Study of U-net, Attention-Unet and SEU-net for Breast Cancer Image Segmentation

Filippo Boni
filippo.boni@studenti.unipd.it

Abstract

In this work we study the performances of deep learning models in the image segmentation of ultra-sound images of breast cancer. In particular we start studying a very common architecture used for this task: the U-net. Then we analyze if adding two particular units, the attention gate and the squeeze-excitation block, enchances the performances of the U-net. In addition we study the effects of a particular preprocessing technique called Histogram Equalization. Finally we perform an analysis of the effects of the combination of the attention gate and squeeze-excitation block in the U-net.

1. Introduction

Breast Cancer is one of the most common type of cancer among the women. Among the countries its incidence and mortality rate vary a lot based on various factors [1]. Thus it is important to develop an automated system that it is able to diagnose the cancer in order to support the medical staff during their work. A way to do this is to develop a model able to detect the area in which the cancer is located in a ultra-sound image. This task in computer vision is called image segmentation. It involves partitioning images into multiple segments or objects performing a pixel-wise classification.

Convolutional neural networks are a type of neural networks that are widely used for image segmentation, expecially in the case of medical images such as ultra-sound, X-ray and magnetic field [5].

The goal of the work is to measure the performances for the image segmentation task of ultra-sound breast cancer images of one of the most famous model: U-net. Moreover we study if two variations of the originial architecture called respectively Attention U-net (Att-Unet) and SEU-net can improve the perfromances of the baseline model. Finally we study the effect of a particular image processing technique called Histogram Equalization on the models performances. This report is structured as follows. In Section **2** we describe the papers related to the models proposed. The characteristics of the dataset used and how its processing is performed are described in Section **3**. The models are detailed in Section **4** and they are evaluated in Section **5**. Concluding remarks are provided in Section **6**.

2. Related Works

The work of Long et al. [4] is one of the foundations of applying the Convolutional Neural Networks for semantic segmentation. They proposed a "$fully\ convolutional$" architecture so with no presence of fully connected layers and trained it pixel-to-pixel. The network takes as input an arbitrary sized image and produces an output of the same size. Their model achieved the state-of-the-art segmentation methods.

Another turning point in the medical image segmentation is represented by the U-net proposed by Ronneberger et al. [8]. They proposed a fully convolutional network formed by a contracting path and an expanding path, resulting in a more or less symmetric u-shaped architecture. The authors won the ISBI cell tracking challenge in 2015 by a large margin. The authors created a network based on an input image of 512×512, thus we adapted the original architecture to our problem where the input images are 128×128 sized.

Since U-net gained a lot of popularity many authors proposed variations and integrations to the original architecture. One of them is represented by the work of Oktay et al. [6] who proposed a novel $attention\ gate$ (AG) to integrate in CNN architectures like U-net. The scope is to give the model the ability to implicitly learn to suppress irrelevant regions in an input image while highlighting salient features useful for a specific task. We have included an AG gate in the U-net model in order to verify if it improves the performances for our task.

Another architectural unit introduced by Hu et al. [2] to enchance the performances of a CNN is the $squeeze-excitation$ block. It has the aim to recalibrate the channel wise features responses by explicitly modelling interdependencies between channels. By including this unit in their model the authors obtained significant performance improvements for existing state of-art deep architectures at minimal additional computational cost. Their SENet

formed the foundation of their ILSVRC 2017 classification submission which won first place and significantly reduced the top-5 error to 2.251%. With the goal of improving U-net performances Rundo et al. [9] tried to introduce the squeeze-excitation in the model, solving the task of prostate tumor segmentation of images from megnetic resonance. Inspired by their work we introduce the squeeze-excitation in our U-net.

3. Dataset

The dataset used is taken from Kaggle [1] and it is made of 780 ultrasound-image. There are three type of images: the ones reporting malignant tumor, the ones reporting benign tumor and the ones reporting normal situations. Since the mask of the last of set of images is a total black image, to build the final dataset we considered only the images having two pixel-classes so the ones regarding benign and malignant tumor. An example of an image coming from the dataset with its mask is reported in Fig 1.

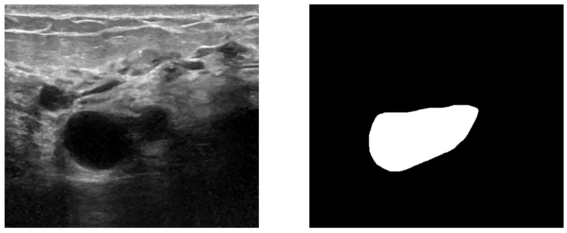

Figure 1. Original ultra-sound image(left) and corresponding mask(right).

3.1. Preprocessing

All the orginal images have different sizes, but we have decided to fix the shape of each image to 128×128. In order to avoid the transformation to change mask pixels, adding shades and thus making impossible the one-hot encoding we have used the cv.INTER_NEAREST. This makes possible to have pixels in the mask that still represent the classes. After having resized the images, we convert them from a RGB to GRAYSCALE.
Then we have also adopted a particular preprocessing technique for the input images called Histogram Equalizer [7]. The aim of this processing is to increase the contrast of the image. We have verified if this process can help the model to perform the task. It is based on considering the histogram of the pixel intensity of the image which ranges from 0 to 255. To enhance the image's contrast, it spreads out the most frequent pixel intensity values or stretches out the intensity range of the image. An example of the application of the histogram equalization on our dataset can be seen in Fig 2. The last step of the preprocessing consists in normalizing the images dividing them for 255.

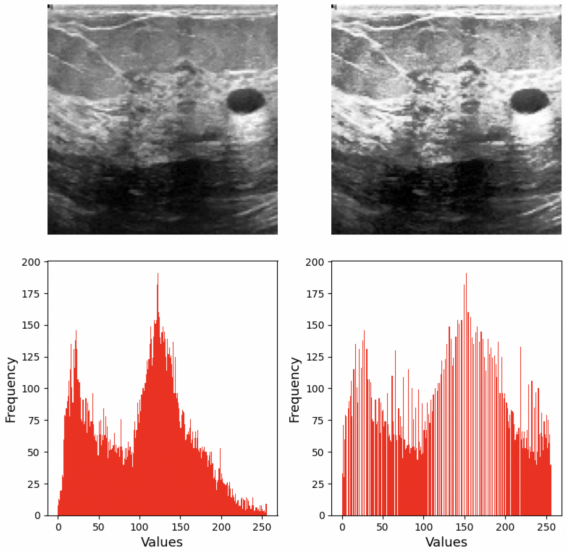

Figure 2. Original image(top-left),histogram of original image(bottom-left), transformed image(top-right),histogram of transformed image(bottom-right).

3.2. Trainig, Validation and Test

Once the dataset has been preprocessed, it has been splitted into training, validation and test set. Once splitted, the dimension of the training set have been increased by randomly choosing 250 images from the the training set itslef and flipping them. In this way the final dataset is composed of 716 images for the training, 83 for the validation set and 98 for the test set.

4. Models

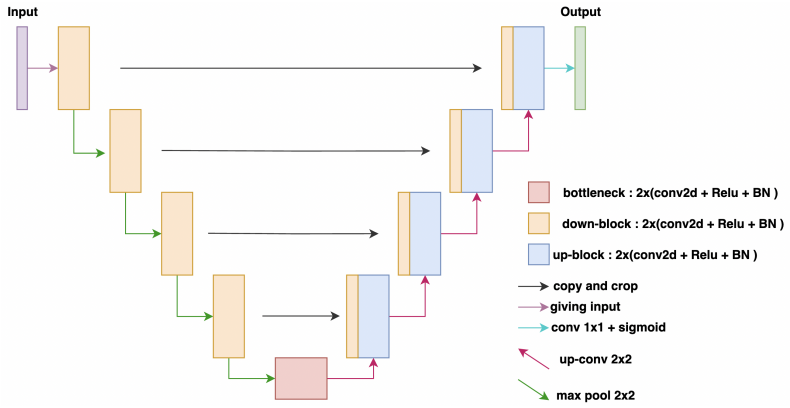

Figure 3. Arhitecture of the U-net used. Since every down block hlaves the dimension of the input the size of the bottleneck is 8×8×1024.

[1] https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset

4.1. U-net

In order to build their network Ronneberger et al. [8] started from the the FCN model [4], which consists of a contracting path built with a series of convolutional layers. These layers increase the resolution of the output. High resolutions filters are combined with an upsampling layer to produce the final ouput. In the case of the U-net the number of feature channels of the usampling part is larger than in the FCN case, creating an expansive path that is symmetric to the contracting path, creating an u-shaped architecture.
The contracting path is formed by a series of convoultional blocks which consist of two 3×3 convolution layers, each followed by a ReLu unit and 2×2 maxpooling layer with stride for the upsampling. At each contracting block (down block) the number of feature maps is doubled. The final output of the contracting path is called bottleneck and it is the input of the expansive path. The expansive blocks (up blocks) instead consist of an upsampling convolutional layer(deconvolution layer) which halves the number of input feature maps. The output of this operation is concatenated with the output of the corresponing feature map of the contracting path using the skip connection. The concatenated result is given as input to two 3×3 convolution layers, each followed by a ReLu. The final layer consist of a 1×1 convolutional, which has the goal to map the input to the desired pixel classes.
We have adapted this architecture to our input size of 128×128. The number of blocks for each path is four. The input with number of features maps equal to 1 is mapped with the first convolutional block to an output of 64 feature maps. Since each contracting block doubles the number of feature maps we end up with a bottleneck of 1024 feature maps. It is than upsampled by the expansive path. Contrarily to the authors of the U-net we used paddded convolutions. In addition each ReLu unit is followed by a Batch-Normalization layer, to speed-up the triaining. The final convolution is followed from a sigmoid layer to allow the classification. The resulting architecture is shown in Fig 3.

4.2. Attention Gate

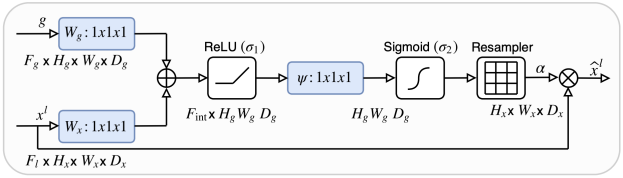

Figure 4. Attention Gate used in the Attention U-net

Attention is an important aspect of any computer vision task where we need to focus on specific regions in an image. Regarding image segmentation, attention is a massive help in highlighting the important regions of the image. The goal of the attention gate is to generate a vector of coefficients used to weight the input. This process prune features respones to preserve only the information relevant to the task that the model is performing. There exist a single scalar attention coefficient(element of the attention coefficient vector) for each pixel of the input $\mathbf{x}^l$ having number of feature maps $\mathbf{F}^l$ (where l is the index of the layer). The other input of the attention gate is the gating vector $\mathbf{g}$ with number of feature maps $\mathbf{F}^g$, which is used to determine the focus regions. This two inputs are linearly transformed. The outputs are than summed and given as input to a ReLu layer:

$$\mathbf{s} = \mathbf{W}_x^T \mathbf{x}^l + \mathbf{W}_g^T \mathbf{g} + \mathbf{b}_g \tag{1}$$

$$\hat{\mathbf{s}} = \Delta(\mathbf{s}) \tag{2}$$

where $\mathbf{W}_x^T$ and $\mathbf{W}_g^T$ are the linear trnsformations and $\mathbf{b}_g$ is the bias. Then another linear transformation is applied, obtaining an output of number of feature maps equal to 1. Once the result is computed, the attention coefficient vector is calculated by applying a sigmoid:

$$\mathbf{a} = (\psi^T(\hat{\mathbf{s}} + \mathbf{b}_\psi)) \tag{3}$$

with ψ^T being the linear transofmation and $\mathbf{b}_\psi$ as bias. The linear transformations are computed using 1x1x1 convolutions for the input tensors. Finally the input channels can be weighted:

$$\hat{\mathbf{x}}_c^l = \mathbf{x}_c^l \cdot \mathbf{a} \tag{4}$$

4.3. Attention U-net

The Attention U-net is a variation of the original U-net which makes use of the AG in the expansive path. In particular it uses as gating vector the output of the corresponding contracting block and as input the output of the deconvolutional layer. The operation is done before the concatenation. The output of the AG is finally concatenated to the output of the deconvolutional layer. The final architecture is summarized in Fig 5.

4.4. Squeeze and Excitation Block

The channel relationships modelled by convolution are inherently implicit and local. The $squeeze - excitation$ block have been introduced to model explicitly channel interdependencies. Hu et al.[2] in their work propose a squeeze and excitation block that acts as a computational unit for transformation of form $\mathbf{F}_{\text{tr}} : X \rightarrow U, \quad X \in \mathbb{R}^{W' \times H' \times C'}, \quad U \in \mathbb{R}^{W \times H \times C}$ (where C and C' are the channels). The outputs of $\mathbf{F}_{\text{tr}}$ are represented as $\mathbf{U} = [\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_C}]$.

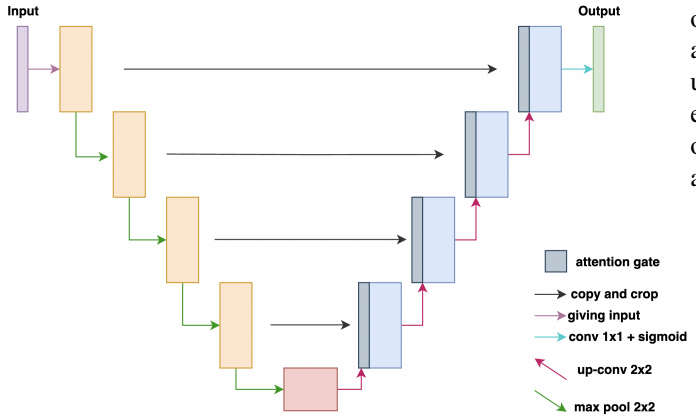

Figure 5. Architecture of the Attention U-net. Each AG receives as input the output of the corresponding down block and the output ot the upsampling operation

The first step is the *squeeze* operation. It utilizes contextual information beyond the local receptive field by employing a global average pooling operation to compute channel-specific statistics. The transformation output $\mathbf{U}$ is shrunk The transformation output, U, is shrunk through spatial dimensions, $W \times H$. The c-th element of the output of the operation $\mathbf{z}$ is calculated as follows:

$$\mathbf{z_c} = \mathbf{F}_{\text{sq}}(\mathbf{u_c}) = \frac{1}{W \times H} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{u_c}(i, j) \qquad (5)$$

The aggregated information obtained from the *squeeze* operation is followed by an *excite* operation, whose objective is to capture the channel-wise dependencies. It is obtained by applying two dense layers.

$$\mathbf{s} = \mathbf{F}_{\text{ex}}(\mathbf{z}, \mathbf{W}) = \sigma(\mathbf{g}(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W_2}(\mathbf{\Delta}(\mathbf{W_1 z}))) \quad (6)$$

where σ is the Sigmoid activation function, Δ is the ReLU activation function, $\mathbf{W_1} \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathbf{W_2} \in \mathbb{R}^{C \times \frac{C}{r}}$ are learnable parameters of $\mathbf{F}_{\text{ex}}$, and $\mathbf{r}$ is the reduction ratio. Finally, the output $\mathbf{u_c}$ is scaled as follows:

$$\tilde{\mathbf{u}}_{\mathbf{c}} = \mathbf{F}_{\text{scale}}(\mathbf{u_c}, \mathbf{s_c}) = \mathbf{s_c} \cdot \mathbf{u_c} \qquad (7)$$

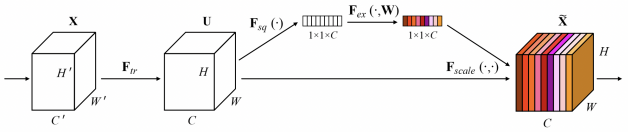

Figure 6. Squeeze and Excitation block

4.5. SEU-net

Inspired by the work of Rundo et al. [9] we have introduced the squeeze-excitation block to enchance the capabilities of the network and expecting an increased representational power from modeling the channel-wise dependencies of convolutional features. Differently from [9], who used a modified squeeze-excitation block vharacterized by residual connections, we use the simple version. The squeeze-excitation blocks are placed after each convolutional block of both the contracting and expanding path. The proposed architecture is show in Fig. 7.

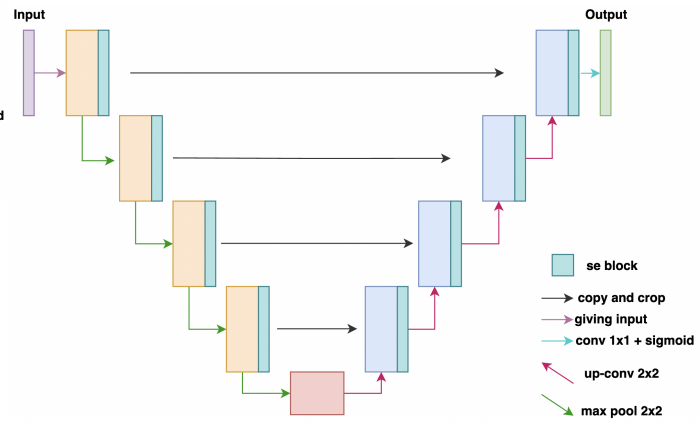

Figure 7. Architecture of the SEU-net. There is a squeeze-excitation block after each down and up block.

4.6. SAU-net

After analyzing the effects of introducing attention gates and squeeze-excitation blocks in the U-net, we prose a novel architecture that combines the two methods. It is possible to notice that the AG is used right before the set of convolutions in the corresponding expansion block while the squeeze-excutation block right after the convolutions. The architecture that we propose is the same of the SEU-net but with the difference of processing the input from the skip connections and the upsampling operation with the attention gate, like it is done in the Att-Unet. We want to study if this model, called SAU-net, can exploit the best of the two methods.

4.7. Dice Loss

A popular loss function for image segmentation tasks is based on the Dice coefficient, which is essentially a measure of overlap between two samples. This measure ranges from 0 to 1 where a Dice coefficient of 1 denotes perfect and complete overlap. The Dice coefficient is calculated using the follwoing formula (8):

$$Dice_{coeff} = \frac{2\mathbf{y}\hat{\mathbf{y}}}{\mathbf{y} + \hat{\mathbf{y}}} \qquad (8)$$

$$Dice_{loss} = 1 - Dice_{coeff} \qquad (9)$$

For the case of evaluating a Dice coefficient on predicted segmentation masks, we can approximate as the element-wise multiplication between the prediction and target mask,

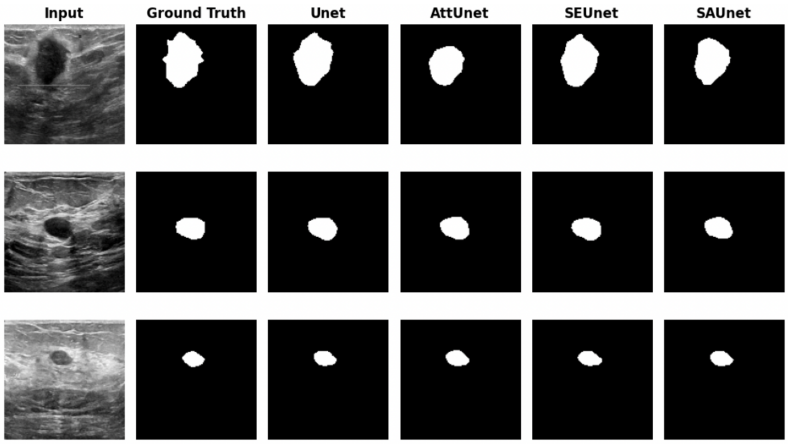

Figure 8. Qualitative comparison of the predicted masks in the case without Histogram equalization

and then sum the resulting matrix. The loss [3] function resulting from the Dice coefficient and which was used to train our models can be described by the equation (9).

5. Results

In order to provide a fair comparison, each model is trained in the same environment thus using a GPU T4 provided by Google Colab. The models are compared using three different metrics: **accuracy**, **precision**(true positives over all predicted positives), **recall**(true positives over all actual positives). In addition we have considered the **MIoU** (Mean Intersection over Union) which is the mean of the correclty classified pixels over the total for each class. It measures how much the predicted mask and the true one overlap. In addition the complexity and the training time of the models have been evaluated.

5.1. With out Histogram Equalization

Here we perform a comparison of the models in the case of data not preporcessed with Histogram Equalization. Since the MIoU represents the degree of overlap between the perdicted and the true mask, we consider it as the major indicator of the performances of the models. It is possible to notice from Table 1 that the three models have very similar performances. This might indicate that with this dataset using more complex model is not very beneficial. On the other hand both Att-Unet and SEU-net achieve better MIoU than the simple U-net, even if the improvement is little. The higher perfomances of the Att-Unet indicate that it is better to use the output of the contracting path to weight the output of the corresponding dilated convolution than just concatenate it to the output of the upsampling layer. This highlights how the AG can extract more informative features from both its input, paying attention to the information that are more relevant for the task. At the same time the enchanced performances of the SEU-net with respect to the U-net demonstrate that finding channel relationships improve the expressive power of the features extracted by the convolutional blocks. Moreover the U-net benefits more from the addittion of the attention gate than the usage of the squeeze and excitation block, which highlights how in order to extract more informative features it is more effective to decide which information should be retained from the output of the skip connection and the output of the upsampling layer reather than just enchance the output of each convolutional block with channel relationships. Thus concentrating on the relationship of the inputs of the expanding convolutional blocks is more beneficial.
It is important to say that the little imporvement of the two variations come with an increased training time and model complexity. Thus in choosing the model it is important to consider the trade-off between the classification performances and the computational complexity.

	U-net	**Att-Unet**	**SEU-net**
Miou%	66.1	66.8	66.3
Acc%	96.2	96.16	95.95
Recall%	71.37	76.15	76.6
Prec%	87.7	82.28	80.94
Tr time(s)	1957.86	2358.03	2063.96
N_p	31.042 M	31.745 M	31.478 M

Table 1. Metrics results of the models without using Histogram Equalizer. N_p is the Number of parameters and $Trtime$ is the training time

5.2. With out Histogram Equalization

Here we perform a comparison of the models in the case of data preporcessed with Histogram Equalization. The results in Table 2 show how the histogram equalization improve the performances of the U-net in all the metrics. This indicates that increasing the contrast of the input image can highlight some of its characteristics which improve the pixel

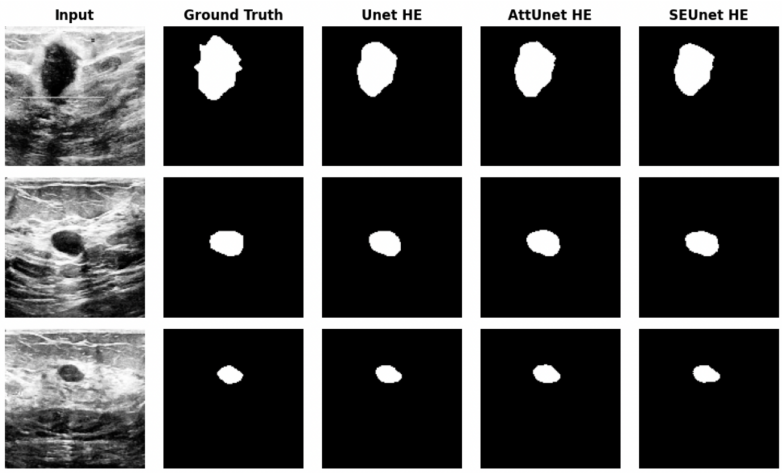

Figure 9. Qualitative comparison of the predicted masks in the case with Histogram equalization

	U-net	Att-Unet	SEU-net
Miou%	66.5	63.6	64.9
Acc%	96.22	96.11	96.15
Recall%	76.11	72.74	71.93
Prec%	83.41	84.54	85.66
Tr time(s)	1978.57	2375.34	2075.77
N_p	31.042 M	31.745 M	31.478 M

Table 2. Metrics results of the models using Histogram Equalizer.

classification of the model. At the same time we can notice a drop of the performances of the two variations of the U-net. It is possible to conclude that by preprocessing the images the features on which the AG and the squeeze-excitation block used to rely are partially lost. This means that the two variations can outperform U-net when the original image is given as input. From a comparison between the two cases we can see that the U-net combined with histogram equalization outperforms the SEU-net without histogram equalization, resulting in a less complex model but with better performances. At the same time it does not achieve the same performances of the Att-Unet without histogram equalization. Again the choice of the best model should consider the classificaation performances and computational complexity trade-off.

5.3. SAU-net

Finally we study the performances of the SAU-net. Since the AG and the squeeze excitation block improve the performances just in the case without the histogram equalization we study only this case. From the results reported in Table 3 it is possible to see that the SAU-net have the lowest MIoU among the models. Thus the combination of the AG with the squeeze-excitation block is not beneficial. It is possble to say that the effects of the two units interfere with each other. A possible future work would be to see weather removing the squeeze and excitation blocks from either the contracting path or the expanding path could remove this interference.

	SAU-net	U-net	Att-Unet	SEU-net
Miou%	64.7	66.1	66.8	66.3
Acc%	96.20	96.2	96.16	95.95
Recall%	74.81	71.37	76.15	76.6
Prec%	83.20	87.7	82.28	80.94
Tr time(s)	2479.39	1957.86	2358.03	2063.96
N_p	32.181 M	31.042 M	31.745 M	31.478 M

Table 3. Metrics results of the models without using Histogram Equalizer and adding SAU-net

6. Conclusion

In this work we have studied the performances of the U-net and its variations in the segmentation of ultra-sound images of breast cancer. We have shown how the introduction of an attention gate or of a squeeze-excitation block improve the MIoU of the U-net. Another way to enchance the perfomances of the U-net is to preprocess the images with the Histogram Equalization technnique, in order to increase the contrast of the input image. We have also shown how this preprocessing technique leads to a drop of the performances of the Att-Unet and SEU-net. Finally we have demonstrated how the combination of the attention gate and the squeeze-excitation block in the U-net do not lead to an improvement of the performances, showing how the two units interfere with each other.

References

[1] Beeravolu, Abhijith Reddy, et al. . Preprocessing of breast cancer images to create datasets for deep-CNN. *IEEE Access*

9 , 2021.

[2] Hu, Jie, Li Shen, and Gang Sun. . Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[3] Jadon, Shruti. A survey of loss functions for semantic segmentation. *IEEE conference on computational intelligence in bioinformatics and computational biology*, 2020.

[4] Jonathan Long, Evan Shelhamer, Trevor Darrell. Fully convolutional networks for semantic segmentation. *Conference on Computer Vision and Pattern Recognition*, 2015.

[5] Kayalibay, Baris, Grady Jensen, and Patrick van der Smagt. CNN-based segmentation of medical imaging data. *arXiv preprint arXiv:1701.03056*, 2017.

[6] Oktay, Ozan, et al. . Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* , 2018.

[7] Patel, Omprakash, Yogendra PS Maravi, and Sanjeev Sharma. A comparative study of histogram equalization based image enhancement techniques for brightness preservation and contrast enhancement. *arXiv preprint arXiv:1311.4033* , 2013.

[8] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention I MICCAI*, 2015.

[9] Rundo, Leonardo, et al. . USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. *Neurocomputing 365* , 2019.