GlimpseNet: Attentional Methods for Full-Image Mammogram Diagnosis

William Hang [*]
Stanford AI Lab
willhang@stanford.edu

Zihua Liu [*]
Stanford AI Lab
zliu19@stanford.edu

Awni Hannun [†]
Stanford AI Lab
awni@stanford.edu

Abstract

Cancer detection is an incredibly interesting problem because of its academic difficulty and importance. There are some, although rudimentary, attempts at using deep learning technology towards radiology imaging. Most of the attempts revolve around using the prelabeled regions of interest sectioned from a patients mammogram to generate predictions on whether the lesion under examination is benign or malignant. In this paper, we present GlimpseNet, along with other various techniques, for full-mammogram diagnosis. GlimpseNet in particular can autonomously extract multiple regions of interest, classify them, and then pool them to obtain a diagnosis for the full image. We obtain state of the art results, including a performance gain of 4.1% compared to previous methods.

1. Introduction

The gravity of the problem of breast cancer to men and women across the globe is clear. The World Cancer Research Fund [5] reported that 1.7 million cases of breast cancer were diagnosed in 2012, accounting for more than 25% of all cancers in women, and contributing to 12% of all new cancer cases that year. Thus, the precise and timely diagnosis of breast cancer is of paramount importance. The most common and least invasive method of breast cancer diagnosis is mammography, which, according to the American Cancer Society [19], reduces the rate of cancer death by 20 to 40%. What is not widely known is the variability in diagnosis. According to the Susan G. Komen Institute [11], mammography may miss anywhere from 16 to 30% of cancers. Furthermore, a study from Stanford University reports that mammogram misdiagnoses translates to a national healthcare cost of $4 billion annually [16]. Improving the accuracy rate of mammography presents itself as a dire and worthy problem.

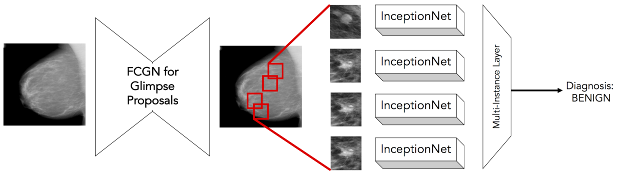


Figure 1. General Overview of GlimpseNet, composed of FCGN and the InceptionNet + MIL layer

Computer assisted diagnosis is a well-researched technique for improving the accuracy of mammograms and reducing human variability. Studies found that although such systems held incredible promise in earlier detection, they also contributed to high false positive rates [4]. Furthermore, many of the existing algorithms either rely on image preprocessing and classical feature extraction techniques to generate diagnoses, or only diagnose on pre-segmented Regions of Interest (ROIs), which are segments of a mammogram already known to contain an abnormality.

The objective of our research is to explore novel deep methods for breast cancer diagnosis on the entire mammogram. This is a significantly more difficult and useful problem because of the increasing promise of convolutional architectures in fine-grained image classification, as well as the fact that entire-mammogram diagnosis operates directly on raw output from the mammography. This means that no human intervention or segmentation is involved in the image to diagnosis pipeline, and that our proposed algorithm will act autonomously on unadulterated input. The techniques we explore are directly applicable to fine-grained image classification and attribute-assisted learning.

2. Related Work

We divide our discussion of related literature into two sections: Deep Methods, which exclusively employ Convolutional Neural Networks, and Feature-Based Approaches, which use supervised feature extraction later fed into a classifier.

[*] denotes equal contribution
[†] not enrolled in CS231N

2.1. Deep Methods

Geras et al. [13] developed multi-view deep convolutional neural networks for mammogram classification, where their approach was to classify on multiple views by feeding each view into a CNN. Each case was composed of four views, which resulted in four feature images. These feature images were flattened and summed, and then fed into fully connected layers for classification. Their primary finding was that classification at a high resolution was essential to accuracy. On their dataset, they did not achieve state of the art performance.

Levy et al. [3] trained existing networks to classify ROIs taken from DDSM. With data augmentation and transfer learning through initializing the weights of proven models like AlexNet or GoogLeNet, they achieved 92.9% accuracy. However, their work classified ROIs that were pre-segmented by humans, which leaves much to be desired on how the pre-segmented ROI is obtained in the first place.

2.2. Feature-Based Approaches

de la Rosa et al. [20] developed a feature based method using texture features such as Haralick, LBP, gray level histograms, and run-length, as well as the Fast Radial Symmetry Transform, which is a point of interest detector for microcalcifications. These features were fed into two algorithms: Citation k-NN, and mi-Grap, both of which are essentially nearest neighbor algorithms. What is notable about this paper is that it is one of the few we found in existing literature that reports performance on whole-mammogram classification on DDSM. They achieved 62.1% accuracy on the three-class classification task, which, as far as we know, is the state of the art.

3. Dataset and Features

In this study, we utilize the Digital Database for Screening Mammography (DDSM) [14][15], a collection of 2620 breast abnormality cases, where each case is composed of two views of each breast, totaling 10480 mammograms. Due to difficulties in obtaining the original dataset, we are using a curated version of DDSM maintained by The Cancer Imaging Archive (TCIA)[18][2], which contains 3800 images over 2000 patients. Each case is also associated with an attribute vector denoting external information about the abnormality and ultimate diagnosis. Each case contains either a BENIGN WITHOUT CALLBACK tumor, a BENIGN tumor, or a MALIGNANT tumor.

In our preliminary experiments, we resize each image to 512 × 512 pixels. In actuality, the images can reach 3000 × 5000 pixels. We split this dataset into a 90% train set and a 10% val set for early evaluation of our methods.

For GlimpseNet, which is a later model that we develop, we work with a smaller portion of DDSM that contains tumor segmentations. Details of the dataset are given in Section 5.

For brevity, examples of the original mammogram images we use for diagnosis are given in Figures 3 and 8, and examples of the tumor segmentation masks are given in Figure 7. We do not reproduce examples from the dataset in a figure here.

4. Preliminary Model Architectures

In previous works, Convolutional Neural Networks are among the most popular methods to approach image classification problems. As suggested in the previous section, extensive work had been done to perfect classifications on ImageNet, a benchmark dataset for image classification tasks. In the following sections, we will present modifications beyond popular methods employed for ImageNet and their corresponding results. These results will be compared to a simple Convolutional Neural Network as baseline and the state of the art performance on 3-way classification.

4.1. Baseline Model

As a baseline, the simple convolutional model resembles the classic VGG networks. The baseline network consists of 3 convolutional blocks with max-pooling in between. Each block contains 2 to 4 convolutional layers of 64 3x3 filters with ReLU activation. A 2-layered Multilayer Perceptron (MLP) Network is attached to the output of the convolutional blocks to create output prediction. The fully connected layers are 512 and 3 neurons in length each. Dropout is applied after each pooling layer and fulling connected layer. The exact architecture is as below.

Layer	Dimensions
conv_1_1	$3 \times 3 \times 1 \times 64$
conv_1_2	$3 \times 3 \times 64 \times 64$
pool_1	2×2
conv_2_1	$3 \times 3 \times 64 \times 64$
conv_2_2	$3 \times 3 \times 64 \times 64$
pool_2	2×2
conv_3_1	$3 \times 3 \times 64 \times 64$
conv_3_2	$3 \times 3 \times 64 \times 64$
conv_3_3	$3 \times 3 \times 64 \times 64$
conv_3_4	$3 \times 3 \times 64 \times 64$
fc_1	4096×512
fc_2	512×3

4.2. Dilated Convolutional Model

The DDSM dataset contains full resolution mammograms, each over 3000 × 4000 by dimension. Standard approaches rely on resizing the images to much lower resolution to ensure sane network training time and memory consumption. However, undoubtedly it sacrifices much of the image features. Thus, we applied dilated convolutions to this task in order to maintain higher resolution while keeping computation and relative receptive field size consistent.

A dilated convolution layer is a variation of regular convolution layer where we include spaces or dilations between each receptive field to increase receptive field size while keeping computation complexity the same size. Figure 2 demonstrates the effect of dilation of 2 on the receptive field.

4.3. Attribute Model

Besides raw images and segmentations of Region of Interest, the DDSM dataset also provides a set of distinct attributes for each of the mammograms in the dataset. These attributes include both quantitative features and qualitative features. Quantitative features includes continuous values such as age and breast density and number of anomalies in the mammogram. On the other hand, qualitative features contains observations for mass margins and mass shape for patients with mass symptom or calcification distribution and type for patients with calcification symptoms. These qualitative features can be further broken down into non-exclusive subcategories. For instance, for mass shape feature, there is up to 10 distinct subcategories such as ”Lobulated”, ”Round”, ”Amorphous”, and more. An example patient then might have ”Lobulated-Oval” as her mass shape feature.

Given these attribute features, we can use them to help provide additional information in our prediction. Surprisingly, these simple features contain strong predictive power. A simple 2-layered MLP Network with dropout achieves up to 81.1% accuracy on 3-way classification. However, we cannot use these features directly as input to our system at test time, for these features are labeled by expert radiologists. Including them as input defeats the purpose of this project.

As a result, we train our network to jointly reconstruct the attributes features and combining these attribute features with image features to make final predictions. The structure of our network is similar to the baseline network. Image features are extracted as the output of the 3 convolutional blocks. A 2-layered- MLP network is used to predict the attribute features $F_{reconstruct}$. At train time, we use the ground truth F_{truth} as attribute features; at test time, we use predicted attribute features instead.Attribute features and image features are concatenated as input to a final 2-layered MLP network for prediction. Loss in this model is composed of 2 parts: a classification cross-entropy loss and a reconstruction loss on the attributes. We use a L_1 loss for reconstruction because feature attribute is sparse given the qualitative features.

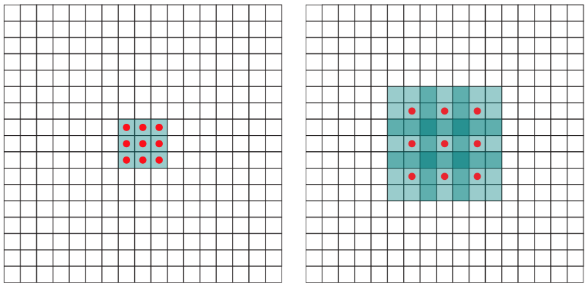


Figure 2. Regular Convolution v. 1-dilated Convolution

$$\mathcal{L}_c = -\frac{1}{N}\sum y_i log \hat{y}_i$$

$$\mathcal{L}_r = \frac{1}{N}||F_{reconstruct} - F_{truth}||_1$$

The final loss is the sum of the classification loss and weighted reconstruction loss.

$$\mathcal{L} = \mathcal{L}_c + \beta\mathcal{L}_r$$

4.4. Multi-view Model

We then established a baseline for classification utilizing multiple views of a case, because cases are composed of separate scans of a single patient, and these multiple scans contribute to the final diagnosis. To build this model, we extracted cases from the available DDSM dataset that are composed of two views. For each training example, we fed both views into our baseline model and extracted both feature images at the pool_2 layer. We then flatten and concatenate both feature images together, and pass this concatenated vector through the fully connected layers.

4.5. Coattentional Model

One of our hypotheses is that coattention between both views of the case will yield higher accuracy because the model can correlate both views together and attend over the most salient regions common to both views instead of view both in isolation.

In our coattentional model, we take the feature images I generated for one view and J generated for the other view from the pool_2 layer, and flatten both along the position dimension to obtain two $\mathbb{R}^{N\times D}$ vectors, where N is the total positions in the feature image and D is the feature dimension along each position. We then generate the coattention matrix:

$$\mathbf{C} = \text{softmax}(I\mathbf{W_c}J^{\intercal})$$

$\mathbf{C}$ can be interpreted as the bilinear covariance between I and J at each coordinate. We use a softmax nonlinearity to widen the distance between maxima and minima and scale the probabilities of the covariance matrix to sum to 1. We use this covariance matrix as an attentional vector to weight the contribution of image features at each position by:

$$I_{att} = \mathbf{C}I^{\intercal}$$

$$J_{att} = \mathbf{C}J$$

and summing I_{att} and J_{att} along the position dimension. These vectors are then concatenated and fed into the fully connected layers.

This coattention model derives inspiration from [8]

5. GlimpseNet

Our next experiment is the automated extraction of ROI bounding boxes from low resolution images. The hope is that these ROI bounding boxes yield the coordinates of tumors or image regions highly salient to accurate diagnosis. Our approach is to train an upstream model called Fully Convolutional Glimpse Network to generate salient region proposals, and glimpses are fed into a downstream model which classifies each glimpse and pools their probability distributions to generate a diagnosis. For this experiment, we used a section of the DDSM dataset where the ground truth of ROIs are provided. This contains 1318 mammograms from 691 patients, summing up to total of 1186 training images and 132 testing images.

As a note, we use the terms ROI, glimpse, and segmentation interchangeably.

5.1. Fully Convolutional Glimpse Network

We propose that extracting salient information from a noisy high resolution mammogram will enable more accurate diagnosis in the downstream model. To train a model to attend on important information (e.g. tumors, calcifications), we adapt Fully Convolutional Networks (FCN), a technique developed by [7], to our task. Fully Convolutional Networks were originally developed to predict which class each pixel in an image belonged to. By doing this, FCNs could perform pixel-wise segmentation of the original image.

We adapt the work of [9] towards applying FCNs to region proposals in images to extract relevant information. Our contribution in this respect is that instead of utilizing FCNs towards detecting discrete objects in an image, we use FCNs to extract relevant information that will lead to a final classification. Our version of the FCN is called a Fully Convolutional Glimpse Network, which convolves and aggressively pools an image until it becomes a one-dimensional hidden code. Transpose convolution and unpooling is applied to the code until it regains the same dimensions as the original input. FCGN is trained as an autoencoder to accept a low resolution mammogram and attempt to reconstruct a greyscale mask with each pixel value indicating its confidence of belonging to a tumor in the low resolution image. We can attempt to train the model this way because the DDSM dataset offers a 0/1 binary mask for each mammogram that indicates where tumors are located within the image. Examples of ground truth and predicted masks are provided in Figure 7. We thus train FCGN to reconstruct the tumor mask for the input image with a pixel-wise Mean Squared Error loss, and penalize heavily on pixels indicated as a tumor in the reference mask, but indicated as not in the prediction mask:

$$MSE = \frac{1}{nm}\sum_{i=0}^{n}\sum_{j=0}^{m}(\hat{I}_{ij} - \lambda I_{ij})^2$$

where n, m are the image dimensions, $\hat{I}, I$ are the predicted masks and reference masks, and λ is a penalty term.

The predicted mask is then processed with non-maximum suppression with IoU threshold set to 0.5 to recover the top non-overlapping regions predicted by FCGN. These regions are then cropped from the high-resolution image and sent to the downstream model. We adapt a code snippet from [6] to perform non-maximum suppression.

One contribution of the FCGN is that it resolves a conflict noted in [13] where high-resolution mammograms contain copious information, but are difficult to fit into memory and thus often need to be resized. The FCGN can recover high resolution regions of saliency while being tractable in memory, as it only performs convolution on the low resolution mammogram.

We are aware that Faster-RCNN [1] offers another approach to region proposals. However, we have chosen FCN methods for this stage of our work due to its high interpretability, because FCN methods yield a mask or attentional image that can be easily evaluated by the human eye and matched with the corresponding ground truth mask, whereas Faster-RCNN yields lists of coordinates.

5.2. Multi-Instance Learning Layer

The outputs of the Fully Convolutional Glimpse Network (FCGN) are then fed into the downstream classification model for each of the segmented region of interest to extract image features. For this case, we choose to use Inception V3 network as image feature extractor because of its relatively short inference time and memory consumption. Formally, we produce a bag of instances of image features $\mathbf{x} = \{x_1, x_2, ..., x_N\}$ where $x_n \in \mathrm{R}^{\mathrm{D}}$ by running each of the segmented ROIs from FCGN through the feature extractor network.

To combine these individual image features from various segmented ROIs, we employ a multi-instance learning (MIL) framework proposed in Kraus et al[12]. In their MIL framework, each class i of 3 can be treated separately as a binary classification task with label $t_i \in \{0, 1\}$. A probability $p_{ij} = P(t_i = 1|x_j)$ is produced for each instance in x_j and each class label t_i. The final probability of the set of in-

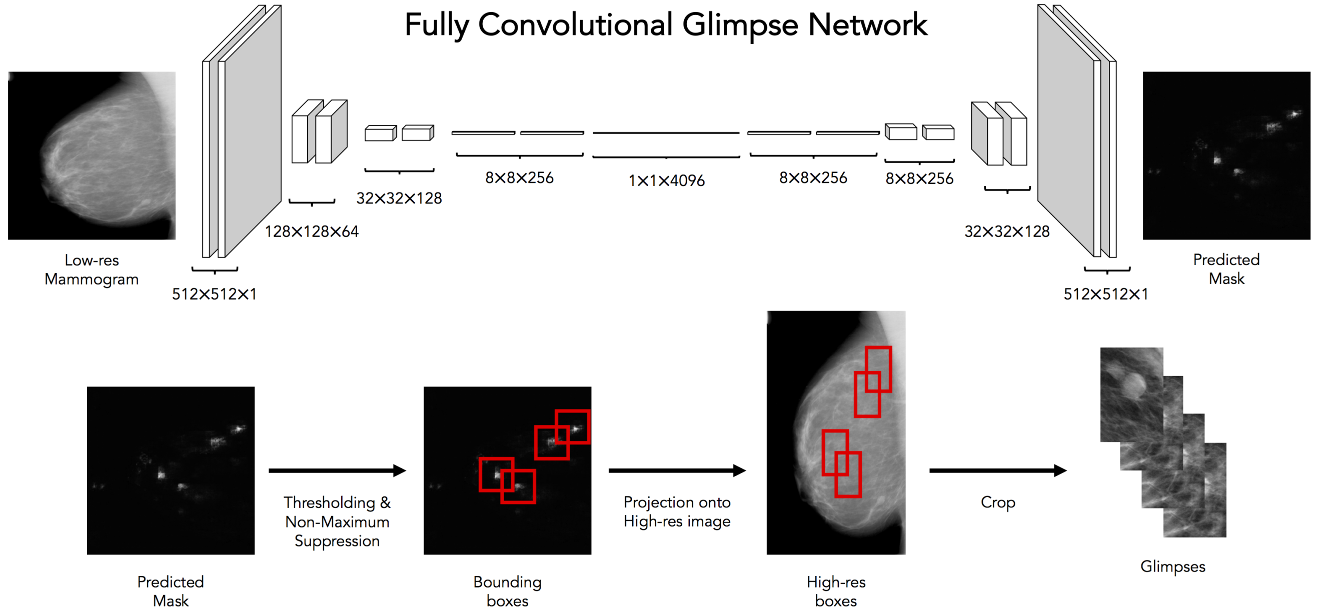


Figure 3. Fully Convolutional Glimpse Net accepts a low resolution image and determines the most tumor-like regions. Thresholding and non-maximum suppression are applied afterwards to generate the most relevant image crops.

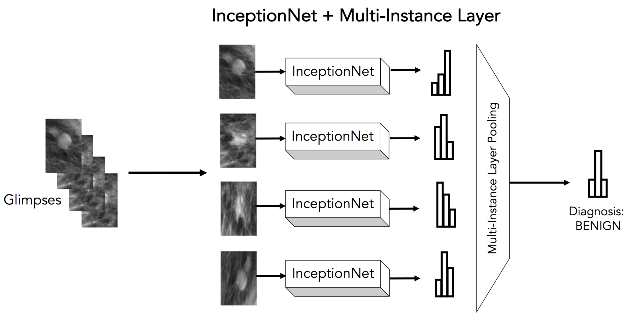


Figure 4. Downstream Classification network with Multi-Instance Pooling layer

stances for a particular class i $P_i = P(t_i = 1|x_1, x_2, ...x_N)$ is computed by applying a pooling function $f(\cdot)$, a MIL layer, on individual probabilities of each instance on class i. The loss for training is computed by a cross entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i}(logP(t_i|P_i))$$

where $P(t_i|P_i)$ is the binary classification prediction from the MIL layer. $P(t_i|P_i) = P_i^{t_i}(1-P_i)^{(1-t_i)}$. The purpose for such a MIL layer is to combine the probability distribution of class generated from each glimpse from FCGN into a full probability distribution for the entire image. For this project, we experimented on three different MIL layer: Top k, Log-Sum-Exponential (LSE), and Noisy-AND (NAND). The principle for each of these is consistent: if even only one glimpse is highly activated in the MALIGNANT category, then we should output the entire image as MALIGNANT.

For LSE pooling as shown in equation (1), r is a hyperparameter. LSE can be seen as a smooth approximation to maximum. The value of r dictates the sharpness of the approximation: as r increases, LSE gets closer to the maximum of the instances. LSE was first proposed by Ramon and Readt in 2000 to approximate max operation [17]

For NAND pooling as shown in equation (2), a and b are hyperparameters. NAND tries to utilize our assumption

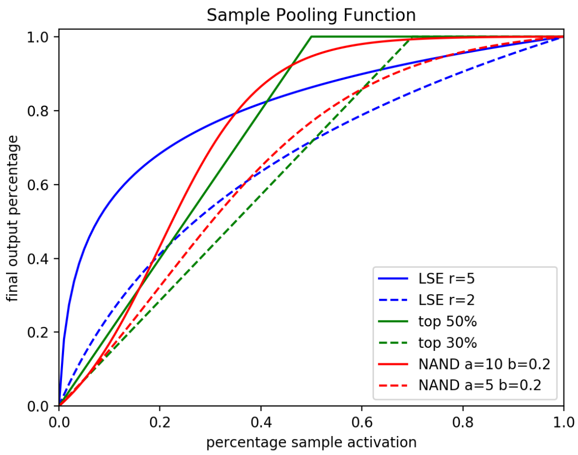


Figure 5. Pooling Function behavior against proportion of activated instances

that if the number of positive instances exceeds a certain threshold (in this case 1), the entire set of instances become positive. Thus equation (2) arises where $p_{i\bar{j}} = \frac{1}{|j|}\sum_j p_{ij}$ and σ is the sigmoid function

For Top-k pooling, we average the probability distribution for k number of instances that are most highly activated in the MALIGNANT class. A sample behavior of all three pooling function against the proportion of activated instances is shown in Figure 5.

$$P_i = f(\{p_{ij}\}) = \frac{1}{r}log(\frac{1}{|j|}\sum_j e^{rp_{ij}}) \tag{1}$$

$$P_i = f(\{p_{ij}\}) = \frac{\sigma(a(p_{i\bar{j}} - b)) - \sigma(-ab)}{\sigma(a(1-b)) - \sigma(-ab)} \tag{2}$$

6. Experiments and Results

In this section, we will present experimental results for all three different methods in both supervised learning stage and reinforcement learning stage. All of our experiments are trained on NVIDIA GeForce GTX TITAN GPUs. In this section, we will present experimental results for each of the methods we presented in the previous sections as well as visualization of the training process.

6.1. Training

In most proposed models an initial learning rate of 4×10^{-4} is used to initialize training. Adam optimizer [10] is used for the full duration of 150 epochs or when the training is seen to converge. For each of the mentioned task, input data is preprocessed by first reshaping to lower resolution of 512×512 and then normalized by reducing channel mean of the images. No data augmentation or variance normalization is applied. 40% dropout is applied to pooling and fully connected layers along with weight decay with coefficient $\lambda = 0.01$ is applied as regularization to all proposed models. Other hyperparameters for models include $\beta = 0.1$ for Attribute Model. All models except for GlimpseNet can be trained within 10 hours and takes up to 5 to 6GB of GPU memory with batch size of 8. As for GlimpseNet, we divide our discussion on training into two parts.

Method	3-class accuracy
Baseline Model	0.490
Dilated Convolution Model	0.522
Multiview Model	0.620
de la Rosa et al.	0.621
Attribute Model, $\beta = 0.1$	0.645
Coattention Model	0.647
GlimpseNet NAND (a=10, b=0.2)	0.589
GlimpseNet topk (k=3)	0.571
GlimpseNet LSE (r=5)	**0.662**

Table 1. Results on 3-way classification on DDSM dataset

FCGN used a learning rate of 10^{-3} on the Adam optimizer. Training was halted after 30 epochs and takes approximately 4 hours on a 4 GB GPU. Each convolution and deconvolution layer was followed by batch normalization with scale 1 and shift 0.01, and with ReLU. Our penalty term in the MSE loss for FCGN is 100 to heavily penalize predicting a tumor pixel as irrelevant.

InceptionNet + MIL takes up to 1 full day to train and takes up to 9 to 10 GB of GPU memories with batch size of 8. Only Inception V3 network of the downstream process of GlimpseNet is initialized with learning rate of 10^{-4} with pretrained weights. This initial learning rate is reduced by a factor of 0.5 every 30 epochs. Hyperparameters include $a = 10$, $b = 0.2$ for MIL with NAND pooling, $r = 5$ for MIL with LSE pooling, and $k = 3$ for MIL with TopK pooling.

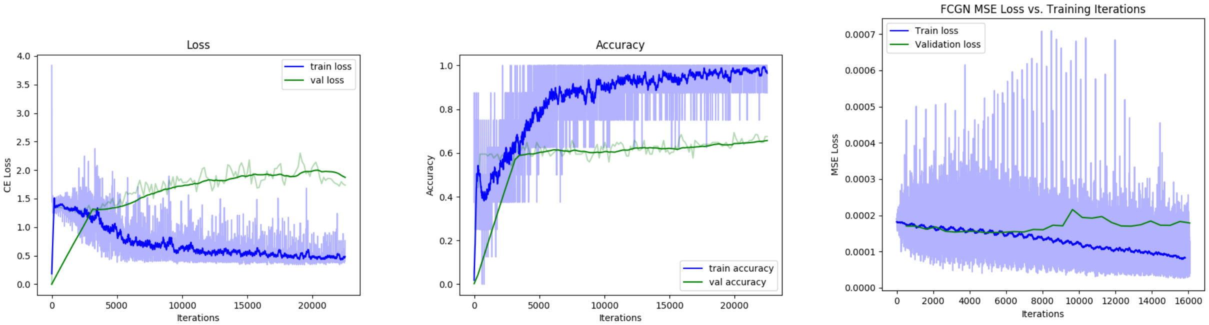
6.2. Model Evaluation

All aforementioned models are evaluated in the task of 3-class diagnosis prediction as suggested in section 3. Our models are compared against work by de la Rosa et al [20], which serve as the state of the art for this study. de la Rosa et al. achieves a 62.1% accuracy on this task. Table 1 shows an ensemble of prediction accuracy for our models in order of increasing accuracy. We acquire these reported accuracy by evaluating the proposed models on validation images. Among all of the model proposed, Attribute Model, Coattention Model and GlimpseNet with LSE pooling was able to surpass the accuracy of de la Rosa et al. by 2.4%, 2.6%, and 4.1% respectively without aggressive hyperparameter tuning or any model ensemble, proving the strength of our proposed models.

6.3. Discussion

To understand the behavior of GlimpseNet, training graphs for both parts of InceptionNet + MIL are provided in Figure 6. To interpret the graph, the translucent blue lines denote the actual training error per iteration. Dark blue lines denote a sliding average of 20 iterations. The same notation is used for validation accuracy and loss. One reason for the very uneven loss and accuracy curve can be attributed to the small batch size used in training. Batch size of 8 is used due to memory constraints and thus contributes greatly to the high variance observed.

The FCGN train/test curve is unusual. The train loss monotonically decreases as is expected, but the test loss spikes in the middle. We attribute this to overfitting as the model trains throughout time, as the test loss does decrease until around 7000 iterations. Furthermore, with a train/test split of 10, our test set is around 131 images, so a single image in the test set that contains more pixels that is not



(a) Classification Loss

(b) Classification Accuracy

(c) FCGN Loss

Figure 6. Loss and Accuracy curve for training of GlimpseNet; includes both upstream FCGN and downstream classification network

(a) Successful Predicted Mask

(b) Unsuccessful Predicted Mask

Figure 7. Examples of successful and unsuccessful mask predictions. The ground truth segmentation mask is on the left of each subfigure, and the predicted segmentation mask is on the right.

predicted properly will contribute a lot more to the overall loss for that epoch. Since the scale is on the order of 10^{-4}, the loss per validation epoch could be sensitive.

We also provide qualitative results for FCGN on some of the predicted masks and image crops that are generated. Such qualitative results are easy to interpret for FCGN because we are provided with a corresponding ground truth mask, but qualitative results for the downstream model are not helpful to our discussion because they require us to attempt to interpret a mammogram. Thus, we do not include them here. Figure 7 shows several results from the FCGN.

We suspect that overall loss in the FCGN might be attributed to overfitting on the train set and incorrectly proposing glimpses to the downstream model, or including a large amount of noncancerous glimpses in addition to cancerous glimpses that might skew the MIL Layer towards incorrect diagnoses.

We will perform further ablation studies where we feed in the segmented ROI images directly into InceptionNet + MIL, bypassing FCGN, to determine the loss contributed by the FCGN.

For the down stream classification network with MIL layer, we observe that training loss quickly drops and training accuracy increases within the first 7,000 iterations. From then on the loss and accuracy slowly buy consistently decreases and increases respectively. In addition, the model almost overfit completely after 15,000 training iterations with close to 100% training accuracy and thus exhibiting a plateauing loss. At the end, we are able to achieve a maximum of 66.18% prediction accuracy

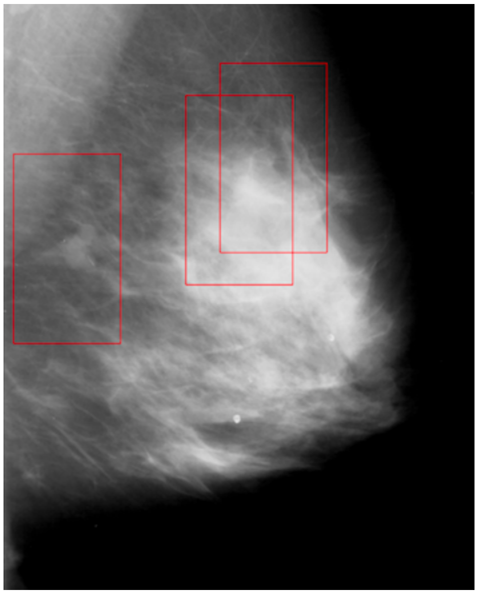


Figure 8. Example of raw scaled bounding boxes drawn over the high-resolution image.

7. Conclusions

We have demonstrated a body of work comparing different methods, some existing, some novel, to full-image breast cancer diagnosis. Previous methods such as dilated convolutions or multiview models are not as effective on the fine-grained image recognition task, whereas attribute models and coattention models either incorporate novel information about the images, or enable correlations to be made between different views, ultimately contributing to better diagnoses.

The model that is superior in our work is GlimpseNet, which we believe to outperform existing methods due to its hard attentional mechanism. FCGN in Glimpsenet is able to attend on regions of the mammogram that are likely to contain tumors. FCGN is not limited to attending on only one region; it can attend on multiple regions that it also thinks belong to tumors. Thus, FCGN is able to recognize regions that are either tumors or contribute to tumor classifications. The InceptionNet + MIL Layer can thus utilize multiple high-resolution glimpses to arrive at a more accurate diagnosis by pooling multiple probability distributions generated by InceptionNet on each glimpse. This pooling allows GlimpseNet to combine multiple diagnoses into a final diagnosis for the full image.

We have demonstrated GlimpseNet to outperform the state of the art by 4.1%, a large gain in full-image mammogram diagnosis, and in a field where much of the state of the art relies on pre-segmented tumor ROIs. We hope this work can be applied towards fine-grained image recognition in other tasks.

8. Future Work

8.1. Data Augmentation

One major challenge for using the DDSM dataset compared to traditional dataset such as ImageNet is that DDSM contains only a very small amount of data. Previous work on this exact dataset had shown that augmenting the dataset helps with the network performance. In addition, since masses does not have a particular spatial orientation, classic augmentation techniques such as rotation or reflection does not change the underlying pathology of the masses. Thus, a sane next step is augmenting the dataset and observe the effects.

8.2. Fully Differentiable Model

Another area of future work lies in devising end-to-end trainable network. Although the model is end-to-end runnable, as of now, GlimpseNet cannot be end-to-end trained. The biggest bottle neck lies in the step between the upstream FCGN and the downstream Inception V3 network with MIL layer. The reason is that to extract the ROI instance required fort the downstream processes, an argmax operation needs to be performed in order to crop out the relevant region of interests. The motivation for an end-to-end trainable network is that we want error information to flow from our end objective, the three way classification, to our input, the full mammogram image. With the currently segmented training objective, the upstream FCGN can only learn to detect regions of interests labeled by radiologist instead of actual patches that aids downstream prediction.

There are a few ways to combat this problem: first, instead of applying hard attention on the original image, we can use soft attention on a resized image and feed an altered original image to the down stream classification network. An alternative to this approach is use reinforcement learning to perform region proposals. The reward for cropping can be simply the cross-entropy error from the final classification.

References

[1] Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS 2015*, 2015.

[2] S. K. F. J. K. J. K. P. M. S. P. S. M. D. P. M. T. L. P. F. Clark K, Vendt B. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013.

[3] A. J. Daniel Levy. Breast mass classification from mammograms using deep convolutional neural networks. *Neural Information Processing Systems*, 30, 2016.

[4] E. J. Fenton JJ, Xing G. Short-term outcomes of screening mammography using computer-aided detection: A population-based study of medicare enrollees. *Ann Intern Med.*, 158(8):580–587, 2013.

[5] W. C. R. Fund. Breast cancer statistics, 2015. http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics.

[6] R. Girshick. https://github.com/rbgirshick/py-faster-rcnn.

[7] B. H. Hyeonwoo Noh, Seunghoon Hong. Learning deconvolution network for semantic segmentation. *ICCV*, 2015.

[8] D. B. D. P. Jiasen Lu, Jianwei Yang. Hierarchical question-image co-attention for visual question answering. *Neural Information Processing Systems*, 30, 2016.

[9] K. H. J. S. Jifeng Dai, Yi Li. R-fcn: Object detection via region-based fully convolutional networks. *NIPS*, 2016.

[10] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] S. G. Komen. Accuracy of mammograms, 2016. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2015-2016.pdf.

[12] O. Z. Kraus, J. L. Ba, and B. J. Frey. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):52–59, 2016.

[13] S. G. K. L. M. K. C. Krzysztof J. Geras, Stacey Wolfson. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv*, 2017.

[14] D. K. R. M. Michael Heath, Kevin Bowyer and W. P. Kegelmeyer. The digital database for screening mammography. *Proceedings of the Fifth International Workshop on Digital Mammography*, pages 212–218, 2001.

[15] D. K. W. P. K. R. M. K. C. Michael Heath, Kevin Bowyer and S. MunishKumaran. Current status of the digital database for screening mammography. *Proceedings of the Fourth International Workshop on Digital Mammography*, pages 457–460, 1998.

[16] M.-S. Ong and K. D. Mandl. National expenditure for false-positive mammograms and breast cancer overdiagnoses estimated at $4 billion a year. *Health Affairs*, 34(4):576–583, 2015.

[17] J. Ramon and L. De Raedt. Multi instance neural networks. In *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*, pages 53–60, 2000.

[18] A. H. D. R. Rebecca Sawyer Lee, Francisco Gimenez. Cbis-ddsm, 2016. Curated Breast Imaging Subset of DDSM. The Cancer Imaging Archive.

[19] A. C. Society. Breast cancer facts & figures 2015-2016, 2016. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2015-2016.pdf.

[20] C. G. C. G. C. M. Q. G. Snchez de la Rosa R, Lamard M. Multiple-instance learning for breast cancer detection in mammograms. *Engineering in Medicine and Biology Society*, 37, 2015.