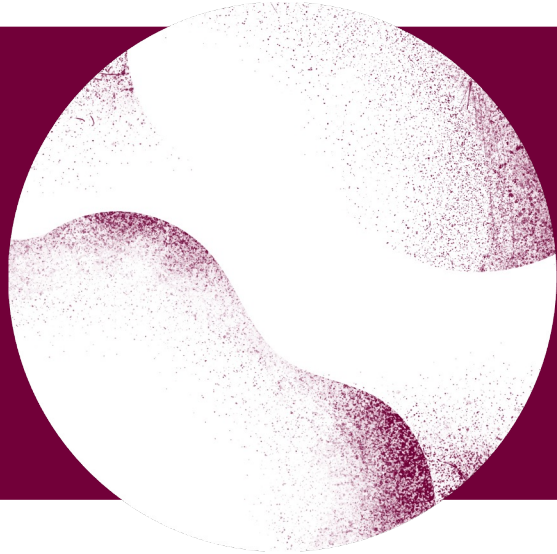
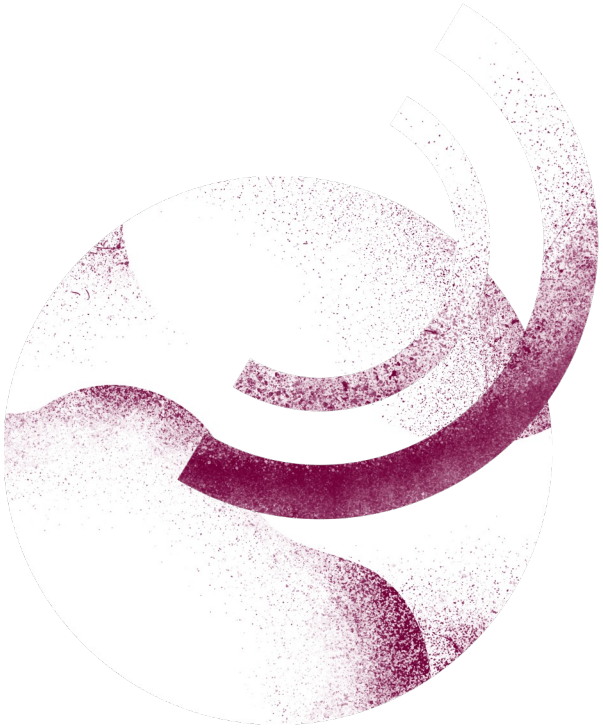
RSNA Screening Mammography Breast Cancer Detection



昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Presenters:
Chenglin Zhang
Lihui Chen
Yijia Xue



01

Problem Introduction

Background

Breast cancer is the most frequently diagnosed cancer in women worldwide with 2.26 million [95% UI, 2.24–2.79 million] new cases in 2020. And in the U.S., breast cancer alone is expected to account for 29% of all new cancers in women (Łukasiewicz et al., 2021)

Table 1. Modifiable and non-modifiable risk factors of breast cancer.

Non-Modifiable Factors	Modifiable Factors
Female sex	Hormonal replacement therapy
Older age	Diethylstilbestrol
Family history (of breast or ovarian cancer)	Physical activity
Genetic mutations	Overweight/obesity
Race/ethnicity	Alcohol intake
Pregnancy and breastfeeding	Smoking
Menstrual period and menopause	Insufficient vitamin supplementation
Density of breast tissue	Excessive exposure to artificial light
Previous history of breast cancer	Intake of processed food
Non-cancerous breast diseases	Exposure to chemicals
Previous radiation therapy	Other drugs

Retrieved from (Łukasiewicz et al., 2021)

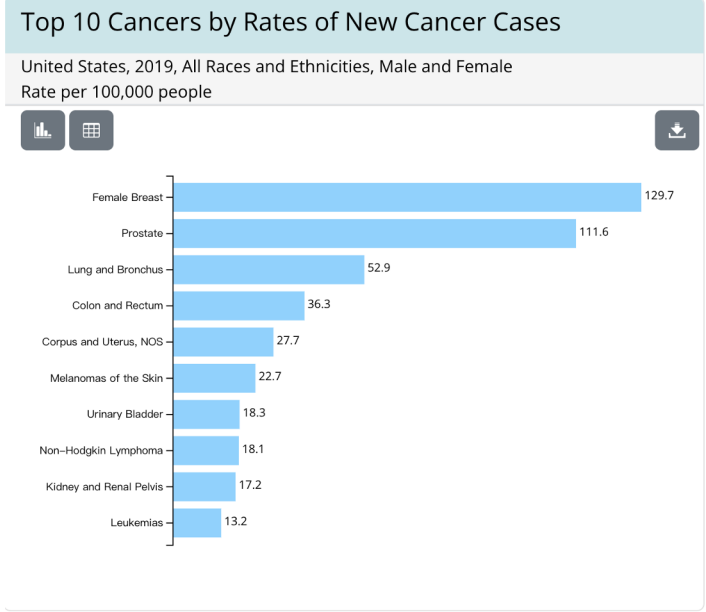


Figure 1. Top 10 Cancer by Rates of New Cancer Cases

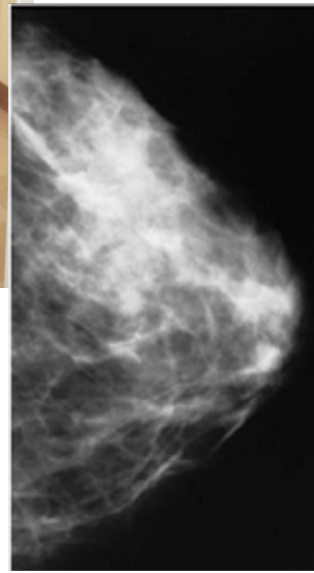
Retrieved from https://health.economictimes.indiatimes.com/news/industry/breast-cancer-lack-of-early-detection-killing-thousands-of-women-every-year/94941188

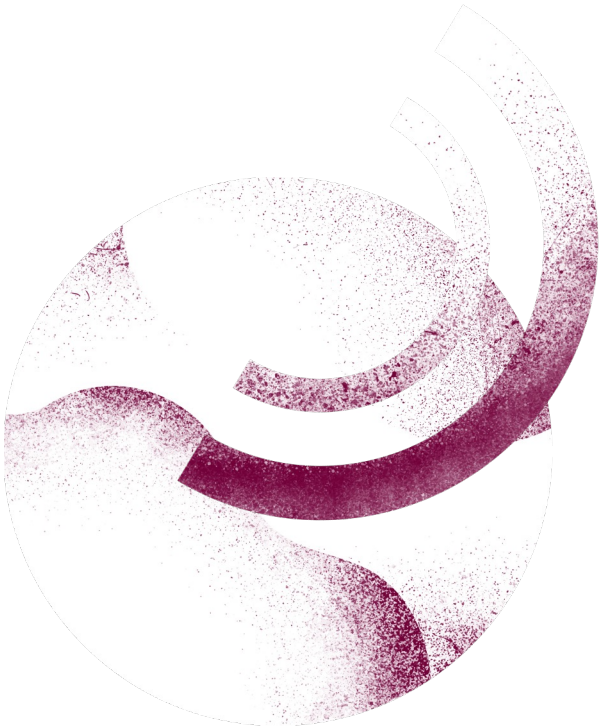
Background

- Breast cancer usually lacks evidence in its early stage, which results in late detection of the disease. Detection at advanced stages of the decease implies the treatment is more difficult and uncertain (Milosevic, 2018).

- Therefore, an early stage detection is proposed in order to prevent the disease from being diagnosed too late. Mammographic screening is a good way of early detection when resource permitted (Anderson, 2003).

- With the advancement of machine learning/deep learning algorithms, and the popularity of mammography in many countries, we aim to incorporate deep learning techniques to “read” the mammography and returns the diagosis of the given mammography.

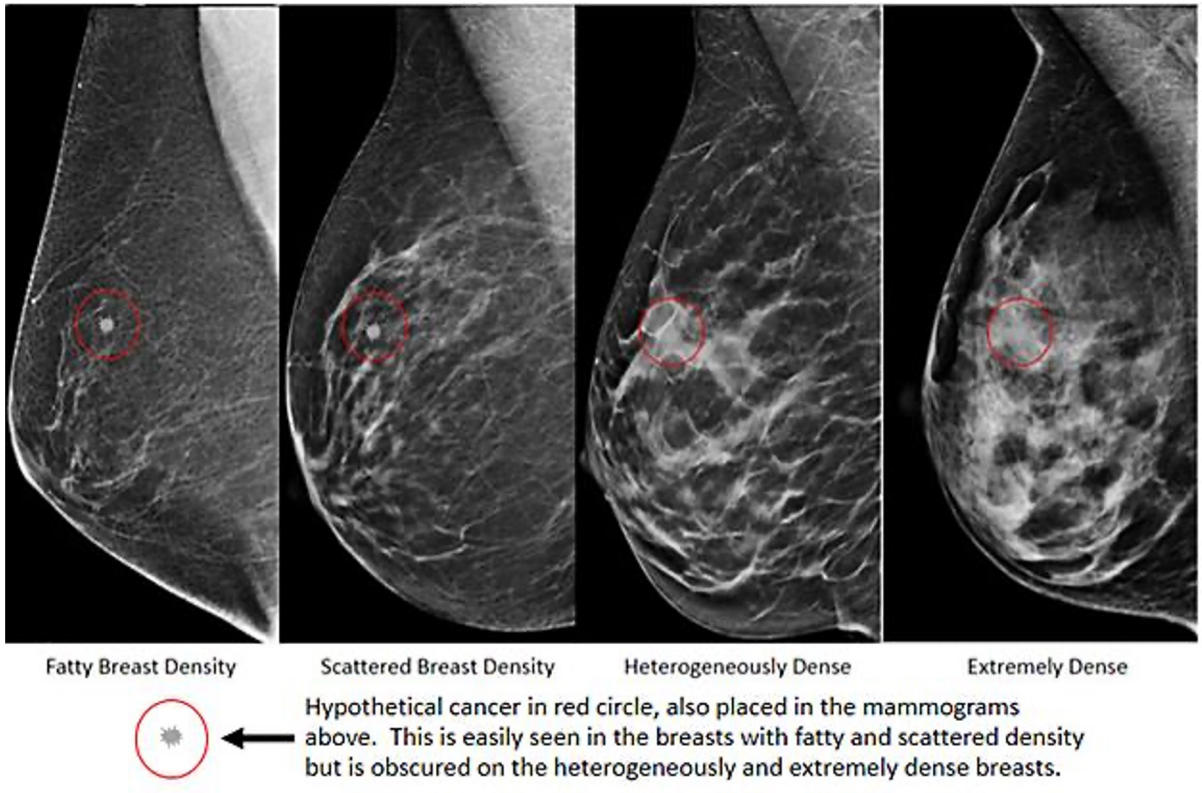




02

Previous Work and Limitations

Closer Look at Mammography



https://www.snmmi.org/Patients/Procedures/Content.aspx?ItemNumber=13977

Composition of breasts

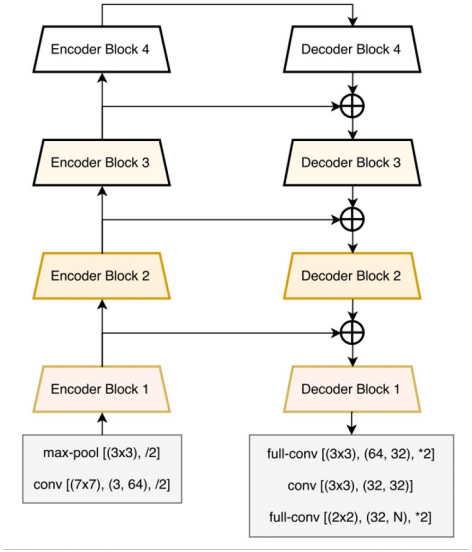
- **Fibrous and glandular tissue**
  - higher density
  - white part on the screening image
- **Fat**
  - lower density
  - black part on the screening image

Cancer cells

- Unusual white dots on the screening image (circled in red)
- This is because of the unusual growth of cells

Previous Work

Chaurasia, A., & Culurciello, E. (2017, December). Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)* (pp. 1-4). IEEE.

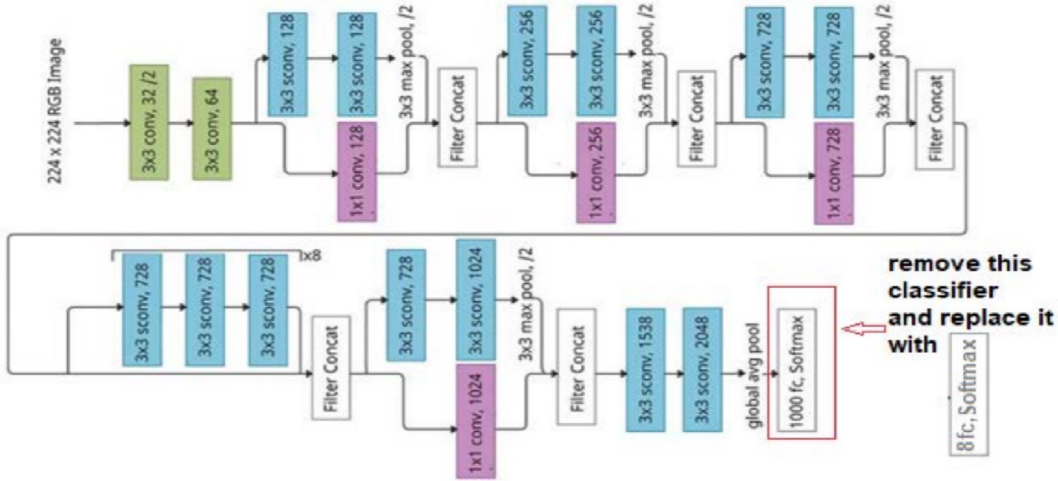


- Link each encoder with decoder
- Input of each encoder layer is bypassed to the output of its corresponding decoder

Retrieved from: https://ieeexplore-ieee-org.proxy.lib.duke.edu/abstract/document/8305148

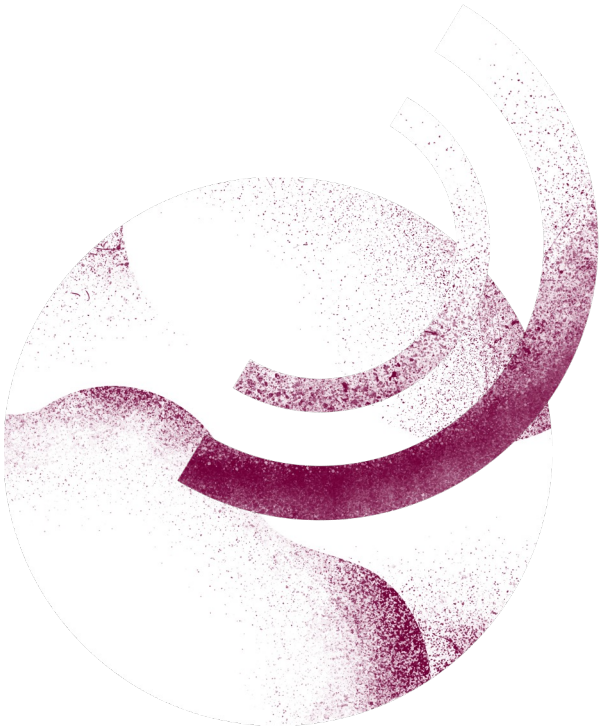
Previous Work

Basem S. Abunasser, Mohammed Rasheed J. AL-Hiealy, Ihab S. Zaqout and Samy S. Abu-Naser, “Breast Cancer Detection and Classification using Deep Learning Xception Algorithm” International Journal of Advanced Computer Science and Applications(IJACSA), 13(7), 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130729



- Filter concat
- Identity map

Retrieved from: http://dx.doi.org/10.14569/IJACSA.2022.0130729



03

Project Research Contents

Work Flow

Dataset from Kaggle competition:

- The screening images (10000+ patients).
- Patients' age.
- Density of the breast
- Imaging machine type (Brand)
- Whether cancer
- whether case difficult to identify

Exploratory Data Analysis and Data preprocessing

- Ratio of the cancer: ~1%
- Mean age: 58.5
- Age range: [26.0, 89.0]
- Preprocessing: Resize, center cropping

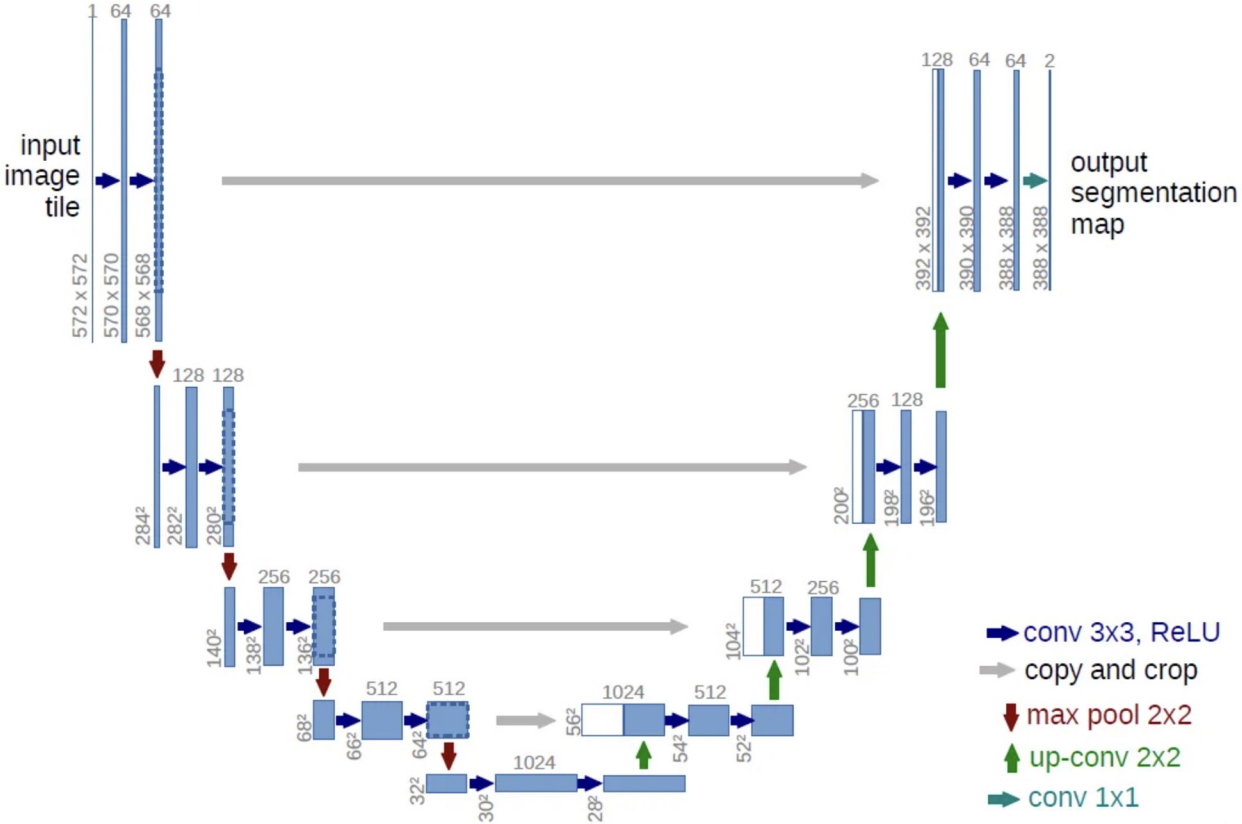
Deep Learning Networks

- UNet
- Vision Transformer (ViT)
- Efficient Net

Model Evaluation and Comparions

- F1 probabilistic score
- Comparisons among models
- Conclusion

Architecture of U-Net

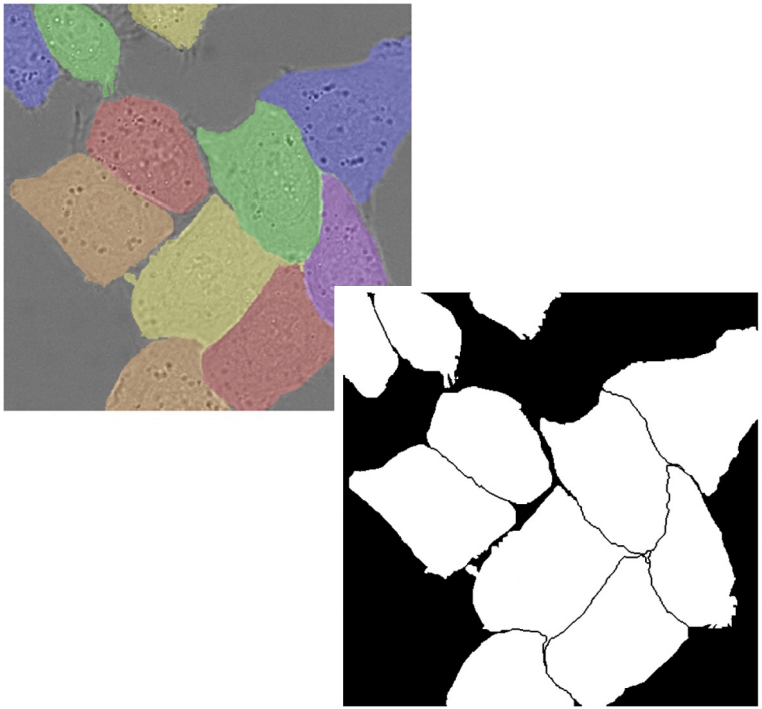


U-Net: Convolutional Networks for Biomedical Image Segmentation https://arxiv.org/abs/1505.04597

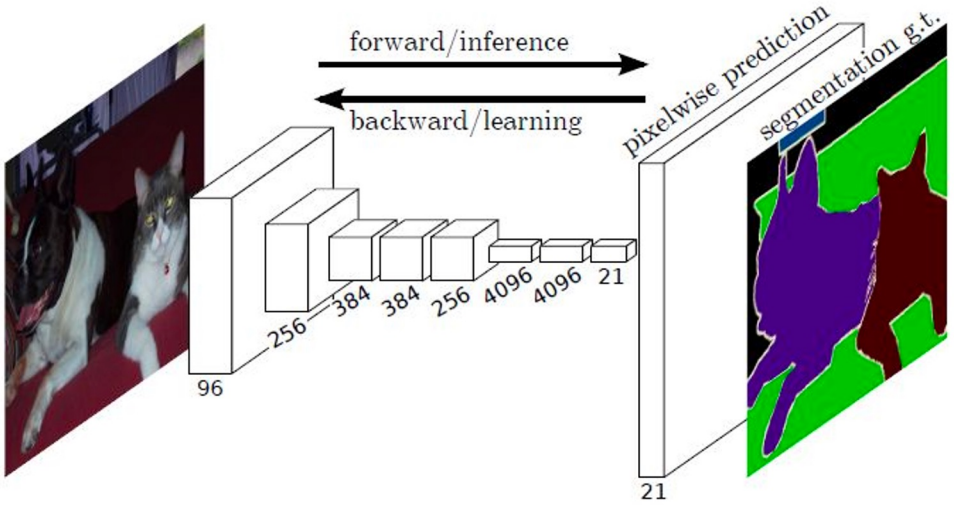
U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

1. **Downsampling (Encoding)**
2. **Upsampling (Decoding)**
3. **Skip-Connection**

Architecture of U-Net



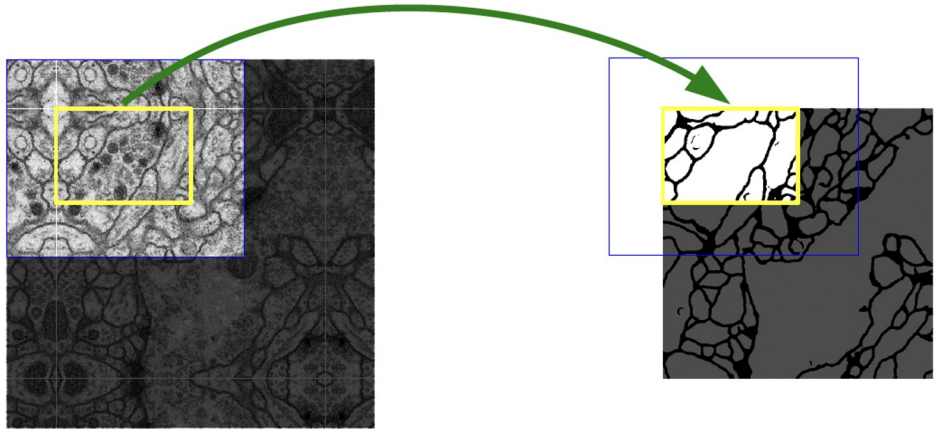
Segmentaion



U-Net: Convolutional Networks for Biomedical Image Segmentation https://arxiv.org/abs/1505.04597

Fully Convolutional Networks for Semantic Segmentation https://arxiv.org/abs/1411.4038

Architecture of 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.

https://arxiv.org/abs/1505.04597

Architecture of U-Net

Upsampling (De-convolution)

Image interpolation

Original image:  x 10

Nearest-neighbor interpolation

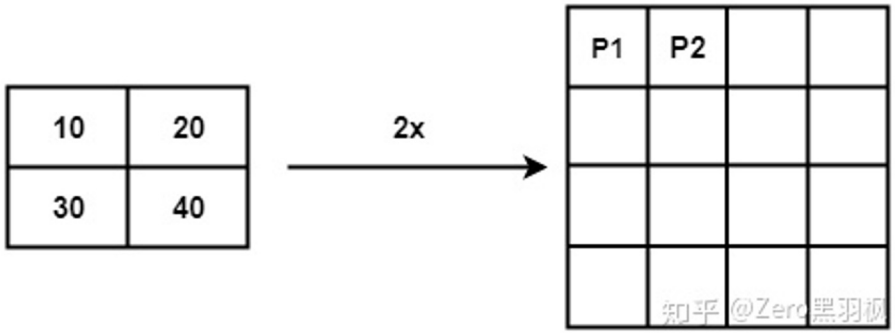
Bilinear interpolation

Bicubic interpolation

https://www.cs.toronto.edu/~guerzhoy/320/lec/upsampling.pdf

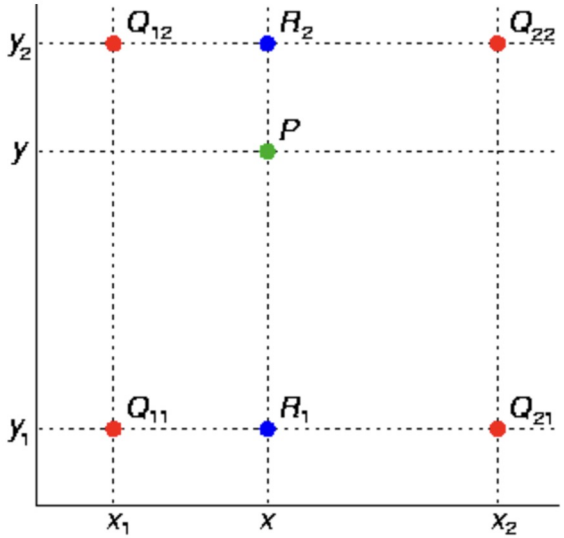
Architecture of U-Net

Upsampling (De-convolution)



$$X_{src} = (X_{dst} + 0.5) * (\frac{Width_{src}}{Width_{dst}}) - 0.5$$

$$Y_{src} = (Y_{dst} + 0.5) * (\frac{Height_{src}}{Height_{dst}}) - 0.5$$



Similar to the overlap-tile strategy

10	10	20	20
10	10	20	20
30	30	40	40
30	30	40	40

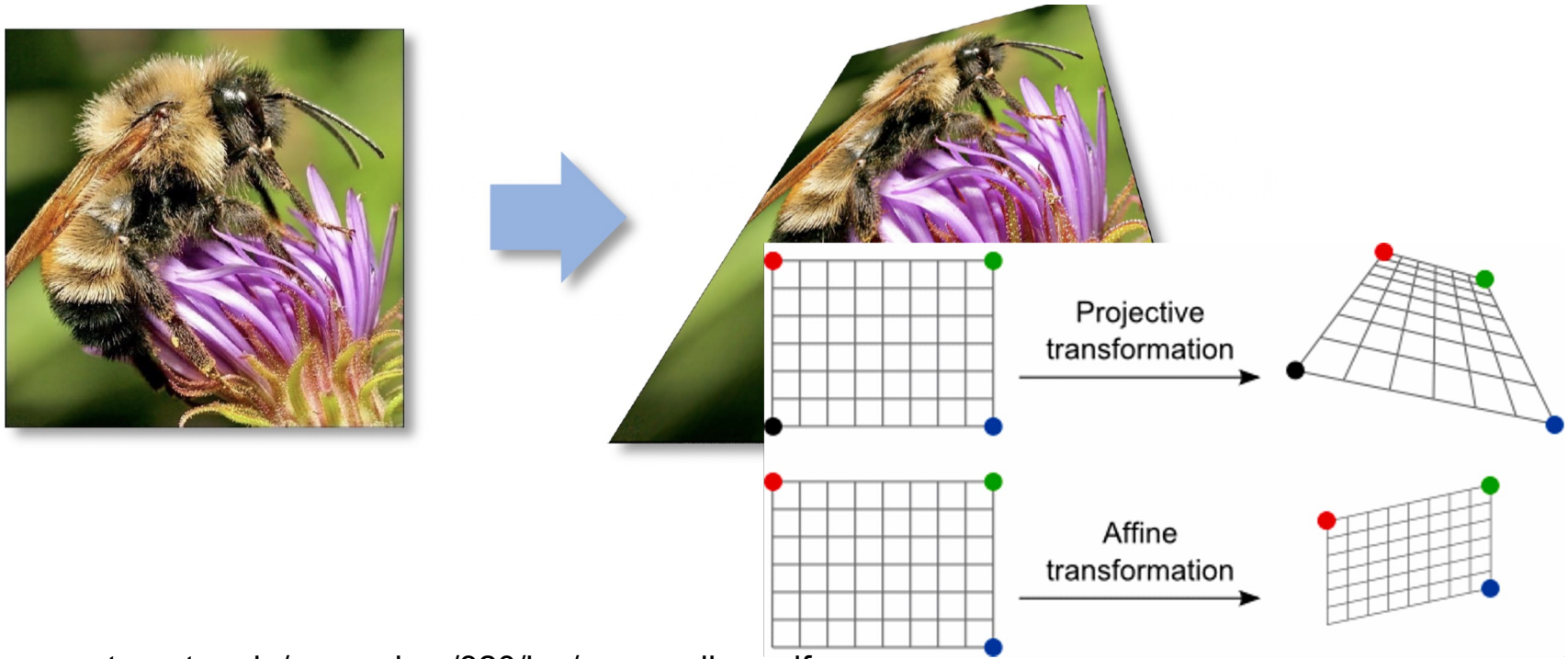
https://www.cs.toronto.edu/~guerzhoy/320/lec/upsampling.pdf

Architecture of U-Net

Image Interpolation

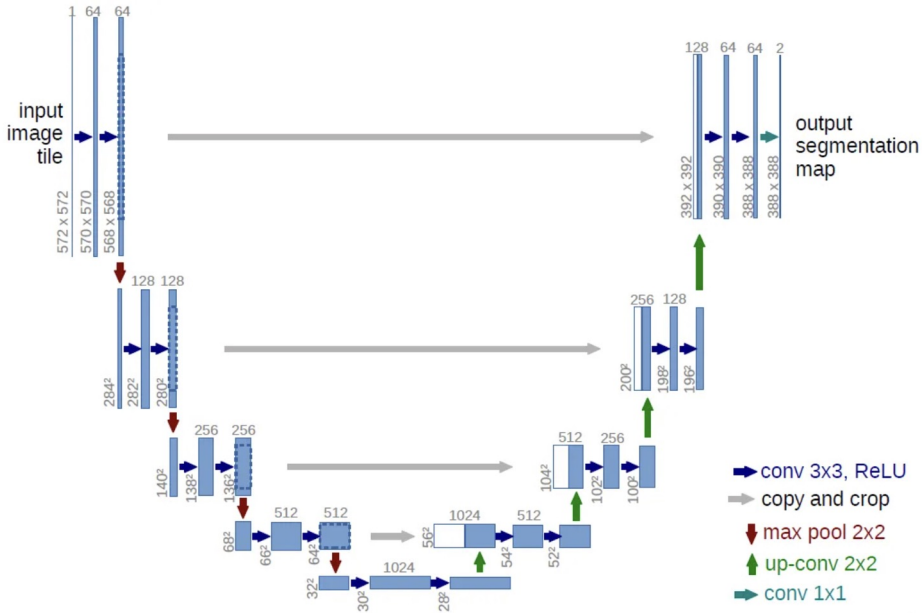
Also used for *resampling*

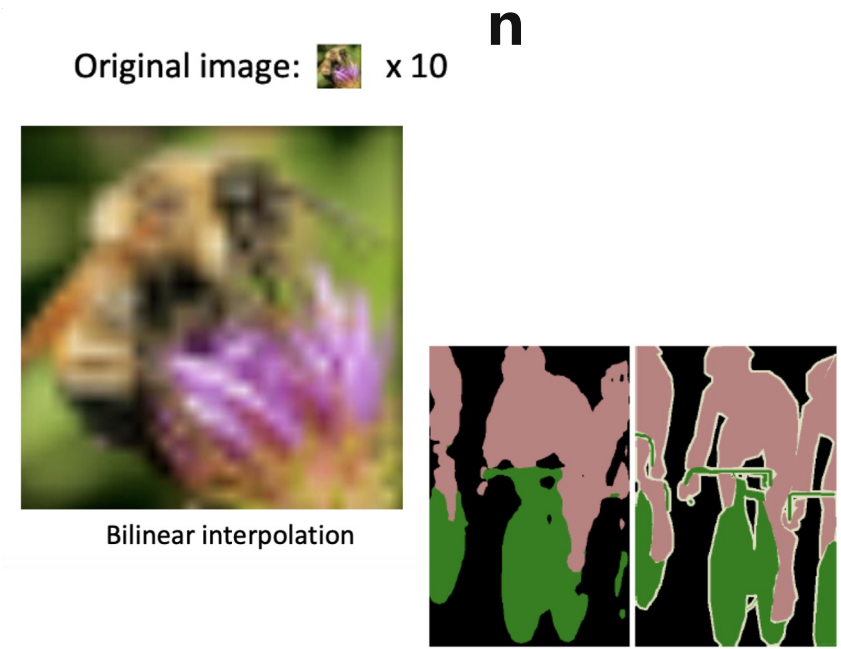
which we will use later



https://www.cs.toronto.edu/~guerzhoy/320/lec/upsampling.pdf

Architecture of U-Net





In order to make up for the information lost in the down-sampling during the encoding stage, between the encoder and the decoder of the network, the U-Net uses the Concat layer to fuse the feature maps. Thus, more high-resolution information from the downsampling phase can be used during upsampling. So, the detailed information (semantic boundaries) in the original image can be recovered better. Thus, the segmentation performance (accuracy) will be improved.

Architecture of U-Net

Loss Function

Pixel-wise Softmax Function

$$p_k(\mathbf{x}) = \exp(a_k(\mathbf{x}))/\left(\sum_{k'=1}^{K}\exp(a_{k'}(\mathbf{x}))\right)$$

where $a_k(\mathbf{x})$ represents the activation value of pixel *x* in the *k*-th channel of the feature map

k represents the *k*-th channel in the feature map

K represents the total number of classes

Architecture of U-Net

Loss Function

The **cross entropy** then penalizes at each position the deviation of $p_{\ell(\mathbf{x})}(\mathbf{x})$ from 1 using

$$E = \sum_{\mathbf{x} \in \Omega} w(\mathbf{x}) \log(p_{\ell(\mathbf{x})}(\mathbf{x}))$$

where $\ell : \Omega \to \{1, \ldots, K\}$ is the true label of each pixel and $w : \Omega \to \mathbb{R}$ is a weight map that we introduced to give some pixels more importance in the training.

Architecture of U-Net

Loss Function

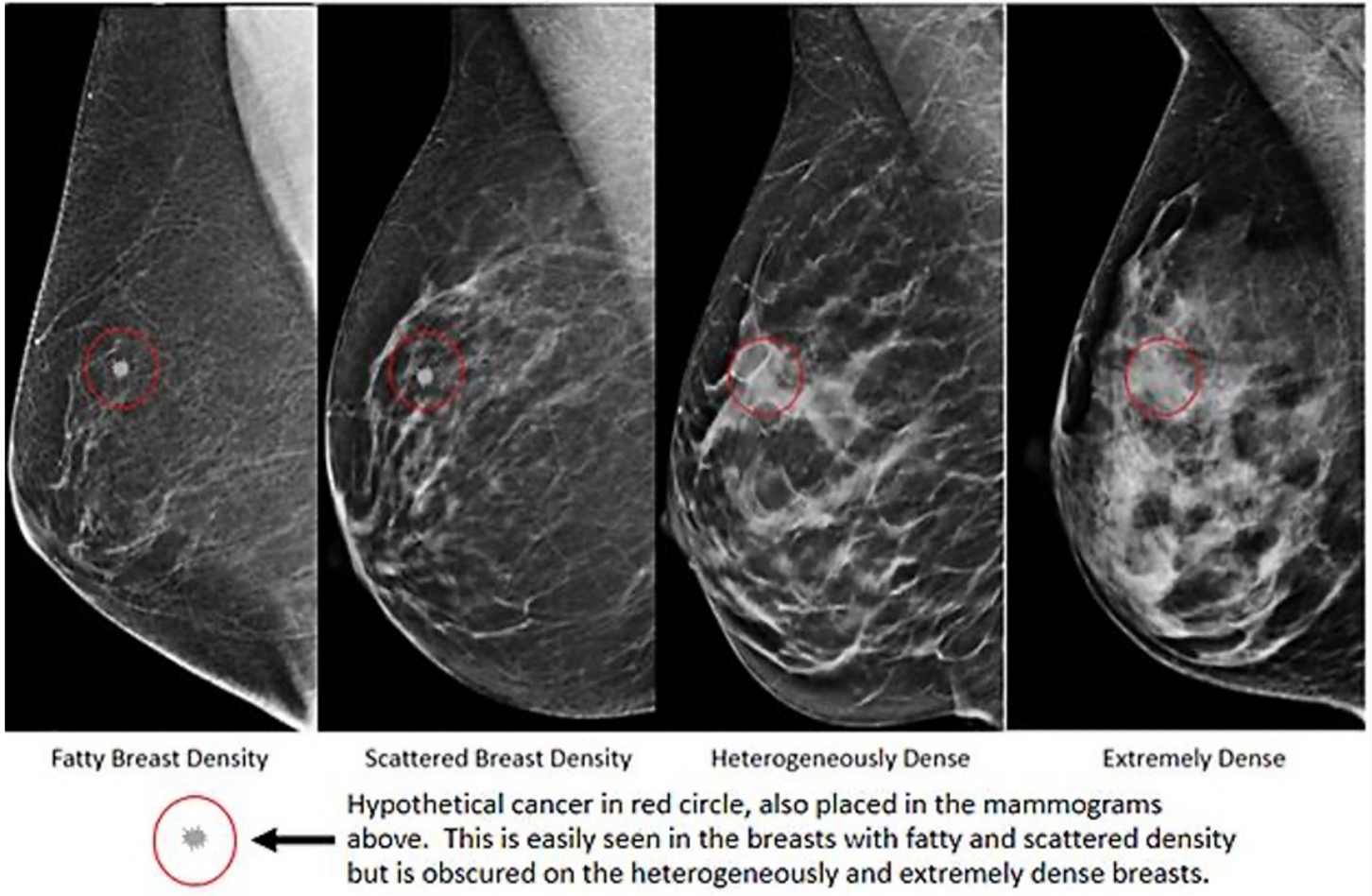
The **weight map is given by**

$$w(\mathbf{x}) = w_c(\mathbf{x}) + w_0 \cdot \exp\left(-\frac{(d_1(\mathbf{x}) + d_2(\mathbf{x}))^2}{2\sigma^2}\right)$$

where $w_c : \Omega \to \mathbb{R}$ is the weight map to balance the class frequencies, $d_1 : \Omega \to \mathbb{R}$ denotes the distance to the border of the nearest cell and $d_2 : \Omega \to \mathbb{R}$ the distance to the border of the second nearest cell. In our experiments we set $w_0 = 10$ and $\sigma \approx 5$ pixels.

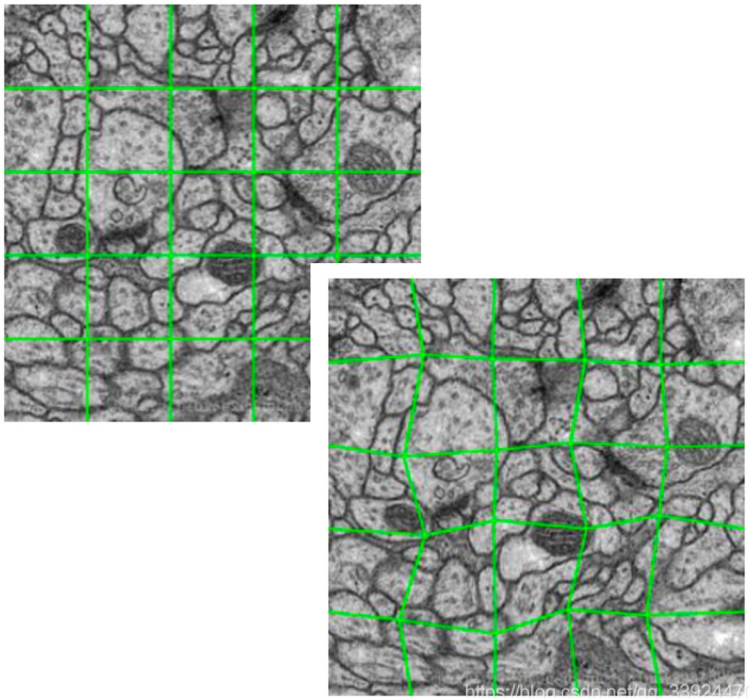
Details can be found at the paper: U-Net: Convolutional Networks for Biomedical Image Segmentation, https://arxiv.org/abs/1505.04597

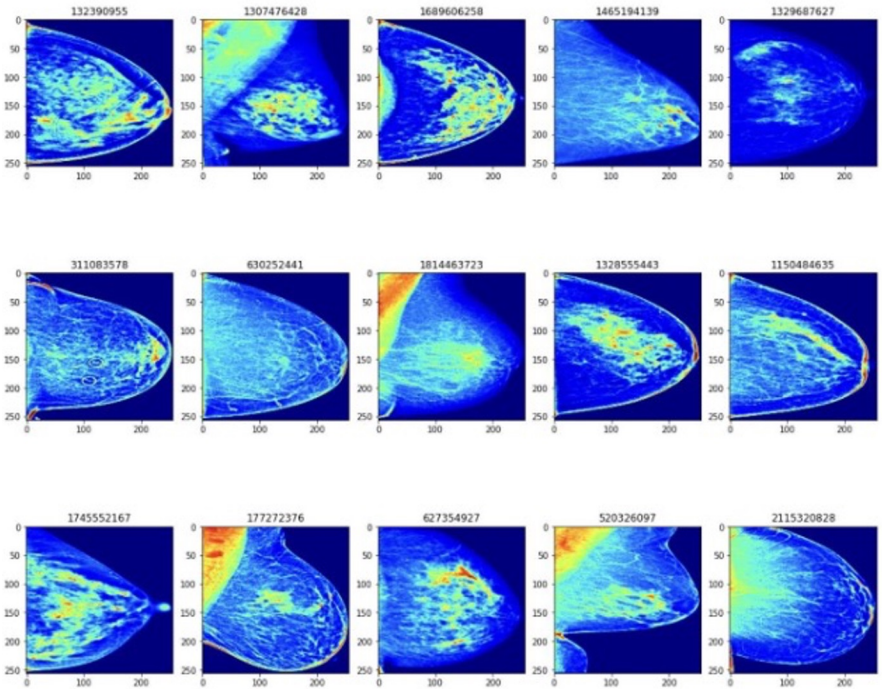
Our project We only have classification data: a 0-1 problem



Data Preparation and Pre-processing Augmentation

Distorting, Flipping, Cropping, Resizing, Brightness, Contrast, Dropout, etc.





https://www.kaggle.com/datasets/paulbacher/rsna-bcd-1024x512-preprocessed

Efficient Net: Motivation

Depth?
Width?
Resolution?

The target of Efficient Net is to maximize the model accuracy for any given resource constraints, which can be formulated as an optimization problem.

$$\max_{d,w,r} \quad Accuracy\big(\mathcal{N}(d,w,r)\big)$$

$$s.t. \quad \mathcal{N}(d,w,r) = \bigodot_{i=1\ldots s} \hat{\mathcal{F}}_i^{d\cdot\hat{L}_i}\big(X_{\langle r\cdot\hat{H}_i, r\cdot\hat{W}_i, w\cdot\hat{C}_i\rangle}\big)$$

$$\text{Memory}(\mathcal{N}) \leq \text{target\_memory}$$

$$\text{FLOPS}(\mathcal{N}) \leq \text{target\_flops}$$

N is the Conv Model
Different Network Width (w),
Depth (d)
Resolution (r) Coefficients
Xi : input tensor, with tensor shape Hi , Wi , Ci ,

Equations from: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

Efficient Net: main contribution

Compound Scaling Method:

Use a compound coefficient to uniformly scales network width, depth, and resolution

$$\begin{aligned} \text{depth: } & d = \alpha^\phi \\ \text{width: } & w = \beta^\phi \\ \text{resolution: } & r = \gamma^\phi \\ \text{s.t. } & \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\ & \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \end{aligned}$$

α, β, γ measures the ratio of depth, width and resolution respectively
β, γ is squared in the constraint because doubling the width or resolution will result in a four times of the computational effort, but doubling the depth will only result in a doubling of the computational effort.

Here, α, β, γ are constants that can be determined by a small grid search

Equations from: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

Efficient Net: main contribution

Compound Scaling Method:

Use a compound coefficient to uniformly scales network width, depth, and resolution

$$\text{depth: } d = \alpha^\phi$$
$$\text{width: } w = \beta^\phi$$
$$\text{resolution: } r = \gamma^\phi$$
$$\text{s.t. } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$$
$$\alpha \geq 1, \beta \geq 1, \gamma \geq 1$$

The magnitude of φ corresponds to the magnitude of resources that consumed.

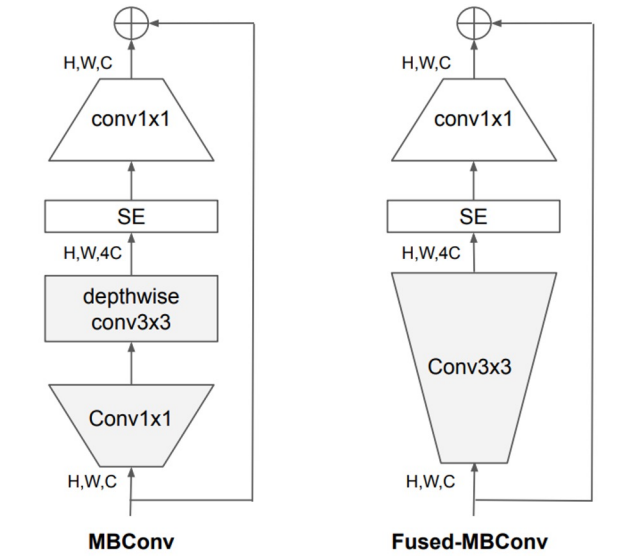
When φ is increased, it is equivalent to extending the three dimensions of the base model at the same time. The larger the model, the performance will be improved, and the resource consumption will also be increased

Equations from: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

Efficient Net: adopted baseline model

Stage	Operator	Stride	#Channels	#Layers
0	Conv3x3	2	24	1
1	Fused-MBConv1, k3x3	1	24	2
2	Fused-MBConv4, k3x3	2	48	4
3	Fused-MBConv4, k3x3	2	64	4
4	MBConv4, k3x3, SE0.25	2	128	6
5	MBConv6, k3x3, SE0.25	1	160	9
6	MBConv6, k3x3, SE0.25	2	256	15
7	Conv1x1 & Pooling & FC	-	1280	1

Compare with original Efficient Net,
Efficient Net V2 adopted new operators:
Fused-MBConv
Better utilize mobile or server accelerators



Structure of MBConv and Fused-MBConv

Model Architecture and Image derived from the paper: EfficientNetV2: Smaller Models and Faster Training

Vision Transformer: Procedure

1.Reshape the image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$

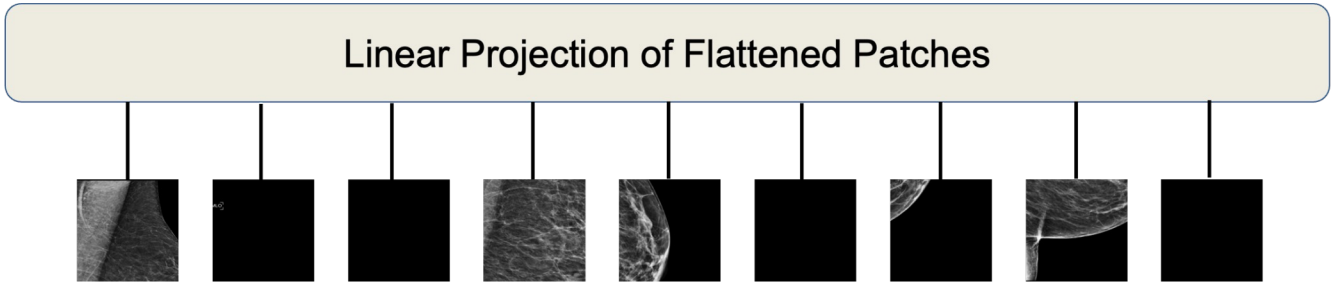
(H, W): resolution of the original image
C: the number of channels
(P, P): is the resolution of each image patch
N = HW/(P^2): the resulting number of patches

Mammography Image derived from the dataset,
image id:1442180348

Vision Transformer: Procedure

2. Flatten the patches and map to D dimensions with a trainable linear projection E

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}};\ \mathbf{x}_p^1\mathbf{E};\ \mathbf{x}_p^2\mathbf{E};\ \cdots\ ;\ \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$$

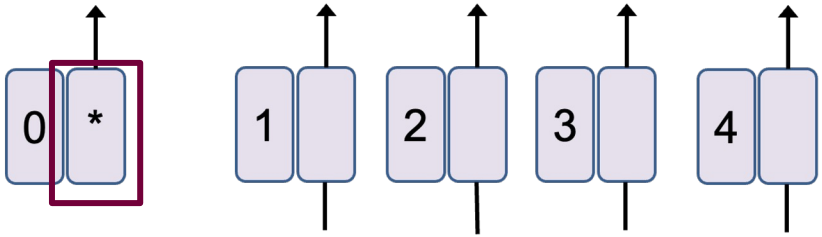


Assuming that the input image is 224x224x3 and the original image shape of a patch is 16x16x3, then the image can be divided into (224/16)^2 =196 patches. Then, each patch is linearly mapped to a one-dimensional vector, and the length of this one-dimensional vector is 16∗16∗3=768. You put 196 tokens on top of each other and the final dimension is [196, 768]

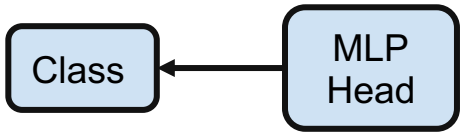
Vision Transformer: Procedure

3. The sequence of the embedded image patches was prefixed with a learnable class embedding $\mathbf{x}_{\text{class}}$ .The $\mathbf{x}_{\text{class}}$ values correspond to the classification outcome Y

* Extra learnable [class] embedding



Both during pre-training and fine-tuning, a classification head is attached to the embedding vectors. The classification head is implemented by a MLP with one hidden layer at pre-training time and by a single linear layer at fine-tuning time.



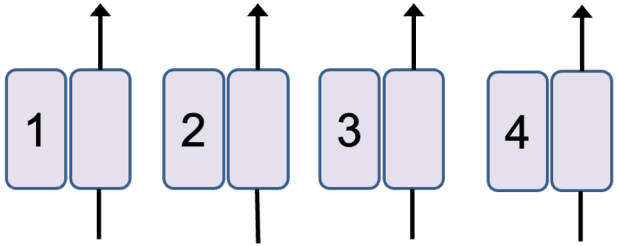
Vision Transformer: Procedure

4. Position embeddings are added to the patch embeddings to retain positional information.

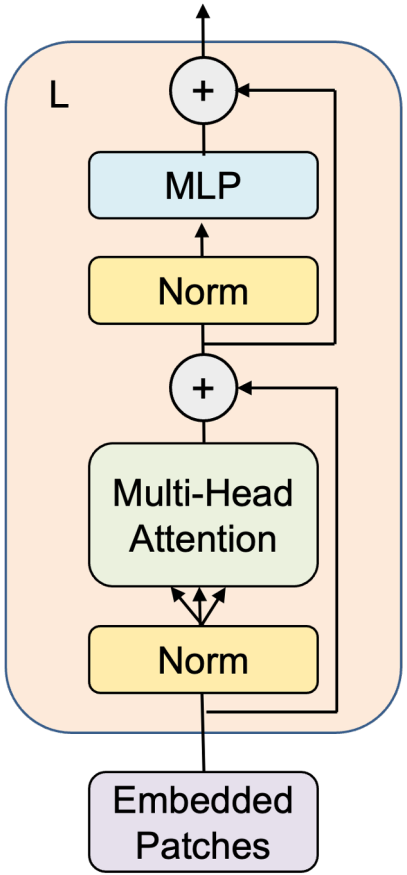
Positional information: 1-dimensional positional embedding:
Considering the inputs as a sequence of patches in the raster order

Raster Order:

Patch + Position Encoding



Vision Transformer: Procedure

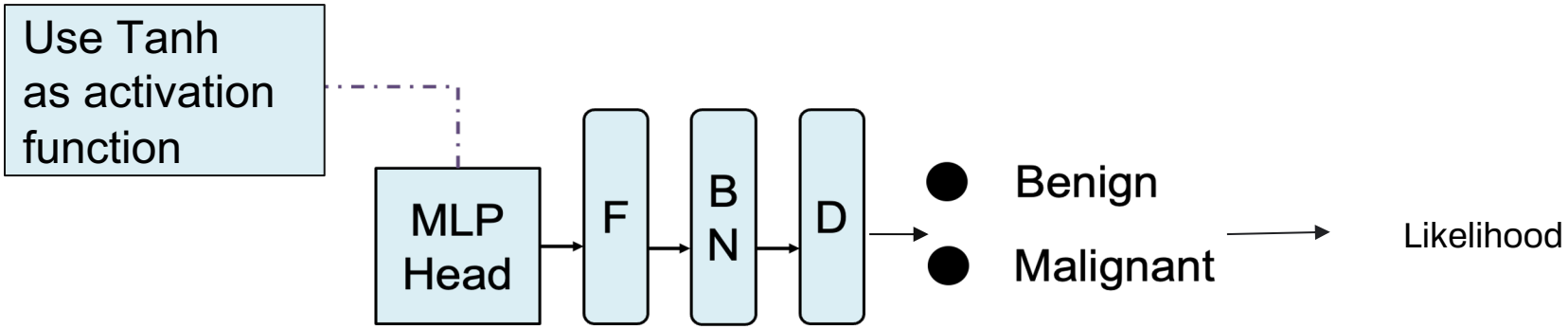


The Transformer encoder consists of alternating layers of multiheaded self-attention (Lecture 7 Page 35-36) and MLP blocks (Equation shown below). Layernorm (LN) is applied before every block, and residual connections after every block.

$$\mathbf{z}'_\ell = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \dots L$$
$$\mathbf{z}_\ell = \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \qquad \ell = 1 \dots L$$

We fed embedding patches to the transformer–encoder network, which is a stack of L identical layers, to conduct the classification(Prediction).

Vision Transformer: Procedure

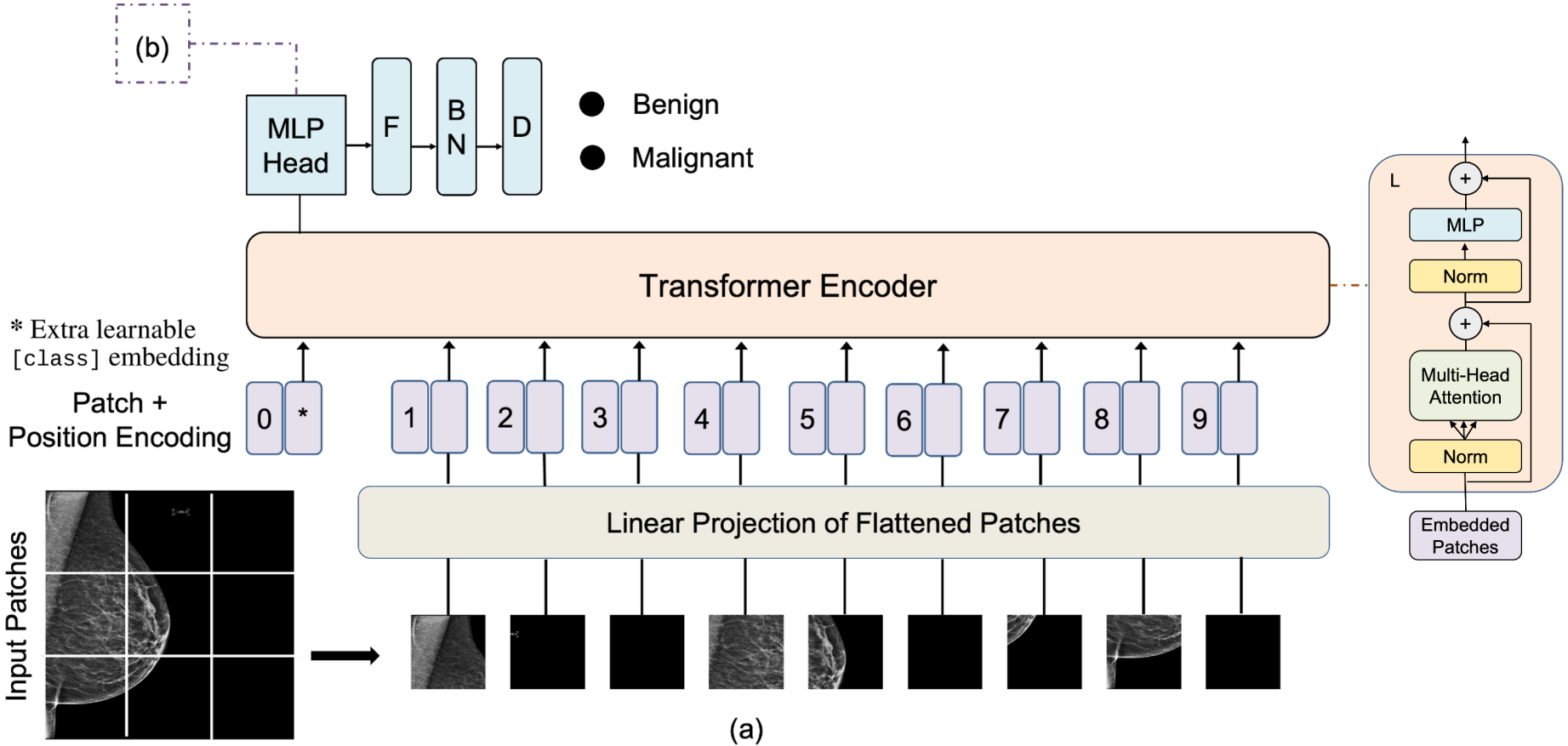


F:flattening; BN:batch normalization; D:dense

The downstream task is the classification model, so the corresponding class token needs to be extracted to obtain the Prediction result.

Compare to the original model, in our task vision transformer model was used the way that the last layer was replaced with a flattening layer followed by batch normalization and an output dense layer.

Architecture of ViT for Breast Cancer Detection



Experimental Setting

Model Backbone:

EfficientNetV2:
https://github.com/d-li14/efficientnetv2.pytorch

Tan, Mingxing, and Quoc Le. "Efficientnetv2: Smaller models and faster training." *International conference on machine learning. PMLR, 2021*.

Vit-base-patch16-224: https://huggingface.co/google/vit-base-patch16-224

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *ICLR 2021*

Experimental Setting

Training Hyperparameters

Hyperparameter	Value
Learning rate	2×10^{-4}
Weight decay	1×10^{-5}
Batch size	64
Epoch number	30
Loss	Binary cross-entropy with logits loss
Optimizer	Adam
Training and testing ratio	4:1

Binary cross entropy with logits loss is actually binary cross entropy loss, but the input doesn't have to be between 0 and 1, and the sigmoid function is automatically adds.

04

Performance Evaluation

Performance Metrics

Probabilistic F1 Score
for More Thorough Evaluation of Classification Models

A data set S: (**x_1**,y_1),...,(**x_n**,y_n)∈R^p× {C_1,...,C_m }
x_i : a vector of p features corresponding to sample i
y_i : the class corresponding to sample i
{C_1,...,C_m }: the set of possible classes

A classification model M :R^p↦ {C_1,...,C_m }. trained to predict label assignment given an input vector xi . The model assigns a confidence score (or probability if the model is probabilistically calibrated) to each possible class C_j for any given input vector **x_i** , signifying the model's confidence that Cj is the true class for the given input vector (which can also be expressed as C_j = yi).

Performance Metrics

Probabilistic F1 Score
for More Thorough Evaluation of Classification Models

confidence score: M(x_i , C_j). The class with the highest confidence score will be the model's predicted class y^ .

The commonly used definition of true positive for class C_j is any model prediction for which y^_i = y_i = C_j .

Confidence true positive: $cTP_{C_j} = M(\mathbf{x_i}, C_j) * I_{C_j = y_i}$
Confidence false positive: $cFP_{C_j} = M(\mathbf{x_i}, C_j) * I_{C_j \neq y_i}$

$$cPrecision = \frac{cTP}{cTP + cFP}$$

$$cRecall = \frac{cTP}{TP + FN}$$

$$cF_1 = \frac{2cPrecision * cRecall}{cPrecision + cRecall}$$

Yacouby, Reda, and Dustin Axman. "Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models." *Proceedings of the first workshop on evaluation and comparison of NLP systems*. 2020.

Results

Model	pF1 Score	pPrecision	pRecall	pAUROC
EfficientNetV2	0.42	0.42	0.43	0.41
Vit-base	0.52	0.53	0.52	0.52

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

Five-fold cross-validation was used to compare the model performances.

The vision-transformer-based transfer-learning model exhibited superior performance on the provided dataset

05

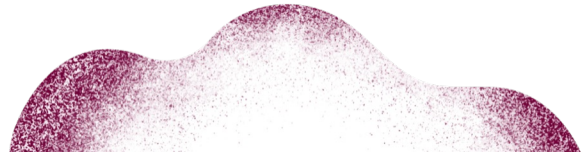
Discussion & Future Work

Discussion

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.

Possible Reason

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

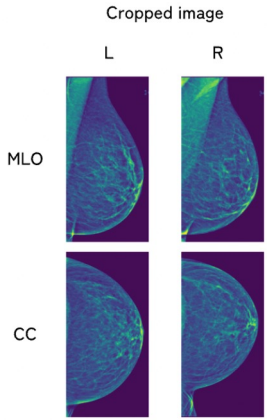


Discussion

Other Findings

Vision transformer models are more effective and efficient(computationally less expensive) when used for transfer learning on the provided dataset than training the models from scratch

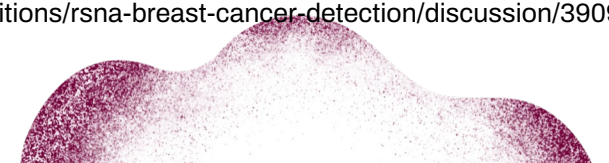
Future Work/Limitations of this project



For each patient, there is four mammography images, so the features may have some inner correlations

Left-Right Differences are important for diagnosis of breast cancer

Image Credit: https://www.kaggle.com/competitions/rsna-breast-cancer-detection/discussion/390974



Discussion

Future Work/Limitations of this project

Multi-View/ Pyramid Fusion Model

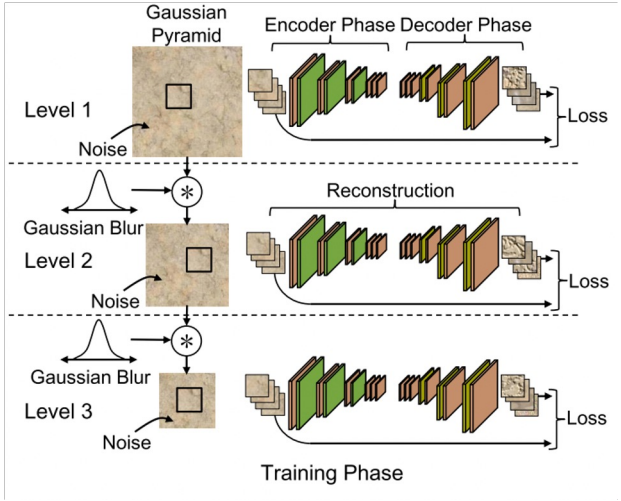


Image Credit: An Unsupervised-Learning-Based Approach for Automated Defect Inspection on Textured Surfaces

Motivated by the paper: An Unsupervised-Learning-Based Approach for Automated Defect Inspection on Textured Surfaces

Reduce/crop the image to different scales to catch both the global features and the local features.

#	Team	Members	Score	Entries	Last
1	Chiral Mistrals		0.69	291	1d
2	Racers		0.69	91	1d
3	H.B.M.F.		0.68	257	1d
4	CDI		0.66	166	1d
5	Team Hydrogen		0.66	180	1d
6	nk35jk		0.65	41	1d
7	Buvinic & Sheoran & Aerlic		0.65	302	1d
8	Q_takka		0.64	325	1d

This leaderboard is calculated with approximately 28% of the test data.

The private leaderboard is calculated with approximately 72% of the test data.

#	△	Team	Members	Score	Entries
1	▲ 21	mr.robot		0.55	17
2	▲ 52	cancerdetectman		0.53	61
3	—	H.B.M.F.		0.53	257
4	—	CDI		0.53	166
5	▼ 3	Racers		0.52	91

Data Preparation and Pre-processing

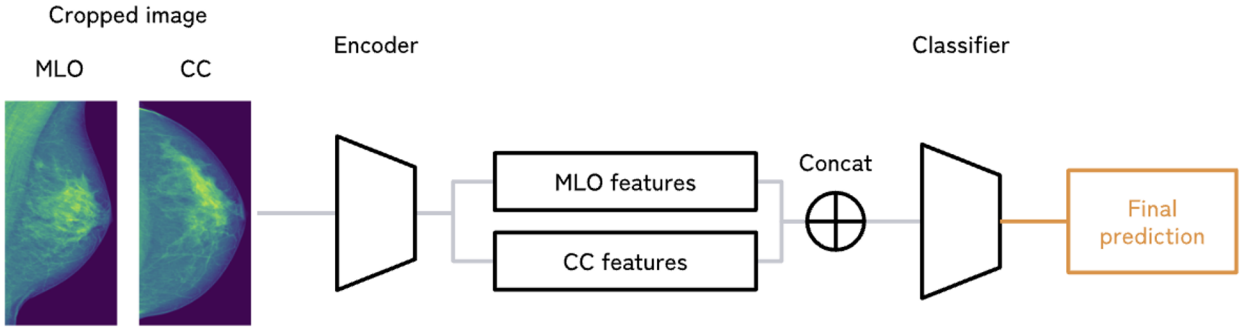
1. Original image arrays were converted into 2048 x 2048 x 1
2. Images were then cropped to exclude blank space
3. YOLO model was trained to generate breast bbox
4. Compared to simple rule-based breast extraction, YOLO cropped images usually have a smaller region, which seemed to prevent our models from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

https://www.kaggle.com/datasets/radek1/rsna-mammography-images-as-pngs

Award-winning architecture

Multi-view model (MV model)

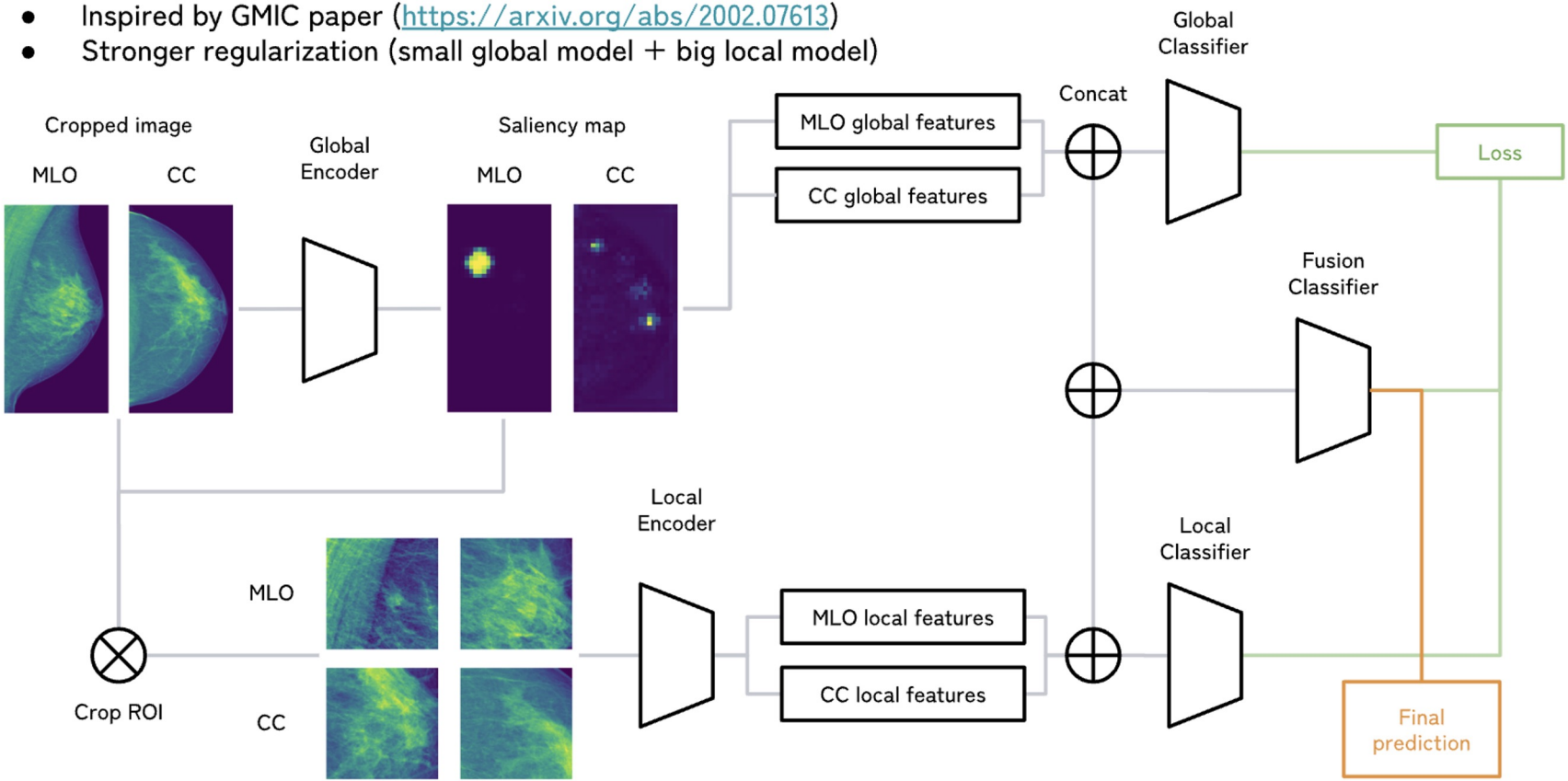
- Cancer-related findings may not be visible on both views for all cancer positive patients (i.e., label noise)
- Multi-view model is a very intuitive idea to mitigate the effect of such label noise



https://www.kaggle.com/competitions/rsna-breast-cancer-detection/discussion/390974

Multi-view fusion model (MVF model)

- Inspired by GMIC paper (https://arxiv.org/abs/2002.07613)
- Stronger regularization (small global model + big local model)

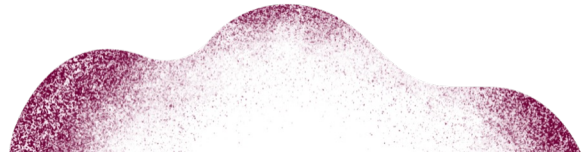


https://www.kaggle.com/competitions/rsna-breast-cancer-detection/discussion/390974

Discussion

What have we learned?

- Grabbing the ideas of some classical medical image analysis model
- Learn how to implement the efficient net, vision transformer-based method in Pytorch framework
- Study how to use Fast AI, Hugging Face library to help set up the experiment framework
- Medical image processing technology: such as how to convert dcm image into png image
- How to make efficient group collaboration



References

1. Łukasiewicz S, Czeczelewski M, Forma A, Baj J, Sitarz R, Stanisławek A. Breast Cancer-Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies-An Updated Review. Cancers (Basel). 2021 Aug 25;13(17):4287. doi: 10.3390/cancers13174287. PMID: 34503097; PMCID: PMC8428369.
2. Milosevic, Marina et al. ‘Early Diagnosis and Detection of Breast Cancer’. 1 Jan. 2018 : 729 – 759.
3. Anderson, B.O., Braun, S., Lim, S., Smith, R.A., Taplin, S. and Thomas, D.B. (2003), Early Detection of Breast Cancer in Countries with Limited Resources. The Breast Journal, 9: S51-S59. https://doi.org/10.1046/j.1524-4741.9.s2.4.x
4. Yacouby, Reda, and Dustin Axman. "Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models." Proceedings of the first workshop on evaluation and comparison of NLP systems. 2020.
5. Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." International conference on machine learning. PMLR, 2019.
6. Tan, Mingxing, and Quoc Le. "Efficientnetv2: Smaller models and faster training." International conference on machine learning. PMLR, 2021.
7. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
8. Mei, Shuang, Hua Yang, and Zhouping Yin. "An unsupervised-learning-based approach for automated defect inspection on textured surfaces." IEEE Transactions on Instrumentation and Measurement 67.6 (2018): 1266-1277.

Thanks!

Do you have any questions?