Project: Classification and Segmentation of Breast Cancer Ultrasound Images

Team members: Lu Qiao, I-Shu Wang

Introduction

Breast cancer, affecting 2.3 million women globally in 2020, remains a significant health concern. Despite advancements in early detection therapies leading to an impressive 99% survival rate, challenges persist in the automatic segmentation of tumors using breast ultrasound. Tumors exhibit variations in size and shape, presenting a hurdle for accurate identification. Additionally, the presence of irregular and ambiguous boundaries in these tumors further complicates the segmentation process. The frequently low signal-to-noise ratios within breast ultrasound images add another layer of difficulty, making it challenging to distinguish tumors from background noise. Our project tackles these challenges head-on by leveraging an encoder-decoder model, with the goal of enhancing the precision and efficiency of breast cancer detection through sophisticated computational approaches.

Dataset

The dataset, sourced from a Kaggle challenge, consists of 780 images classified into three distinct classes: normal, benign, and malignant. More precisely, there are 437 normal images, 210 benign images, and 133 malignant images. With an average size of 500x500 pixels, some images may contain two tumors, each accompanied by a corresponding mask, which we consolidated into a single representation using *Numpy* package. We then partitioned the data into training (70%), validation (15%), and test (15%) sets, maintaining a balanced distribution of the three classes in each split. After importing the original image in the grayscale form, we then addressed variations in image sizes, by resizing the image and mask uniformly to 224x224 pixels. Different interpolation methods were used for images and masks during resizing. Bilinear interpolation is used for images, which smooths images by averaging pixel values, ideal for maintaining gradual transitions and detail when resizing photographic or continuous-tone images. Nearest neighbor interpolation was used for masks, which preserves distinct categorical values in masks by simply copying the nearest pixel value, avoiding the creation of new, unintended categories during resizing. Therefore, the final input format to the model is (batch size, 1, 224, 224).

Additionally, various combinations of augmentations were applied to diversify the dataset:

1. CLAHE (Contrast Limited Adaptive Histogram Equalization): CLAHE enhances image contrast by dividing the image into small blocks and applying histogram equalization to each, limiting amplification to reduce noise.
2. Median Filter: A median filter is a non-linear digital filtering technique, often used in image processing, that replaces each pixel's value with the median value of the intensities in the neighborhood of that pixel.
3. Gaussian Filter: A Gaussian filter is a linear smoothing filter that reduces noise and detail in images using a Gaussian function, which averages the pixels based on their spatial closeness and intensity similarity, resulting in a smoothing or blurring effect.

Approach and Methods

1. Pre-train model

In this project, we used pre-trained models provided by *segmentation_models_pytorch*, which is an open-source library on GitHub. This library focuses on providing pre-trained state-of-the-art segmentation models for PyTorch, simplifying the implementation of image segmentation tasks in computer vision.

First of all, we adopted the encoder-decoder architecture, which is widely used for image segmentation because it effectively captures both high-level semantic information and low-level details. Below are the specific selections of architectures. We used pre-trained weights from Imagenet for these models.

- Encoder
    1. VGG

       VGG's architecture, with its sequential layers and use of small filters, provides a good starting point for feature extraction in image segmentation tasks. Its simplicity and depth make it possible to learn hierarchical features, which is effective for identifying regions of interest in breast ultrasound images.
    2. ResNet

       ResNet's architecture utilizes skip connections to allow gradients to flow through the network without vanishing, enabling deeper networks. This is beneficial for breast ultrasound images where subtle features may determine pathology, and deeper networks can capture more complex patterns.
    3. DPN68

       DPN offers the combined advantages of both ResNet and DenseNet, providing a balance between feature re-usage and feature exploration. This can be particularly advantageous for breast ultrasound segmentation by improving the model's ability to discriminate between tissue types and enhancing detail preservation.
- Decoder - UNet

  UNet, with its U-shaped architecture, is ideal for image segmentation. In the decoding phase, it utilizes transposed convolutions for upsampling, reconstructing high-resolution features. Skip connections concatenate encoder and decoder feature maps, preserving fine details lost during downsampling. This makes UNet well-suited for tasks like breast ultrasound segmentation, emphasizing pixel-level accuracy.

2. Evaluation Metrics

- Intersection over Union (IoU)

  IoU, derived from the ratio of intersection to union areas of pixel sets, gauges the alignment between segmented regions and actual object boundaries. Higher IoU values signify superior segmentation accuracy, crucial for fine-tuning models and ensuring precise pixel-level delineation.
- Pixel Accuracy

  Pixel Accuracy assesses image segmentation model performance by measuring the ratio of correctly classified pixels to the total number of pixels. It calculates the percentage of accurately assigned pixels, offering an overall measure of pixel-level accuracy. However, it may not be

sensitive to class imbalances and is often used with other metrics for a comprehensive evaluation of segmentation models.

- Dice Coefficient
  The Dice Coefficient, also known as the Sørensen-Dice coefficient, quantifies the similarity between the predicted and ground truth segmentations by measuring the overlap of the two sets of segmented pixels. The Dice Coefficient yields values between 0 and 1, where 1 indicates perfect overlap and 0 indicates no overlap. It is a particularly useful metric for assessing segmentation accuracy, especially when dealing with imbalanced class distributions in images.
- F1 Score
  F1 Score, a common metric in classification, balances precision and recall, especially effective in imbalanced class scenarios. Precision is the ratio of true positives to predicted positives, while recall is the ratio of true positives to actual positives. F1 Score ranges from 0 to 1, with 1 indicating perfect precision and recall, and 0 indicating poor performance. In this project, sklearn.metrics.f1_score is employed with average=weights to handle multiple classes.

Results and Discussion

- Tendency for Overfitting

One important observation is a high tendency for pre-trained models to overfit, possibly because the ultrasound dataset has characteristics differing from those they were originally trained on. To address this, a variety of methods aimed at reducing overfitting were implemented, alongside extensive hyperparameter tuning, to optimize model performance.

Initially, the pre-trained model exhibited signs of overfitting, as evidenced by a substantial gap between training and validation loss. To combat overfitting, several regularization techniques were used. These included dropout layers and L2 weight regularization, which significantly mitigated the overfitting issue by introducing a level of randomness and penalizing complex models respectively. Data augmentation methods discussed in the previous section can also expand the dataset artificially and help the model generalize better to new, unseen data.

- Hyperparameter Tuning

Extensive hyperparameter tuning was conducted by experimenting with different learning rates, batch sizes, and different layers in the network. The optimal set of hyperparameters includes a smaller batch size, a lower initial learning rate, and replacing max-pooling layers with average pooling, which resulted in a marked improvement in balancing training and validation performance. The model achieved a more consistent accuracy across both datasets, indicating a reduction in overfitting. Adjusting the learning rate over time proved particularly effective. A scheduler was chosen to gradually reduce the learning rate as the training progressed, leading to more stable and gradual convergence. In addition, the Squeeze-and-Excitation (SE) attention mechanism was applied. It has been proven to enhance model performance by recalibrating channel-wise feature responses. It works by first 'squeezing' global spatial

information into a channel descriptor using global average pooling, and then 'exciting' specific channels by applying a learned gating mechanism, thereby enabling the network to emphasize informative features while suppressing less useful ones. Even though the model with Resnet50 as encoder had higher validation loss across epochs, it performed better according to a higher IoU score with more stable increasing trends (Figure 1, Appendix). Therefore, SE attention was used in all models for the later comparison studies.

- **Model Evaluation**

Figure 1 shows the models' performance during the training stage. In the top left, the training loss graph indicates that all models show a decrease in loss over time, with DPN and ResNet50 outperforming VGG16 and VGG19, suggesting better learning efficiency. ResNet50 shows the lowest training loss, indicating a good fit to the training data without signs of underfitting. The top right panel illustrates the validation loss. Initially, all models exhibit volatility in loss, but as epochs increase, their validation loss tends to stabilize. VGG16 shows higher loss variability, which might indicate overfitting or a model struggling to generalize. The rest models maintain a lower and more stable validation loss, while DPN has the lowest validation loss, suggesting better generalization when compared to VGG16. The Train IoU graph (bottom left) shows that DPN and ResNet50 achieve higher IoU scores, implying a more accurate model on the training set. In the Validation IoU graph (bottom right), all models' IoU scores improve over time, with DPN and VGG19 performing better, especially towards later epochs, reinforcing their superior generalization ability.

Overall, DPN consistently shows better performance across both training and validation phases, implying it is likely more effective for the tasks at hand.

The final evaluation of models was performed using different metrics explained in the previous section. A separate test dataset was used and the results are shown in Table 1.

- **Best Model Selection**

The test dataset results reveal that DPN68 outperforms the other models with the highest IoU score of 0.59 and Dice Score of 0.74, indicating better overlap and agreement between the predicted and actual tumor regions (Table 1, Appendix). VGG16, while achieving the highest F1 Score of 0.877 for classification, reflects lower IoU and Dice Scores, suggesting it may not be as precise in tumor segmentation. Upon visual inspection, the DPN68 model performed well for segmenting normal and benign tumors, but not as well for malignant tumors (Figure 2, Appendix).

- **Discussion**

This project faced limitations, such as the size and quality of our dataset. The limited number of images and occasional low quality of both the images and masks posed challenges in training a robust model. To improve the model performance, a broader range of data augmentation techniques can be

applied to enhance the diversity of training data. Moreover, improved results may be possible through conducting more comprehensive hyperparameter tuning, and investigating the potential benefits of ensemble methods to determine if a collective approach may exceed the performance of individual models.

In conclusion, this project studied the capabilities of various deep learning models for the classification and segmentation of breast ultrasound images. The insights gained from the results have the potential to substantially enhance medical imaging analysis, thereby offering promising avenues for improving the early detection and treatment of breast cancer.

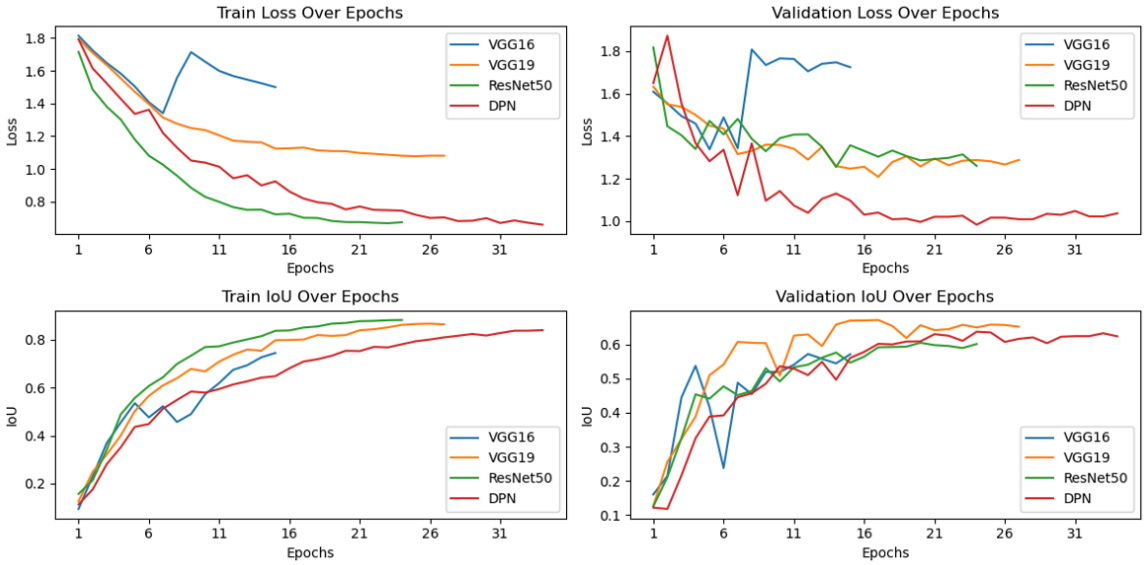

Fig. 1 Comparison of VGG16, VGG19, ResNet50, and DPN68 Architectures with Attention Mechanisms as Encoders

	IoU score	Dice Score	Pixel Score	F1 Score
VGG16 with Attention	0.41	0.58	0.95	0.877
VGG19 with Attention	0.51	0.67	0.95	0.82
ResNet50 with Attention	0.50	0.66	0.95	0.82
ResNet50 without Attention	0.43	0.60	0.95	0.80
DPN68 with Attention	0.59	0.74	0.96	0.81

Table 1 Classification and segmentation results for test dataset using various convolutional neural network architectures with attention mechanisms

Reference:

1. Baccouche, A., Garcia-Zapirain, B., Castillo Olea, C., & Elmaghraby, A. S. (2021). Connected-unets: A deep learning architecture for breast mass segmentation. Npj Breast Cancer, 7(1). https://doi.org/10.1038/s41523-021-00358-x
2. Dar, M. F., & Ganivada, A. (2023). Efficientu-net: A novel deep learning method for breast tumor segmentation and classification in ultrasound images. *Neural Processing Letters*. https://doi.org/10.1007/s11063-023-11333-x
3. Cho, S. W., Baek, N. R., & Park, K. R. (2022a). Deep learning-based multi-stage segmentation method using ultrasound images for breast cancer diagnosis. *Journal of King Saud University - Computer and Information Sciences*, *34*(10), 10273–10292. https://doi.org/10.1016/j.jksuci.2022.10.020
4. Sivanandan, R., & Jayakumari, J. (2021). A new CNN architecture for efficient classification of ultrasound breast tumor images with activation map clustering based prediction validation. *Medical & Biological Engineering & Computing*, *59*(4), 957–968. https://doi.org/10.1007/s11517-021-02357-3
5. https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset/data