ABSTRACT

Majority of deep learning methods are developed for RGB imagery. However, for many applications such as detecting objects underwater other types of sensors such as sonar or radar are required. One of the most precise sensors to map the seagrass disturbance is side scan sonar. Here we developed a new deep learning framework based on dilated convolution, dense module, and inception to perform semantic segmentation for automatic extraction of potholes in underwater sonar imagery. We tested our proposed approach on a collection of underwater sonar images taken from Laguna Madre in Texas. Experimental results in comparison with the ground-truth and state-of-the-art semantic segmentation methods show the efficiency and improved accuracy of our proposed method.

1. INTRODUCTION

In recent years Convolutional Neural Networks (CNN) have been widely used in computer vision research including classification [1, 2], object recognition [3], counting [4, 5] and semantic segmentation [6]. Majority of deep learning methods are developed for RGB imagery. However for many applications such as detecting objects underwater [7, 8] or under ice [9, 10] other types of sensors such as sonar or radar are required. This research investigate developing a novel semantic segmentation technique based on dense inception network for identifying pothole in seagrass.

The widespread loss of seagrass beds is largely caused by the rapid expansion of human populations around coastal waterways. Detection of seagrass with optical remote sensing (both satellite imagery and aerial photography) is complicated by the fact that light is attenuated as it passes through the water column and reflects back from the benthos.

Underwater acoustics mapping produces a high definition, two-dimensional sonar image of seagrass ecosystems. The intensity and contours of the image are then determined by the amount of time a sound wave takes to return to the transducer and how the waves were reflected.

Several pattern recognition techniques have been applied on sonar images mainly for detecting concrete objects on sandy sea floor based on hierarchical MRF model [11], active contour [12], Bayesian classifier [13], Gauss-Markov random field model and level-set [14].

Side scan sonar (SSS) has been recently used for detecting the extent of seagrass beds and mapping its disturbance [7, 8] based on mathematical morphology and level set. All of the aforementioned techniques are based on hand-crafted feature engineering.

Here we developed a new deep learning framework based on Dilated convolution, DenseNet, and Inception to perform semantic segmentation for automatic extraction of potholes in underwater sonar imagery. We tested our proposed approach on a collection of underwater sonar images taken from Laguna Madre in Texas. Experimental results in comparison with the ground-truth and state-of-the-art semantic segmentation methods show the efficiency and improved accuracy of our proposed method.

2. METHODOLOGY

The overall architecture of our method is depicted in Figure 1. It includes three different modules: 1) Atrous block, 2) Dense module, 3) DeconvXY block. The specific parameters of the network is shown on Table1.

Dilated (Atrous) Convolutions: Atrous convolutions help to build scale invariance and provide the network with a larger viewing window. We used this module to extract multiscale information about our original image. This assists the early level dense modules with contextual information. Our Atrous Block is comprised of a series of convolutions with different dilations that share kernel weights. Gradients are calculated only for the first convolution. Each output is concatenated together, and then a final convolution is performed in order to help disrupt the gradient shifts from our unconventional method and compress the information to prevent parameter explosion in our coming Dense Blocks.

Dense module: Dense convolutional networks (DenseNet) is build with feed-forward connections between each layer to every other layers [15]. Such network is designed to alleviate the vanishing-gradient problem while maximizing the information flow between network layers. It requires fewer parameters by reducing redundant feature map learning. Moreover, each layer has direct access to the gradients from the loss function and the original input signal which will provide an...
implicit formulation between layers. This is also shown to perform some regularization effect which may be helpful for over-fitting problems.

**DeconvXY**: Our proposed Inception-Deconvolution, referred to as DeconvXY due to the way channels perform their deconvolutions can be seen on Figure 2. This module uses three different channels of transposed convolutions which up-sample the network at different rates. The output feature map of each transposed convolution is the inverse of how much upsampling occurs. As an example, a $1 \times 1$ network with 4 feature maps, $(1 \times 1 \times 4)$, being upsampled by a rate of 2 would become $(2 \times 1 \times 2)$ and $(1 \times 2 \times 2)$ after the first transposes. The final shape would be $2 \times 2 \times 1$ at the end of each channel. These outputs are concatenated together and then have dropout performed with a keep percentage of .85, then fed to the next module in the network. This type of data representation ensures that these modules are limited in scope to translating information into different spatial dimensions, rather than creating new inferences.

"Fig. 1. Our Network Architecture"

<table>
<thead>
<tr>
<th>Network State</th>
<th>Shape of State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Images</td>
<td>(440,200,1)</td>
</tr>
<tr>
<td>Output of Atrous Block</td>
<td>(440,200,12)</td>
</tr>
<tr>
<td>Output of DenseBlock1 (Pre-Pooling)</td>
<td>(440,200,92)</td>
</tr>
<tr>
<td>Output of DenseBlock1 (Post-Pooling)</td>
<td>(220,100,92)</td>
</tr>
<tr>
<td>Output of DenseBlock2 (Pre-Pooling)</td>
<td>(220,100,172)</td>
</tr>
<tr>
<td>Output of DenseBlock2 (Post-Pooling)</td>
<td>(110, 50,172)</td>
</tr>
<tr>
<td>Output of DenseBlock3 (Pre-Pooling)</td>
<td>(110, 50,300)</td>
</tr>
<tr>
<td>Output of DenseBlock3 (Post-Pooling)</td>
<td>(22, 10,300)</td>
</tr>
<tr>
<td>Output of DeconvXY3</td>
<td>(110, 50, 36)</td>
</tr>
<tr>
<td>Output of DeconvXY2</td>
<td>(220,100,150)</td>
</tr>
<tr>
<td>Output of DeconvXY1</td>
<td>(440,200,160)</td>
</tr>
</tbody>
</table>

| Table 1. Network State Sizes |

3. EXPERIMENTAL RESULTS

Data was collected from the seagrass beds of the Lower Laguna Madre in southern Texas in 2016 from an average depth of 75 cm. A specialized side scan sonar unit was constructed consisting of a towfish with two Lowrance Structure Scan HD LSS-2 transducers, a Dual Beam 200 kHz down-imaging transducer connected to a Humminbird 998C HD SI control unit. A total of six transects approximately 5000 x 60000 pixels and overlapping by 50% were processed individually to image an area of approximately 88,000 m². The images used in this experiment were large transects. Each transect was 600 meters long, and spaced 20 meters apart with a horizontal swath of approximately 50 m. Total area covered between all six transects was 88,000 square meters (150mX600m).

To apply our proposed method on these images, they were re-sized to a uniform width of 4400 then chopped into varying amounts of $4400 \times 2000$ images, resulting in 167 total images, with 117 for training and 50 for testing. These images are augmented by flipping along the horizontal and vertical axes, increasing the number to 468 and 200.

For training, we evaluate our loss using cross entropy and optimize using the Adam optimizer, a learning rate of 1e-3, and a batch size of 8. The normal training, validation, testing format is replaced with training, testing for each epoch until the error begins to increase. The optimal amount of training on this dataset for our architecture was 14 epochs. For evaluating metrics we use accuracy, intersection over union (IOU), precision, recall, and F1 score for each class then calculate the mean per class metrics.

Detailed quantitative results can be viewed on Table 3,
In this project we developed a novel architecture for semantic segmentation of pothole in sonar imagery. Our method outperformed the state-of-the-art techniques in terms of both accuracy and efficiency. While our proposed network outperforms previous networks in this segmentation task, feature extraction of argumentative features could be improved. Other avenues of improvement are further optimization of the network to decrease parameters, increase portability, and speed. The intended goal of this system is to be able to deploy it on a mobile platform in order to generate real time sensor analysis for environmental and fishery preservation.

5. REFERENCES


Fig. 3. Comparison between our result (second column), Ground Truth, and FCN [16]


