STATS 403 Final Project Report: RSNA Screening Mammography Breast Cancer Detection

Chenglin Zhang
cz155@duke.edu

Yijia Xue
yx179@duke.edu

Lihui Chen
@duke.edu

***K*eywords**

Abstract

Contents

1 Problem Introduction

1.1 Background Information, Importance & Motivation

1.2 Difficulties of the problem

2 Literature Review

2.1 Medical Images Related Task Setting Taxonomy

2.2 Mammography Breast Cancer Detection

3 Project Research Contents

3.1 Dataset and Class Imbalance

3.2 Data Preprocess Techniques

3.3 Implenmented Model and Methodology

3.3.1 U-Net

The typical use of convolutional networks is on classification tasks, where the output of an image is a single class label. However, in many visual tasks, especially in biomedical image processing, the desired output should include localization, i.e., a class label is supposed to be assigned to each pixel. Moreover, thousands of training images are usually beyond reach in biomedical tasks.

introduction related work explanation and quality of approach explanation and quality of experiments / theoretical arguments and results description of what each team member did

The U-Net is widely known for medical image segmentation. new line

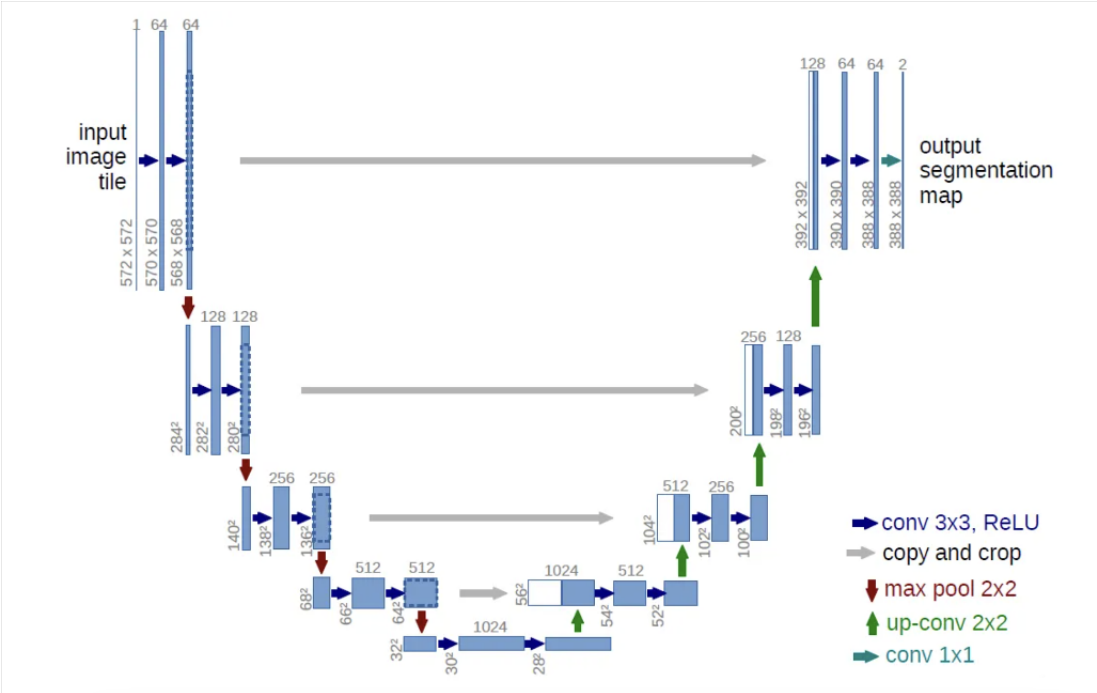

Figure 1: Basic Structure of the U-Net

Overlap-tile strategy for seamless segmentation of arbitrary large images. Prediction of the segmentation in the yellow area, requires image data within the blue area as input. Missing input data is extrapolated by mirroring.

Because the edge part is no more continuous after cutting (which made the final concatenation difficult). The overlap-tile strategy is adopted.

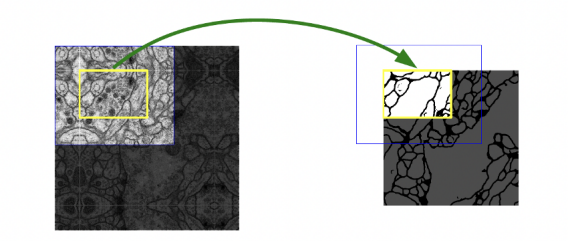

Figure 2: Basic Structure of the U-Net

Skip-tile Strategy

Image Interpolation

$$X_{src} = (X_{dst} + 0.5) \times \frac{Width_{src}}{Width_{dst}} - 0.5$$
$$Y_{src} = (Y_{dst} + 0.5) \times \frac{Height_{src}}{Height_{dst}} - 0.5$$

Bilinear Interpolation

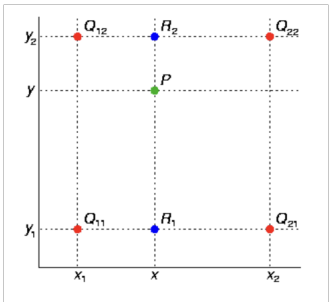

Figure 3: Basic Structure of the U-Net

Loss Function Firstly, the U-Net uses

$$p_k(x) = \exp a_k(x) / \left(\sum_K k' = 1 \exp(a_{k'}(x)) \right)$$

The implications of the U-Net in our project

3.3.2 Image Preprocessing

- Original image arrays were converted into 2048 x 2048 x 1
- Images were then cropped to exclude blank space
- YOLO model was trained to generate breast bbox
- Compared to simple rule-based breast extraction, YOLO cropped images usually have a smaller region, which seemed to prevent our models from overfitting
- In order to shorten inference time, we used simple rule-based crop during inference
- Affine transform, V/H flip, brightness/contrast, blur, CLAHE, distortion, dropout

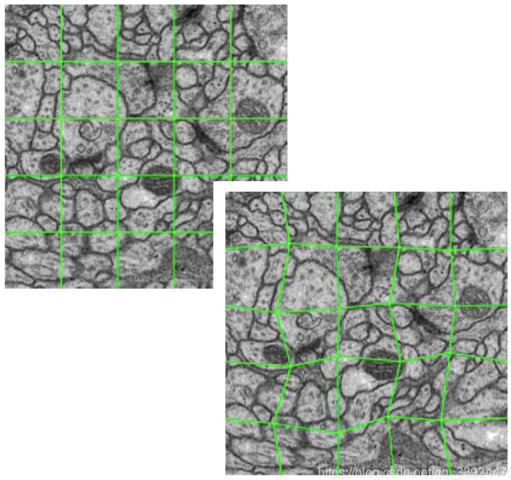

Figure 4: Basic Structure of the U-Net

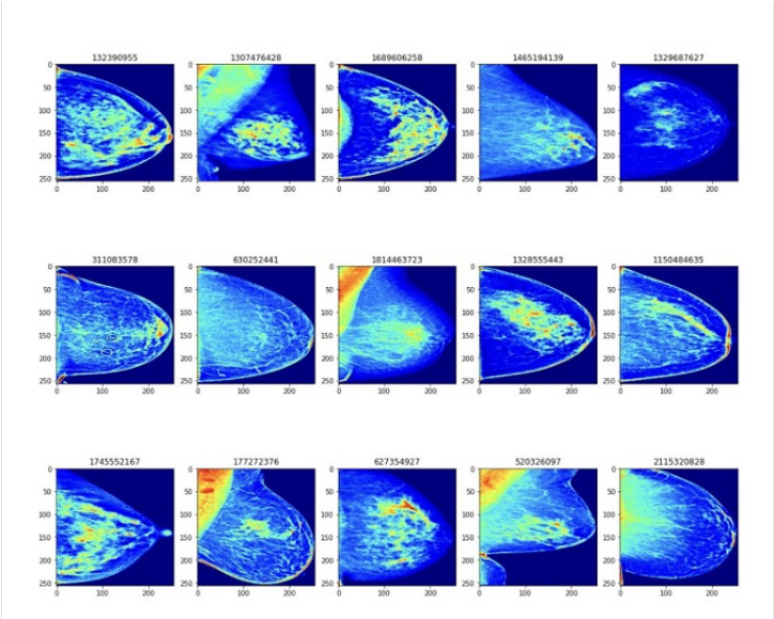

Figure 5: Basic Structure of the U-Net

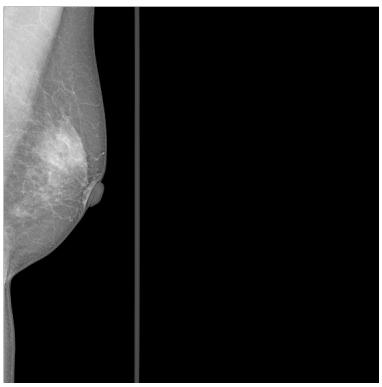

Figure 6: Basic Structure of the U-Net

3.3.3 Award-winning Architecture

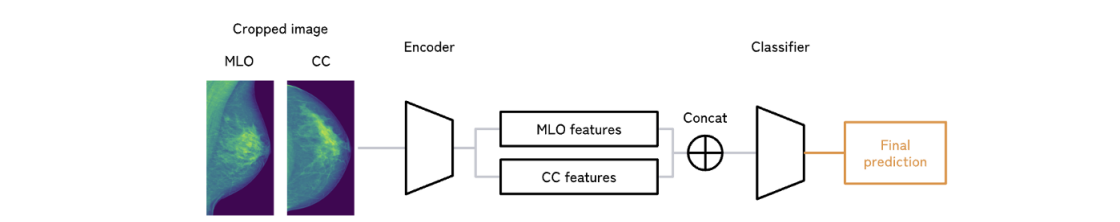

Figure 7: Basic Structure of the U-Net

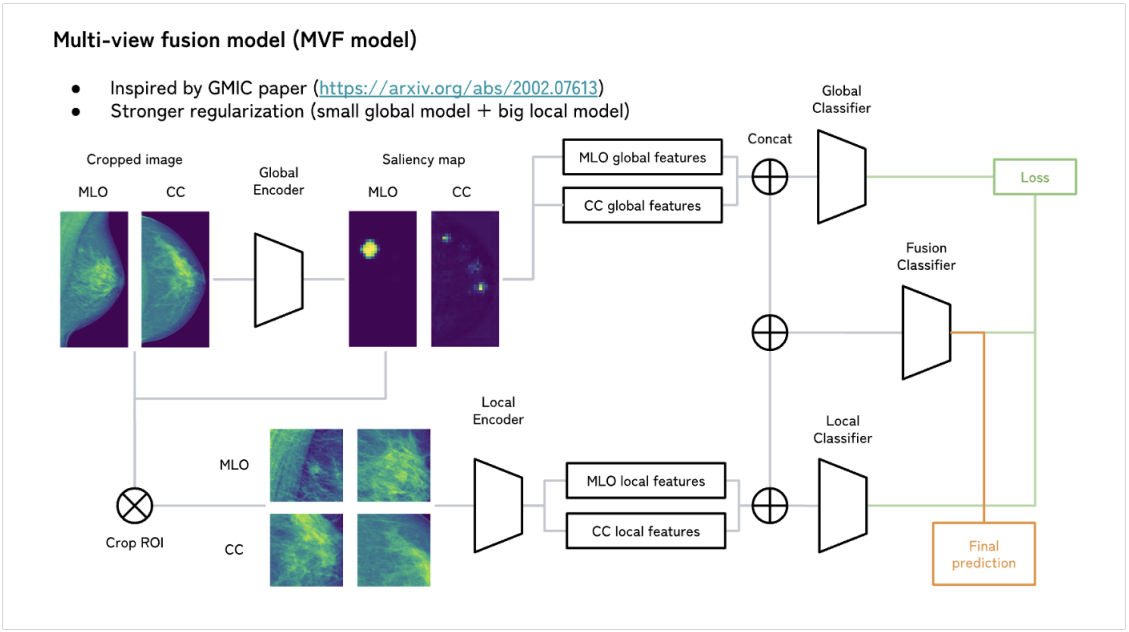

Figure 8: Basic Structure of the U-Net

3.3.4 Efficient Net

Efficient Net[**?**] mainly introduces a new scaling method that balances network depth, width, and resolution to achieve better performance. The compound scaling method proposed in this paper is based on the observation that simply increasing the width, depth, or resolution of a ConvNet can lead to diminishing returns in terms of accuracy and efficiency. Instead, the authors propose a more principled approach that balances all three dimensions using a single compound coefficient ϕ.

The method involves scaling the depth, width, and resolution of the network according to the following equations:

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi} \tag{1}$$

where d, w, and r represent the depth, width, and resolution of the network respectively. The constants α, β, and γ are determined by a small grid search and subject to the constraint that $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$.

The authors show that this compound scaling method outperforms other single-dimension scaling methods on image classification tasks such as ImageNet. They also demonstrate that this approach can be used to efficiently scale up existing ConvNets without sacrificing accuracy.

Overall, this paper provides a novel methodology for scaling ConvNets that achieves better accuracy and efficiency by balancing all dimensions of network width, depth, and resolution in a principled way using a single compound coefficient. Since this the model is commonly implemented in various competitions in Kaggle, we also adopt it as the baseline model.

3.3.5 Vision-Transformer-Based Model

Architecture of Vision Transformer

Vision transformers[**?**] are based on the transformer model originally used in natural language processing (NLP), which takes one-dimensional sequences of word tokens as input. However, since images are two-dimensional, vision transformers partition them into smaller two-dimensional patches, which are then treated as word tokens in the transformer model. The input image, with height H, width W, and C channels, is divided into $N = \frac{HW}{P^2}$ patches of size $P \times P$ to align with the input structure used in NLP [**?**]. Prior to feeding the patches to the transformer encoder, they undergo flattening, sequence embedding, learnable embedding, and patch embedding in the order shown below:

- Flattening each patch into a vector X_{np}, where $n = 1, \ldots, N$, with a length of $P^2 \times C$ was performed.
- The flattened patches were then mapped to D dimensions using a trainable linear projection E, producing a series of embedded image patches.
- Adding to the input sequence, the sequence of embedded image patches was prefixed with a learnable class embedding X_{class} that corresponded to the classification outcome Y.
- Positional information was incorporated into the input by adding one-dimensional positional embeddings E_{pos}, which were also learned during training, to the patch embeddings.

The embedding vectors that are obtained from the operations mentioned above are expressed by z_0:

$$z_o = \left[X_{\text{class}}\,; X_p^1 E; \ldots; X_p^N E\right] + E_{pos} \tag{2}$$

The embedding vectors z_o are input to the transformer-encoder network, which is a stack of L identical layers, to conduct the classification task.At the L-th layer of the encoder output, the classification head is fed with the value of X_{class}. In the pretraining stage, a multilayer perceptron (MLP) with a single hidden layer implementing the GELU nonlinearity is utilized as the classification head. In the fine-tuning stage, a single linear layer is utilized as the classification head.

The vision transformer utilizes the encoder components of the original NLP transformer architecture. The input consists of a sequence of embedded image patches of size 16 x 16, along with positional data and a learnable class embedding. A patch size of 16 x 16 was chosen to strike a balance between performance and computational cost. The learnable class embedding is fed to a classification head connected to the output of the encoder, which produces a classification output based on its state. Figure **??** illustrates the model architecture based on the vision transformer. In this work, the original vision transformer model pre-trained on the ImageNet dataset was modified by replacing its last layer with a flattening layer followed by batch normalization and an output dense layer.

3.3.6 Transfer Learning

Transfer learning was used to train vision transformer models on the mammography dataset using pre-trained examples from the vast ImageNet natural image dataset. The goal was to categorize breast mammograms into two classes—those from benign and malignant tissues—using the vision transformer's expertise from the substantial natural image collection. To do this, we removed the pre-trained prediction head and substituted a D $\times$ K feedforward layer, where K = 2 represents the total number of classes in the downstream direction. Our goal in this application of transfer learning was to improve our ability to learn the target function $f_t(\cdot)$ in the target domain D_t by drawing on our prior understanding of the source domain D_s and the learning task T_s. There are m training examples $\left\{\left(x^1, y^1\right), \ldots, \left(x^i, y^i\right), \ldots, \left(x^m, y^m\right)\right\}$ in the ImageNet dataset, with x^i denoting the ith input and y^i the ith label. By then minimizing the objective function in equation**??**, where $\left\langle y^{ij} \mid x^{ij}, W_0, W_1, b\right\rangle$ is the Softmax output probability function, and b is the bias, W_1 was generated using the weights of the ImageNet pre-trained vision transformer model W_0.

$$J\left(\langle W_1, b \mid W_0\rangle\right) = \frac{-1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{m} y^{ij}\log\left(P\left\langle y^{ij} \mid x^{ij}, W_0, W_1, b\right\rangle\right) \tag{3}$$

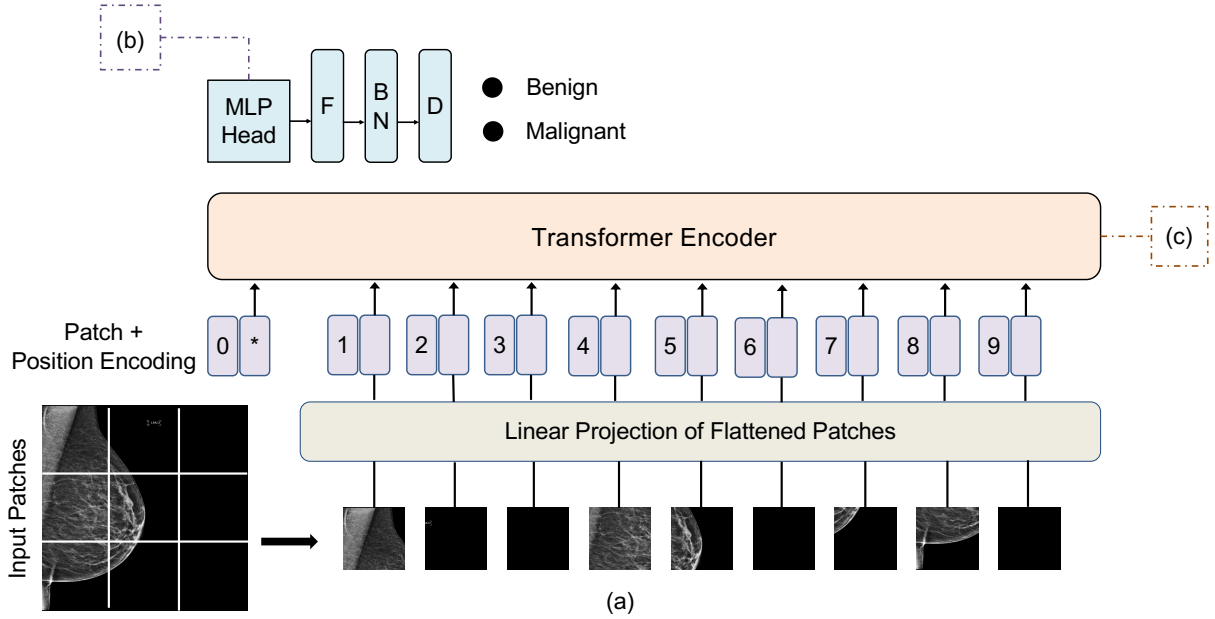

Figure 9: (a) Architecture of Vision Transformer applied to mammography breast cancer detection

4 Results

4.1 Experimental setup

In this task, what we predict is the likelihood of each breast of each patient that getting cancered.We compared the performance of the Vision Transformer pretrained on ImageNet[**?**] with the Efficient Net on the same dataset: RSNA screening mammography breast cancer detection dataset. The training hyperparameters and model implenmentation details can see Table.**??**.

Hyperparameter	Value
Learning rate	2×10^{-4}
Weight decay	1×10^{-5}
Batch size	64
Epoch number	20
Loss	Binary cross-entropy with logits loss
Optimizer	Adam
Training and validating ratio	4:1

Table 1: Training hyperparameters of the implemented model.

4.2 Evaluation Metric

In this task, we adopt the probabilistic F1 score [**?**] for performance evaluation. Probabilistic F1 score is an extension of the traditional F1 score that takes into account the uncertainty of a model's predictions. It is a metric used to evaluate classification models in natural language processing. The Probabilistic F1 score considers both precision and recall, as well as the confidence level of the model's predictions. This allows for a more thorough evaluation of the model's performance, especially in cases where there are razor-thin margins and low-resource test sets. The mathematical formulation of probabilistic F1 score is shown below.

$$pTP_{C_j} = M\left(\mathbf{x_i}, C_j\right) * I_{C_j = y_i} \tag{4}$$

$$pFP_{C_j} = M\left(\mathbf{x_i}, C_j\right) * I_{C_j \neq y_i} \tag{5}$$

$$
\begin{aligned}
p\,\text{Precision} &= \frac{pTP}{pTP + pFP} \\
p\,\text{Recall} &= \frac{pTP}{pTP + pFN}
\end{aligned}
\tag{6}
$$

Here $\mathbf{x_i}$ denotes the ith sample, C_j denotes the jth possible class, y_i denotes the ith label. pTP, pFP, pPrecision, pRecall represents the probabilistic true positive,probabilistic false positive,probabilistic precision, probabilistic recall.

$$
pF_1 = \frac{2pPrecision * pRecall}{pPrecision + pRecall} \tag{7}
$$

4.3 Experimental Results

The experiment results are shown in Table**??**.Five-fold cross-validation was used to compare the model performances. From the table, we can see that the vision-transformer-based transfer-learning approach provided the highest quantitative and statistical measures for predicting the likelihood of breast mammograms as being from benign or malignant tissues. This proves the effectiveness and quality of the vision-transformer-based transfer-learning approach for detecting breast cancer from mammograms. The possible reason is that vision transformer has the ability to capture global information from the early layers and the deep self-attention mechanism that enables features in each patch to be carefully analyzed for decision making.

Model	pF1 Score	pPrecision	pRecall	pAUROC
EfficientNetV2[**?**]	0.45	0.45	0.46	0.44
Vit-base	0.52	0.53	0.52	0.52

Table 2: Testing Result on RSNA Screening Mammography Breast Cancer Detection Dataset

5 Future Work & Conclusion

5.1 Limitations

Lack of in-depth Feature Analysis
For each patient, there is four mammography images(Left/Right with two kinds of image-forming condition), so the features may have some inner correlations. However, in the project, we omit this information and does not come up with a method to extract more useful information from the inner correlation between these four images.

Multi-View Fusion Model
In the recent year, multi-view fusion model is really a hot topic. For example, in the Mei et al.[**?**]'s work, they proposed the pyramid image fusion method to train the model. The pyramid structure allows for the extraction of features at different scales, which are then synthesized using a multimodal strategy to improve the robustness and accuracy of the method. The idea of cropping the image to different scales to catch both the global features and the local features may also be suitable for this project.

5.2 Further Study

6 Author contributions

Chenglin Zhang:
Lihui Chen:
Yijia Xue: Implementing the Vision Transformer. Report Writing: includes the section related to Efficient Net/ Vision Transformer/Transfer Learning/Results/Model Limitations

7 Acknowledgement

We thank Prof. Dongmian Zou for the generous and insightful guidance in this session's Deep Learning course, that it will empower our future learning greatly.