Localization models trained on high-resolution breast X-ray images provide **accurate** and **interpretable** predictions for **breast cancer screening exam classification**

Improving localization-based approaches for breast cancer screening exam classification

Thibault Févry, Jason Phang, Nan Wu, S. Gene Kim, Linda Moy, Kyunghyun Cho, Krzysztof J. Geras

INTRO

- Breast cancer is the 2nd leading cause of cancer-related death in women.
- Can we use localization methods for accurate and interpretable predictions that assist radiologists?

METHODS

Base: Faster-RCNN with RoIAlign

- **Backbone**: ResNe(X)t (50/101) pretrained on ImageNet (**Figure 1**).
- **Fine-tuning**:
  - We use high-resolution images.
  - We use non-annotated images as negatives for the RPN and classifier.
  - We over-sample annotated images.
- **Inference**: Malignant prediction for a breast: max prediction per view (CC/MLO), then mean over views.
- **Hyperparameters and evaluation details**: see right column.

THE NYU DATASET

- 1m+ images, 200k+ exams.
- 5.8k exams from patients who had a biopsy annotated at pixel-level.
- 985 had malignant tumors, 5.6k benign and 234 both.
- AUC computed on malignant tumor detection over all test exams.

RESULTS

Model	AUC
Single (Wu *et al*, 2019)	0.886±0.003
Ensemble (Wu et al, 2019)	0.895
ResNet-50	0.891±0.005
ResNet-101	0.887±0.011
ResNeXt-101	0.908±0.014
Ensemble	0.919
Ours + Wu Ensemble	0.930

See more details in **Table 1**

See examples predictions **in Figure 2**

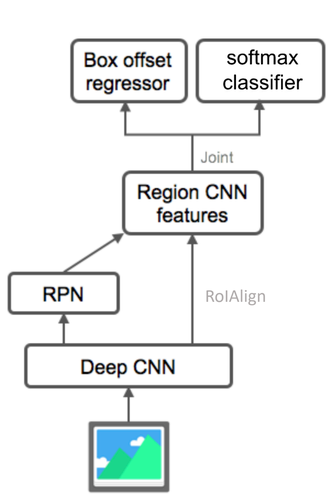

Fig. 1: Our architecture: Faster-RCNN with the RoIAlign pooling from Mask-RCNN (Diagram taken and modified from Lilian Weng)

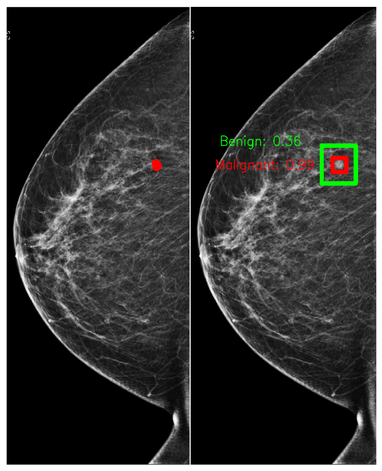

Fig. 2: Sample prediction from a model with the ResNeXT-101 backbone. This showcases the interpretability and high accuracy of our method. Note a benign tumor is also incorrectly predicted, albeit with a low probability

Changes to default Faster-RCNN parameters:

- **Resolution**: Our ResNet-50 backbone uses a resolution of 2200×3000 (1700×2700 for ResNet 101, 1300×2100 for ResNeXT 101). This was crucial for good performance and it is likely increasing it further would contribute. However, decreasing batch size to less than 4 (on 2 32GB GPUs) proved problematic for optimization.
- **Use of non-annotated images:** Images with no bounding-boxes are used as negatives for both the RPN patch selection classifier and the final classifier. We over-sample exams with annotations to help convergence
- **Intersection Over Union (IoU) thresholds**: To account for (i) noisier annotations and (ii) less overlapping in boxes, we decrease the IoU threshold of the RPN from 0.7 to 0.5 and of the final non-maximum suppression to 0.1.
- **Bounding box augmentation**: We resize the bounding boxes by a factor chosen uniformly in [0.8, 1.2] in our bounding-box scaling experiment.
- **Inference**: Our training objective (bounding box localization and classification) is disconnected from our final objective (breast-level tumor prediction). To obtain predictions, we take the maximum prediction over the boxes for each view, and then the mean over the views. To ensure we account for the tail behavior of the distribution which is important for AUC computation, we decrease the minimum score threshold at inference time from 0.05 to 0.001.

model	test set		reader study	
Wu et al. (2019a) single model	0.886	± 0.003	-	
Wu et al. (2019a) ensemble	0.895		0.876	
Base setup	0.891	± 0.005	0.845	± 0.007
+ biopsy ratio 0.75	0.895	± 0.004	0.855	± 0.012
+ biopsy ratio 1	0.887	± 0.011	0.855	± 0.013
+ bounding box scaling	0.890	± 0.006	0.841	± 0.010
+ classifier*5	0.903	± 0.007	0.859	± 0.013
+ bb scaling, classifier*5, ratio 0.75	0.888	± 0.006	0.855	± 0.006
+ recommended lr schedule	0.897	± 0.007	0.858	± 0.007
+ R-101 backbone	0.887	± 0.011	0.834	± 0.011
+ X-101 backbone	**0.908**	± 0.014	**0.866**	± 0.020
Ensemble	**0.919**		**0.879**	
Ensemble + Wu et al. (2019a) ensemble	**0.930**		**0.895**	

Table 1: Detailed results on the test set of our models and comparison with *Wu et al.*. Standard deviations computed on three runs based on taking the checkpoint with the best validation performance.

NYU

CAI²R Center for Advanced Imaging Innovation and Research

Download the full abstract!