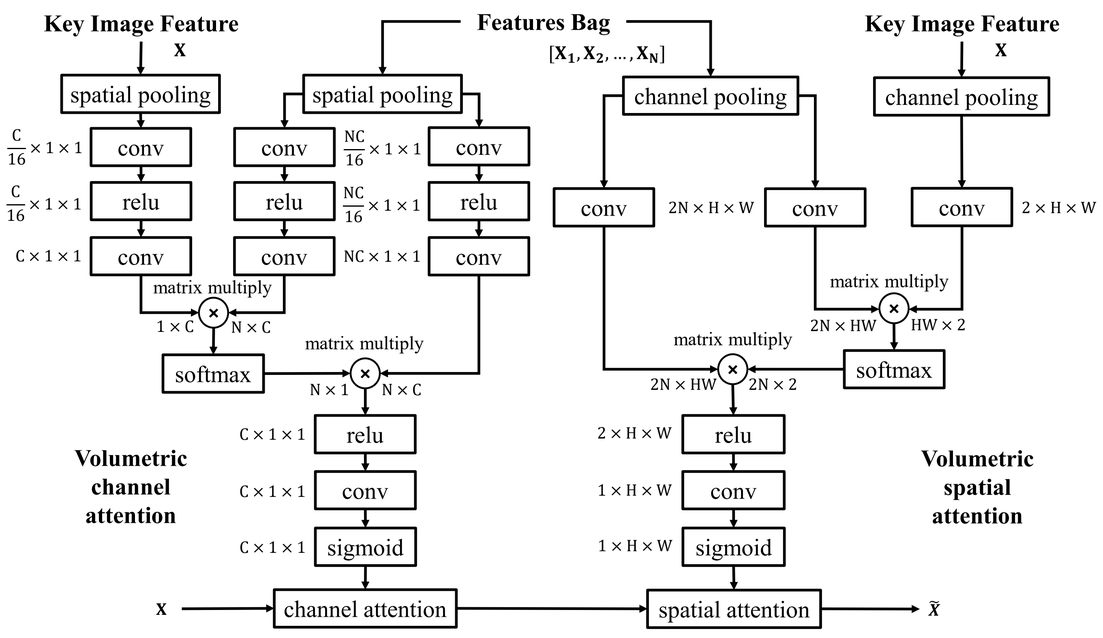

Figure. 1: Volumetric spatial and channel attention module. N is the bag size, C, H, W the feature map channel size, height and width, respectively. Spatial and channel pooling are used to reduce computation.

Volumetric Attention Components

1. Bags of Long Range Features

$$\mathbf{X}^i_{long} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N] \in \mathbb{R}^{N \times C^i \times H^i \times W^i}, \quad (1)$$

Where i is the pyramid levels and $C^l \times H^l \times W^l$ its dimensions (channel, height, and width, respectively), X^i_{long} is the corresponding bag of long-range features, and N the number of contextual images. The features X_K are sorted by the order of the corresponding images along the z direction of the 3D volume.

2. Volumetric Channel Attention

$$\mathbf{F}^c_{emb}(\mathbf{X}) = W_2\delta(W_1\mathbf{F}^c_{avg}(\mathbf{X}))$$

$$\mathbf{S}^c_{att} = \text{softmax}(\mathbf{F}^c_{emb}(\mathbf{X}_{tgt}) \cdot \mathbf{F}^c_{emb}(\mathbf{X}_{long})) \in \mathbb{R}^{1 \times N} \quad (2)$$

Where $\mathbf{F}^c_{avg}$ is the global average pooling operator, $\mathbf{X}_{tgt}$ corresponding target image feature map and $\mathbf{X}_{long}$ the bag of long range features. The slice attention signal is then applied to $\mathbf{F}^c_{emb}(\mathbf{X}_{long})$ followed by a relu layer, a 1×1 conv layer and a sigmoid layer, to learn a nonlinear interaction $S_c \in \mathbb{R}^{C \times 1 \times 1}$ between channels.

3. Volumetric Spatial Attention

$$\mathbf{F}^s_{pool}(\mathbf{X}) = [\mathbf{F}^s_{max}(\mathbf{X}), \mathbf{F}^s_{avg}(\mathbf{X})] \in \mathbb{R}^{2 \times H \times W}$$

$$\mathbf{S}^s_{att} = \text{softmax}(\mathbf{F}^s_{emb}(\mathbf{X}_{tgt}) \cdot \mathbf{F}^s_{emb}(\mathbf{X}_{long})) \in \mathbb{R}^{1 \times N} \quad (3)$$

Where the volumetric spatial attention module uses max and average pooling to shrink feature maps along the channel dimension, concatenating them into two channel feature maps as in above equation. A spatial attention map, $S_s \in \mathbb{R}^{1 \times H \times W}$, is generated similar to channel attention.

Architecture of VA Mask-RCNN

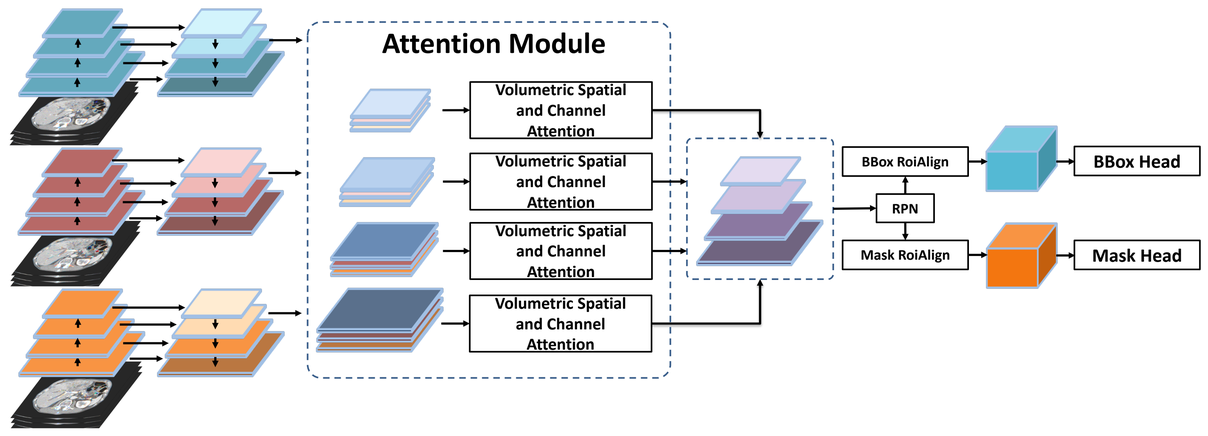

Figure. 2: Architecture of the Volumetric Attention (VA) Mask-RCNN. Three continuous 2.5D images, each composed of 3 adjacent slices, are shown as example.

Results

1. Experiments on Liver Tumor Segmentation Challenge

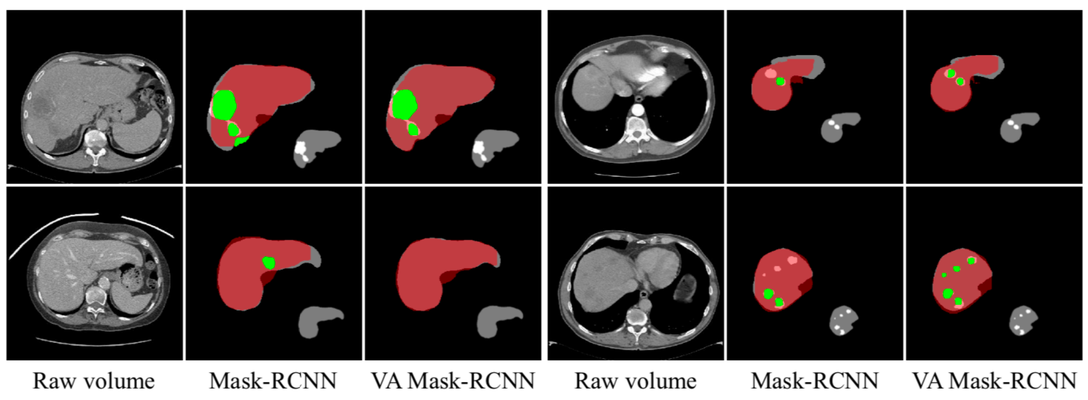

Figure. 3: 2D visualization of segmentations by Mask-RCNN and VA Mask-RCNN on LiTS val set. Segmented liver is shown in red and lesions in green. Zoomed out ground truth masks are shown on bottom right, with liver in gray and lesions in white. The VA Mask-RCNN produces smoother segmentation boundaries and lower FP and miss rates. In the top left, the gallbladder area is easily confused with the lesion area. VA Mask-RCNN leverages contextual slices to remove this FP.

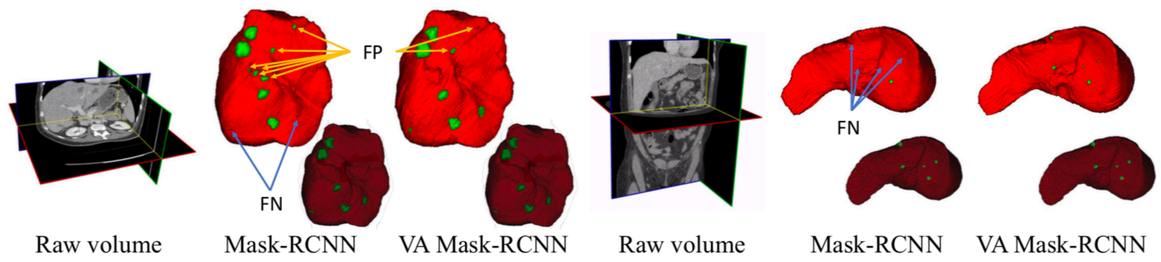

Figure. 4: Comparison of 3D segmentations by the Mask-RCNN and the proposed VA Mask- RCNN on the LiTS val set. Red denotes segmented liver, green segmented lesions. 3D ground truth is shown on the bottom right, with liver in dark red and lesions in dark green. These examples illustrate how the VA module both enhances small lesion prediction and enables the network to avoid false positives.

Team	Model	Dice per case
3D U-Net(Ours) [5]	3D U-Net	55.0
G. Chlebus [3]	2D U-Net	65.0
E. Vorontsov et al. [16]	2D + 3D FCN	65.0
Y. Yuan [20]	Deconv-Conv Net	65.7
X. Han [9]	2D U-Net	67.0
LeHealth	-	70.2
Mask-RCNN(Ours)[10]	Mask-RCNN	70.3
X. Li et al.[13]	H-DenseUNet	72.2
VolumetricAttention	VA Mask-RCNN	**74.1**

Table 1: Comparison with LiTS Challenge leaderboard, as of July 1st, 2019

# Slices	Dice	Dice_s	Dice_m	Dice_l
$9(3 \times 3)$	61.7	52.2	71.6	79.5
$21(3 \times 7)$	62.5	52.6	72.2	79.8
$27(3 \times 9)$	**63.3**	**54.3**	**73.7**	80.3
$33(3 \times 11)$	63.1	53.6	73.4	**80.6**

Table 2: Influence of number of slices on LiTS val set..

Scale	Dice	Dice_s	Dice_m	Dice_l
512	50.2	35.8	65.1	77.9
800	61.1	52.1	71.6	79.3
1024	**63.3**	**54.3**	**73.7**	**80.3**
1333	**63.5**	**54.8**	**73.5**	**80.4**

Table 3: Influence of image scales.

2. Experiments on DeepLesion(Lesion detection datasets)

Model	Backbone	1 FPs	AP_{50}
Faster-RCNN[8]	ResNet152	77.4	64.9
Faster-RCNN[8]	ResNet101	75.1	61.8
Faster-RCNN[8]	ResNet50	73.4	60.0
Deformable Faster-RCNN[7]	ResNet50	76.3	62.4
Faster-RCNN+VA	ResNet50	75.6	63.0
Deformable Faster-RCNN+VCA	ResNet50	76.8	63.8
Deformable Faster-RCNN+VSA	ResNet50	76.9	64.1
Deformable Faster-RCNN+VA	ResNet50	**77.9**	**65.0**

Table 4: Sensitivity (%) at 1 FPs/image and AP_{50} on the DeepLesion test set.

Model	Backbone	0.5	1
Faster-RCNN[8]	VGG-16	56.9	67.3
R-FCN[6]	VGG-16	55.7	67.3
Improved R-FCN [6]	VGG-16	56.5	67.7
Data-level fusion, 11 slices	VGG-16	58.5	70.0
3-DCE,9 Slices[18]	VGG-16	59.3	70.7
3-DCE,27 Slices[18]	VGG-16	62.5	73.4
Faster-RCNN+VA, 9 Slices	ResNet50	**67.6**	**75.6**
Deformable Faster-RCNN+VA	ResNet50	**69.1**	**77.9**

Table 5: Sensitivity(%) at 0.5 and 1 FPs per image on the DeepLesion test set.

Conclusion

In this paper, we proposed a volumetric attention module that enables 2.5D methods to leverage contextual information along the z direction and the use of pretrained 2D detection models when training data is limited, as is often the case for medical ap- plications. VA can be combined with any CNN architecture, including one-stage and two-stage detectors and segmentation networks. It was shown that 2.5D networks with VA achieve state of the art results for both lesion segmentation and detection.