Endometriosis Classification and Segmentation Using CNN and U-Net

Brandon Herrera[1]

[1]University of Minnesota, United States

herre350@umn.edu

Index Terms: Endometriosis, Biomedical Image Classification, Biomedical Image Masking, Deep Learning

1. Introduction

Endometriosis is a chronic condition that affects about 10% of women and is characterized by their uterine tissue growing outside of the uterus [1]. This can cause a multitude of issues including but not limited to painful menstrual cycles, pelvic floor issues, and decreased fertility. It ranks among the top reasons for gynecological hospitalizations as well as hysterectomies [2]. The only way to have a confirmed endometriosis diagnosis is to get a minimally invasive Laparoscopy done by a specialist. This surgery involves getting a small cut to the abdomen and having a small camera inserted so that endometriosis can be removed. Due to endometriosis often being misattributed to other problems such as irritable bowel syndrome and interstitial cystitis [3], the limited schedules of specialists, as well as the amount of effort it takes to schedule and go through surgery, it takes people an average of 7 years to actually get diagnosed. If a specialist were not required to complete this surgery, then people may be able to get this surgery sooner and drive this 7 year average down.

This paper proposes using deep learning methods to allow for any surgeon to be able to complete this surgery. This involves two steps.

1. Given a surgical image identify if endometriosis is present
2. If endometriosis is present, show where it is so that it can be removed.

Given these two things, it should be possible for a surgeon to be able to remove the endometriosis with minimal expertise. This paper utilizes a standard CNN for classification as well as a U-Net for image segmentation.

2. Data

2.1. Data Overview

The data comes from frames of videos from over 400 gynecologic laparoscopy surgeries [4]. There are 5 possible classes that each image may belong to. The classes are based off of where the endometriosis was found in the body, or if none was found. The different classes are peritoneum, ovary, uterus, deep infiltrating endometriosis, and no pathology (no endometriosis). There are about 350 image samples containing endometriosis (from about 100 surgeries) and about 13k containing no endometriosis (from about 20 surgeries). Along with the surgery images, the data also contains annotations which are images meant to be overlaid over the original image specifying where the endometriosis is. The color of the annotation overlay specifies which class the image belongs to. Figure 1 is an example of an annotation. Since the color of the annotation is yellow, it specifies that the endometriosis is in the uterus. My goal is to do binary classification on the surgical images to see if endometriosis is present as well as use the annotation images to segment where in the image the endometriosis is.

Figure 1: *Example image annotation*

2.2. Data Preprocessing for Classification

Since I could not find any previous code using this dataset I had to do all of the data preprocessing myself. I used the os library to go through my directories and collect all of the image paths, then I used scikit image to convert all of the images to readable matrices. Then I wrote my own pytorch Dataset which reshapes the images into 256x256x3 and transforms the images to tensors. Since none of the images have standard class labels, I also had to label the images by checking if there was a corresponding annotation. If there was then the image had endometriosis (class=1) and if it didn't then it was class=0.

2.3. Data Preprocessing for Segmentation

I used most of the same preprocessing techniques as I did with the classification task however I had to make some adjustments. Given that there are only about 350 images with pathology, I knew it would be difficult to learn the annotation color (where the endometriosis was in the body) and instead I chose to grey scale the annotation and only focus on where the endometriosis was relative to the picture. I split my training and testing by .8 and .2 respectively. This meant that I was only training with about 280 images. My peer reviews warned me about unbalanced data, so to circumvent the limited amount of training data I had, I decided to augment my images. For every image, I rotated it by 90 degrees and added it to my training set.

3. Methodology

3.1. Image Classification

To do image classification I created a standard convolutional neural network from scratch. I had one convolutional layer along with 4 feed forward layers with ReLU activation and a sigmoid output. Dropout was used after each layer with a probability of .5. To train the CNN I used the Adam optimizer with a learning rate of .001 for 10 epochs. The loss over time for this model can be seen in figure 2.

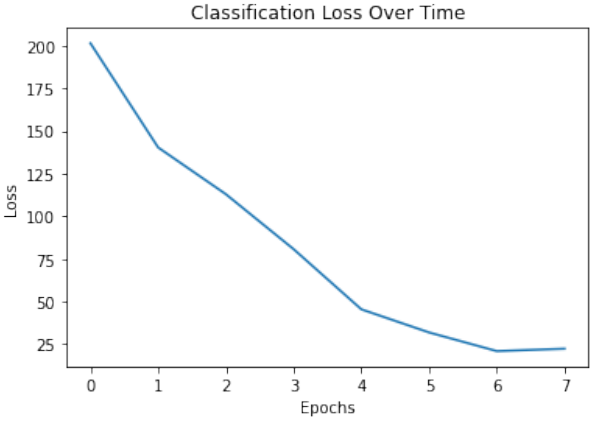

Figure 2: *Loss for CNN Classification*

3.2. Image Segmentation

Image annotation required a more specialized network in order to work. I decided to utilize U-Net as this was suggested to me during my peer reviews and is designed specifically for medical image segmentation. The way that U-net works is that the data takes a path down an encoder that contracts the images and learns what is in the image, and then up through a decoder that expands the image and learns spatial information [5]. The specific architecture can be found in figure 2. I found an implementation of U-Net here which I modified slightly to include a sigmoid on the output. U-Net takes an input of 256x256x3 and outputs a 256x256x1 image highlighting where it thinks the endometriosis is. I trained the U-Net with the Adam optimizer with a learning rate of .0001 for 52 epochs. The loss over time for this model can be seen in figure 4.

4. Results

4.1. Image Classification

Classical disease testing uses two metrics to determine if they are allowed to be used in practice. These are sensitivity and specifcity. These are defined as follows

$$sensitivity = \frac{TruePositive}{TruePositive + FalseNegative} \quad (1)$$

$$specificity = \frac{TrueNegative}{TrueNegative + FalsePositive} \quad (2)$$

A clinically relevant accuracy is given as having a sensitivity of .94 and a specificity of .79 [3]. My results are shown in table 1 with a testing dataset of 129 samples. The results shown

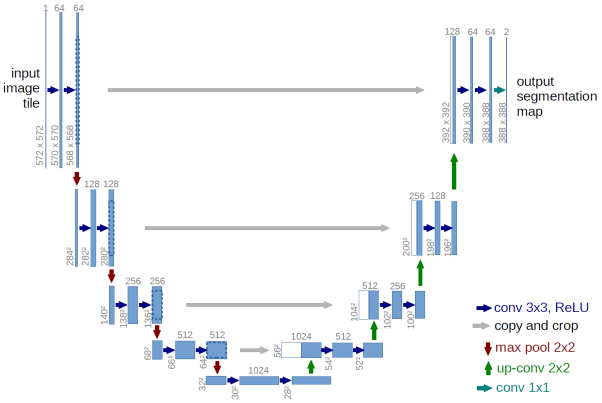

Figure 3: *U-Net Architecture*

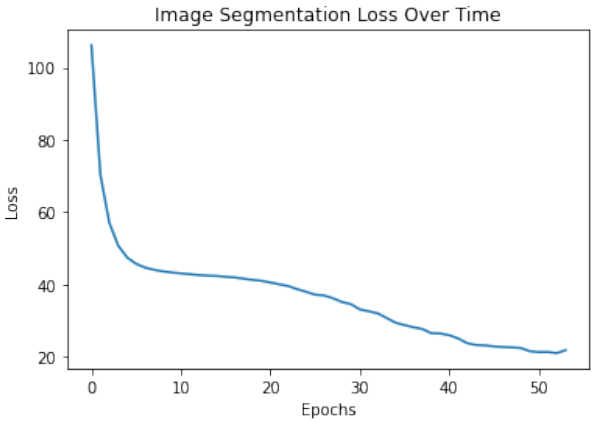

Figure 4: *Loss for U-Net Segmentation*

are above the clinical threshold, but with a limited dataset more research must be conducted. More on this in the discussion section.

Table 1: *Classification Metrics*

Sensitivity	.958
Specificity	.947
Accuracy	.953

4.2. Image Segmentation

Since the image segmentation task does not involve classification, I created my own metrics to be used to evaluate my results. My metrics will be referred to as agreement, disagreement and rank. These are defined as follows.

4.2.1. Agreement

Let I be a flattened 256x256 ground truth annotation and P be the predicted annotation. Let j be the subset of indices in I where $I_i > 0$. The variable n is defined to be the length of the true positive region, j. We can define the agreement for a

single annotation prediction pair as

$$agreement = \frac{1}{n} \sum_{q \in j} P_q \tag{3}$$

This value shows how well the predicted annotation put pixels in the places where endometriosis actually exists, basically true positives. This value also takes into the account the brightness of the pixels. If a pixel is in the correct place but is not very bright, the agreement will weight it accordingly.

4.2.2. Disagreement

Let I be a flattened 256x256 ground truth annotation and P be the flattened predicted annotation. Let t be the subset of indices in i where $I_i = 0$ and $P_i > 0$. The variable m is the length of the true negative region. We can define the disagreement for a single annotatoin prediction pair as.

$$disagreement = \frac{1}{m} \sum_{q \in t} P_q \tag{4}$$

This value shows how much the predicted annotation showed endometriosis where none exists.

4.2.3. Rank

Rank is what determines how ”good” an image is. This is defined for a single annotation prediction pair as

$$rank = \frac{agreement}{disagreement} \tag{5}$$

This value favors values with large agreement and small disagreement.

I wrote algorithms to compute these values. A summary of my results can be seen in table 2 and the top 5 ranked images can be seen in figure 5.

Table 2: *Segmentation Metrics*

Average Agreement	0.0280
Average Disagreement	0.0132
Average Rank	1.723
Rank Stdev	6.92
Max Agreement	0.3455
Min Disagreement	0.0027
Max Rank	14.97

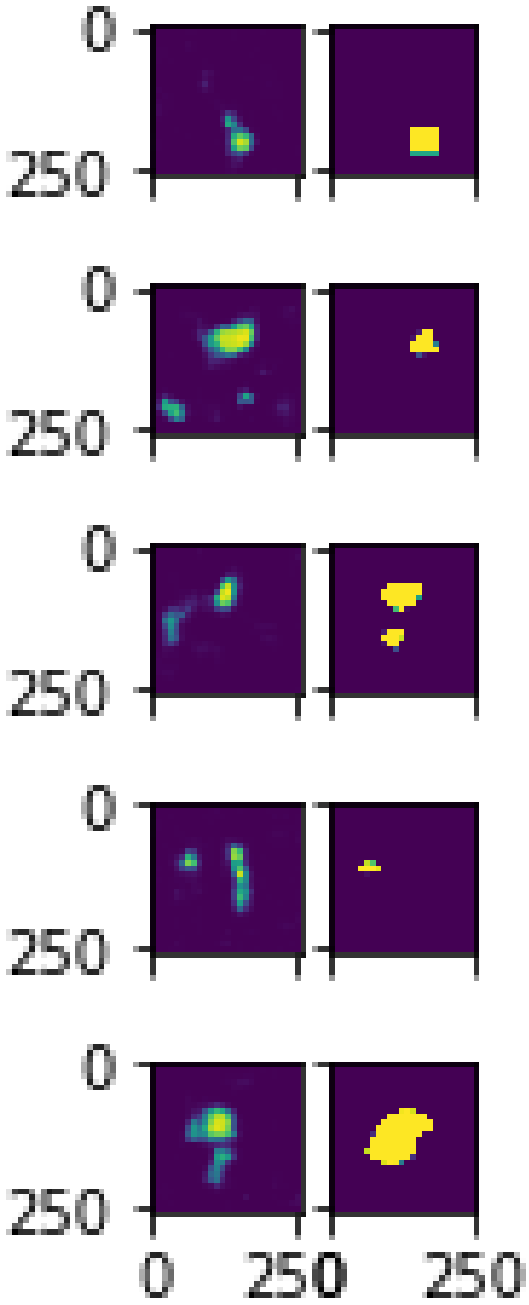

Figure 5: *Top 5 Images.*
Left: Prediction
Right: Ground Truth

5. Discussion

5.1. Image Classification

The results that were found for the image classification task are incredibly promising. Both the sensitivity (.958) and the specificity (.947) cleared the clinically relevant threshold, however this was only conducted from images taken from about 120 surgeries. This shows good evidence that machine learning can be used in clinical settings for endometriosis detection, but I believe that if I want to take this further, I will need to obtain more samples to make my model even better and confirm results.

5.2. Image Segmentation

My results found that the average rank of a prediction is 1.723 which is not very good compared to the best result of 14.97. My standard deviation was high at 6.92 showing that my model was incredibly variable in its results. Despite the average result not being good, I believe that the top ranked images show potential that this annotation system can work. The biggest issues that

this model faces is the fact that the amount of data was relatively small for a rather complicated task. U-Net does not require a ton of data, however the images that I am using vary significantly as endometriosis looks different depending on where it is in the body. This means that although I did have a decent amount of data, I might not have had enough data for each body location. Seeking out new image samples as well as augmenting my images more would be the next steps in improving the results.

6. Future Work

6.1. Improving Results

I would like to continue this work in the future. The main issue that I had was a lack of diverse data so that would be the first thing that I would like to investigate. I would like to figure out which augmentation techniques would be appropriate for this task. The original U-Net paper gives examples of augmentations, so implementing these would be my go to. I would also like to do more research to see if there is any improved version of U-Net that might improve the segmentation performance.

6.2. Extending Research

During my research I learned about the rasrm score which assigns endometriosis lesions points based on their size [6]. The amount of points a person has determines which stage of endometriosis a person has (1-4). I think that it could be possible to automate this process using machine learning tools. This may be a possible next step in my research.

7. References

[1] K. T. e. a. Zondervan, *Nat. Rev. Dis. Primer*, 2018.

[2] B. Eskenazi and M. L. Warner, "Epidemiology of endometriosis," *Obstetrics and Gynecology Clinics of North America*, vol. 24, no. 2, pp. 235–258, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0889854505703028

[3] P. A. S. S. e. a. Bendifallah, S., "Machine learning algorithms as new screening approach for patients with endometriosis," *Sci Rep*, 2020.

[4] A. Leibetseder, S. Kletz, K. Schoeffmann, S. Keckstein, and J. Keckstein, "GLENDA: gynecologic laparoscopy endometriosis dataset," in *MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II*, ser. Lecture Notes in Computer Science, vol. 11962. Springer, 2020, pp. 439–450. [Online]. Available: https://doi.org/10.1007/978-3-030-37734-2_36

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[6] D. HAAS, O. SHEBL, A. SHAMIYEH, and P. OPPELT, "The rasrm score and the enzian classification for endometriosis: their strengths and weaknesses," *Acta Obstetricia et Gynecologica Scandinavica*, vol. 92, no. 1, pp. 3–7, 2013. [Online]. Available: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1111/aogs.12026