Breast Mass Detection and Classification via YOLOv5 Models

Summer Research Internship 2021

Krutik M. Parmar - 201801199
Information and Communication Technology, DAIICT

Supervisor:
Prof. Pankaj Kumar

Abstract

With recent breakthroughs in deep learning, the use of Deep convolution neural networks (CNNs) in healthcare, especially in medical imaging has achieved remarkable success. Besides that, accurate breast mass detection and classification in mammograms is still a critical challenge because of several reasons, for instance, the heterogeneity of breast masses and the complexity of surrounding environment. In this report, an end-to-end system for localization and classification of suspicious breast lesions from entire mammograms is presented. The system is based on You-Only-Look-Once (YOLO) object detection model. The suggested system converts mammograms from DICOM format to pictures without sacrificing data. Then, without any human intervention, it detects masses in full-field digital mammograms and distinguishes between MALIGNANT and BENIGN tumours. In this project, YOLOv5, which is the latest version of YOLO algorithm is used. The model is evaluated on two publicly available datasets, with 1127 mammograms from Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) and 106 mammograms from INbreast database. We have experimented with all 4 different models of YOLOv5 and our best results reached an mAP@IoU=0.5 36.4% on unseen breast mammograms.

Acknowledgement

I would like to take the opportunity to thank and express my deep sense of gratitude to my supervisor Prof. Pankaj Kumar. I am greatly indebted to him for providing his valuable guidance at all stages of this project, his advice, constructive suggestions, positive and supportive attitude and continuous encouragement, without which it would have not been possible to complete this prject.

Krutik Parmar,
DAIICT,
Gandhinagar.

Contents

List of Tables

List of Figures

1 Introduction

According to recent studies, Breast Cancer is the most frequent cancer in women, accounting for almost one-third of newly diagnosed tumours in the United States. Breast cancer is also associated with a significant mortality rate, accounting for 17% of all cancer-related fatalities. Besides that, as expected by WHO, in 2025 the number of cases that will be diagnosed with breast cancer will be nearly 19.3 million cases. Early detection of breast cancer has been found to improve the survival rate of patients in clinical trials. Hence, research into early detection of breast cancer, as well as precise diagnosis and therapy, is critical in order to improve the survival rate. Breast cancer can be detected by abnormalities such as lumps, micro-calcifications, and areas of asymmetry and distortion in the breast.

Masses are the most characteristic and common sort of abnormality among these. Masses, on the other hand, can be easily disguised by overlapping breast tissues, making detection difficult. Furthermore, some breast tissues physically resemble masses and are therefore mistaken as masses. An undetected mass can delay patient's diagnosis and a misidentified mass can lead to additional testing causing unnecessary anxiety and pain to the patient. These issues limit mammography's effectiveness and utility. As a result, ra-

diologists consider this duty to be a daily challenge since there are so many mammograms to check that it takes time and effort to evaluate each view of a mammogram.

Mammography has remained the most effective method for general population screening to this day. Mammography is a procedure that involves using low-energy X-rays to examine the breast for any suspicious malignancies. Alternatives to mammograms include Magnetic Resonance Imaging (MRI), which is used by radiologists to confirm the presence of any mass in the human breast. However, one disadvantage of MRI is that the patient may develop an allergic reaction to the contrasting agent or claustrophobia as a result of a skin infection at the injection site. As a result of the disadvantages of most screening technologies, x-ray mammography is the most popular imaging technology used to screen breasts and identify cancer due to its speed of acquisition and cost-efficiency. However, detecting and diagnosing a breast lesion simply based on mammography findings is difficult and heavily reliant on the radiologist's competence, resulting in a high proportion of false positives and extra examinations. As a result, the development of precise Computer-Aided Diagnose (CAD) systems is seen as the second key opinion that physicians can utilise to aid and support their decisions about mass detection and classification.

Deep learning has recently received a lot of interest in the field of machine learning. It has also been used in the field of breast cancer CAD. The usage of handcrafted features has recently been superseded by deep learning-based CAD systems that can learn automatically which picture features are more relevant to be utilised to perform a diagnosis, thanks to tremendous break-

throughs in the development of deep learning technologies. Deep learning methods are distinguished by their capacity to learn data representations utilising several layers of representation, beginning with raw input and progressing through higher levels of representation until completing a complete learning process. The most common type of deep learning architecture is the convolution neural network (CNN). CNN consists of a number of convolution layers which can extract features that represent the various contexts of images without feature engineering. Many models have been evolved from a simple CNN in order to accurately detect object from the image, such as R-CNNs, Fast CNNs and Faster R-CNNs. These popular models have overcome many limitations of deep learning such as computational time, redundancy, overfitting and parameters size. However, training and implementing most of these models is often time-consuming and requires a high computational memory. Therefore, another variation called You-Only-Look-Once (YOLO), which is characterized with a low-memory dependence, has been recognized as a fast object detection model and suitable for CAD systems. We have used the latest version of YOLO algorithm which is YOLOv5 for detection and classification. It has very low execution time and very high mean Average Precision value on COCO dataset.

2 Dataset

2.1 Datasets

In this project, the CBIS-DDSM and INbreast public datasets were used to train and evaluate the proposed methodology.

CBIS-DDSM (Clark et al. (2013)) is an updated and standardized version of the Digital Database for Screening Mammography (DDSM) dataset, where images were converted from Lossless Joint Photographic Experts Group (LJPEG) to Digital Imaging and Communications in Medicine (DICOM) format. It was reviewed by radiologists after eliminating inaccurate cases and confirmed with the histopathology classification. It contains 2907 mammograms from 1555 patients and it is organized in two categories of pathology which are Mass and Calcification. We have taken only 1127 mammograms which contains masses. For each breast, mammograms were taken from two distinct angles (i.e., MLO and CC). The images have an average size of 3000 x 4800 pixels and are linked to their pixel-level ground-truth for the location and type of suspicious spots.

INbreast (Moreira et al. (2012)) is a public dataset of images acquired using the MammoNovation Siemens fullfield digital mammography (FFDM) that are stored in DICOM format. The database contains 410 mammograms

where 235 cases include abnormalities in both MLO and CC views from 115 patients, and thus normal mammograms were excluded. Images have an average size of 3328 x4084 pixels and come with their annotated ground-truth. In this dataset also, we have taken only those mammograms which contains mass which are total 107 mammograms.

2.2 Data Preparation

For the CBIS-DDSM dataset, the images were already classified in BENIGN and MALIGNANT. But for the INbreast dataset, the cases were classified in Bi-Rads category from 0 to 6. The images are originally in DICOM format and each patient case has an XML file containing all pertinent information, such as pixel-level ground truth annotations and histology verification for cancers.

The Bi-Rads categories 2 and 3 are assigned to the benign cases, while the categories 4, 5, and 6 are assigned to the malignant cases. So, total, there are 41 masses with Bi-Rads belong to 2, 3 are classified as benign, while 75 masses with Bi-Rads belong to 4, 5, 6 are classified as malignant, in overall 107 mammograms.

Because the pixels in the mammograms have a 14-bit contrast, their values ranged from 0 to $2^{14} - 1$, representing 16384 different colours. The mammograms are scaled from 0 to 255. The mass ground truth coordinates are taken from the XML files and are normalised relative to the width and height of the image to be between [0.0 and 1.0] to make the experiments easier to run with different resolutions and sizes. Both the classes are recorded, together with its coordinates in x, y, width, and height, to be used as an in-

put to the model, indicating the case annotation. Finally, the mammograms are reduced to 640x640 resolution in order to fit into the model.

Deep learning models require a huge quantity of labelled data to be trained, which aids in their generalisation. Most of the collected datasets for medical applications have a small number of instances and often have an uneven distribution, which makes training deep learning models difficult. To overcome this problem, we have employes the solution which is used in many studies and that is Data Augmentation.

Data augmentation offers a process of increasing experimentally size of the dataset. In this project and for the particular detection task, we augmented the training images which are 80% of total images, eight times. First we rotated original images with eight different angles, each angle can be a random angle between 0 and 180° and we transformed them using Contrast Limited Adaptive Histogram Equalization (CLAHE) method with two variations, clip limit 1 and clip limt 2 with the same grid size of (8,8) and then we merged both of these images. Thus, a total of 7208 and 680 mammograms were respectively collected for CBIS-DDS and INbreast to train the proposed model. We have only applied CLAHE method on testing data.

The initialization of the trainable parameters is the first step in deep learning models (i.e., weights, bias). Random initialization and transfer learning are two widely used approaches for doing so. We used the transfer learning technique in our work, which involved applying the weights of a pre-trained model on a larger annotated dataset (i.e., ImageNet, MSCOCO, etc.) and then re-training and fine-tuning the new weights on the augmented dataset.

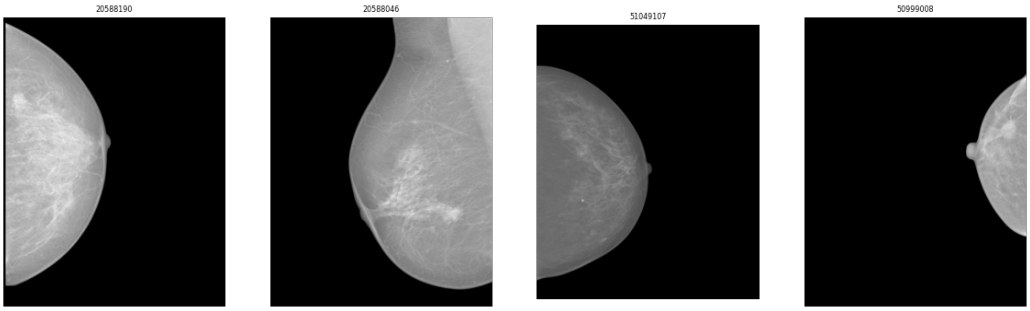

Figure 2.1: Original INbreast Images

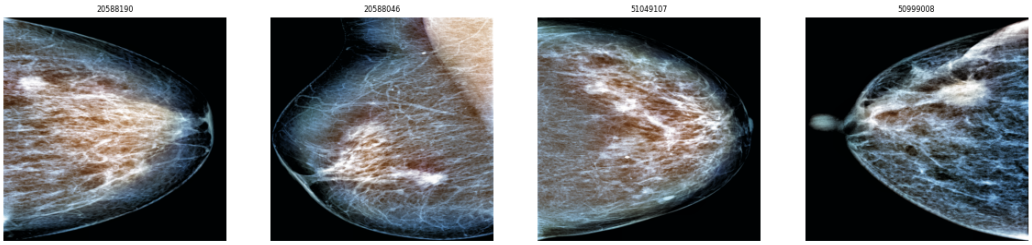

Figure 2.2: INbreast Images after preprocessing

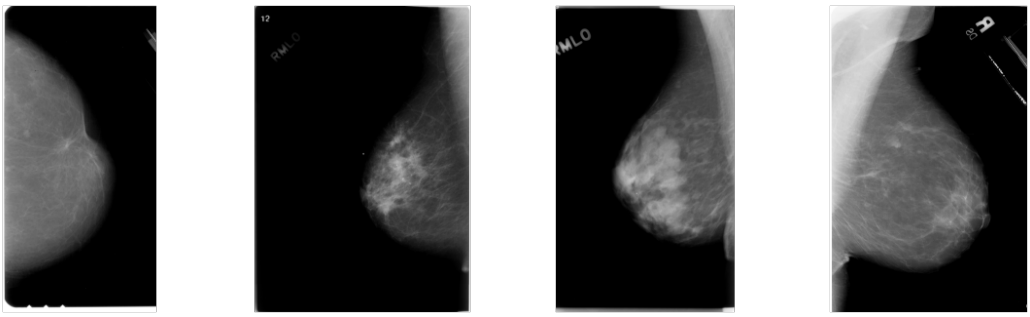

Figure 2.3: Original CBIS-DDSM Images

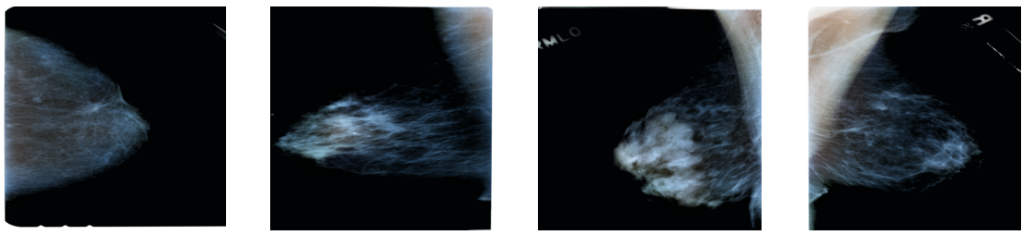

Figure 2.4: CBIS-DDSM Images after preprocessing

3 Methodology

The goal of this project is to detect the presence of a tumour in a mammogram and characterise its pathology as malignant or benign without the use of humans. The main building block of our model is a state-of-the-art object detection framework, You Only Look Once (YOLO).

3.1 YOLO structure

Object detection is a regression problem in which the right coordinates of an image's pixels are translated into a bounding box that surrounds a specific object. Popular regional-based neural network models predict multiple bounding boxes and use regions to localize objects within images after being fed into a CNN that generates a convolutional feature map. This method uses a selective search to identify the most suitable regions from photos, after which it predicts the offset values for the final bounding boxes. Because this technique is typically sluggish and memory intensive, a YOLO deep learning network was proposed, in which a single CNN predicts bounding box allocation and class label probabilities from an entire image. The fact that YOLO does not involve extracting features from sliding windows contributes to its low computing cost. In fact, to immediately detect each bounding box and

its class label probability, it just uses information from the entire image.

YOLO is simply based on the fully convolution neural network design. It starts with dividing an input mammogram into $N \times N$ grid cells and for each grid cell, bounding boxes are returned with a confidence score and class probabilities. Confidence score is computed by multiplying the probability of an existing class object with the intersection over the union (IoU) score as detailed in Eq. (3.1).

$$Confidence = Prob(object) \times IoUscore \tag{3.1}$$

In addition, the detected object is classified as Benign or Malignant class according to its class probability and its confidence score for that specific class label as explained below in Eq. (2).

$$ClassProbability = Prob(Class_i|object) \times IoUscore \tag{3.2}$$

YOLO is trained with the entire breast image and ROIs and understands the characteristics of ROIs. For training, we prepare the training data with the ROI position and size information: the information of training data contains the center position (x, y), width (w), height (h), and class label of the masses.

The YOLO network consists of three main pieces.

1. **Backbone :** A convolutional neural network that aggregates and forms image features at different granularities. YOLOv5 uses CSP Bottleneck to formulate image features. The CSP model is based on DenseNet. The CSP solves vanishing gradient problems in larger ConvNet back-

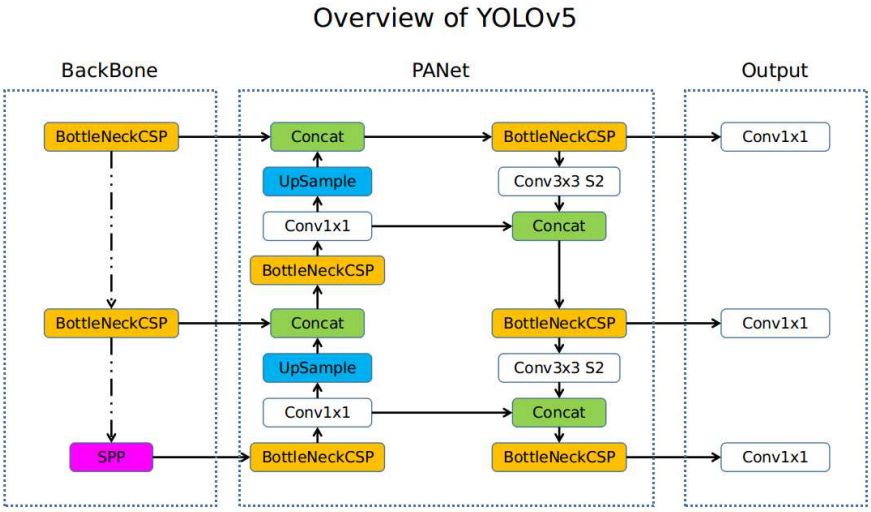

Figure 3.1: YOLO structure

bones, resulting in fewer parameters and FLOPS for the same amount of work. This is critical for the YOLO family, which places a premium on inference speed and compact model size.

2. **Neck :** A series of layers to mix and combine image features to pass them forward to prediction. YOLOv5 uses PA-NET for feature aggregation.

3. **Head :** Consumes features from the neck and takes box and class prediction steps.

[1]

[1] *Image Source* : https://github.com/ultralytics/yolov5/issues/280

3.2 Evaluation Metrics

We have used mean Average Precision (mAP) to evaluate both the classification and localization of the model. To guarantee that mammograms accurately detected breast lesions, we first calculated the intersection over union (IoU) score between each detected box and its ground-truth. The IoU score formula is detailed in Eq. (3).

$$IoUscore = \frac{AreaOfIntersection}{AreaOfUnion} \tag{3.3}$$

Average Precision is calculated as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight.

$$Precision = \frac{TP}{TP + FP} \tag{3.4}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.5}$$

where, TP and FN denote the true positive and false negative cases, respectively. TN and FP contain the true negative and false positive cases, respectively. There are three conditions which needs to be satisfied in order to make a detection true positive(TP). First, confidence score of the detection has to be greater than a specific threshold. Second, the predicted class should match the class of ground truth. Third, the predicted bounding-box should have IoU greater than a threshold.

The calculation of Average Precision(AP) only involves one class. However, in our case, there are two classes. Mean average precision (mAP) is defined as the mean of AP across both the classes:

$$mAP = \frac{AP_{BENIGN} + AP_{MALIGNANT}}{2} \tag{3.6}$$

4

Experiments and Results

In this experiment, models were trained on Google Colab which has Cloud GPU Tesla-T4 with 16 GB RAM. Both the datasets CBIS-DDSM and INbreast were divided into 80% training test, 10% Validation set and 10% testing set. We have trained all 4 types of YOLOv5 models and compared the results. We have trained the models for 100 epochs. The Adam optimizer was used for training of the models. Also the IOU used in the experiment is set to be 0.5 and the confidence probability threshold is set to 0.25 to accept all objects the model confident from them by more than 25%. We have tested our models on 134 different unseen mammogram of which 21 are from INbreast dataset and remaining 113 are from CBIS-DDSM dataset. Results for CBIS-DDSM and INbreat dataset are shown in the Table 4.1 and Table 4.2, respectively. Here, in both the tables the execution time is given per image.

Table 4.1: CBIS-DDSM Results

YOLOv5 Model	Pretrained	Batch Size	Benign AP	Malignant AP	mAP(%) @IoU=0.5	Execution Time(mS)
Small	COCO yolov5s	64	21.4%	41.2%	31.3%	12.3
Medium	COCO yolov5m	32	29.5%	43.3%	36.4%	25.3
Large	COCO yolov5l	20	22.9%	34.7%	28.8%	39.5
Extra Large	COCO yolov5x	16	8.53%	38%	23.3%	68.8

For training the models on CBIS-DDSM dataset, we have intialized the YOLOv5 models with default models' weights which were pre-trained on COCO dataset and then we fully trained all four YOLOv5 models. As we can observe from results of CBIS-DDSM dataset that as we go from small to medium model the mAP is increasing but if we go from medium to large and extra large sized models, the mAP is decreasing. This is because we the amount of data that we have is very less so bigger sized models are overfitting on this data. The execution time per image is increasing from small to large image because the size of the model is increasing.

Table 4.2: INbreast Results

YOLOv5 Model	Pretrained	Freezed Layers	Batch Size	Benign AP	Malignant AP	mAP(%) @IoU=0.5	Execution Time(mS)
Small	COCO yolov5s	None	64	5.68%	9.05%	7.36%	16.2
Medium	COCO yolov5m	None	32	2.88%	11.7%	7.28%	23.4
Large	COCO yolov5l	None	16	5.28%	7.84%	6.56%	42.6
Extra Large	COCO yolov5x	None	16	3.75%	12.4%	8.07%	66
Small	CBIS-DDSM Small	Backbone	128	3.86%	20%	11.9%	13.1
Medium	CBIS-DDSM Medium	Backbone	114	1.04%	18.6%	9.82%	25.4
Large	CBIS-DDSM Large	Backbone	64	4.11%	24.1%	14.1%	41.8
Extra Large	CBIS-DDSM Extra Large	Backbone	28	4.82%	11.5%	8.18%	68.7

For INbreast dataset, same as previously describes, we initialized the all four YOLOv5 models with default models' weights which were pre-trained on COCO dataset. If we observe the INbreast dataset results, first four models

which are fully trained without any fine tuning are giving very poor results. For the next four models, we froze the backbone of pre-trained CBIS-DDSM models and then finetune the remaining models with INbreast dataset. It gave us slightly better results than the previously fully trained models' results.

5

Conclusions and future work

In present work, we trained YOLOv5 models on breast mass detection and classification. When we performed experiments, we found that the mAP(%) @IoU=0.5 is highest for YOLOv5 medium model which is 36.4% with execution time per image 25.3ms. From this result, it can be concluded that we can use YOLOv5 Medium model for real-time Breast Mass Detection and Classification because of its low average execution time and high mAP. We can definitely increase the mAP by training model for more number of epochs on a Physical GPU and increasing the amount of data.

References

M. A. Al-masni, M. A. Al-antari, J. M. Park, G. Gi, T. Y. Kim, P. Rivera, E. Valarezo, S.-M. Han, and T.-S. Kim. Detection and classification of the breast abnormalities in digital mammograms via regional Convolutional Neural Network. IEEE, 7 2017. doi: 10.1109/embc.2017.8037053.

Ghada Hamed Aly, Mohammed Marey, Safaa Amin El-Sayed, and Mohamed Fahmy Tolba. YOLO Based Breast Masses Detection and Classification in Full-Field Digital Mammograms. page 105823. Elsevier BV, 3 2021. doi: 10.1016/j.cmpb.2020.105823.

Haichao Cao, Shiliang Pu, Wenming Tan, and Junyan Tong. Breast mass detection in digital mammography based on anchor-free architecture. page 106033. Elsevier BV, 6 2021. doi: 10.1016/j.cmpb.2021.106033.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. pages 1045–1057. Springer Science and Business Media LLC, 7 2013. doi: 10.1007/s10278-013-9622-7.

Ghada Hamed, Mohammed Abd El-Rahman Marey, Safaa El-Sayed Amin, and Mohamed Fahmy Tolba. Deep Learning in Breast Cancer Detection and Classification. pages 322–333, Cham, 2020. Springer International Publishing. doi: 10.1007/978-3-030-44289-7\{_}30.

Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei,

Doug, Durgesh, and Francisco Ingham. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations, April 2021. URL `https://doi.org/10.5281/zenodo.4679653`.

Ins C. Moreira, Igor Amaral, Ins Domingues, Antnio Cardoso, Maria Joo Cardoso, and Jaime S. Cardoso. INbreast. pages 236–248. Elsevier BV, 2 2012. doi: 10.1016/j.acra.2011.09.014.

Lazaros Tsochatzidis, Lena Costaridou, and Ioannis Pratikakis. Deep Learning for Breast Cancer Diagnosis from MammogramsA Comparative Study. page 37. MDPI AG, 3 2019. doi: 10.3390/jimaging5030037.