RSNA Screening Mammography Breast Cancer Detection

赛后讲解

主讲人：H老师

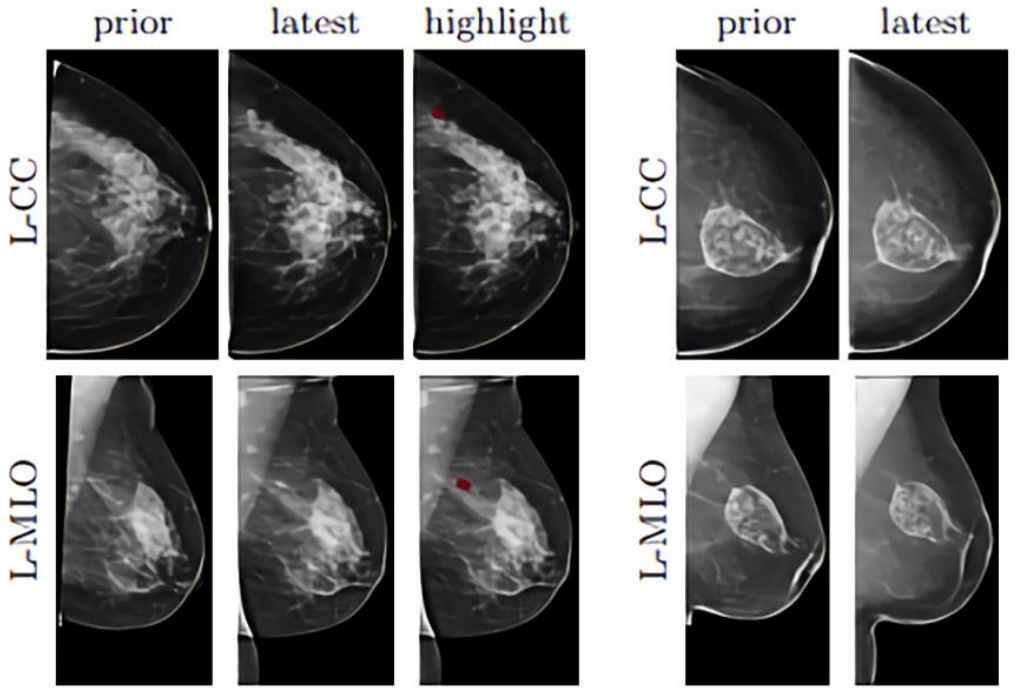
赛题背景

比赛目标

本次比赛的目标是识别乳腺癌。您将使用从定期筛查中获得的筛查性乳房 X 线照片来训练您的模型。

您在提高乳腺 X 线筛查检测自动化方面的工作可能使放射科医生更加准确和高效，从而提高患者护理的质量和安全性。它还可以帮助降低成本和不必要的医疗程序。

- 该竞赛为医学图像二分类竞赛

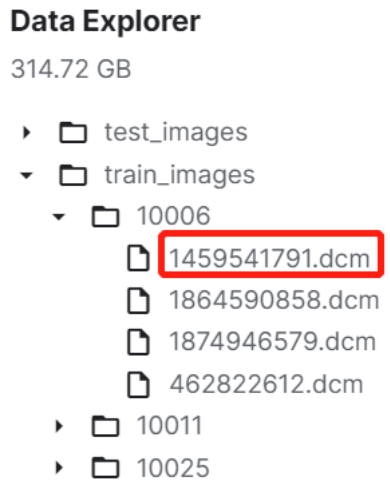


数据概览

Dicom格式

医疗数位影像传输协定（DICOM，Digital Imaging and Communications in Medicine）是一组通用的标准协定，在对于医学影像的处理、储存、打印、传输上。它包含了档案格式的定义及网络通信协定。DICOM是以TCP/IP为基础的应用协定，并以TCP/IP联系各个系统。两个能接受DICOM格式的医疗仪器间，可借由DICOM格式的档案，来接收与交换影像及病人资料。

DICOM可以整合不同厂商的医疗影像仪器、服务器、工作站、打印机和网络设备，使它们都能整合在PACS系统中。许多不同厂商的仪器、服务器、工作站都根据DICOM的标准，来制造支援DICOM机器。DICOM已经广泛地被医院所采用，并且在牙医和一般的诊所中获得小规模的运用。

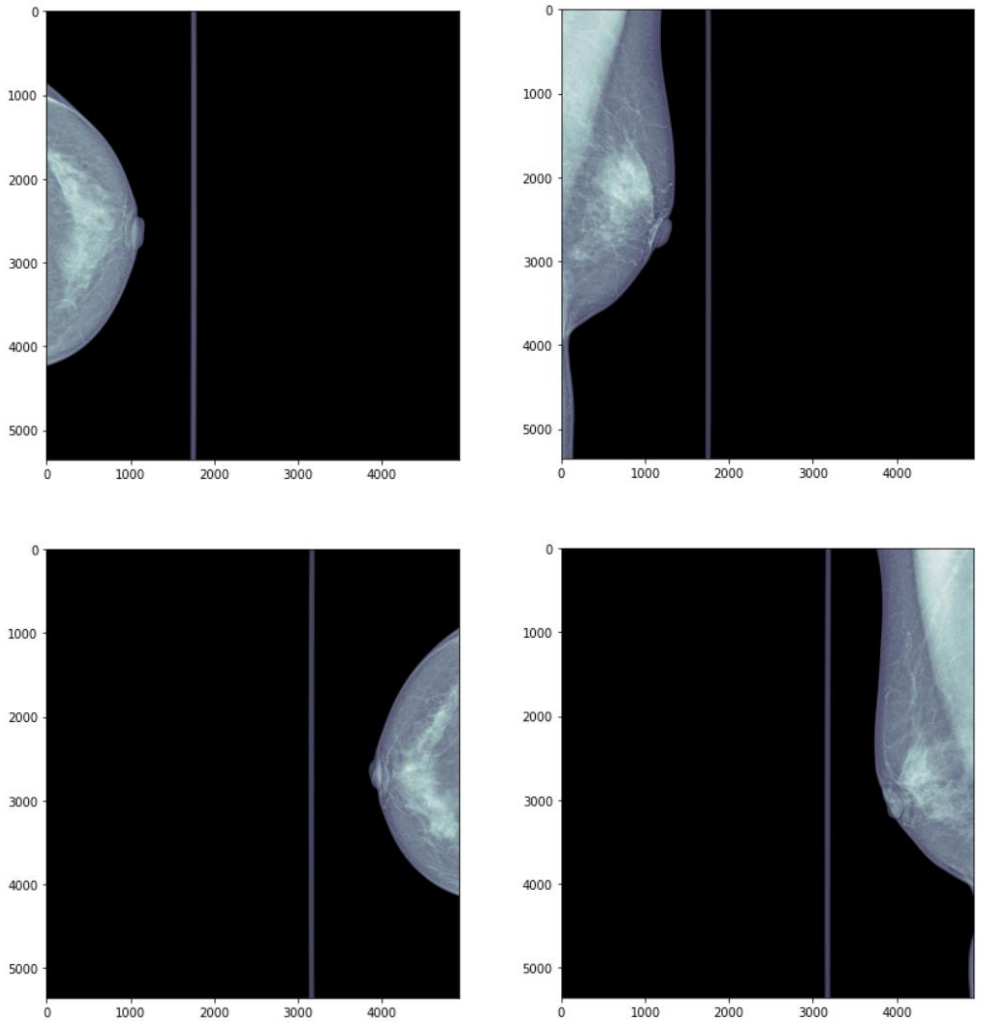



```
example = '../input/rsna-breast-cancer-detection/train_images/10006/1459541791.dcm'
pydicom.dcmread(example)
```

```
Dataset.file_meta -------------------------------
(0002, 0001) File Meta Information Version       OB: b'\x00\x01'
(0002, 0002) Media Storage SOP Class UID         UI: Digital X-Ray Image Storage - For Presentati
on
(0002, 0003) Media Storage SOP Instance UID      UI: 1.2.840.10009.1.2.3.10006.1.1459541791
(0002, 0010) Transfer Syntax UID                 UI: JPEG 2000 Image Compression (Lossless Only)
(0002, 0012) Implementation Class UID            UI: 1.2.840.113654.2.3.1995.2.12.0
(0002, 0013) Implementation Version Name         SH: 'PYDICOM 2.3.0'
-------------------------------------------------
(0008, 0018) SOP Instance UID                    UI: 1.2.840.10009.1.2.3.10006.1.1459541791
(0008, 0023) Content Date                        DA: '20221118'
(0008, 0033) Content Time                        TM: '183901.792591'
(0010, 0020) Patient ID                          LO: '10006'
(0020, 000d) Study Instance UID                  UI: 1.2.840.10009.1.2.3.10006
(0020, 000e) Series Instance UID                 UI: 1.2.840.10009.1.2.3.10006.1
(0020, 0013) Instance Number                     IS: '1459541791'
(0020, 0062) Image Laterality                    CS: 'L'
(0028, 0002) Samples per Pixel                   US: 1
(0028, 0004) Photometric Interpretation          CS: 'MONOCHROME1'
(0028, 0010) Rows                                US: 5355
(0028, 0011) Columns                             US: 4915
(0028, 0100) Bits Allocated                      US: 16
(0028, 0101) Bits Stored                         US: 16
(0028, 0102) High Bit                            US: 15
(0028, 0103) Pixel Representation                US: 0
(0028, 0120) Pixel Padding Value                 US: 3044
(0028, 1040) Pixel Intensity Relationship        CS: 'LOG'
(0028, 1041) Pixel Intensity Relationship Sign   SS: 1
(0028, 1050) Window Center                       DS: [1802.310000, 1802.310000, 2020.704000, 158
3.916000]
(0028, 1051) Window Width                        DS: [1091.970000, 1091.970000, 1091.970000, 109
1.970000]
(0028, 1052) Rescale Intercept                   DS: '0.0'
(0028, 1053) Rescale Slope                       DS: '1.0'
(0028, 1054) Rescale Type                        LO: 'US'
(0028, 1056) VOI LUT Function                    CS: 'SIGMOID'
(0028, 1350) Partial View                        CS: 'NO'
(0028, 2110) Lossy Image Compression             CS: '00'
(7fe0, 0010) Pixel Data                          OW: Array of 4362044 elements
```

数据概览

图像数据案例 – Patient ID 10006



[train/test]_images/[patient_id]/[image_id].dcm: 训练集/测试集病例图像数据，由dicom格式给出。训练集有54706张图片。线上测试集大约有8k个病例，每个病例约有4张样本图像。

- 正如我们从图中看到的那样，图像的**尺寸相当大**。
- 我们看到乳房只占图像的一小部分。
- 因此（特别是考虑到有限的计算预算！）找到**裁剪掉不包含任何信息的图像部分**的方法很可能是个好主意！分隔线在这里可以发挥重要作用。

数据概览

csv数据

train.csv

	site_id	patient_id	image_id	laterality	view	age	cancer	biopsy	invasive	BIRADS	implant	density	machine_id	difficult_negative_case
0	2	10006	462822612	L	CC	61.0	0	0	0	NaN	0	NaN	29	False
1	2	10006	1459541791	L	MLO	61.0	0	0	0	NaN	0	NaN	29	False
2	2	10006	1864590858	R	MLO	61.0	0	0	0	NaN	0	NaN	29	False
3	2	10006	1874946579	R	CC	61.0	0	0	0	NaN	0	NaN	29	False
4	2	10011	220375232	L	CC	55.0	0	0	0	0.0	0	NaN	21	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
54701	1	9973	1729524723	R	MLO	43.0	0	0	0	1.0	0	C	49	False
54702	1	9989	63473691	L	MLO	60.0	0	0	0	NaN	0	C	216	False
54703	1	9989	1078943060	L	CC	60.0	0	0	0	NaN	0	C	216	False
54704	1	9989	398038886	R	MLO	60.0	0	0	0	0.0	0	C	216	True
54705	1	9989	439796429	R	CC	60.0	0	0	0	0.0	0	C	216	True

54706 rows × 14 columns

test.csv

	site_id	patient_id	image_id	laterality	view	age	implant	machine_id	prediction_id
0	2	10008	736471439	L	MLO	81	0	21	10008_L
1	2	10008	1591370361	L	CC	81	0	21	10008_L
2	2	10008	68070693	R	MLO	81	0	21	10008_R
3	2	10008	361203119	R	CC	81	0	21	10008_R

数据概览

字段含义

[train/test].csv Metadata for each patient and image. Only the first few rows of the test set are available for download.

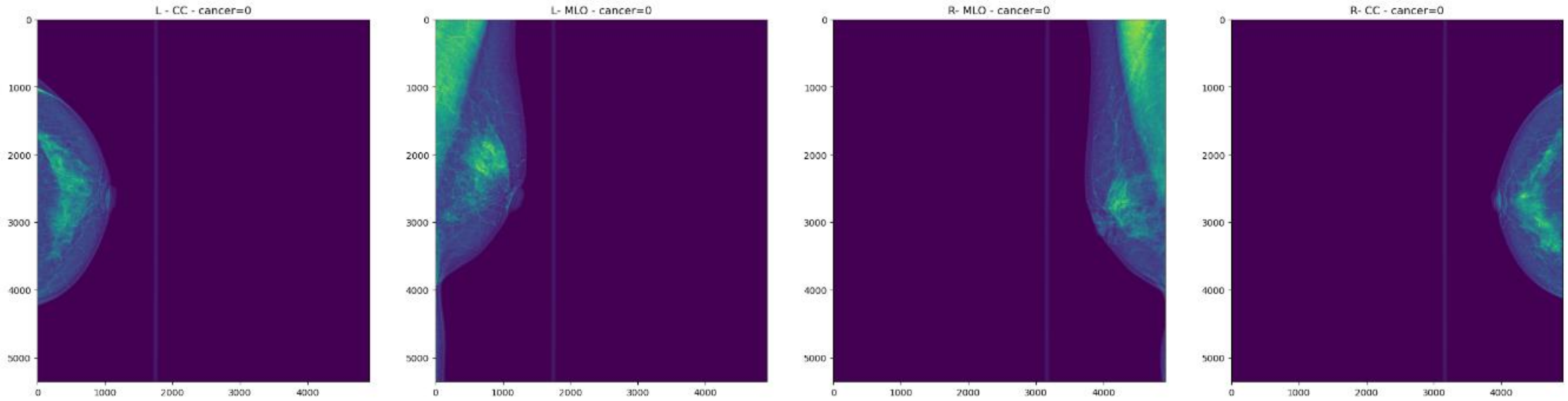
- `site_id` - ID code for the source hospital.
- `patient_id` - ID code for the patient.
- `image_id` - ID code for the image.
- `laterality` - Whether the image is of the left or right breast.
- `view` - The orientation of the image. The default for a screening exam is to capture two views per breast.
- `age` - The patient's age in years.
- `implant` - Whether or not the patient had breast implants. Site 1 only provides breast implant information at the patient level, not at the breast level.
- `density` - A rating for how dense the breast tissue is, with A being the least dense and D being the most dense. Extremely dense tissue can make diagnosis more difficult. Only provided for train.
- `machine_id` - An ID code for the imaging device.
- `cancer` - Whether or not the breast was positive for malignant cancer. The target value. Only provided for train.
- `biopsy` - Whether or not a follow-up biopsy was performed on the breast. Only provided for train.
- `invasive` - If the breast is positive for cancer, whether or not the cancer proved to be invasive. Only provided for train.
- `BIRADS` - 0 if the breast required follow-up, 1 if the breast was rated as negative for cancer, and 2 if the breast was rated as normal. Only provided for train.
- `prediction_id` - The ID for the matching submission row. Multiple images will share the same prediction ID. Test only.
- `difficult_negative_case` - True if the case was unusually difficult. Only provided for train.

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

数据概览

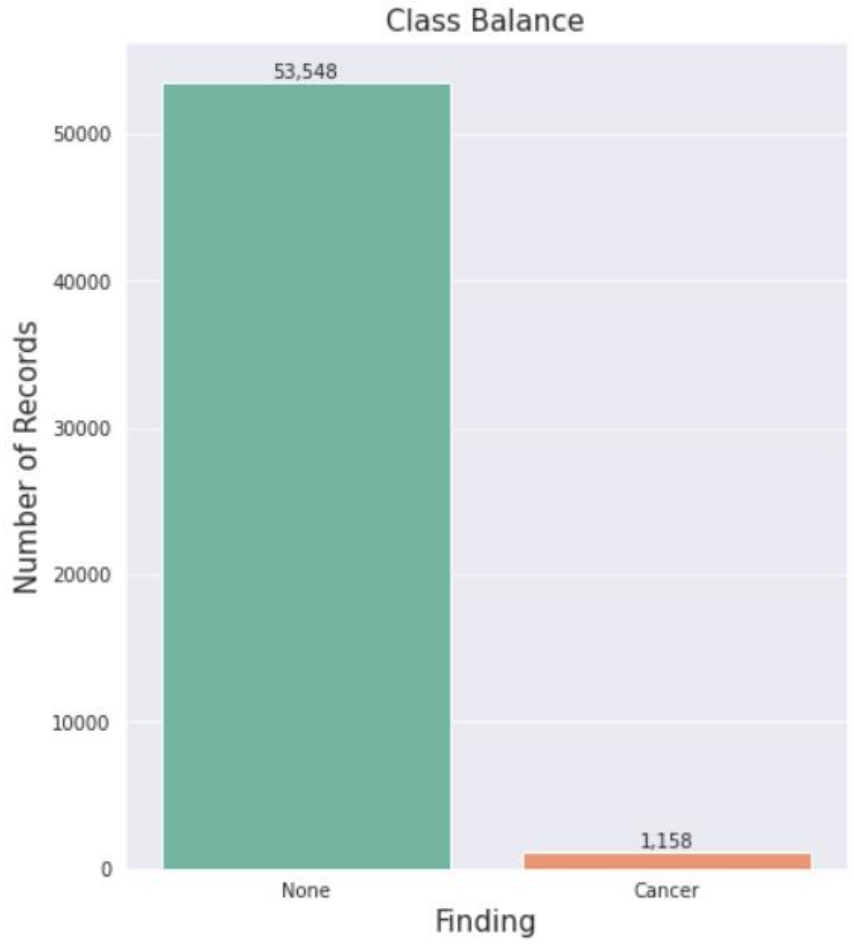
字段含义

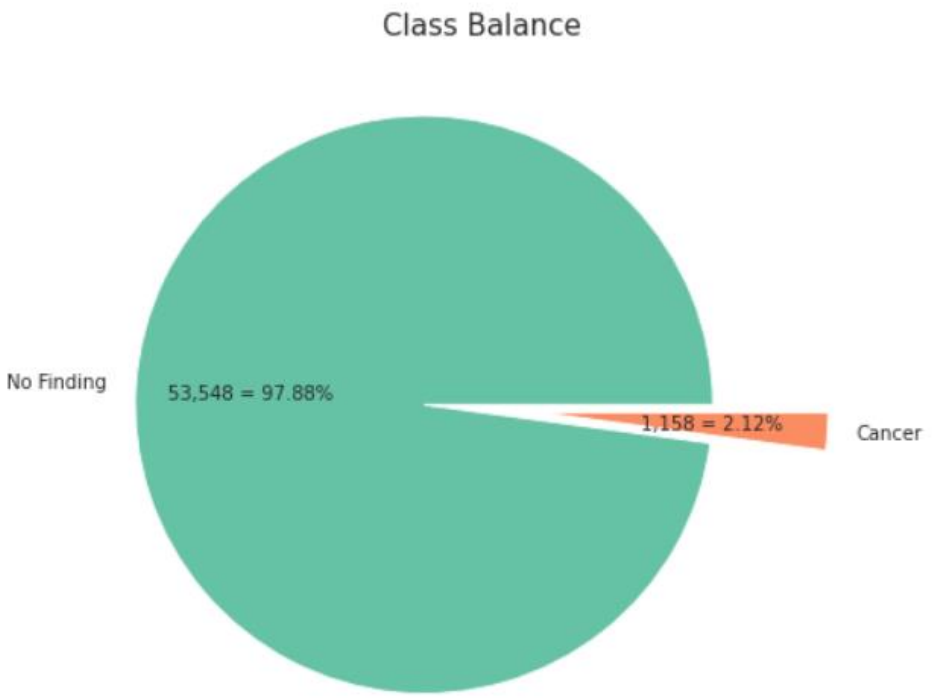
来自某病例的4张样本图片，L/R-左/右边乳腺部位，CC/MLO代表不同视角，cancer代表是否癌变



数据概览

标签分布

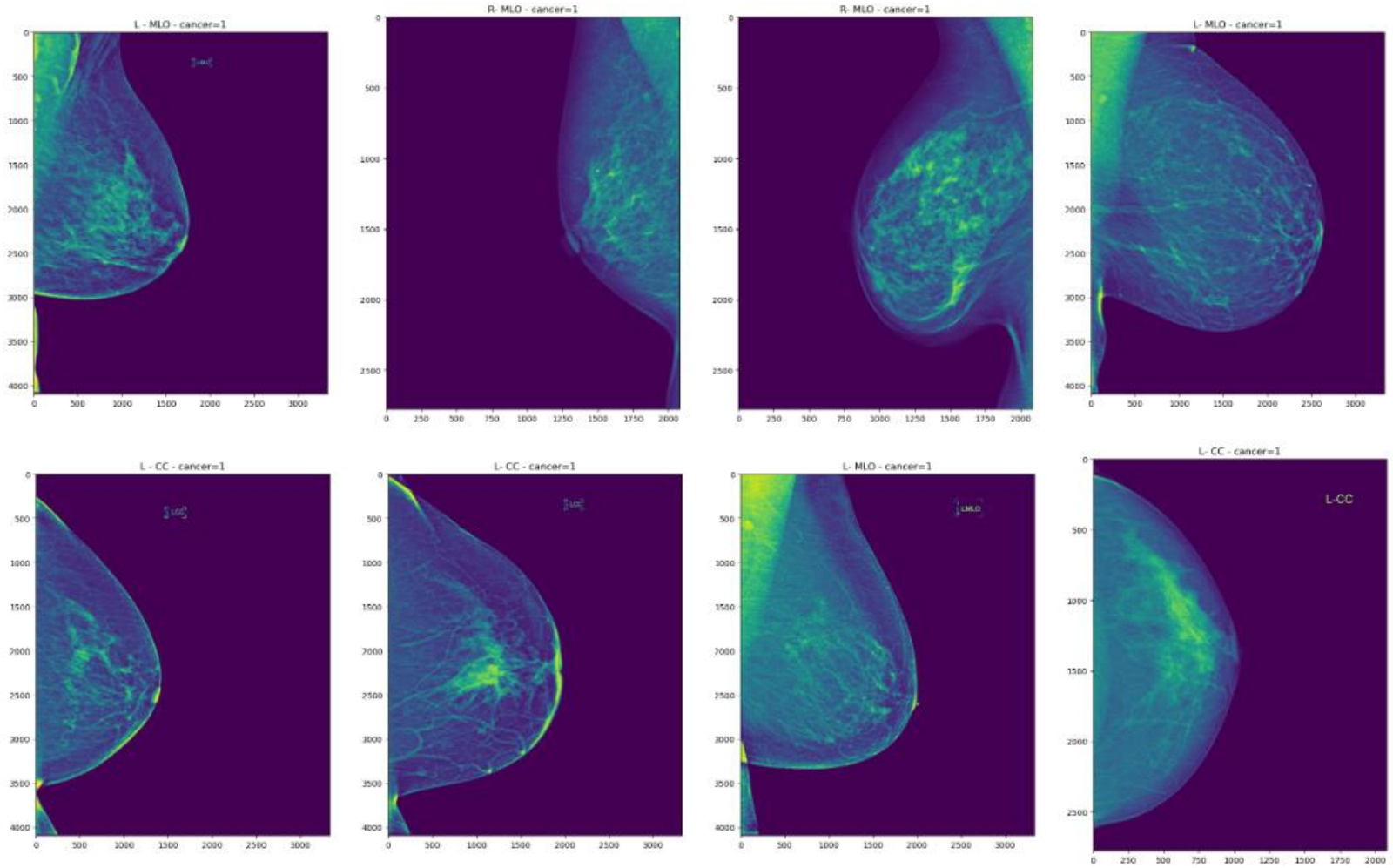




样本**标签分布极不平衡**的分类问题

数据概览

来自8个癌变病例的样本图片



评价指标

评价指标 – probabilistic F1 score (pF1)

$$pF_1 = 2\frac{p\text{ Precision } \cdot p\text{ Recall}}{p\text{ Precision } + p\text{ Recall}}$$

其中：

$$p\text{ Precision } = \frac{pTP}{pTP + pFP}$$

$$p\text{ Recall } = \frac{pTP}{TP + FN}$$

```python
def pfbeta(labels, predictions, beta):
    y_true_count = 0
    ctp = 0
    cfp = 0

    for idx in range(len(labels)):
        prediction = min(max(predictions[idx], 0), 1)
        if (labels[idx]):
            y_true_count += 1
            ctp += prediction
        else:
            cfp += prediction

    beta_squared = beta * beta
    c_precision = ctp / (ctp + cfp)
    c_recall = ctp / y_true_count
    if (c_precision > 0 and c_recall > 0):
        result = (1 + beta_squared) * (c_precision * c_recall) / (beta_squared * c_precision + c_recall)
        return result
    else:
        return 0
```

Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models

Reda Yacouby
Amazon Alexa
redaya@amazon.com

Dustin Axman
Amazon Alexa
dax@amazon.com

Abstract

In pursuit of the perfect supervised NLP classifier, razor thin margins and low-resource testsets can make modeling decisions difficult. Popular metrics such as Accuracy, Precision, and Recall are often insufficient as they fail to give a complete picture of the model's behavior. We present a probabilistic extension of Precision, Recall, and F1 score, which we refer to as confidence-Precision (cPrecision), confidence-Recall (cRecall), and confidence-F1 (cF1) respectively. The proposed metrics address some of the challenges faced when evaluating large-scale NLP systems, specifically when the model's confidence score assignments have an impact on the system's behavior. We describe four key benefits of our proposed metrics as compared to their threshold-based counterparts. Two of these benefits, which we refer to as *robustness to missing values* and *sensitivity to model confidence score assignments* are self-evident from the metrics' definitions; the remaining benefits, *generalization*, and *functional consistency* are demonstrated empirically.

is because the mentioned loss functions are differentiable convex functions, enabling optimization algorithms such as gradient descent to find minima with reasonable computational cost. In contrast, the test-set evaluation metrics are often required to be easy to relate to the real-world problem the classifier is designed to help solve, in order to give a concrete idea of performance or success to the stakeholders.

Essentially, all the criteria mentioned serve the same underlying purpose of driving modeling decisions. The heterogeneous nature of model evaluation illustrates how there could be no universal criteria for driving model decisions, or so-called "best metric", as each criterion could be advantageous under specific operating conditions (Hernández-Orallo et al., 2012), or even preferred by stakeholders for reasons that do not need to be scientifically driven (such as interpretability and business purposes).

In the Natural Language Processing (NLP) industry, new challenges have risen in the past few years in terms of performance evaluation, due to

https://www.kaggle.com/competitions/rsna-breast-cancer-detection/overview/evaluation

算法思路

■ 竞赛概览

该竞赛为一个**样本标签分布极不平衡**的医学影像二分类竞赛，竞赛优胜关键在于图像数据的预处理与样本不平衡问题的处理。

■ 所用算法

1. 本方案首先针对dicom图像进行关键兴趣区域提取（ROI extract），之后删除图像多余空白部分并根据原始比例适当填充图像区域。

2. 模型训练时，将图片放缩至适当尺寸，并选用ConvNeXtV2Tiny的ImageNet预训练模型进行迁移学习。

3. 同时采用undersample（欠采样）和引入额外数据集的方法解决样本标签不平衡问题，并设置不同负样本比例得到多个泛化模型进行集成学习。

4. 最后根据原始样本分布比例设置阈值得到最终的预测结果，并根据Kaggle公榜分数与本地验证集分数选择最佳模型。

■ 所用模型

ConvNeXtV2Tiny

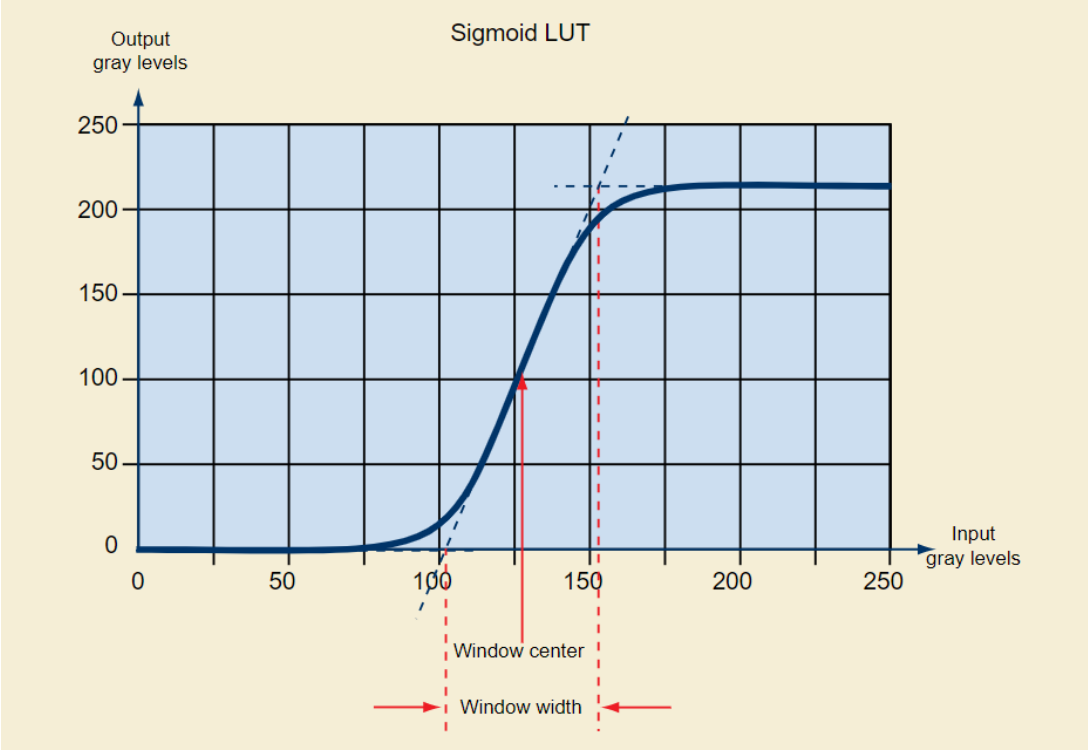
■ 额外数据集

https://www.kaggle.com/datasets/pourchot/ddsm-mammography-positive-case

算法思路

数据预处理：VOI-LUT转换

由于数字探测器的性能，数字投影 X 射线图像通常具有非常高的动态范围。为了显示这些图像，可以对图像应用各种感兴趣值 (VOI) 变换以促进诊断解释。针对dicom图像数据，VOI-LUT处理可以增强图片纹理和线条。

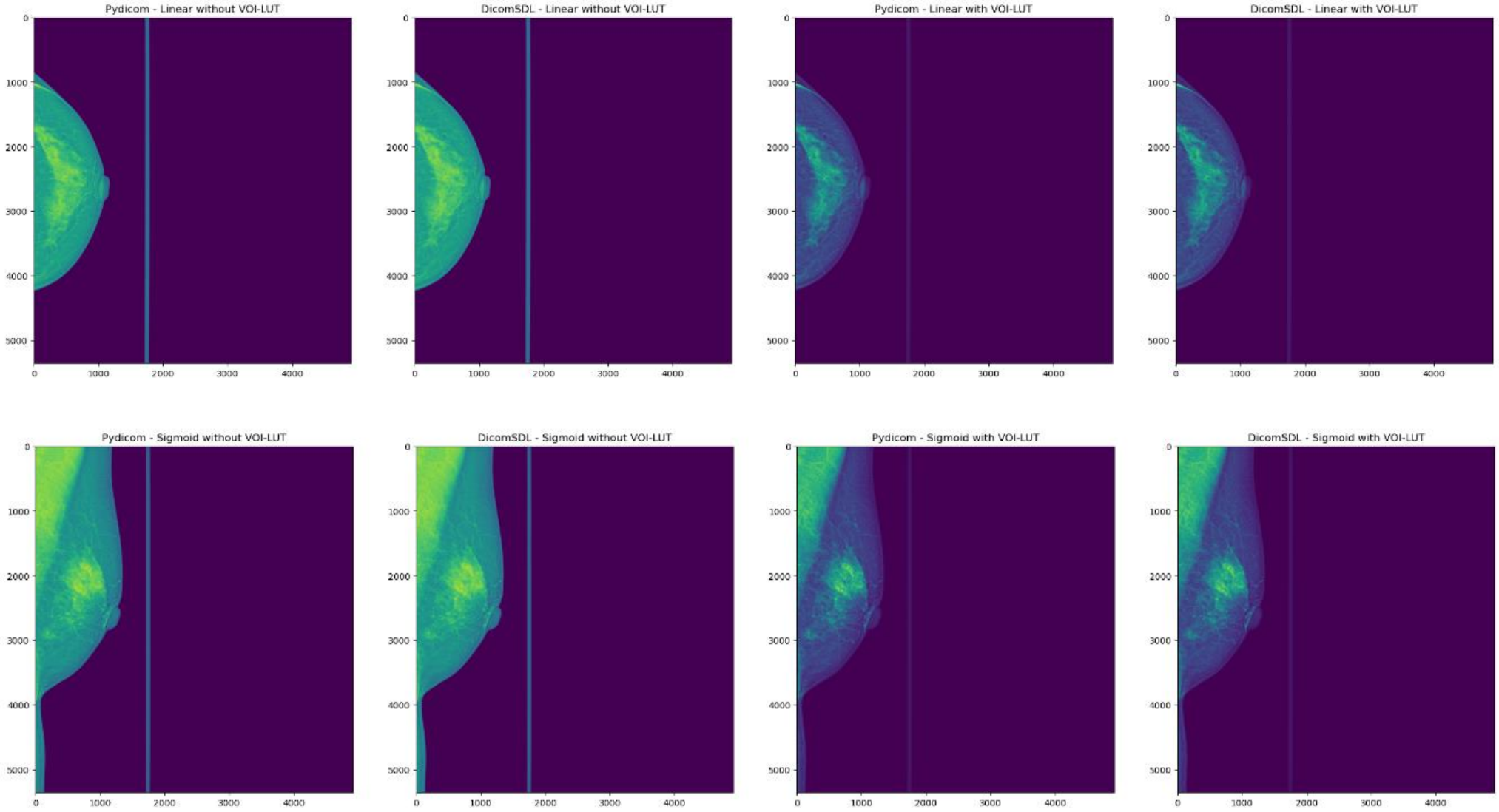


VOI_LUT

```python
# Source: https://www.kaggle.com/code/bobdegraaf/dicomsdl-voi-lut
def voi_lut(image, dicom):
    # Load only the variables we need
    center = dicom['WindowCenter']
    width = dicom['WindowWidth']
    bits_stored = dicom['BitsStored']
    voi_lut_function = dicom['VOILUTFunction']

    # For sigmoid it's a list, otherwise a single value
    if isinstance(center, list):
        center = center[0]
    if isinstance(width, list):
        width = width[0]

    # Set y_min, max & range
    y_min = 0
    y_max = float(2**bits_stored - 1)
    y_range = y_max

    # Function with default LINEAR (so for Nan, it will use linear)
    if voi_lut_function == 'SIGMOID':
        image = y_range / (1 + np.exp(-4 * (image - center) / width)) + y_min
    else:
        # Checks width for < 1 (in our case not necessary, always >= 750)
        center -= 0.5
        width -= 1

        below = image <= (center - width / 2)
        above = image > (center + width / 2)
        between = np.logical_and(~below, ~above)

        image[below] = y_min
        image[above] = y_max
        if between.any():
            image[between] = (
                ((image[between] - center) / width + 0.5) * y_range + y_min
            )

    # Normalize to have 0 as background, some images are reversed where 0 is max intensity
    if dicom['PhotometricInterpretation'] == 'MONOCHROME1':
        image = np.max(image) - image

    return image
```

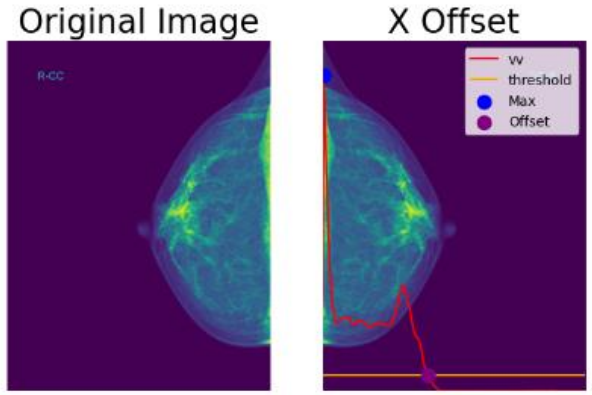
算法思路

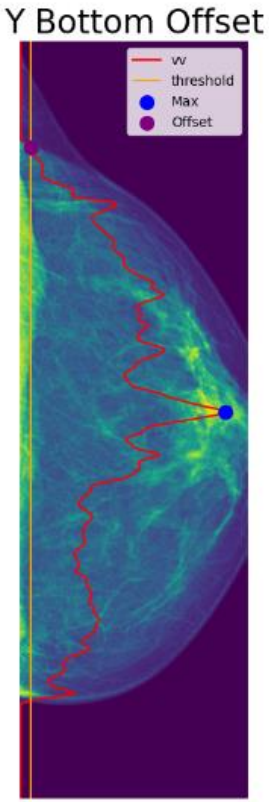
VOI-LUT转换

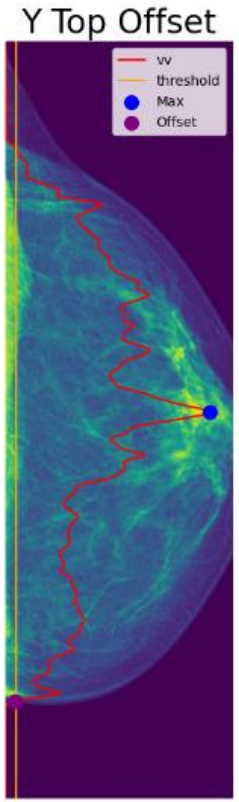


算法思路

数据预处理：ROI Extract







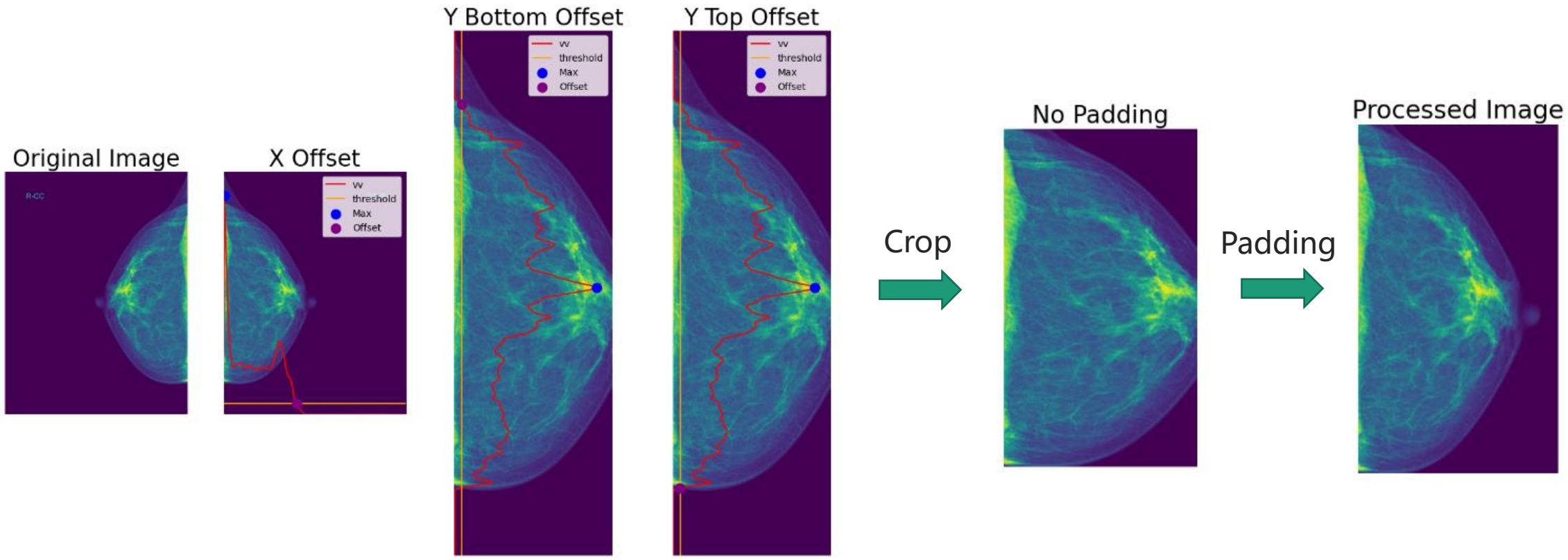
Crop Image

```python
# Smooth vector used to smoothen sums/stds of axes
def smooth(l):
    # kernel size is 1% of vector
    kernel_size = int(len(l) * 0.01)
    kernel = np.ones(kernel_size) / kernel_size
    return np.convolve(l, kernel, mode='same')

# X Crop offset based on first column with sum below 5% of maximum column sums*std
def get_x_offset(image, max_col_sum_ratio_threshold=0.05, debug=None):
    # Image Dimensions
    H, W = image.shape
    # Percentual margin added to offset
    margin = int(image.shape[1] * 0.00)
    # Threshold values based on smoothed sum x std to capture varying intensity columns
    vv = smooth(image.sum(axis=0).squeeze()) * smooth(image.std(axis=0).squeeze())
    # Find maximum sum in first 75% of columns
    vv_argmax = vv[:int(image.shape[1] * 0.75)].argmax()
    # Threshold value
    vv_threshold = vv.max() * max_col_sum_ratio_threshold
```

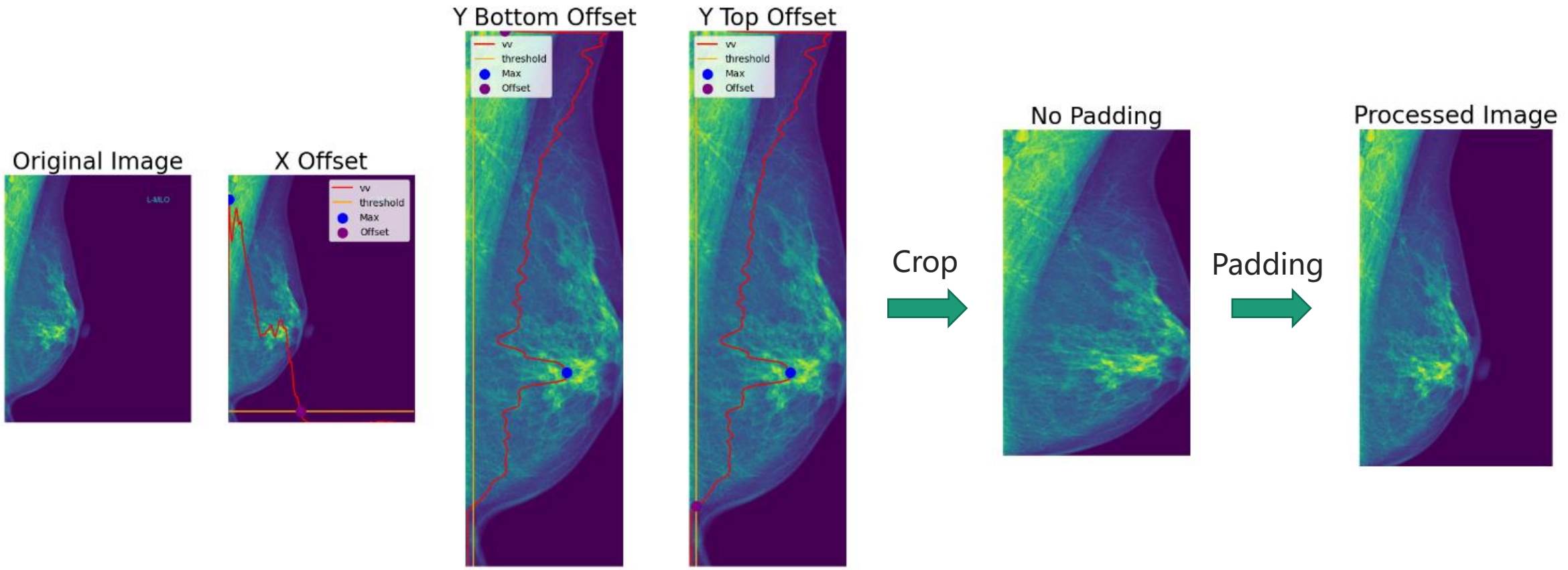
```python
# Y Crop offset based on first bottom and top rows with sum below 10% of maximum row sum*std
def get_y_offsets(image, max_row_sum_ratio_threshold=0.05, debug=None):
    # Image Dimensions
    H, W = image.shape
    # Margin to add to offsets
    margin = 0
    # Threshold values based on smoothed sum x std to capture varying intensity columns
    vv = smooth(image.sum(axis=1).squeeze()) * smooth(image.std(axis=1).squeeze())
    # Find maximum sum * std row in inter quartile rows
    vv_argmax = int(image.shape[0] * 0.25) + vv[int(image.shape[0] * 0.25):int(image.shape[0] * 0.75)].argmax()
    # Threshold value
    vv_threshold = vv.max() * max_row_sum_ratio_threshold
    # Default crop offsets
    offset_bottom = 0
    offset_top = H
```

算法思路

Padding



算法思路


Original Image
X Offset
vv
threshold
Max
Offset
Y Bottom Offset
Y Top Offset
Crop
No Padding
Padding
Processed Image

算法思路

样本标签不平衡的解决办法

1. Oversampling: 过采样通过增加少数类的大小来平衡数据集。通过使用例如重复、自举或SMOTE （合成少数过采样技术）生成新的稀有样本，而不是去除大量样本。

2. **Undersampling**: 欠采样通过减少多数类的大小来平衡数据集。当数据量足够时，使用此方法。通过保持稀有类中的所有样本，并随机选择丰富类中相同数量的样本，可以检索到平衡的新数据集，用于进一步建模。

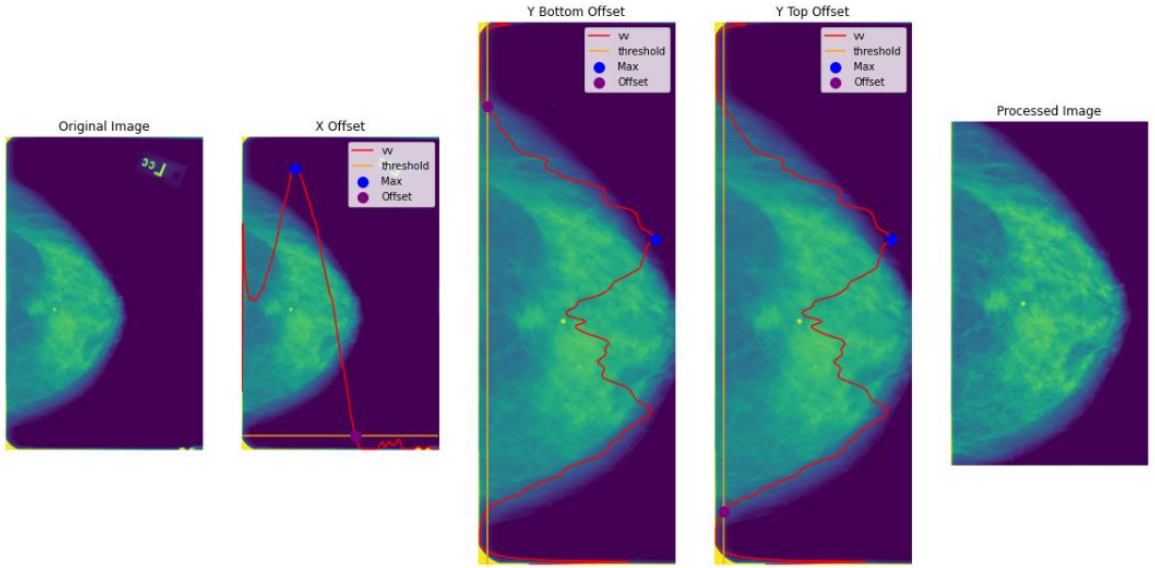
```
def undersample_majority(X, y):
    # Filter 2/3 of negative samples to upsample positive samples by a factor 3
    return y == 1 or tf.random.uniform([ ]) > 0.66
```

3. **External data**: 收集更多的少数类数据。

算法思路

额外数据集

将额外数据集的所有癌变样本（正样本）加入到训练集数据，采用统一数据处理格式，同时调整负样本undersample比例参数，以适应原始类别标签分布。由于额外数据样本与训练数据样本像素统计分布有差异，可通过数据增强\转换（augment）混淆样本，达到减小分布差异的效果。



https://www.kaggle.com/datasets/pourchot/ddsm-mammography-positive-case

DDSM-mammography-positive-case

University of South Florida Digital Mammography

Data Card | Code (0) | Discussion (0)

About Dataset

No description available

additional_train.csv (27.52 kB)

Detail | Compact | Column

4 of 4 colum

patient_id	laterality	view	cancer
752 unique values	LEFT 53% RIGHT 47%	MLO 53% CC 47%	1 1
P_00127	RIGHT	CC	1
P_00127	RIGHT	MLO	1
P_00150	RIGHT	MLO	1
P_00164	RIGHT	CC	1
P_00202	RIGHT	CC	1
P_00202	RIGHT	MLO	1
P_00244	RIGHT	MLO	1

算法思路

模型搭建 (训练：Kaggle TPU环境)

CNN backbone – ConvNeXtV2Tiny

```
def get_model():
    with STRATEGY.scope():
        image = tf.keras.layers.Input(INPUT_SHAPE, name='image', dtype=tf.uint8)
        image_norm = normalize(image)
        x = convnext.ConvNeXtV2Tiny(
            input_shape=(IMG_HEIGHT, IMG_WIDTH, 3),
            pretrained='imagenet21k-ft1k',
            num_classes=0,
        )(image_norm)
        x = tf.keras.layers.GlobalAveragePooling2D()(x)
        x = tf.keras.layers.Dropout(0.30)(x)
        outputs = tf.keras.layers.Dense(1, activation='sigmoid')(x)
        optimizer = tfa.optimizers.AdamW(learning_rate=LR_MAX, weight_decay=LR_MAX*WD_RATIO, epsilon=1e-6)
        loss = tf.keras.losses.BinaryCrossentropy(from_logits=False)
        metrics = [pF1()]
        model = tf.keras.models.Model(inputs=image, outputs=outputs)
        model.compile(optimizer=optimizer, loss=loss, metrics=metrics)
        return model

history = model.fit(
        ...
        class_weight = {
            0: 1.0,
            1: 5.0,
        },
    )
```

ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders

Sanghyun Woo[1*] Shoubhik Debnath[2] Ronghang Hu[2]
Xinlei Chen[2] Zhuang Liu[2] In So Kweon[1] Saining Xie[3†]
[1]KAIST [2] Meta AI, FAIR [3]New York University

Code: https://github.com/facebookresearch/ConvNeXt-V2

Abstract

Driven by improved architectures and better representation learning frameworks, the field of visual recognition has enjoyed rapid modernization and performance boost in the early 2020s. For example, modern ConvNets, represented by ConvNeXt [52], have demonstrated strong performance in various scenarios. While these models were originally designed for supervised learning with ImageNet labels, they can also potentially benefit from self-supervised learning techniques such as masked autoencoders (MAE) [31]. However, we found that simply combining these two approaches leads to subpar performance. In this paper, we propose a fully convolutional masked autoencoder framework and a new Global Response Normalization (GRN) layer that can be added to the ConvNeXt architecture to enhance inter-channel feature competition. This co-design of self-supervised learning techniques and architectural improvement results in a new model family called ConvNeXt V2, which significantly improves the performance of pure ConvNets on various recognition benchmarks, including ImageNet classification, COCO detection, and ADE20K segmentation. We also provide pre-trained ConvNeXt V2 models of various sizes, ranging from an efficient 3.7M-parameter Atto model with 76.7% top-1 accuracy on ImageNet, to a 650M Huge model that achieves a state-of-the-art 88.9% accuracy using only public training data.

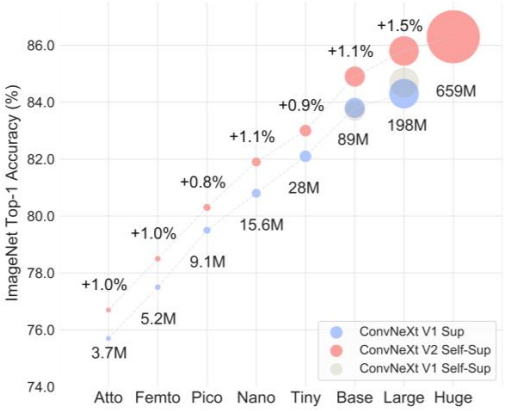


Figure 1. **ConvNeXt V2 model scaling**. The ConvNeXt V2 model, which has been pre-trained using our fully convolutional masked autoencoder framework, performs significantly better than the previous version across a wide range of model sizes.

used for training the network, and the data used for training. In the field of visual recognition, progress in each of these areas contributes to overall improvements in performance.

Innovation in neural network architecture design has consistently played a major role in the field of representation learning. Convolutional neural network architectures

竞赛总结

- 计算机视觉类对于图像尺寸比较敏感，一般而言，尺寸越大，效果越好。
- 图像数据的处理-ROI区域提取以及原始比例的保持。
- 极度类别不平衡数据的解决方法-undersample，引入额外数据。
- 根据标签分布来设置后处理阈值。