CapsuleNet: A Deep Convolutional Network For Lesion Detection in Wireless Capsule Endoscopy.

Jacob Nye
Columbia University
MS Candidate, Biomedical Eng.

Drew Afromsky
Columbia University
MS Candidate, Biomedical Eng.

Alicia Dagle
Columbia University
MS/PhD Candidate, Mechanical Eng.

***Abstract*—Accurate detection and localization for lesions in the small intestine is important for early stage diagnostics of gastrointestinal bleeding, inflammation, and anemia. The gold-standard for screening of small bowel diseases is wireless video capsule endoscopy. This pill-like device is able to produce thousands of high resolution images during one passage through gastrointestinal tract. While capsule endoscopy technology has led to increased visibility of the small intestine, analyzing capsule endoscopy data remains largely inefficient. Here, we present DeepLabV4 among other neural networks that can improve diagnostic outcome by classifying and identifying inflammatory and vascular regions. The proposed semantic segmentation network labels every pixel in an image as one of three classes: vascular lesions, inflammatory lesions, or background. Based on this analysis we developed a classifier and a network that segments vascular lesions, inflammatory lesions, and healthy background.**

***Index Terms*—Segmentation, classification, SegNet, DeepLab, FCN, endoscopy, disease detection.**

I. Introduction

Capsule endoscopy has proven a great advancement in medical technology and remains the current standard of care, enabling clinical care professionals to visualize the small intestine with a minimally invasive tool. This procedure can be used to detect inflammatory mucosa in patients with Crohn's disease or vascular lesions in patients with ulcerative colitis. However, endoscopy videos contain on average 50,000 still frames, posing a tedious task for technicians who may take on average 45-90 minutes to analyze this footage [1]. These factors limit diagnostic yield and can also lead to many missed lesions or diagnoses due to human error [2]. There is an unmet need for an accurate, automated tool to help physicians locate and diagnose lesions in the small bowel using wireless capsule endoscopy technology.

Some work has been done to try to address this problem. Polyp detection was added as a computational method to conventional video endoscopy for increased detection [1]. There are wide efforts to apply deep learning to diagnostic imaging, often consisting of classification and segmentation networks. A pixel-based semantic segmentation network is well-posed to identify vascular and inflammatory lesions in an otherwise healthy patient. This is done using pixel-by-pixel semantic segmentation, labeling every pixel in the image, [3] rather than the simplified classification problem which assigns one value (in this case, normal, vascular, or inflammatory) as the output. [3]

In order to complete the learning algorithm, it is essential to optimize hyper-parameters, thereby minimizing generalization error [4]. This is a complex problem because there is an iterative process involved by optimizing the network design while simultaneously optimizing the associated hyper-parameters. Both of these factors can affect the performance accuracy of a network.

In this work, we consider multiple semantic segmentation network backbones for the transfer learning task such as Fully Convolutional Networks (FCN-8), SegNet, and DeepLabV3+.

II. Materials and Methods

A. Data Cleaning & Augmentation

The dataset consisted of 1,800 WCE images that were downloaded from the WCE subchallenge of the Gastrointestinal Image Analysis (GIANA), which itself was a subchallenge of the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference 2018. The dataset consisted of 600 images with vascular lesion(s), 600 images with inflammatory lesion(s), and 600 images without any lesions (healthy/background). Prior to preprocessing, each image had a resolution of 576 x 576 pixels and were 24-bit JPEG images. Each image had a corresponding mask image. White pixels in the mask images corresponded to lesion localization for both the inflammatory and vascular lesions, and black corresponded to background or healthy tissue. We preprocessed our data by cropping them to 509 x 495 pixels to remove the text annotations inside the image and minimize the empty black space in each image. We also converted the white pixels in the masks for vascular lesions to a gray value of 125 since MATLAB requires each class to have a unique pixel value and this allowed us to train our models for each lesion type concurrently.

In order to expand our training dataset we augmented our 1,800 VCE images to over 9,000 images. Our augmentation techniques included shearing with a min and max of 8 pixels, skew (perspective) with a magnitude of 0.4, image rotation by 25 degrees left and right, and left and right image flipping [5]. We performed this augmentation after we split our data into 85 % (1,540 images) training, 10 % (181 images) validation, and 5 % (91 images) for testing.

B. Training

In this work we evaluate 3 different deep architectures for segmenting WCE images: DeepLabV3+ [3], Segnet [6], and FCN-8 [7]. Between these three architectures, we trained a total of 12 models successfully. All models trained used the stochastic gradient descent optimizer with 0.9 momentum in MATLAB and were trained using 2 NVIDIA 1080 TI. For training our best Segnet model, we trained the model on 1540 images and used an initial learning rate of 1e-3, L2 regularization factor of 0.001, mini-batch size of 8, validation patience of 4 epochs, and it trained for the max epoch amount of 20 epochs. Our best Fully Convolutional Network (FCN-8) model (CapsuleNet-FCN) was trained on 9240 images with an initial learning rate of 1e-3, L2 regularization factor of 0.05, mini-batch size of 8, and L2-norm gradient threshold of 0.05, validation patience of 4, and it trained until it met validation patience criteria at epoch 16 after its loss had not improved for 4 epochs. Our best DeepLabV3+ network (CapsuleNetV7) was trained on 9240 images with an initial learning rate of 1e-3, L2 regularization factor of 0.005, mini-batch size of 8, validation patience of 4 epochs, and it was trained until it met validation patience criteria at epoch 10. In total, we trained 6 DeepLabV3+ models because they were the easiest and most stable networks to train, 5 SegNet models, and 1 FCN-8 model because the FCN-8 took much longer to train than the other two model types and used the most VRAM during training. We modified each network by replacing the last fully connected layer and pixel classification layer with a class-weighed pixel classification layer and a new fully connected layer preceding it so that we could use these backbone networks for transfer learning. The classes were given the following weights for inflammatory, vascular, and healthy regions: 1, 8.3, 0.3.

Bayesian optimization may be used to complete hyper-parameter tuning for a semantic segmentation network. This was considered and remains an area for future work. However, hyperparameters to produce sufficient results were found manually rather than using a grid search for the network trained in this paper.

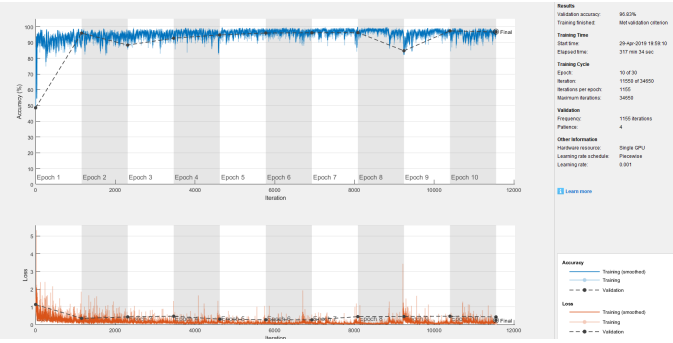

Fig. 1. Example training plot from our DeepLabV3+ network.

TABLE I

TEST ACCURACY COMPARISONS FOR GLOBAL DATASETS COMPARING DEEPLAPV3+, SEGNET, AND FCN AS WELL AS PER-CLASS ACCURACY COMPARISONS FOR CAPSULESEGNET V5, CAPSULENET V7 AND CAPSULENET-FCN ARE SHOWN IN THE TABLE BELOW.

CapsuleNet Backbone	Global Accuracy (Acc.)	Mean Acc.	Mean IoU	BF Score
Segnet	91.92%	**85.48%**	0.4129	0.6166
DeepLabV3+	**95.83%**	83.94%	**0.5321**	**0.8347**
FCN	94.12%	76.19%	0.3910	0.7028

	Per-Class Accuracy		
CapsuleNet Backbone	Inflammatory Accuracy	Vascular Accuracy	Background Accuracy
CapsuleSegNet V5	**83.16%**	81.22%	92.07%
CapsuleNet V7	69.91%	86.72%	**96.20%**
CapsuleNet-FCN	44.21%	**89.58%**	94.77%

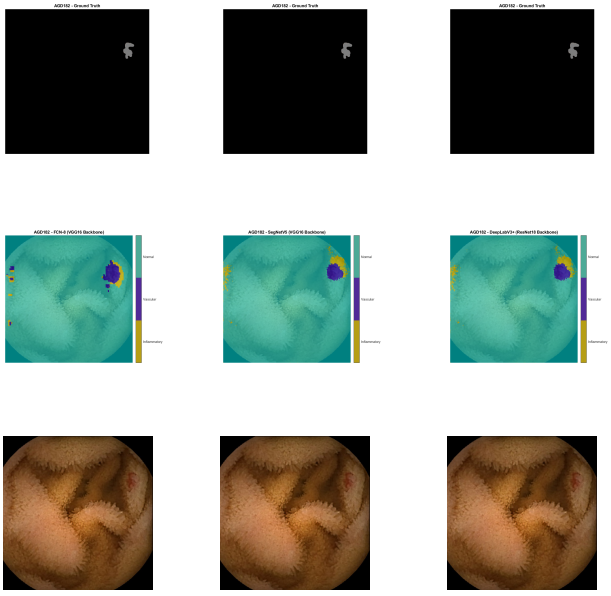

Fig. 2. The above figure shows comparisons for the FCN (Column 1), SegNet (Column 2) and DeepLab (Column 3) network results on the same vascular test image. Row 1 consists of masks showing vascular lesions in gray. Row 2 consists of pixel segmentation results identifying vascular (violet) and inflammatory (yellow) lesions. Row 3 consists of the original test image.

III. RESULTS AND CONCLUSIONS

A. Testing

A learning rate of 10^{-3} was found to produce optimal results. While each network was able to achieve a top accuracy for one of the three classes, we found that qualitatively the DeepLabV3+ network performed much better than the other two networks, largely as a result of its superior specificity. Over 98% of the pixels in the dataset were considered healthy or normal, which resulted in small changes in accuracy for the healthy class corresponding to qualitatively large changes in segmentation quality. There was also an apparent trade off we observed when training our models between sensitivity for the inflammatory and vascular lesions and the specificity. An increase in either vascular or especially with inflammatory

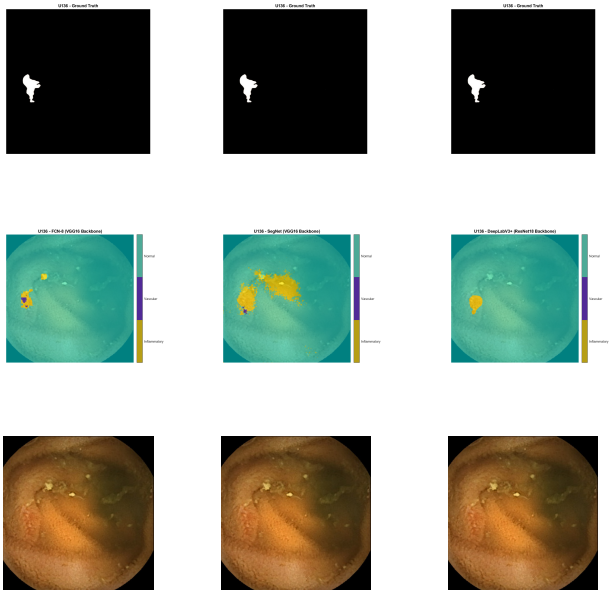

Fig. 3. The above figure shows comparisons for the FCN (Column 1), SegNet (Column 2) and DeepLab (Column 3) network results on the same inflammatory test image. Row 1 consists of masks showing inflammatory lesions in gray. Row 2 consists of pixel segmentation results identifying vascular (violet) and inflammatory (yellow) lesions. Row 3 consists of the original test image. Row 4 consists of the class activation maps.

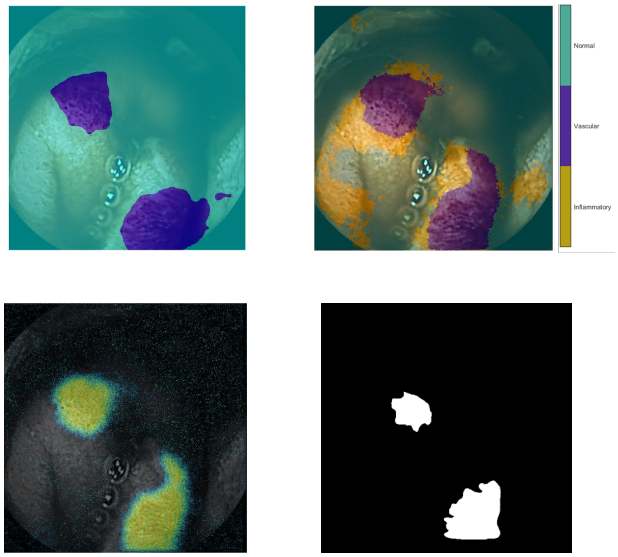

Fig. 4. The figure at the top left is the segmentation output from DeepLabV3+ on a vascular lesion. The top right image is a segmentation output for the same vascular lesion from our Segnet-based model. The image at the bottom left is a class activation map (CAM) of the vascular lesion. The CAM shows which areas in the region are important features and led to the network's class prediction. The majority of pixels mapped are from the vascular lesion, but some healthy background pixels are also captured. The image at the bottom right is the binary ground truth mask of the vascular lesion

sensitivity or accuracy (seen in our Segnet model) resulted in significantly more false positive pixels that degraded output image quality and rendered the segmentation less effective in the clinical sections. The metric we tracked that best correlated with segmentation quality was mean IoU, with the DeepLabV3+ model achieving the best in that metric. In future iterations of our model, we believe that using a DICE loss layer for our pixel classification layer would be improve our performance in this metric over our current class-weighted cross-entropy loss layer.

B. Class Activation Maps

In order to better understand how our network makes its predictions and to address the "black box" problem with neural networks, we implemented class activation mapping (CAM). This technique outlines the decision making process of our model because CAMs indicate the discriminate regions of images used by the convolutional neural network to identify a class or make a prediction. We can use CAM to identify bias in our training set and increase model accuracy [8]. The CAM for a particular class is the activation map of the ReLU layer that follows the final convolutional layer [9]. CAM provided us with unique insight into which features were being learned by our network and were important to classification and localization.

IV. Discussion

In the future, these results can be improved by post-processing. Further data collection may also be performed to improve model robustness. Additional methods of data augmentation such as GAN can be used to synthesize new data, thereby vastly increasing the sample size. When training a network with a larger dataset, the results become more generalizable and less prone to overfitting. While our network did not achieve a mean intersectional reunion (IoU) as good as the winners of the MICCAI 2017 challenge, it is reasonably to believe that this network would achieve comparable performance, given the addition of similar postprocessing. Moving forward, it may also be worthwhile to implement a deepresnet for improved feature extraction and therefore higher accuracy. This deep network may be able to learn richer features, also leading to increased lesion identification.

Acknowledgment

The authors would like to thank the professors: Paul Sajda and Andrew Laine as well as the teaching assistant Arunesh Mittal for their guidance and lectures throughout the semester.

References

[1] D. K. Iakovidis and A. Koulaouzidis, "Software for enhanced video capsule endoscopy: challenges for essential progress," *Nature Reviews Gastroenterology & Hepatology*, vol. 12, no. 3, p. 172186, 2015.

[2] S. K. Lo, "How should we do capsule reading?," *Techniques in Gastrointestinal Endoscopy*, vol. 8, no. 4, p. 146148, 2006.

[3] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *CoRR*, vol. abs/1802.02611, 2018.

[4] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[5] M. D. Bloice, "Augmentor," Mar 2019.

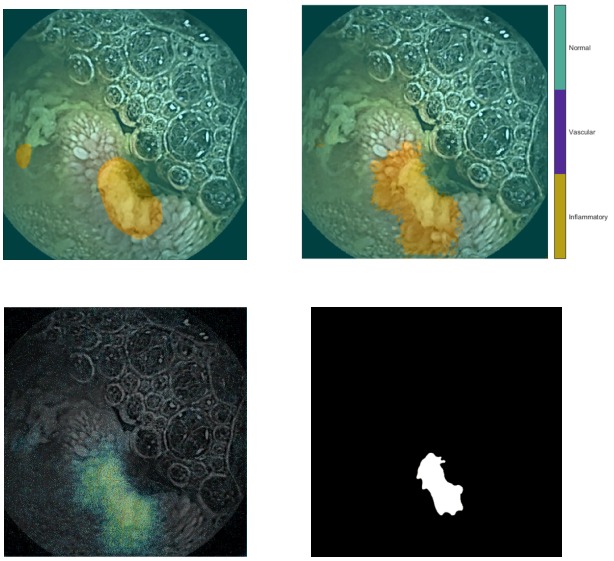

Fig. 5. The figure at the top left is the segmentation output from DeepLabV3+ on an inflammatory lesion. The top right image is a segmentation output for the same inflammatory lesion. The image at the bottom left is a class activation map (CAM) of the inflammatory lesion. The CAM shows which areas in the region are important features and led to the network's class prediction. The majority of pixels mapped are from the inflammatory lesion, but some healthy background pixels are also captured. The image at the bottom right is the binary ground truth mask of the inflammatory lesion

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *CoRR*, vol. abs/1511.00561, 2015.
[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, June 2015.
[8] "Investigate network predictions using class activation mapping."
[9] J. Pingel, "Deep learning visualizations: Cam visualization," Jan 2019.