Project Report - ECE 285

Yang Zhang
1st year CSE master
A59016268

Abstract

In this project, we aimed to address the task of pixel-wise lesion segmentation on the ISIC (International Skin Imaging Collaboration) dataset, which is an important question for skin cancer early detection. We modified the famous U-Net architecture and used Pytorch to implement the training process. The testing IoU could reach 82% accruacy. We will walk through the preparation of the dataset, the model architecture, our training details, and an evaluation of the model's performance in this paper.

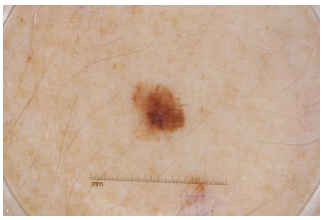
1 Introduction

Skin diseases hold significant importance in the field of healthcare due to their prevalence, impact on individuals' well-being, and potential risks they pose to public health. Segmentation of skin lesions is a crucial step in the diagnosis and monitoring of skin diseases, especially melanoma, one of the deadliest forms of skin cancer. Unfortunately, there are instances where doctors may fail to identify potential skin diseases based on the skin pictures because of two reasons:

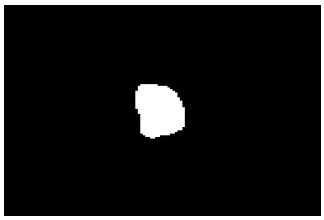
- A doctor's prediction power may get affected by his mood, current body condition, etc.
- Diseases or cancer at early stages are not visible enough for doctors to detect.

With these motivations, and also with the motivation that AI can help boost the medical care efficiency, we would like to utilize deep learning techniques to help doctors in their decision making process.

In this paper, we aim for solving the task that **given a picture of skin, provide each pixel a binary mask of whether it belongs to a lesion region (i.e. region of unusual skin)**. We believe that this task serves as a "first-step" task, meaning that doctors or researchers can utilize our masked region to perform further analysis based on their own needs, such as determining the type of disease this region might refer to. Essentially, this task is a system to filter out those healthy part of the skin so that doctors can look better into the skin of interest, as Figure 1 shows.



(a) Sample input



(b) Sample output

Figure 1: Example of task

With this task, we utilized and modified U-Net architecture, a model highly effective for biomedical image segmentation. U-Net is particularly well-suited for this task due to its capability to capture

both local and global information about the image. The training and validation losses, along with Intersection over Union (IoU) scores, were monitored throughout the training process where we used the BCEWithLogitsLoss criterion and Adam optimizer. Our results showed promising performance in segmenting skin lesions from the dermatoscopic images, in a sense that we raodomly tested 5 images of skin and achieved 82% IoU on these unseen pictures.

This paper is devoted into the following sections: Section 2 discusses the related work, Section 3 shows the detailed method including network architecture, testing algorithm, etc. Section 4 proposes the experiments and results. Section 5 includes a video demo of this project.

2 Related Work

Previous work about medical segmentation, also the foundation of our approach, is grounded in the seminal work of Ronneberger et al., titled "U-Net: Convolutional Networks for Biomedical Image Segmentation"[2]. This landmark paper, published in 2015, has been at the forefront of biomedical image segmentation due to the introduction of the U-Net architecture. U-Net, characterized by its "U" shape and efficient use of trainable parameters, is specifically designed to excel in tasks requiring precise localization, such as medical image segmentation. It achieves this by implementing a series of contracting (downsampling) layers followed by an equal number of expanding (upsampling) layers. In our paper, we use the U-Net architecture as the backbone of our model due to its excellent performance in semantic segmentation tasks. We did some improvements based on this model including batch normalization, dropout, etc. Also, our dataset is purely skin pictures instead of the biomedical image used in this paper.

Another relavant past paper is "Skin Lesion Segmentation from Dermoscopic Images Using Convolutional Neural Network" [3]. In this study, the authors proposed an approach that incorporates both ResNet and U-Net architectures for the task of skin lesion segmentation. Their model employs ResNet as an encoder for its ability to extract complex patterns and U-Net for its superior image reconstruction capability. In contrast, our work solely harnesses the power of the U-Net model. While ResNet's strength in learning intricate patterns is acknowledged, we found that U-Net, with its unique design emphasizing precise localization, suffices for the task at hand. This paper uses exactly the same dataset, but in terms of preprocessing, the authors chose to exclude blurred images, such as those containing hair, from the dataset to create a cleaner and more uniform set of images for training. While this step could potentially ease the learning process, we intentionally chose not to follow this practice. We believe that preserving the original integrity of the dataset, including its imperfect elements, is crucial as it reflects the variety of images we might encounter in real-world situations. Additionally, they converted the original RGB images into grayscale before feeding them into the model. In our project, we decided to keep the original RGB channels intact. We hypothesized that the color information, particularly variations in skin tones and the color of the lesion itself, could be valuable features for the model to learn.

An additional study that is closely aligned with our work is "Dermoscopic Image Segmentation via Multistage Fully Convolution Networks" by Bi et al[1]. This paper presents a novel approach to skin lesion segmentation that employs a multistage fully convolutional network. The authors argue that the multistage nature of their model allows for increasingly refined segmentations to be produced at each stage, enhancing the overall precision of the segmentation. The key difference between this study and our work lies in the complexity of the models employed. While Bi et al.'s multistage fully convolutional network represents a highly sophisticated approach to segmentation, our work leverages the power and simplicity of the U-Net architecture. We believe that the U-Net, despite its comparatively simpler structure, provides an excellent balance between performance and complexity, making it a highly effective tool for this task.

3 Method

3.1 Network Architecture

Our network is designed to input a $64 \times 64 \times 3$ RGB based skin picture and output a $64 \times 64 \times 1$ mask picture with only 1 or 0. Details are included in the next section. Now we start to describe our network design.

The network consists of an encoder path (downsampling) and a decoder path (upsampling). The encoding path begins with a double convolutional layer with 3 input channels (representing RGB image) and 64 output channels. Each double convolutional layer includes a 3x3 convolutional operation with padding 1, batch normalization, ReLU activation, and a dropout of 0.1 after the first ReLU activation. The output is then passed through another double convolutional layer, where the number of input channels matches the number of output channels from the previous layer, and the number of output channels is increased to 128. This process is repeated one more time, resulting in a final feature map with 256 channels. Also, after each double convolutional layer, a max pooling operation with a kernel size of 2 is applied to downsample the spatial dimensions by a factor of 2.

The decoding path starts with an upsampling operation that doubles the spatial dimensions of the feature map. The upsampled feature map is then concatenated with the corresponding feature map from the encoding path, resulting in an input with 256 + 128 channels. This concatenated feature map is passed through another double convolutional layer, which reduces the number of channels to 128. The same upsampling and concatenation process is repeated once more, with the output of the previous layer being upsampled and concatenated with the corresponding feature map from the encoding path. This time, the double convolutional layer reduces the number of channels to 64. Finally, a 1x1 convolutional layer is used to map the 64-channel feature map to the desired output channels, which in this case is 1 (representing the mask image with values of 1 or 0). To get the final output, we used sigmoid function on the last layer and set a threshold of 0.5 to classify 1s and 0s.

Our model is similar to U-net so it captures the advantages of U-net, as shown in Figure 2. For example, the skip connection in the second half of the network enables the model to capture both low-level and high-level features. However, looking back the traditional U-Net when it is first proposed, there are a few points that we have modified. First, we added batch norm and a small portion of dropout in the during the two convolutions at each stage. We believe this will not only train the network faster, but more importantly reduce the chance of over-fitting. We have a relatively small dataset and the network parameter is much larger, which is a sign of potential over-fitting. The second point is the original paper upsampled to 1024 dimension but we only reached 256 here. The primary reason is to boost the training process and reduce the complexity of network as well.

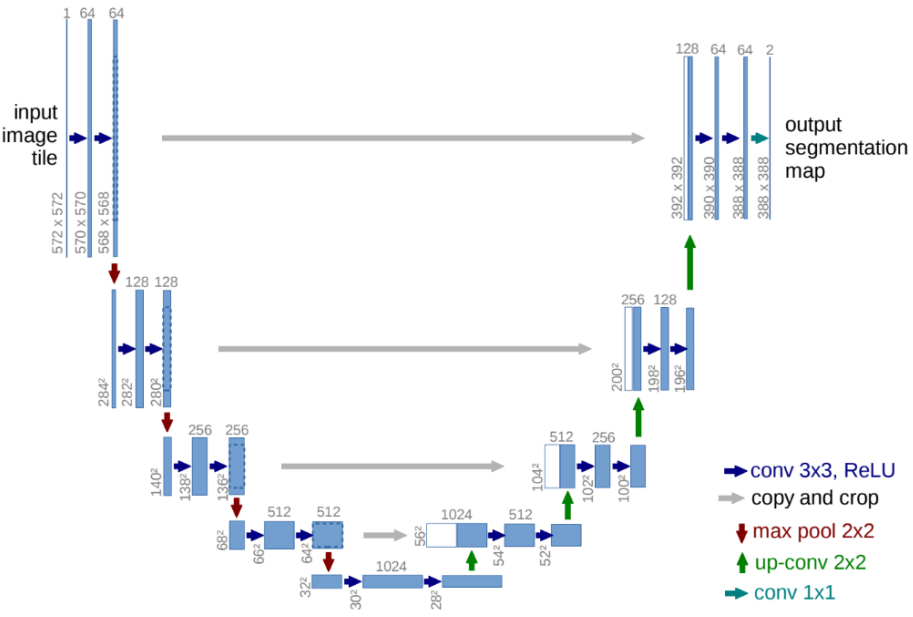


Figure 2: Traditional U-net architecture (adapted from paper)

3.2 Training and Evaluation Algorithm

Our approach to train the proposed U-Net model involves using a binary cross-entropy with logits loss function and Adam optimizer with a learning rate of 1e-4 using PyTorch. The training set is

batch-fed into the model with each batch containing a set of skin images and corresponding ground truth masks with batch size 40, prepared by DataLoader. The loss function is formulated as below:

$$\text{Binary Cross-Entropy Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

The model is then evaluated on a separate validation set. The model's performance is quantified using the Intersection over Union (IoU) score shown below, a common metric for evaluating the quality of an image segmentation. The IoU score measures the overlap between the predicted mask and the ground truth mask as a ratio of their combined area. We used IoU because we think it is a straightforward metric to determine whether the predicted mask is precise.

$$IoU = \frac{\text{Intersection area}}{\text{Union area}}$$

Finally, we also calculated the average training loss and IoU, as well as the validation IoU for each epoch. These recorded metrics provide insights into the model's learning dynamics and are used to plot a graph of loss and IoU over time. Results are shown in the next section.

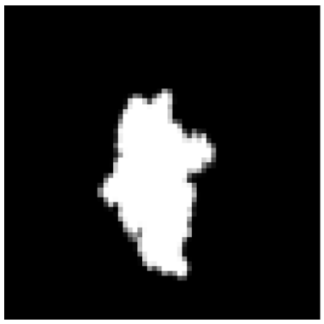
4 Experiments

4.1 Dataset and Data Preprocessing

The dataset we utilized is called ISIC, which stand for the International Skin Imaging Collaboration. It is an organization which produces widely used medical dataset. We downloaded its testing dataset which contains 2000 different sized 3-channel RGB pictures and correspondingly different sized 1-channel binary mask. They are in jpg and png forms so we simply used PTL library to load them. We also simply ignored some metadata attached to the images.

For data preprocessing, we resized all images and masks to 64×64 because we encountered memory issues when handling larger-sized images in our UCSD Datahub Environment. The largest original picture could reach 2000×2000 which is too big. One example of resized input and output is shown in Figure 3. The images are then prepared in DataLoader for training. For the validation set, we used 200 extra samples from ISIC and resized in the same way. For the testing example, we randomly sampled 5 representative pictures from ISIC as well, calculated the IoU, but more importantly visualize the mask result. Details are shown in the next subsections.

(a) Resized input



(b) Resized output

Figure 3: Example of resized input and output

4.2 Training Result

We trained the model for 20 epochs, and recorded the average IoU and loss for each epoch. The trends are plotted in Figure 4 below. We can observe that the validation loss shows similar trend with training loss, indicating that our model is not overfitting the data.

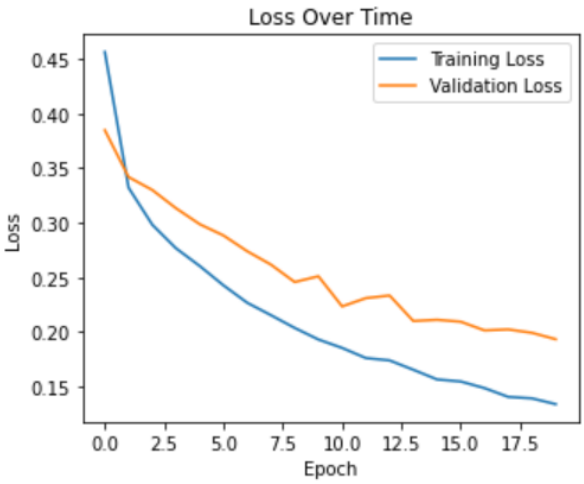


Figure 4: Training versus Validation loss

Numerically, as shown in Table 1, at epoch 20/20, the average training loss is 0.1336, and the average training IoU is 0.8107. The average validation loss is 0.1932, and the average validation IoU is 0.6954.

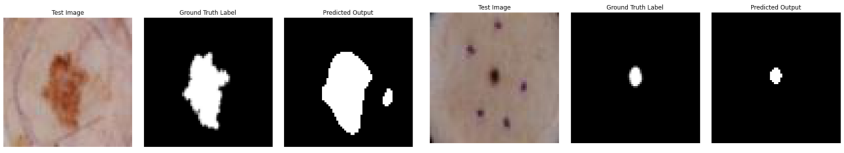
	Loss	IoU
Testing	0.1336	0.8107
Validation	0.1932	0.6954

Table 1: Loss and IoU results

As a sidenote, we initially tried classical U-net and the validation acuuracy was about 55%. Thus, our improvement of using batch norm and dropout is pretty effective.

4.3 Testing Result

We randomly picked 5 images from ISIC dataset, passed into our model, and visualized the result. The average IoU is 82% which is pretty good. We show two of the five here and more could be found in the code. In the first prediction, the major parts are correct but there is a little white dot on the right. This is understandable because in the raw input, the right part is a bit yellow-ish. In the second prediction, the model did almost a perfect job. This is where color plays an important role. Without color channel, the middle illness part is the same as the other four dots so that the model may not learn it. However, the middle dot is more black, and so the color information is fully used in the prediction.



5 Supplementary Material

A 5-minute presentation, including a testing sample demo, can be accessed at the following URL: `https://drive.google.com/file/d/13FgcBhQ-8wPmDXJztQzrClX8e38XRcDx/view?usp=sharing`.

References

[1] Lei Bi, Jinman Kim, Euijoon Ahn, Ashnil Kumar, Michael Fulham, and Dagan Feng. Dermoscopic image segmentation via multistage fully convolutional networks. *IEEE Transactions on Biomedical Engineering*, 64(9):2065–2074, 2017.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[3] Kashan Zafar, Syed Omer Gilani, Asim Waris, Ali Ahmed, Mohsin Jamil, Muhammad Nasir Khan, and Amer Sohail Kashif. Skin lesion segmentation from dermoscopic images using convolutional neural network. *Sensors*, 20(6), 2020.