2023

Detecting Overlapping Gene Regions Using the U-Net Attention Mechanism

Samuel Lemma
Minnesota State University, Mankato

DETECTING OVERLAPPING GENE REGIONS USING THE U-NET ATTENTION MECHANISM

A Dissertation
Presented to
The Academic Faculty

by

Samuel Lemma
Dr. Naseef Mansoor PhD

In Partial Fulfillment
of the Requirements for the Master's Degree of
Data Science in the
Minnesota State University, Mankato

Minnesota State University, Mankato
December 2023

Approved by:

Dr. Naseef Mansoor, Advisor
College of Science, Engineering & Technology
Department of Computer Information Science
Minnesota State University Mankato

Dr. John Burke
College of Science, Engineering & Technology
Department of Computer Information Science
Minnesota State University Mankato

Date Approved: [11/17/2023]

TABLE OF CONTENTS

Detecting overlapping gene regions using the U-Net attention mechanism **A**

List of tables **v**

List of figures **vi**

Abstract **viii**

Chapter 1. Introduction **1**
1.1 Scope of the Project and Contribution **2**
1.2 Organization of the Report **3**

CHAPTER 2. LITERATURE REVIEW **4**

CHAPTER 3. Background **8**
3.1 Convolution **8**
3.2 Pooling **9**
3.3 Attention **11**
3.4 Upsampling and downsampling **12**
3.5 Encoder-Decoder Architecture **13**

CHAPTER 4. Methodology **15**
4.1 Dataset **15**
4.2 Image Preprocessing **16**
4.3 Background for the Proposed Model **16**
4.4 Our Proposed Model **20**

CHAPTER 5. RESULTs **23**
5.1 Metrics for Evaluation **23**
5.2 Evaluation of Model Performance over Epoch **24**

CHAPTER 6. Conclusion **29**
6.1 Future Work and limitation **29**

REFERENCES **30**

LIST OF TABLES

Table 1	Breakdown of each Hyper parameters from our model	22
Table 2	The outcome of our model compared to previous work that was done in the same dataset	28

LIST OF FIGURES

Figure 1	A convolution in a 4x4 image and a filter sized 3x3 and bias 1x1	**9**
Figure 2	shows how max pooling work in a matrix way	10
Figure 3	An image example to describe attention mechanism in a real-world way	11
Figure 4	An original image on the left and up-sampled image on the right	13
Figure 5	Original image on the left and down-sampled image on the right	13
Figure 6	The Encoder_ decoder Architecture	14
Figure 7	Sample overlapping Image and ground truth from the data set.	15
Figure 8	Example of 2D Attention U-Net architecture	17
Figure 9	A deep explanation of Attention Gate	17
Figure 10	Mean Squared Error	24

Figure 11	Training and validation IoU for different epochs.	25
Figure 12	Training and testing loss per epoch.	26
Figure 13	The input image, the predicted image, and the ground truth from left to right.	27

ABSTRACT

The current issue of locating, diagnosing, and treating cancer and other diseases linked to specific target genes necessitates the creation of a reliable system for precisely identifying target genes that are initially extracted from a human chromosome. Current methodologies often suffer from overlapping gene regions in the target gene that occurs during the analysis process, which can have a substantial impact on the accuracy of the results. Our recommended approach, which was the appropriate model to apply for this particular problem, is set to enhance the analytical process by utilizing neural networks' U-Net with an attention mechanism. We were able to extract a result with 97.8% Validation accuracy from our proposed model by streamlining the process and generating more precise and timely results.

CHAPTER 1. INTRODUCTION

A probe is a single-stranded genetic material or sequence of RNA that is used to look for its complementary sequence in a sample genome[1]. Researchers can use the probe for testing different diseases and conducting research Researchers, for example, can employ probes to scan the genome for extra copies, which are typically found in tumors, or missing copies of specific areas of the genome, which are present in inherited diseases and cancers. Probes are used in various laboratories for multiple purposes. The first step in the process is culturing the DNA to find the target gene. The probing method is utilized to locate the target gene because genes reside in specific locations of cells and tissues. When retrieved using the culturing technique, a significant amount of useful information is lost. Once this procedure is completed, the target gene will be transferred to a lab for further research. Upon receiving the target genes, lab technicians follow specific steps for analysis. Before a better methodology was available, technicians analyzed each gene separately using the challenging manual microscopy method. This method was then replaced with on-screen analysis of digital images, providing substantial practical advantages [2]. This made it simple to complete the process and get an outcome that was better than the manual method. It was an effective approach to enhance productivity temporarily. However, subsequent issues arose after its implementation. As the images were scanned for analysis, an overlap in the probes began to occur, posing a significant setback in the healthcare industry. Analyzing the digital images simplified the process and improved outcomes compared to the manual method. Furthermore, this approach enhanced productivity. Furthermore, this solution was a good approach to increase productivity. However, due to overlapping gene

regions, certain complex images appear during the scanning of the target genes that are more complex than those that are common. For example, when the target genes are extracted using the Fluorescence in Situ Hybridization (FISH)[3],method some gene regions overlap with other regions complicating the gene identification process. Furthermore, these overlaps require manual analysis to ensure accurate segmentation of the physical properties of each gene region is preserved. This hinders the efficiency of the current gene analysis process which is optimized for non-overlapping gene regions only. Thus, detecting these overlapping gene regions and understanding the characteristics of each of these regions is of pivotal importance in gene research and thus impacts the healthcare industry in a broader sense.

1.1 Scope of the Project and Contribution

This work aims to develop an enhanced and optimized technique utilizing deep learning techniques for the identification of overlapping gene regions that often go unnoticed. To do this, we've developed a convolutional neural network (U-Net) model with an attention mechanism. The attention allows the model to concentrate on the most important pixels in the input image, increasing prediction accuracy. The proposed model substantially improves the accuracy of detecting overlapping gene regions and thus, improves the performance of gene analysis. To build the proposed model, we used the overlapping image dataset [4] containing images of both overlapping and non-overlapping genes. We also conducted a comparative performance study with existing methods. From our experimentation, we observe that the proposed model outperforms the existing models and has an accuracy of 97.8% in detecting overlapping gene regions.

1.2 Organization of the Report

The report is organized in the following manner. In chapter 2, we present an intensive literature review that discussed various approaches to detect overlapping gene regions. In chapter 3 we describe the U-Net architecture background of our model Then, in chapter 4, we go over our proposed model which was trained on the overlapping image dataset. In chapter 5, we present the results from our proposed model in detecting overlapping gene regions. Finally, chapter 6 presents the concluding remarks.

CHAPTER 2. LITERATURE REVIEW

The world is rapidly embracing the machine learning era. Numerous studies have been conducted on image segmentation and overlapping images using deep learning and other techniques. While most of the studies currently in use that we looked at do not use Gene region images, they still assure that the model that we used outperform models. Rather, this study, for instance, uses photos of disasters to automate and visualize the procedure. The goal of the research is to localize and classify damaged buildings at one time. To achieve this goal the author adopts the U-Net attention mechanism for model construction. A broad concept, the attention mechanism is now most frequently used in combination with the encoder-decoder framework but is independent of any one framework [5]. Essentially, this mechanism enables a neural network to focus on a specific subset of its inputs or features. Because of the attention mechanism, when information is chosen, it computes the weighted average of the N inputs before moving on to the following block. The U-Net's component includes an attention mechanism. After the experimentation, the author concludes that what channels are more crucial for the current task can be communicated to the network using the U-Net SE module. The data doesn't seem to have as much of a need to choose significant channels as photos with hundreds of channels do. The attention mechanism on the channels will work more effectively as the number of channels rises. After it was used to train various networks, and the F1 results on the verification set were compared. The U-Net model with the attention mechanism on two dimensions performs the best of them all [6], [7].

Furthermore, some papers suggest combining the U-Net model with different methods such as Test Time Augmentation. In order to adapt the model to various data sources, including street view data and lower resolution remote sensing data, the author plans to explore multiple techniques. The main proposed method is using U-Net architecture created by

This architecture comprises two paths that are joined by concatenating the respective up-sampling and down-sampling layers on each path, resulting in a stable and accurate segmentation outcome [8]. Because of its form, it produces a segmentation result that is steady and accurate. The U-Net in this research paper is used with test time augmentation distinguishing it from the proposed method. During training, test time augmentation multiplies and alters the training data. There are a total of 15 training epochs employed, and the Keras Python library with TensorFlow is used for deep learning for this model [8]. The author concludes the paper by emphasizing on when test time augmentation is used for the training dataset, the model has a higher likelihood of accurately capturing the shape of overlapping chromosomes, which enhances prediction accuracy and test image quality. This is how it affects the result in terms of increasing the semantic segmentation accuracy.

The U-Net model is still the favorite one for biological image segmentation. There are also some other models that can perform well for image segmentation purposes. Alex-net with attention mechanism is the other model that some researchers recommended for biological image segmentation. Combining attention mechanism with Neural networks is really one of the best ways to have a better segmentation for our image. For example, one of the research projects primarily analyzes how the attention mechanism works and how it might be used to identify micronuclei in a cell picture. Which cell pictures or micronuclei are very small and need to have an attention weight toward the target part of the image [9]. To extract cell image elements and create attention maps highlighting the area of interest, two attention modules are used. Researchers use data augmentation and focus loss to lessen the impact of the issues in the data set. The author compares Alex-net and other convolutional Neural networks and claims that AlexNet has outperformed all the other CNNs by comparing it based on different data sets. We have seen throughout the literature review that authors are trying to prove that their models are accurate and better performers than other models. Using AlexNet to detect biomedical images since it is more effective and

sufficient to complete the task. Even though the outcome was achieved by combining Alex-net and the attention mechanism. While the classification performance is improved by the attention mechanism, the strategy produces better outcomes using the same inputs.

In the field of image segmentation, combining attention mechanisms with other models produces a lot more fruitful outcome. A strategy based on improved U-Net is suggested by several researchers to lessen the segmentation loss of insulators in aerial photographs which needs a better view in the image. The enhanced process requires combining the attention mechanism with U-Net., In the coding stage of the U-Net network model, the method embeds ECA-Net in each preliminary effective feature extraction layer [10]. The U-Net SE module's fully connected layer operations are optimized in a 1-dimensional sparse convolution algorithm. The technique significantly lowers the number of parameters while maintaining an impressive level of performance. The F1-Score is presented as a more thorough assessment of the classifier's performance [11]. In order to evaluate the algorithm's effectiveness objectively, the author of this work uses three widely used indicators. For a total of 200 Epoch iteration cycles, the model in this study is trained [12]. This model freezes a portion of the neural network for training in the first 50 epochs of training to expedite training and prevent weights from being lost during the early stages of training. A method has been created by a research team to increase the precision with which various insulator kinds may be separated against complex backdrops. The approach involves embedding ECA-Net into the original U-Net model's downsampling section [12]. The proposed method aims to enhance the precision with which various insulator types can be identified against complex backgrounds.

Although combining attention mechanisms with other models appeared to be the preferable strategy, the model in question was still ambiguous. U-Net appeared to be the most

trustworthy image segmentation model after all the reviews. A lot of the research paper explains deeply how U-Net is more effective than other models. From the review what we learnt is that U-Net has become more applicable to any image segmentation problem. The model includes a symmetric expanding path that allows exact localization and a contracting path to capture context [13]. The fundamental idea of the model is to replace a typical contracting network with layers as a supplement, with upsampling operators taking the role of pooling operators. This model can be combined with an object detection algorithm to create a system that can run on whole microscopy images. Then, using the U-Net as a foundation, an image segmentation algorithm can segment overlapping gene regions.

In conclusion, the comprehensive literature review highlighted a variety of strategies for biological image segmentation, ultimately confirming the efficacy of the U-Net and attention mechanism approach. We initially gathered the models that are applicable for biological image segmentation based on the research papers in order to determine which technique would be the best. The U-Net has proven to be versatile and applicable to a wide array of image segmentation challenges. It has been demonstrated that the U-Net is adaptable and suitable for a broad range of image segmentation problems. The solid performance of the model is demonstrated by its ability to accurately detect gene regions using a combination of U-Net and the attention mechanism. To detect overlapping gene regions a crucial area that has not received much attention, more research is required to examine the application of the U-Net and attention mechanism.

CHAPTER 3. BACKGROUND

3.1 Convolution

Convolution plays an essential role in the machine learning world that originated from mathematics and is used commonly in convolutional neural networks (CNN). Convolution is best described in mathematical way as a combination process of two functions to create a third one [14]. There are multiple types of convolution operation, each serving different purposes. There are three main types of convolutions: 1D convolution, which is applied to sequences or time series data; 2D convolution, usually used in image processing; and 3D convolution, frequently used for video data or high-dimensional medical pictures. Each serves specific purposes. [15]. In our research paper we use a 2d convolution layer which is common for a biomedical image. Convolution is used to extract features from an image which makes it vital in several deep learning techniques. Here we have used a simple analogy to explain this concept. In our everyday life, we look at things and learn what they are and can recognize them in the next glance. This is achieved by separating some features in one section of the image from another based on brightness, color, and shape differences. The same concept applies in convolution, it extracts useful features from an image. In a CNN, these features are fed to a classification layer which is trained to recognize the image based on these features. This procedure allows the network to learn structures that are hierarchical by capturing specific patterns. This process involves the use of 2D convolution in image segmentation and classification tasks. The example in Figure 1 shows how useful features are extracted from an image using matrix. The 3x3 matrix is multiplied by the

input 4x4 matrix to produce the convolution output. The feature map displays the strengths at each location in the images the stronger the feature the brighter the pixel [16].

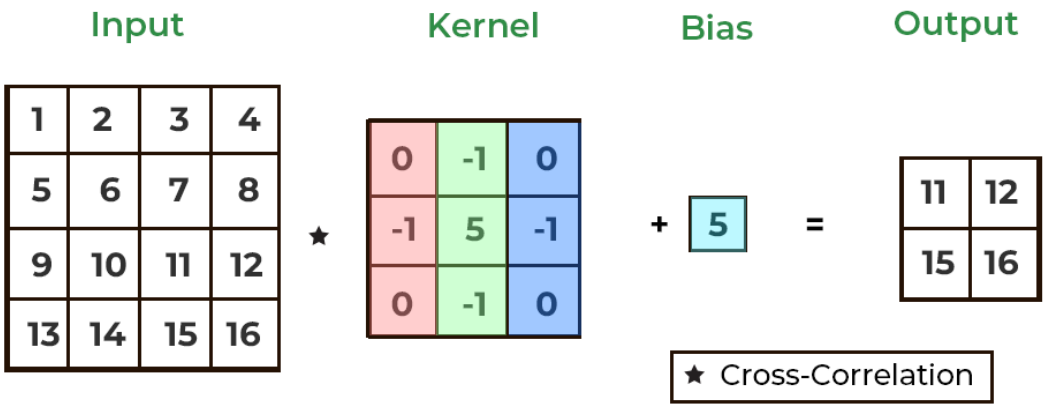

Figure 1 A convolution in a 4x4 image and a filter sized 3x3 and bias 1x1 [16]

3.2 Pooling

Pooling is a common method used on CNN. It offers a way for down sampling feature maps by adding the features in feature map patches. There are two common techniques to pooling, the first being average pooling and the second being max pooling. The average presence of a feature is determined by average pooling, whereas max pooling selects the most activated features or the greatest value in the matrix.[17].

Max pooling is a type of procedure that is commonly used in CNN to downsample the feature map from a given image while maintaining useful information. In our model, we used max pooling, which is applied after each convolution layer and takes the maximum

value from each window and passes it to the next layer of the network as shown in Figure 2. Max pooling discards some information from the feature map, reducing the risk of the network overfitting to the training data. Moreover, it offers another advantage: the loss of certain features during downsampling diminishes the number of network parameters, leading to a reduction in the computational resources required for training and testing the model.

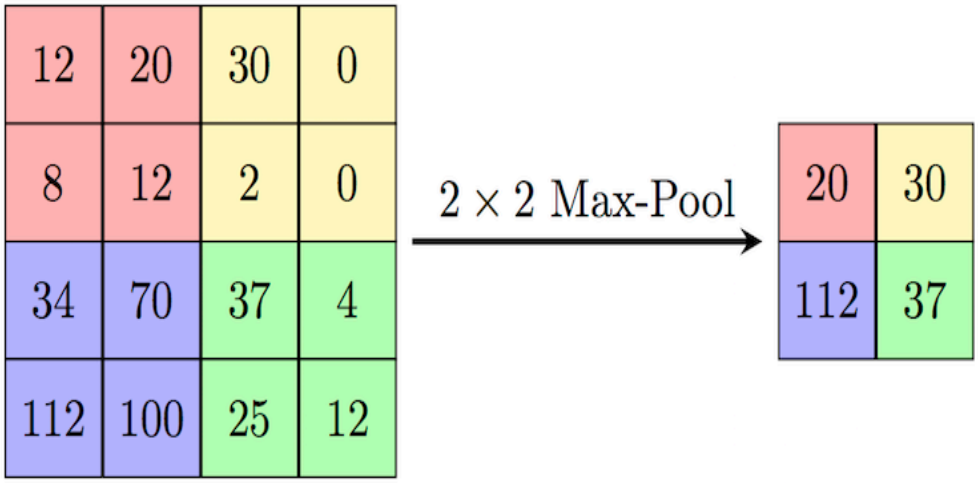

Figure 2 shows how max pooling work in a matrix way [31]

3.3 Attention

As the term indicates, attention is a strategy used in many neural networks to focus on the most significant parts of the input data[18]. It has evolved into a fundamental element in deep learning models, particularly prevalent in natural language processing (NLP) and image recognition applications. Attention really implies what the definition stands for; it's only used in diverse ways depending on the problem. For example, if you are looking at Figure 3 on a street and someone says, "The yellow taxi is mine," your attention will be directed to the photograph in search of a yellow cab, ignoring everything else in the photograph.

Figure 3 An image example to describe attention mechanism in a real-world way [32]

There are different kinds of attention mechanisms like Soft attention (Each input element is assigned a probability, allowing the model to pay attention to numerous elements at the same time), Hard attention (makes a hard judgment on which input is crucial for each

position.), Self-attention (this allows the model to focus in different features in the same input), Additive attention (this combines the attention vectors using element-wise addition), Multiplicative attention (it combines the vectors using multiplication) and Concatenative attention (in this attention the vectors are concatenated before being applied to the transformers) are some of the many types of attention that are all used for different purposes [18].

When it comes to image-based works, there are two sorts of attention mechanisms that are applied with various models: spatial attention and channel attention. The emphasis of spatial attention is on the exact regions that are most relevant to the task. It figures out how to create a weighted map for each pixel in the feature. Weights are assigned to pixels that are more relevant to the task. Channel attention focuses on specific channels rather than the features therefore whichever channel has more relevance to the task will be given more weight.

3.4 Upsampling and downsampling

Upsampling and downsampling are operations that are frequently used processes, especially in relation to CNN and image processing. upsampling involves increasing the spatial dimensions of an input signal or image. Upsampling techniques used in CNNs include transposed convolution (also known as deconvolution) and bilinear interpolation. Transposed convolution upsamples the input using learnable filters, whereas bilinear interpolation computes new pixel values based on the weighted average of nearby pixels, resulting in smoother transitions. In contrast, downsampling is the process of lowering the spatial dimensions of an input signal or image. Max pooling is a typical downsampling

approach used in CNNs that retains the largest value in a small region while discarding the others. This procedure shrinks the size of the feature maps while still capturing important information and increasing computing performance [19].

Figure 4 An original image on the left and up-sampled image on the right [19]

Figure 5 Original image on the left and down-sampled image on the right

3.5 Encoder-Decoder Architecture

The Encoder-Decoder architecture is widely used in deep learning for tasks like image segmentation, natural language processing, and speech recognition.[20] As shown in Figure 6, the Encoder receives the input and processes it, after which the Decoder produces the output after several stages.

The architecture is made up of two blocks: encoders and decoders. The encoder block is in charge of processing the input and converting it into a representation that shows only the important parts of the input. The Encoded state is the name given to this representation. Each has elements such as self-attention and feedforward. The self-attention allows the Encoder to concentrate on various aspects of the input features. The feedforward layer transforms the processed input using nonlinear transformations, allowing Increased level of representation obtained from the input.[18]

The Decoder on the other hand is in charge of generating the output from the encoded representation of the Encoder layer. It begins with the state from the Encoder layer as a starting point and generates the output one feature at a time. The Decoder layer also contains masked self-attention and feed-forward in Decoder block. The masked self-attention ensures that it only attends to the present features so that it doesn't look at the features that are going to come. The Encoder-Decoder architecture has proven to be effective in wide range of deep learning problems.

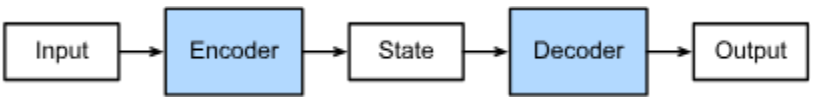

Figure 6 The Encoder_ decoder Architectu [35]

CHAPTER 4. METHODOLOGY

4.1 Dataset

In this study, we used a dataset previously employed in related research works [4]. The dataset primarily focuses on microscopic images of gene regions within the domain of biology and genetics. The data set was collected using thousands of semi-synthetically generated overlapping chromosomes [4]. The dataset consists of 13,434 microscopy images of target genes or gene regions. The validation of these images and their corresponding regions was conducted by multiple researchers [21]. A sample picture from the dataset demonstrating the microscopic image and the ground truth is shown in Figure 6. The red segment is the overlap between the two genes regions. similar to this example, all images within the dataset are accompanied by a ground truth label that identifies distinct areas within the picture.

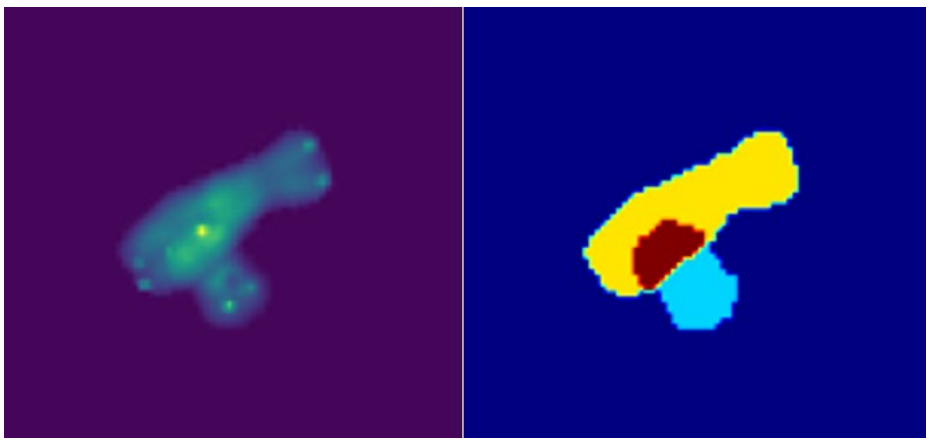

Figure 7 Sample overlapping Image and ground truth from the data

4.2 Image Preprocessing

Data augmentation was used to improve the accuracy and diversity of the training data. We cropped and zoomed in on the images to improve the data. Cropping an image to a specific size, which enables image resizing to a common dimension. These functions work to prepare and improve our labeled image datasets, making them suitable for training and evaluation for the image segmentation task we will be performing. This helped in generating a more diverse set of images, mitigating the risk of overfitting. This contributed to the generation of a more diverse set of images, reducing the risk of overfitting. We were able to generate 13484 images for training the model using augmentation. An example image with this improvement is shown in Figure 6.

4.3 Background for the Proposed Model

We were able to identify various approaches to the detection of an overlapping gene region based on the literature reviews. Most of the suggested models had a rather low impact and low accuracy. According to the current state the problem still exists, and it is causing a lot of problems for the people who analyze the gene region. The analysis typically fails when separating the image from the background, where segmentation is a key component. Failing to detect the overlapping gene regions, which is a major setback for the procedure because of the gaps that are available here.

Detecting the overlapping regions using U-Net and attention mechanism is a better approach to achieve the desired outcome according to research on previous work. The

architecture that we are going to use is U-Net. U-Net is a biomedical image segmentation convolutional neural network[22], [23]. The segmentation map generated by the model has been constructed to match the dimensions of the input image to shorten storage and processing [22]. By adding the attention gate to the model for a better result we can be able to obtain an accurate prediction compared to the ground truth. The purpose of combining the two, is images of the gene regions will have focus on some of the useful features using the attention mechanism [18]. It helps to enhance the model performance; The attention can automatically identify and highlight image segments that are the most valuable in the scanned image. For medical image analysis, single-layer features extracted from the image were used. Single layer features reflect the cell pictures just partially and have a limited amount of information [24]. By integrating deep and shallow characteristics, the attention mechanism in this network creates more relevant features [9]. After focusing on the part where the gene region tends to overlap, we use an image segmentation method to detect the signals.

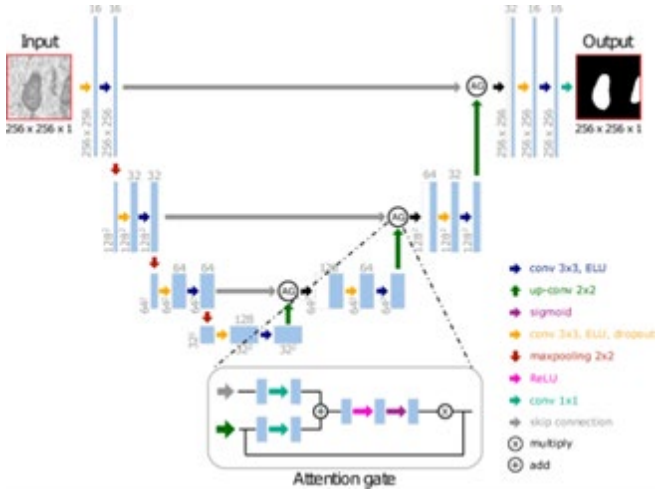

Figure 8 Example of 2D Attention U-Net architecture [33]

Figure 8 shows a basic 2D Attention U-Net architecture with three levels of downsampling and a detailed description of the skip connections' attention gate. Attention gates are an effective method to constantly Improve the U-Net in a variety of datasets without incurring significant computational costs. As training progresses, the network learns to concentrate on the desired area. The attention coefficients get better at highlighting important regions because of the differentiable nature of the attention gate, which enables training during backpropagation.

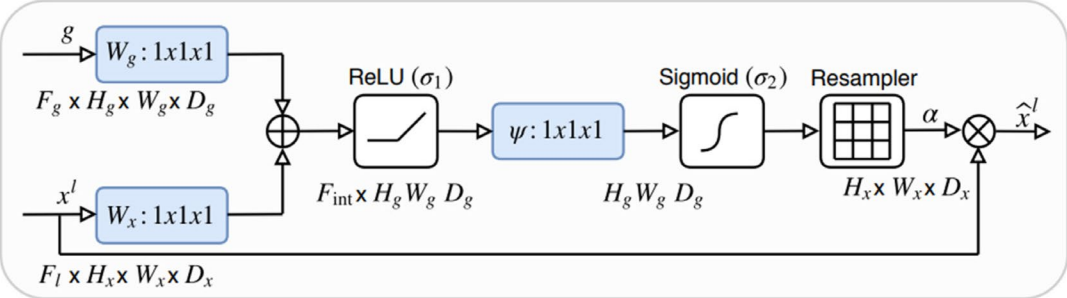

Figure 9 A deep explanation of Attention Gate [34]

In Figure 9, the vectors 1x1 and the skip connection are the two inputs that the attention gate accepts. The network's next-lowest layer yields the vector, skip connection. Because the vector comes from a deeper section of the network, it has fewer dimensions and better feature representation. In Figure 9, vector 1x1 would be 64x64x64 (filters 1x1 height 1x1 width), and vector g would be 32x32x32. [25]. In order for vector 1x1 dimensions to become 64x32x32 and vector skip connection dimensions to become 64x32x32, they each undergo a stride convolution. The two vectors' elements are combined. As a result of this process, aligned weights grow while unaligned weights shrink. The resulting vector is subjected to a 1x1 convolution and a ReLU activation layer, which reduces the dimensions to 1x32x32. The attention coefficients (weights) for this vector are produced via a sigmoid

layer that scales the vector between [0,1], with coefficients closer to 1 representing more relevant information. Our model takes the U-Net architecture and combines with the attention mechanism to make the feature maps better during the decoding process. Applying this technique as discussed above will help to focus on the relevant features from the encoder when generating the output which increases the accuracy of our model. Our structure consists of different functions:

The first function, the attention mechanism, operates on the encoder's features by giving some features more weight than others.

The following two blocks are the decoder and encoder blocks, which both follow the standard U-Net structure[26]. In the encoder block we reduce spatial dimensions and increase the number of filters (depth). In the decoder block an up-sampled layer followed by the attention mechanism and it is the reverse of the encoder as discussed above.

To ensure that the network benefits from the encoder's high-resolution features, the attention gate output is concatenated with the up-sampled features. Two convolutional layers with ReLU activations are applied to this combined feature map.

Finally, by using the functions we construct a U-Net with attention model which consists of encoder block, decoder with an attention gate, bottleneck, and final layer.

4.4 Our Proposed Model

The attention U-Net that we used for this task is made up of various functions and hyperparameters. The Attention U-Net architecture is made up of several encoder and decoder blocks that incorporate attention methods to improve feature representation. The encoder is made up of three blocks, each with two convolutional layers with 3x3 kernels and ReLU activation functions. These encoder blocks' aim is to gradually reduce spatial dimensions and increase the number of filters used to capture hierarchical information in the input image. The input dimensions for each block are determined by the output of the previous block, with the initial block taking the dimensions of the input image size (88,88,1) as shown in Table 1. The number of filters supplied in the code determines the output dimensions of each block.

The decoder is an exact duplicate version of the encoder, with two decoder blocks. The attention gate module and transpose convolutions for upsampling are the two primary components of the decoder block. Before combining with the upsampled data, the attention gate module selectively filters and accentuates relevant features. Convolutional layers are used in the attention mechanism, with the 'inter_channel' parameter determining the number of channels. The attention gate output is concatenated with the upsampled features, and the resultant tensor is passed through two convolutional layers using ReLU activation functions. This procedure assists the model in recovering high-level spatial information that was lost during downsampling.

The final layer is a 1x1 convolutional layer with four output channels on for each segmentation job.

Experimenting with different hyperparameter values in machine learning models, such as adjusting filter numbers, kernel sizes, and activation functions, allows you to evaluate the model's adaptability and performance across a range of complexities. Systematic experimentation assists in the identification of optimal patterns for specific tasks that balance accuracy, efficiency, and generalization, which is critical for improving model performance. Therefore, we decided to use the following hyperparameter that are shown in Table 1 for our model after trying different hyperparameters.

Table 1 Break down of each Hyper parameter from our model.

Hyperparameter	**Value**	**Description**
Number of Classes (Output Channels)	4	The number of output channels in the final convolutional layer
Encoder Filters	[64, 128, 256]	Number of filters in each convolutional layer of the encoder blocks.
Decoder Filters	[128, 64]	Number of filters in each convolutional layer of the decoder blocks.
Convolutional Kernel Size	(3, 3)	Size of the convolutional kernels used in the encoder and decoder blocks.
Attention Mechanism Channels	inter_channel	Number of channels used in the attention mechanism when calling the attention gate
Deconvolution Kernel Size	(2, 2)	Deconvolution kernel size used in decoder blocks.
Pooling Size	(2, 2)	Size of the max pooling used in the encoder blocks.
Activation Functions	'relu' and 'sigmoid'	ReLU activation is applied in convolutional layers for gating, while sigmoid activation is used in the attention mechanism.

CHAPTER 5. RESULTS

To evaluate the proposed model, we used different metrics that tell us the precision comparing the output image with the original image as shown in Figure 13. The weighted loss, accuracy, and Intersection over Union (IoU) [27] score, was utilized to evaluate the proposed model. The section below discusses how the metrics are calculated for the given segmentation problem.

5.1 Metrics for Evaluation

The one metric we used to evaluate the result is the IoU which is by dividing the area of overlap with the area of union. We can be able to obtain a score for it and evaluate it with the ground truth. The IoU measures the overlap between the ground truth and the predicted image[28]. It evaluates the overlapping region between the two to determine how well the predicted segmentation output fits with the actual ground truth. It computes the area where the anticipated and ground truth masks overlap in relation to the overall area of overlap and non-overlap. A high IoU indicates that the model's predicted segmentation matches well with the ground truth and vice versa. If the IoU is low it indicates that potential issue with the model performance.

The other metric used to evaluate the result of the model was loss. This quantifies the inconsistency of the prediction with the custom target [29]. These metrics must go down as the epoch changes. The decrease of the loss function indicates that the model is improving as it progresses through the specified number of epochs. The loss is calculated using the Mean Squared Error (MSE) loss function in our model. MSE assesses the average

squared deviations between predicted and true pixel values across all samples in image segmentation where the model predicts pixel values. This option tries to reduce the average squared error between the predicted segmentation and the ground truth labels.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Figure 10 Mean Squared Error

The third and main metric used was accuracy. As we all know, in various machine learning models' accuracy is the main target we look for to justify if the model has performed well or not. The accuracy indicates the percentage of correctly classified instances. More than simply improving model capacity (uniformly) across all network layers, the addition of Attention Gate makes a big contribution to the score we obtained from the model. The standard accuracy metric is used to calculate accuracy, which is the ratio of correctly classified instances to the total number of occurrences in the dataset.

5.2 Evaluation of Model Performance over Epoch

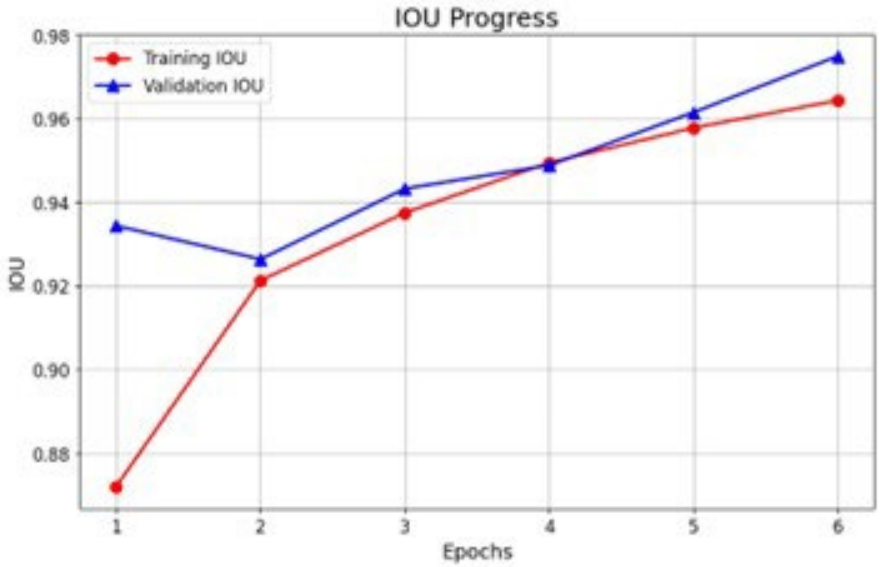

Figure 11 Training and validation IoU for different epochs.

Based on our experimentation when we were performing the training and validation over 6 epochs our model performed well throughout the training process, the rising trend in both the accuracy and IoU metrics highlights the model's improved ability to discern and precisely outline objects within images as shown in Figure 11.The model appears to be learning complex patterns and features as it goes through multiple epochs, enhancing its ability to accurately identify and classify the overlaps in the dataset. The concurrent advancement of the accuracy and IoU metrics demonstrates the model's ability not only to recognize objects correctly but also to precisely their boundaries, demonstrating its potential for detailed image analysis and segmentation tasks. Furthermore, the model's outstanding results on validation data, which it did not encounter during training, demonstrate its ability to generalize well to previously unseen instances. This ability to perform well on previously unseen data indicates that the model has a solid understanding of the underlying structures and features of the things, allowing it to make accurate

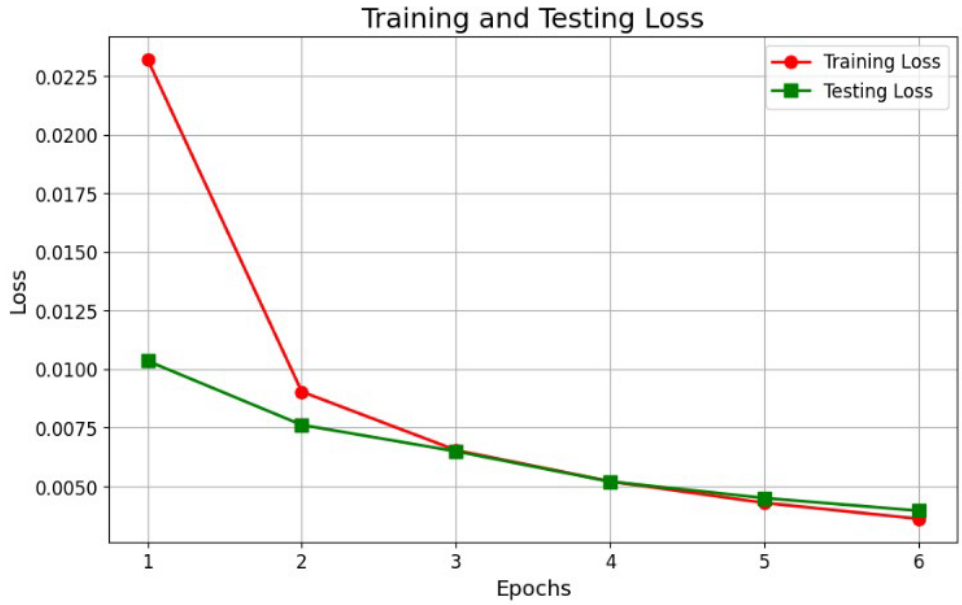

Figure 12 Training and testing loss per epoch.

predictions even in contexts that are unfamiliar. The model's continuous learning and refinement highlight its dependability and effectiveness in precisely categorizing, detecting, and delineating objects within various images, demonstrating its potential for practical real-world applications in image analysis and classification. The observed improvement in the accuracy metric, which reached 97.84% during training and 97.71% during validation, demonstrates the model's capacity to properly categorize and identify items within the images. Simultaneously, with a validation score of 96.33%, the IoU measure shows a consistent climb, suggesting the model's developing competence in precisely distinguishing object boundaries. These measures demonstrate the model's robust learning and potential for dependable performance in complex image segmentation tasks.

The model's capacity to reduce the variance between the expected output and the actual target labels is demonstrated by the improvement in loss throughout the training phase. Within the framework of our model, the loss function is an essential tool for measuring the difference between the ground truth labels and the predicted segmentation map [30]. The model learns to modify its parameters as training proceeds through each epoch to minimize the overall error between the true labels and the predicted segmentation outputs as shown in Figure 12. As a result, a decrease in the loss value suggests that the model is approaching a point where its predictions are more in line with the labels of the ground truth. Figure 13 shows some examples of the prediction from our model.

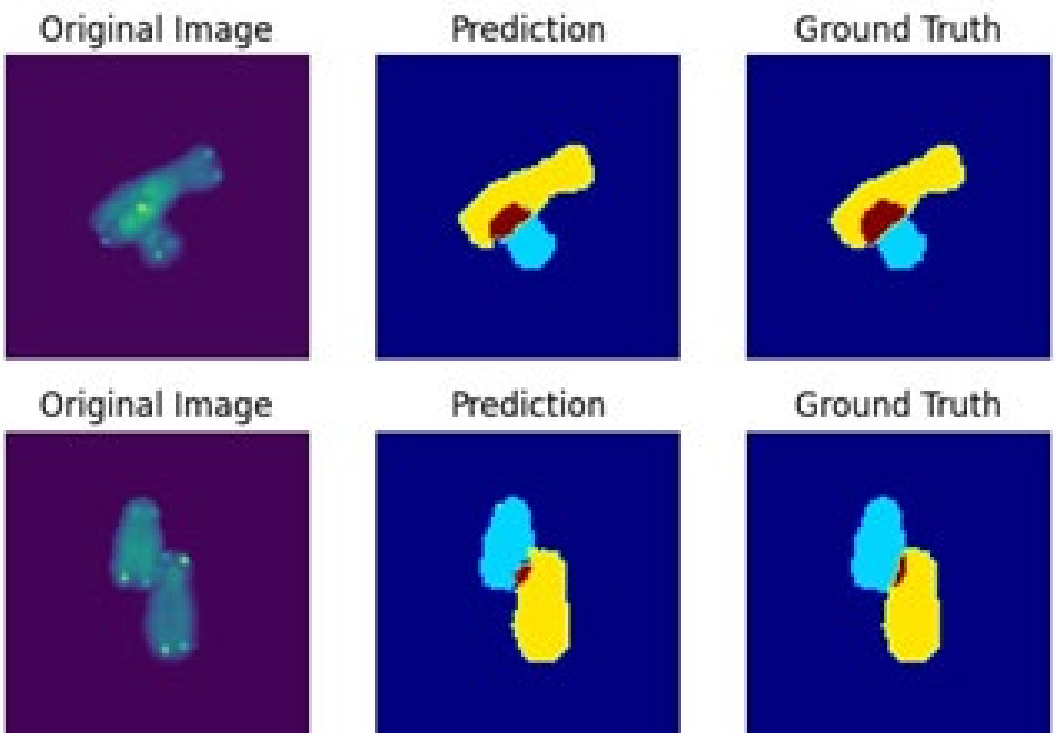

Figure 13 The input image, the predicted image, and the ground truth from left to right.

Table 2 Comparative study of the proposed model with existing models

Research	**Train Accuracy**	**Train IoU**	**Validation Accuracy**	**Validation IoU**
Our Model (**U-Net** with Attention)	0.9903	0.9708	0.9896	0.9723
Former model (**U-Net**)	N/A	0.9470	N/A	0.8820

We also compared our model to earlier work on this exact problem, which used a U-Net model without an attention mechanism. The current model performs better in terms of IoU metrics and both training and validation accuracy as shown in Table 2 Result. Our model shows higher performance with a validation IoU of 97.23% and accuracy of 98.6%. The previous paper attained an IoU of 94.7% for the overlapping and 88.2% and 94.4% for the gene regions [21]. This shows that our suggested mode (U-Net with attention) has a stronger ability to perform image segmentation better.

CHAPTER 6. CONCLUSION

In conclusion, the large improvements in both the accuracy and IoU metrics highlight our model's significant success in effectively segmenting complicated and overlapping gene regions, as demonstrated in Figure 5. By combining the improved U-Net with attention, as previously discussed, we have outperformed past techniques, delivering a significant increase in accuracy shown in Table 1 Result. This significant step underscores the critical role our model may play in genetic research and analysis, providing a more comprehensive and precise understanding of complex genomic landscapes. Our model has the potential to catalyze breakthroughs in overlapping gene region segmentation, leading to larger advances in genomic research and stimulating fresh insights into the intricacies of genetic structures due to its increased precision and capabilities.

6.1 Future Work and limitation

Despite the accomplishments, it is needed to see some limitation from the work. The model's performance may be influenced by variances in input data quality, and it may encounter difficulties in cases with a wide range of overlapping gene region patterns. Future research should investigate approaches to improve the model's robustness to different genomic images, as well as the scalability of the proposed approach to bigger datasets.

REFERENCES

[1] A, “Probe”, Accessed: Oct. 08, 2022. [Online]. Available: https://www.genome.gov/genetics-glossary/Probe

[2] Ab MetaSystems, “Automated slide scanning and imaging | ”, Accessed: Nov. 11, 2023. [Online]. Available: https://metasystems-international.com/en/products/solutions/automatic-imaging-scanning/

[3] A. R. Shakoori, “Ac Fluorescence In Situ Hybridization (FISH) and Its Applications,” *Chromosome Structure and Aberrations*, p. 343, Jan. 2017, doi: 10.1007/978-81-322-3673-3_16.

[4] Jen Pat, “AD Overlapping chromosomes.” Accessed: Nov. 11, 2023. [Online]. Available: https://www.kaggle.com/datasets/jeanpat/overlapping-chromosomes/data

[5] stefania cristina, *The Attention Mechanism from Scratch - MachineLearningMastery.com*. Accessed: Nov. 06, 2023. [Online]. Available: https://machinelearningmastery.com/the-attention-mechanism-from-scratch/

[6] J. Z. Xu, W. Lu, Z. Li, P. Khaitan, and V. Zaytseva, “Building Damage Detection in Satellite Imagery Using Convolutional Neural Networks,” Oct. 2019, Accessed: Nov. 11, 2022. [Online]. Available: http://arxiv.org/abs/1910.06444

[7] E. Weber and H. Kané, “Building Disaster Damage Assessment in Satellite Imagery with Multi-Temporal Fusion,” Apr. 2020, Accessed: Nov. 11, 2022. [Online]. Available: http://arxiv.org/abs/2004.05525

[8] H. M. Saleh, N. H. Saad, and N. A. M. Isa, “Overlapping Chromosome Segmentation using U-Net: Convolutional Networks with Test Time Augmentation,” *Procedia Comput Sci*, vol. 159, pp. 524–533, Jan. 2019, doi: 10.1016/J.PROCS.2019.09.207.

[9] W. Wei, H. Tao, W. Chen, and X. Wu, “Automatic recognition of micronucleus by combining attention mechanism and AlexNet,” *BMC Med Inform Decis Mak*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12911-022-01875-w.

[10] G. Han *et al.*, “Improved U-Net based insulator image segmentation method based on attention mechanism,” *Energy Reports*, vol. 7, pp. 210–217, Nov. 2021, doi: 10.1016/J.EGYR.2021.10.037.

[11] “F-Score Definition | DeepAI.” Accessed: Nov. 07, 2023. [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/f-score

[12] G. Han *et al.*, “Improved U-Net based insulator image segmentation method based on attention mechanism,” *Energy Reports*, vol. 7, pp. 210–217, Nov. 2021, doi: 10.1016/J.EGYR.2021.10.037.

[13] Babina Banjara, “Demystifying UNet and Learning Image Segmentation”, Accessed: Nov. 07, 2023. [Online]. Available:

https://www.analyticsvidhya.com/blog/2023/05/demystifying-unet-and-learning-image-segmentation/

[14] Ph. D. By Steven W. Smith and Copyright © 1997-2011 by California Technical Publishing, “The Scientist and Engineer’s Guide to Digital Signal Processing,” 1997. Accessed: Nov. 11, 2023. [Online]. Available: https://www.dspguide.com/ch24/1.htm

[15] Illarion Khlestov, “Different types of the convolution layers | Illarion’s Notes,” 2019, Accessed: Nov. 11, 2023. [Online]. Available: https://ikhlestov.github.io/pages/machine-learning/convolutions-types/

[16] Isitapol, “Apply a 2D Convolution Operation in PyTorch,” *GeeksforGeeks*, 2023, Accessed: Nov. 11, 2023. [Online]. Available: https://www.geeksforgeeks.org/apply-a-2d-convolution-operation-in-pytorch/

[17] Jason Brownlee, “A Gentle Introduction to Pooling Layers for Convolutional Neural Networks ,” in *MachineLearningMastery*, 2019. Accessed: Nov. 11, 2023. [Online]. Available: https://machinelearningmastery.com/pooling-layers-for-convolutional-neural-networks/

[18] A. Vaswani *et al.*, “Attention Is All You Need,” *Adv Neural Inf Process Syst*, vol. 2017-December, pp. 5999–6009, Jun. 2017, Accessed: Nov. 11, 2023. [Online]. Available: https://arxiv.org/abs/1706.03762v7

[19] M. Chris, “ Image Downsampling & Upsampling”, Accessed: Nov. 11, 2023. [Online]. Available:

https://www.researchgate.net/publication/359772964_Image_Downsampling_Upsampling

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation", Accessed: Nov. 20, 2023. [Online]. Available: https://github.com/tensorflow/models/tree/master/

[21] "Chromosomes with Deep Learning | by Lily Hu | Insight." Accessed: Oct. 09, 2022. [Online]. Available: https://blog.insightdatascience.com/separating-overlapping-chromosomes-with-deep-learning-based-image-segmentation-22f97afd3283

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", Accessed: Oct. 09, 2022. [Online]. Available: http://lmb.informatik.uni-freiburg.de/

[23] "U-Net - Wikipedia." Accessed: Oct. 09, 2022. [Online]. Available: https://en.wikipedia.org/wiki/U-Net

[24] M. Puttagunta and S. Ravi, "Medical image analysis based on deep learning approach," *Multimed Tools Appl*, vol. 80, no. 16, p. 24365, Jul. 2021, doi: 10.1007/S11042-021-10707-4.

[25] "A detailed explanation of the Attention U-Net | by Robin Vinod | Towards Data Science." Accessed: Nov. 07, 2023. [Online]. Available: https://medium.com/towards-data-science/a-detailed-explanation-of-the-attention-u-net-b371a5590831

[26] by Jeremy Zhang, “UNet — Line by Line Explanation. ”, Accessed: Nov. 07, 2023. [Online]. Available: https://towardsdatascience.com/unet-line-by-line-explanation-9b191c76baf5

[27] Y.-J. Cho, “Weighted Intersection over Union (wIoU): A New Evaluation Metric for Image Segmentation,” Jul. 2021, Accessed: Nov. 07, 2023. [Online]. Available: https://arxiv.org/abs/2107.09858v4

[28] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression.” pp. 658–666, 2019.

[29] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, “A Comprehensive Survey of Loss Functions in Machine Learning,” *Annals of Data Science*, vol. 9, no. 2, pp. 187–212, Apr. 2022, doi: 10.1007/S40745-020-00253-5/METRICS.

[30] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, “Asymmetric Loss Functions and Deep Densely-Connected Networks for Highly-Imbalanced Medical Image Segmentation: Application to Multiple Sclerosis Lesion Detection,” *IEEE Access*, vol. 7, pp. 1721–1735, 2019, doi: 10.1109/ACCESS.2018.2886371.

[31] “Max Pooling Explained | Papers With Code.” Accessed: Nov. 11, 2023. [Online]. Available: https://paperswithcode.com/method/max-pooling

[32] “Yellow Cab - TLC.” Accessed: Nov. 11, 2023. [Online]. Available: https://www.nyc.gov/site/tlc/businesses/yellow-cab.page

[33] Daniel Franco-Barranco, “Example of 2D Attention U-Net architecture with 3 downsampling levels... | Download Scientific Diagram”, Accessed: Nov. 12, 2023. [Online]. Available: https://www.researchgate.net/figure/Example-of-2D-Attention-U-Net-architecture-with-3-downsampling-levels-and-detailed_fig1_350750052

[34] J. Nodirov, A. B. Abdusalomov, and T. K. Whangbo, “Attention 3D U-Net with Multiple Skip Connections for Segmentation of Brain Tumor Images,” *Sensors 2022, Vol. 22, Page 6501*, vol. 22, no. 17, p. 6501, Aug. 2022, doi: 10.3390/S22176501.

[35] Dive into Deep Learning 1.0.3 documentation, *The Encoder–Decoder Architecture*. Accessed: Nov. 19, 2023. [Online]. Available: https://d2l.ai/chapter_recurrent-modern/encoder-decoder.html