Building a Breast-Sentence Dataset: Its Usefulness for Computer-Aided Diagnosis

Hyebin Lee[1], Seong Tae Kim[2], and Yong Man Ro*[1]

[1]Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea
{machipoe,ymro}@kaist.ac.kr
[2]Computer Aided Medical Procedures, Technical University of Munich, Germany
seongtae.kim@tum.de

Abstract

In recent years, it is verified that the deep learning network is able to process not only images but also time-series information. Since breast image analysis plays a big role in the diagnosis of breast cancer, there have been a large number of attempts to apply the deep learning method for an accurate diagnosis. With the advance of deep learning approaches, the possibility of using medical reports (in natural language) has been increased. However, there is no public medical report dataset associated with the breast image. Instead, in the conventional public breast mammography datasets, the characteristics of breast cancer are annotated according to the standardized term (Breast Imaging-Reporting and Data System). In this study, a breast-sentence dataset is proposed to investigate its usefulness in computer-aided diagnosis.[1] *Based on the conventional breast mammography datasets, we annotated sentences in the natural language according to the standardized terms (defined in Breast Imaging-Reporting and Data System) in conventional breast mammography datasets. In the experiments, we show three use cases to verify the usefulness of the breast-sentence dataset: 1) CAD framework with radiologist's input, 2) the use of sentence dataset in training a CAD, and 3) visual pointing guided by sentence.*

*Corresponding Author

[1]The breast-sentence dataset based on the public breast mammography dataset is released in the form of caseID-sentence pair. (http://ivylabdb.kaist.ac.kr/base/dataset/breast_sentence.php)

This work was partially supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2017-0-01779, A machine learning and statistical inference framework for explainable artificial intelligence).
This work was partially supported by KAIST grant funded by Ministry of Science and ICT.

1. Introduction

Since the deep learning mechanism has shown outstanding performance in the image classification, it has applied to various types of computer vision tasks. Furthermore, in diverse studies, it has been demonstrated that deep learning approaches are suitable for analyzing not only images but also data with diverse semantics such as natural language. In recent years, large datasets such as MS-COCO [2], Flickr 30K [17] which contain numerous images and corresponding captions have been publicly accessible. The deep learning based recurrent neural network (RNN) module such as gated recurrent unit (GRU) [3] also developed. Based on these previous research, it became possible for the joint learning of the model, which learns images and other types of data together such as image captioning.

In the medical area, image analysis is widely used for diagnosis and treatment. However, expert knowledge of radiologists is needed to make a diagnosis using a medical image, and each radiologist's proficiency is different. Therefore it is required to develop computer-aided diagnosis (CAD) systems. The deep learning approaches have also got successful results in CAD research and shown remarkable performance. In addition, there are plenty of studies which use medical image and associated report or annotation simultaneously [20, 28, 18, 29, 27]. [20] utilize CNN-RNN based structure which is widely used for image captioning to generate annotation of chest X-ray image in the form of a series of terms including information about disease, organ, and location. [28] proposed network architecture, which generated medical reports and classification results by overcoming the limitation of small chest X-ray image and report dataset. [18] reported enhanced accuracy of diagnostic prediction by training classifier to predict semantic information in the medical reports. [29, 27] aimed to generate an artificial report similar to the report made

by the radiologist and improved accuracy of the diagnostic decision. They added the attention module to the existing CNN-RNN model to make spatial attention map pointing specific regions correlated with diagnostic prediction.

For the breast image analysis such as mammography, ultrasound and MRI, Breast Imaging-Reporting and Data System (BI-RADS) [13] have been widely used. It contains six categories to standardize diagnosis in terms of cancer risk. Additionally, it also contains commonly accepted lexicons among the radiologists. Each lexicon describes the appearance of findings in medical image and is closely correlated with malignancy.

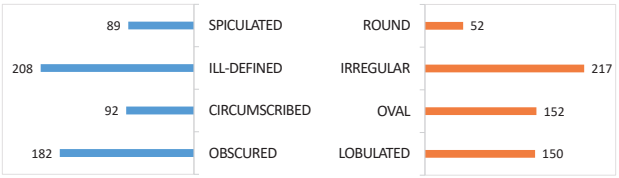
Aforementioned, additional information such as medical reports can be used to convey prior knowledge to the network explicitly and implicitly, which is effective for enhancing network performance. However, due to privacy issues related to patients, there are few medical reports datasets open to the public. In the case of breast images, mammography datasets such as the Digital Database for Screening Mammography (DDSM) dataset [8] and the mini-Mammographic Image Analysis Society (mini-MIAS) dataset [22] are publicly accessible, whereas there is no published medical report data even though there is standard reporting guideline such as BI-RADS.

In this paper, we propose a new sentence dataset which contains information of margin and shape included BI-RADS lexicons portraying the breast mass. In order to give variation to sentences, we collect word and phrase called *visual word* defining and explaining margin and shape lexicon and compose sentence based on mammography with its BI-RADS annotation. All sentences in the dataset are unique. In particular, we demonstrate the efficiency of this dataset in the case of three different situations where the CAD system is used. In each situation, the proposed dataset provides the semantic information which can increase the performance of the diagnostic system or the meaningful visual justification to explain output decision. To embed our sentence data into a vector and fuse with the visual feature, we utilize Bidirectional Encoder Representations from Transformers (BERT) [6] which shows state-of-the-art performance on various natural language processing task.

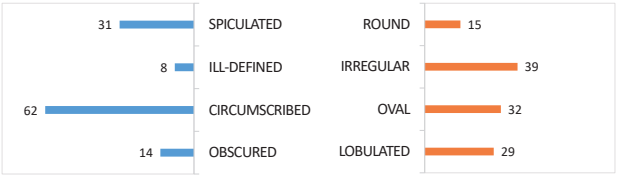
The rest of the paper is organized as follows. In the Section 2, we introduce the related works. In Section 3, our processes to compose the sentence in the proposed datasets (*i.e*. DDSM-sentence dataset and FFDM-sentence dataset) are described in detail. Next, in Section 4, we introduce three useful cases of the proposed dataset with the benchmark results. Finally, Section 5 concludes the paper.

2. Related works

Mammography is known as a standard tool for detecting breast cancer in screening. A lot of research efforts have been devoted to developing CAD system on breast



(a) The DDSM dataset



(b) The FFDM dataset

Figure 1. BI-RADS lexicon distribution of the DDSM dataset and the FFDM dataset.

mammographic images. Two public mammographic image datasets are widely used, the DDSM dataset and the mini-MIAS dataset. The DDSM dataset contains approximately 2,500 studies which include two images of each breast. The associated information (*i.e*. age at the time of the study, ACR breast density rating, subtlety rating for abnormalities, ACR keyword description of abnormalities) with studies also provided. The mini-MIAS dataset contains 322 mass images from digitized films and annotations associated with each image such as the character of background tissue and class of abnormality present. These two datasets provide not only types of suspicious regions but also addictive annotations. However, there is no dataset containing annotation in natural language, such as a medical report in breast mammography. In some medical imaging area such as chest X-rays, many studies have utilized medical report in a standard format to generate a report automatically and enhance the interpretability of the diagnosis system. A prerequisite for many of the studies to deal with medical report with the medical imaging is the existence of a publicly accessible report dataset. In the case of chest X-rays, [4] proposed chest X-ray dataset and reports through Open Access Biomedical Image Search Engine (OpenI), containing 7,470 chest x-rays with 3,955 radiology reports. Each report is divided into subsection such as indication and finding.

3. Construction of the senctence datasets

3.1. Data collection

For constructing sentence dataset, we selected two mammogram datasets. Firstly, the DDSM dataset which is fully available for mammographic image analysis was selected. It contains breast X-ray mammographic images with malignancy annotations and information on the BI-RADS lex-

OBSCURED	CIRCUMSCRIBED	ILL-DEFINED	SPICULATED
hidden	well-defined	Indistinct	characterized by lines
hidden by superimposed or adjacent normal tissue	sharply-defined	poor definition of the margins	lines radiating
cannot be assessed	sharply demarcated	may be infiltration	Radiating lines
be masked by the adjacent gland	abrupt transition between the lesion and the surrounding tissue	no clear demarcation between a mass and its surrounding tissue	stellar
	well marginated	uneven margin	short peripheral spicules
		ambiguous boundary	sharp lines projecting from the mass
		vague	stelliform
		not evident	hairlike projections radiating away from the lesion
		blurred	a small slender pointed structure
			lines are scattered throughout
			outgrowth
			longish structure

Table 1. Collected visual words explaining margin lexicons of BI-RADS.

LOBULATED	OVAL	IRREGULAR	ROUND
undulated contour	elliptical	cannot be characterized	spherical
wavy contour	egg-shaped	neither round nor oval	ball-shaped
concave borders and convex borders	convex borders	complex shape	circular
small lobes can be seen	ovoid	not uniform	globular
bulging and sunken	including two or three gentle undulations	indefinite	convex borders
curly contour		uneven	
winding		non-uniform	
bumpy		complicated	
embossed		intricate	
scalloped parts			
chunks on its body			

Table 2. Collected visual words explaining shape lexicons of BI-RADS.

icon. To make our DDSM-sentence dataset, the images from scanner HOWTEK were selected, and 571 mass ROI images were made by cropping the masses with surrounding parts from the whole breast image based on the annotation in the original DDSM dataset. Secondly, a clinical Full-Field Digital Mammogram (FFDM dataset) containing mammographic images of 67 patients collected from the hospital was also utilized. 115 mass ROI images of the FFDM dataset were collected. Images labeled as *microlobulated* in both datasets were excluded since the number of that image is insufficient. The margin and shape distributions of ROI images collected for the DDSM-sentence dataset and the FFDM-sentence dataset are shown in the Figure 1.

3.2. Sentence annotation

BI-RADS is introduced to standardize the reading of mammography and to avoid confusion in the interpretation of mammography. There have been a number of papers which report predictive values of mammographic features specified in BI-RADS [16, 12, 24, 7]. Therefore it is recommended for radiologists to use BI-RADS terminology to write a report. Likewise, we composed the sentence based on BI-RADS margin and shape information, in order to objectively annotate the sentence of the dataset. To elaborately describe the breast masses, we utilized words and phrases describing the lexicons called *visual words*. The visual words were collected from existing medical papers and books [15, 10, 14, 25, 19, 11, 1, 23], which have words or phrases explaining BI-RADS lexicon. There were 4-12 visual words corresponding to each lexicon shown as Table 1 and Table 2. We made the sentences to describe mass ROI images using combinations of visual words and their synonyms. Since visual word contains various words which are widely used to describe the state or shape of objects in common sentence, a natural language processing model trained with large corpus can embed the proposed sentences more abundantly and easily through this process. According to [9], each sentence has at least 10 words and does not contain BI-RADS lexicon as it is. Every sentence necessarily contains visual words corresponding to the margin and shape each, and most sentences also include phrases describing the individual appearance of each mass ROI image. We made three sentences per a mass ROI image and every sentence is unique. In other words, the DDSM-sentence contains 1,713 sentences and the FFDM-sentence contains 345 sentences. The sample sentences of the DDSM-sentence and associated mass ROI image is shown in Figure 2.

4. Experiments

To provide useful use cases of the proposed dataset, we have conducted comprehensive experiments. Our experiments consist of 1) CAD framework with radiologist's input, 2) the use of sentence dataset in training a CAD, and 3) visual pointing guided by sentence. The main purpose of these experiments is to validate the usefulness of the proposed breast sentence datasets in the CAD use cases.

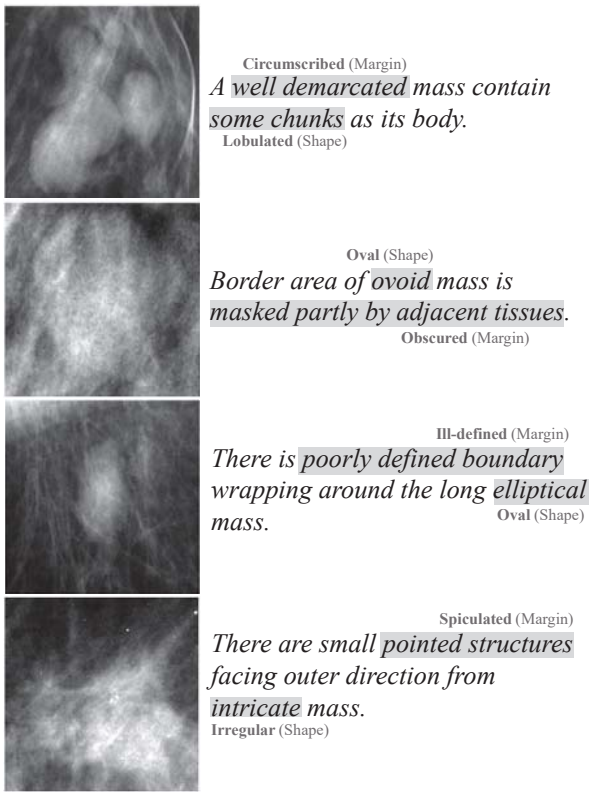


Figure 2. Examples of the DDSM-sentence.

4.1. Sentence vector encoding using BERT

To use the sentence information in our medical image analysis, we firstly encoded sentences using BERT. In other words, the sentence was embedded as a vector in order to make the CAD network utilize the knowledge which the sentence has. The BERT model was used for effective embedding since it showed high performance for a variety of natural language processing tasks. BERT uses a single pre-trained model and is fine-tuned for specific tasks. To focus on the information on the margin and shape of the mass in the sentence, we fine-tuned the BERT model to classify the margin and shape of the sentences. We used BERT-base uncased version model with whole word masking, pre-trained by English Wikipedia (2,500M words) and BooksCorpus [30]. We fine-tuned the model with 1,365 sentences in the DDSM-dataset. All hyperparameter settings were same as that of the pre-trained BERT-base model in [6] except batch size and learning rate. Batch size and learning rate were 256 and 0.00002. In fine-tuning stage, we validated the classification accuracy of the BERT sentence classification model

with 348 sentences in the DDSM-sentence, which were not overlapped with sentences used in fine-tuning stage. We chose the fine-tuned model which showed the highest classification performance in validation. The classification accuracy of the fine-tuned BERT model was 90.49% for the margin and 96.54% for the shape. In the following experiments, we used the special classification token [6] in the fine-tined BERT model with fixed parameters as sentence embedded vector $\mathbf{f}_{\mathrm{sent}} \in \mathbb{R}^{768}$.

4.2. Experiment 1: CAD framework with radiologist's input

In this experiment, we assumed the situation where the CAD network was used for diagnosis and a sentence or margin/shape label associated with the input mass ROI image was provided by radiologists. We considered sentence annotation associated with mass ROI image in the test dataset as provided radiologist's report.

4.2.1 Network architecture

We compared three network models which were 1) trained with only the mass ROI image (*i.e*. baseline), 2) trained with the mass ROI image and margin and shape one-hot vector (*i.e*. one-hot), and 3) trained with the mass ROI image and sentence embedded vector (*i.e*. sentences).

We used VGG16 [21] for benign/malignant diagnosis as baseline model. Its initial weights were set from pre-trained model with ImageNet [5]. For the situation where a sentence is also provided as input of the network, we modify the fully connected layers in baseline model. In VGG16, the mass ROI image feature $\mathbf{f}_{\mathrm{mass}} \in \mathbb{R}^{m \times n \times k}$ after conv 5_3 layer passes through three fully-connected layers $(\mathrm{F}_1, \mathrm{F}_2, \mathrm{F}_3)$ and becomes the diagnostic prediction vector $\hat{\mathbf{y}}_{\mathrm{diag}} \in \mathbb{R}^2$. In the middle of this process, $\mathbf{f}_{\mathrm{mass}}$ combines with $\mathbf{f}_{\mathrm{sent}} \in \mathbb{R}^{768}$ as

$$\hat{\mathbf{y}}_{\mathrm{diag}} = \mathrm{F}_3(\mathrm{F}_2(\mathrm{F}_1(\mathbf{f}_{\mathrm{mass}})) \oplus \mathbf{f}_{\mathrm{sent}}), \tag{1}$$

where $\oplus$ represents vector concatenation and $\hat{\mathbf{y}}_{\mathrm{diag}}$ denotes diagnostic prediction vector. For the situation where a margin and shape labels are provided with the mass ROI image, we used 6 dimensional concatenated vector of one-hot vectors $\{\mathbf{f}_{\mathrm{margin}}, \mathbf{f}_{\mathrm{shape}}\}$ in place of $\mathbf{f}_{\mathrm{sent}}$.

4.2.2 Experimental setting

For the experiment, the DDSM dataset was split into training set (455 images) and test set (116 images). In the case of the FFDM dataset, two-fold cross-validation was conducted. Since the number of image in the FFDM dataset is small, the pre-trained model with the DDSM dataset was used as an initial model for training with the FFDM dataset. According to [26], three margin types were considered in the experiment (circumscribed, speculated, ill-defined-obscured). *Obscured* and *ill-defined* were merged into a single class and as we mentioned, *microlobulated* was excluded because of the small amount. Referring to [15], the *round* and *oval* shapes belong to BI-RADS 3 both. Therefore the shape classes in this paper were divided into three classes (round-oval, irregular, and lobulated).

We conducted data augmentation to increase the number of training data. The two sizes of patches were cropped from the original image at five locations (top left, top right, center, bottom left, bottom right). Flip and rotation (0°, 90°, 180°, and 270°) were also used. The size of mini-batch was set to 64 and an Adam optimizer was used with learning rate 0.0001 in the case of the DDSM dataset and 0.00005 in the case of the FFDM dataset.

	DDSM	FFDM
Baseline	0.962	0.924
One-hot	0.945	0.911
Sentences	**0.975**	**0.966**

Table 3. Results of experiment 1 (AUC).

	DDSM	FFDM
Baseline	0.956	0.851
Multitasking	0.947	0.845
Sentences	**0.962**	**0.856**

Table 4. Results of experiment 2 (AUC).

4.2.3 Results

As shown in Table 3, the information in the sentence has positive value on increasing performance of the CAD network. We also confirmed that margin and shape information, which are closely related to the diagnosis result, could be delivered more effectively than simply using a one-hot vector. Besides, the network trained with the FFDM dataset, which has an extremely small number of image, and sentence embedded vector showed higher performance improvement rate than that of the DDSM dataset. Since the margin and shape information which is important for diagnosis were directly input to the CAD network, this performance can be considered as the upper bound of performance utilizing sentence information.

4.3. Experiment 2: The use of sentence dataset in training a CAD

In this experiment, we assumed the more general situation where we could not get the sentence and get only the mass ROI image for input to the CAD network.

4.3.1 Network architecture

We compared three network models which were 1) trained with only the mass ROI image (*i.e.* baseline), 2) constructed for multi-task classification and trained with the mass ROI image and margin and shape one-hot vector (*i.e.* multitasking), and 3) trained with the mass ROI image and sentence embedded vector (*i.e.* sentences).

We constructed the baseline network base on VGG16 initialized with weights of the network pre-trained with ImageNet. We did not use the sentence directly as an input in training stage because we cannot use the sentence corresponding to the ROI image in the inference process of this case. Instead, we adjusted the output vector dimension of the fully-connected layer in the baseline model and the sentence model so that the sentence embedded vector directly guides the intermediate feature during training stage. We modified second fully-connected layer F_2 to make intermediate feature $\mathbf{f}_{\mathrm{imd}} \in \mathbb{R}^{768}$ and constructed the network as

$$\hat{\mathbf{y}}_{\mathrm{diag}} = \mathrm{F}_3(\mathrm{F}_2(\mathrm{F}_1(\mathbf{f}_{\mathrm{mass}}))) = \mathrm{F}_3(\mathbf{f}_{\mathrm{imd}}). \tag{2}$$

In order to make $\mathbf{f}_{\mathrm{imd}}$ resemble $\mathbf{f}_{\mathrm{sent}}$, we also modified loss function $\mathcal{L}$ as

$$\mathcal{L}_{\mathrm{sent}} = ||\mathbf{f}_{\mathrm{imd}} - \mathbf{f}_{\mathrm{sent}}||_2, \tag{3}$$

$$\mathcal{L} = \lambda_{\mathrm{sent}}\mathcal{L}_{\mathrm{sent}} + \lambda_{\mathrm{class}}\mathcal{L}_{\mathrm{class}}, \tag{4}$$

$\mathcal{L}_{\mathrm{class}}$ is cross entropy loss function for binary diagnosis and λ_{sent} and λ_{class} are balancing parameters. For comparison with efficiency of margin and shape one-hot vector, we constructed the multitasking model. It utilized three sets of fully-connected layers as

$$\begin{aligned} \mathrm{FC}^{\mathrm{diag}} &= \{\mathrm{F}_1^{\mathrm{diag}}, \mathrm{F}_2^{\mathrm{diag}}, \mathrm{F}_3^{\mathrm{diag}}\} \\ \mathrm{FC}^{\mathrm{margin}} &= \{\mathrm{F}_1^{\mathrm{margin}}, \mathrm{F}_2^{\mathrm{margin}}, \mathrm{F}_3^{\mathrm{margin}}\} \\ \mathrm{FC}^{\mathrm{shape}} &= \{\mathrm{F}_1^{\mathrm{shape}}, \mathrm{F}_2^{\mathrm{shape}}, \mathrm{F}_3^{\mathrm{shape}}\} \end{aligned} \tag{5}$$

to predict benign/malignant, margin and shape. $\mathrm{F}_2^{\mathrm{diag}}, \mathrm{F}_2^{\mathrm{margin}}$ and $\mathrm{F}_2^{\mathrm{shape}}$ have 768 dimentional vector as its output. The whole network was constructed as

$$\hat{\mathbf{y}}_{\mathrm{diag}} = \mathrm{F}_3^{\mathrm{diag}}(\mathrm{F}_2^{\mathrm{diag}}(\mathrm{F}_1^{\mathrm{diag}}(\mathbf{f}_{\mathrm{mass}}))) \tag{6}$$

$$\hat{\mathbf{y}}_{\mathrm{margin}} = \mathrm{F}_3^{\mathrm{margin}}(\mathrm{F}_2^{\mathrm{margin}}(\mathrm{F}_1^{\mathrm{margin}}(\mathbf{f}_{\mathrm{mass}}))) \tag{7}$$

$$\hat{\mathbf{y}}_{\mathrm{shape}} = \mathrm{F}_3^{\mathrm{shape}}(\mathrm{F}_2^{\mathrm{shape}}(\mathrm{F}_1^{\mathrm{shape}}(\mathbf{f}_{\mathrm{mass}}))) \tag{8}$$

where $\hat{\mathbf{y}}_{\mathrm{margin}}, \hat{\mathbf{y}}_{\mathrm{shape}}$ are margin and shape prediction vector each. Loss function $\mathcal{L}$ of the network is

$$\mathcal{L} = \lambda_{\mathrm{diag}}\mathcal{L}_{\mathrm{diag}} + \lambda_{\mathrm{margin}}\mathcal{L}_{\mathrm{margin}} + \lambda_{\mathrm{shape}}\mathcal{L}_{\mathrm{shape}} \tag{9}$$

where $\mathcal{L}_{\mathrm{diag}}, \mathcal{L}_{\mathrm{margin}}$ and $\mathcal{L}_{\mathrm{shape}}$ denote cross-entropy loss between prediction (malignancy, margin, shape) and ground truth annotation. $\lambda_{\mathrm{diag}}, \lambda_{\mathrm{margin}}$ and λ_{shape} are balancing parameters.

4.3.2 Experimental setting

We used the DDSM dataset and the FFDM dataset with same split and augmentation process with subsection 4.2. We also utilized three margin classes and three shape classes mentioned in subsection 4.2. We set $[\lambda_{\mathrm{sent}}, \lambda_{\mathrm{class}}]$ to [0, 1.0] for the baseline, [1.0, 0.5] for the sentence model and set $[\lambda_{\mathrm{diag}}, \lambda_{\mathrm{margin}}, \lambda_{\mathrm{shape}}]$ as [1.0, 0.5, 0.5] for multitasking model. The size of mini-batch was set to 64 and an Adam optimizer was used with learning rate 0.0001 in case of the DDSM dataset and 0.00005 in cased of the FFDM dataset. For training with the FFDM dataset, we initialized network parameters with parameters of the model trained with the DDSM dataset.

4.3.3 Results

The results are shown in Table 4. In the case of simply predicting the margin and shape using a fully-connected layer, the CAD performance was reduced, whereas the performance of the CAD network trained with the sentence embedded vector was increased.

4.4. Experiment 3: Visual pointing guided by sentence

In this experiment, we assumed that the situation where the pre-trained CAD network existed. We utilized the image feature extracted from mass ROI image via the pre-trained CAD network to generate a visual attentive map which points suspicious part of mass ROI image. In training stage to generate the visual attentive map, parameters in the pre-trained model were fixed and we utilized the pre-trained CAD model only for extracting $\mathbf{f}_{\mathrm{mass}}$. Therefore its diagnostic performance can be maintained.

4.4.1 Network architecture

We compared two network models which 1) guide visual attentive map with the sentence associated with the mass ROI image (*i.e.* sentence), and 2) guide visual attentive map with the margin and shape one-hot vector associated with the mass ROI image (*i.e.* one-hot).

The general CAD network can be divided into two part which is image feature encoder $\mathrm{F}_{\varphi_{\mathrm{mass}}}(\cdot)$ containing multiple convolutional layers and predictor $\mathrm{F}_{\varphi_{\mathrm{diag}}}(\cdot)$ containing fully-connected layer (or multiple fully-connected layers), respectively. Mass ROI image $\mathbf{I}(n, m)$ becomes a diagnostic prediction vector $\hat{\mathbf{y}}_{\mathrm{diag}} \in \mathbb{R}^2$ through the process as following:

$$\mathbf{f}_{\mathrm{mass}}(n, m, k) = \mathrm{F}_{\varphi_{\mathrm{mass}}}(\mathbf{I}(n, m)), \tag{10}$$

$$\hat{\mathbf{y}}_{\mathrm{diag}} = \mathrm{F}_{\varphi_{\mathrm{diag}}}(\mathbf{f}_{\mathrm{mass}}), \tag{11}$$

where $\mathbf{f}_{\mathrm{mass}}(n, m, k)$ is mass ROI image feature which is output of $\mathrm{F}_{\varphi_{\mathrm{mass}}}(\cdot)$. To reflect diagnostic result to $\mathbf{f}_{\mathrm{mass}}$

in the form of channel attention, $\hat{\mathbf{y}}_{\text{diag}}$ is embedded into k-dimensional attention weight α_{channel} via function with learnable parameter $\mathrm{E}_{\varphi_{\text{emb}}}(\cdot)$ and we apply channel attention on $\mathbf{f}_{\text{mass}}$ as

$$\mathbf{f}_{\text{mass+d}}(n,m,k) = \mathbf{f}_{\text{mass}}(n,m,k) \cdot \alpha_{\text{channel}}(k) \quad (12)$$

where $\mathbf{f}_{\text{mass+d}}$ is diagnosis attentive feature. $\mathbf{f}_{\text{mass+d}}$ is used as input of function $\mathrm{F}_{\varphi_{\text{vis}}}(\cdot)$ corresponding multiple convolutional layers to make 2D map $\alpha_{\text{map}}(n,m)$. Then, α_{map} is normalized through softmax function as

$$\alpha_{\text{map}}^{\text{softmax}}(n,m) = \frac{\exp(\alpha_{\text{map}}(n,m))}{\sum_n \sum_m \exp(\alpha_{map}(n,m))}, \quad (13)$$

where $\alpha_{\text{map}}^{\text{softmax}}$ is 2D visual attentive map. In order to guide $\alpha_{\text{map}}^{\text{softmax}}$ using sentence associated with $\mathbf{I}(n,m)$, a feature for generating sentence $\mathbf{f}_{\text{text}}$ is obtained via the process as

$$\begin{aligned} &\mathbf{f}_{\text{mass+d+map}}(n,m,k) \\ &= \mathbf{f}_{\text{mass+d}}(n,m,k) \cdot \alpha_{\text{map}}^{\text{softmax}}(n,m), \end{aligned} \quad (14)$$

$$\mathbf{f}_{\text{text}} = \mathrm{F}_{\varphi_{\text{text}}}(\mathbf{f}_{\text{mass+d+map}} + \mathbf{f}_{\text{mass+d}}), \quad (15)$$

where $\mathbf{f}_{\text{mass+d+map}}$ is result of applying spatial attention using $\alpha_{\text{map}}^{\text{softmax}}$ and $\mathrm{F}_{\varphi_{\text{text}}}(\cdot)$ is function implemented by multiple convolutional layers. $\mathbf{f}_{\text{text}}$ passed through two-hidden-layer-stacked LSTM and became generated sentence $\mathbf{W} = [w_1, w_2, \cdots]$. All parameters in the pre-trained CAD network ($\mathrm{F}_{\varphi_{\text{mass}}}(\cdot)$, $\mathrm{F}_{\varphi_{\text{diag}}}(\cdot)$) are fixed and other parameters are trained with cross-entropy loss between $\mathbf{W}$ and ground truth sentence. To demonstrate effectiveness of sentence as guidance of visual pointing, we compare results of the network with sentence and the network with margin and shape one-hot vector. The latter network contain two set of multiple fully-connected layers $\mathrm{F}_{\varphi_{\text{margin}}}(\cdot)$ and $\mathrm{F}_{\varphi_{\text{shape}}}(\cdot)$ for predicting margin and shape of input mass ROI image as

$$\hat{\mathbf{y}}_{\text{margin}} = \mathrm{F}_{\varphi_{\text{margin}}}(\mathbf{f}_{\text{text}}), \quad (16)$$

$$\hat{\mathbf{y}}_{\text{shape}} = \mathrm{F}_{\varphi_{\text{shape}}}(\mathbf{f}_{\text{text}}), \quad (17)$$

where $\hat{\mathbf{y}}_{\text{margin}}$ and $\hat{\mathbf{y}}_{\text{shape}}$ denote margin and shape prediction vector. The whole network except $\mathrm{F}_{\varphi_{\text{mass}}}(\cdot)$ and $\mathrm{F}_{\varphi_{\text{diag}}}(\cdot)$ is trained with cross entropy loss function between $\hat{\mathbf{y}}_{\text{margin}}, \hat{\mathbf{y}}_{\text{shape}}$ and one-hot margin/shape vector as ground truth. That is, we compare visual attentive map guided by DDSM-sentence and margin/shape one-hot vector.

4.4.2 Experimental setting

For this experiment, we used mass ROI images in the DDSM dataset as input and sentences in DDSM-sentence as ground truth to calculate loss function. For input mass ROI images, data augmentation process of subsection 4.2 was applied in the same way. Baseline network in subsection 4.2 was utilized as the pre-trained CAD network. $\mathbf{f}_{\text{mass}}$ corresponded feature after conv 5_3 of the baseline network. The size of mini-batch was 64 and Adam optimizer with learning rate 0.0005 was used.

4.4.3 Results

As shown in Figure 3, sentence guided map pointed the mass in the image more accurately. In the case of the image containing noise such as bright dot, one-hot vector guided map tended to point the noised part only. However, sentence guided map pointed the exact part located the mass even with the noised image. In the case of the large mass which fills most part of the image, one-hot vector guided map only pointed parts of the mass, whereas sentence guided map covered almost the whole mass. Additionally, we note that the diagnosis performance of the network did not deteriorate during the training stage since the parameters of the pre-trained CAD network were fixed.

5. Discussion and conclusion

In this study, we presented the sentence annotation process and constructed new datasets called the DDSM-sentence and the FFDM-sentence which are describing collected image from breast mammography datasets. We constructed the sentence dataset using the terminology (*i.e.* BI-RADS lexicons) provided by the conventional datasets. We also proposed three use cases where sentence annotation can be utilized for the CAD network. In each case, it was demonstrated that the proposed datasets were useful to improve diagnostic performance compared with cases where margin and shape lexicons were used as one-hot vector. In addition, the DDSM-sentence was also useful to generate visual pointing on mass ROI images.

For further work, the sentence composing approach with the visual words in this paper will be applied other annotation of breast mammography dataset (e.g. calcification) to make the performance of the CAD network much better by providing additional information as a form of a sentence. Furthermore, if it is shown that the sentence dataset with proposed process is effective not only on diagnostic performance but also on the network in other previous studies such as generating report automatically and making cad network interpretable, our annotation process would be applied to pre-collected different disease imaging dataset which has standardized terms to describe the pathognomonic signs.

References

[1] H. Berment, V. Becette, M. Mohallem, F. Ferreira, and P. Chérel. Masses in mammography: What are the underlying anatomopathological lesions? *Diagnostic and interventional imaging*, 95(2):124–133, 2014.

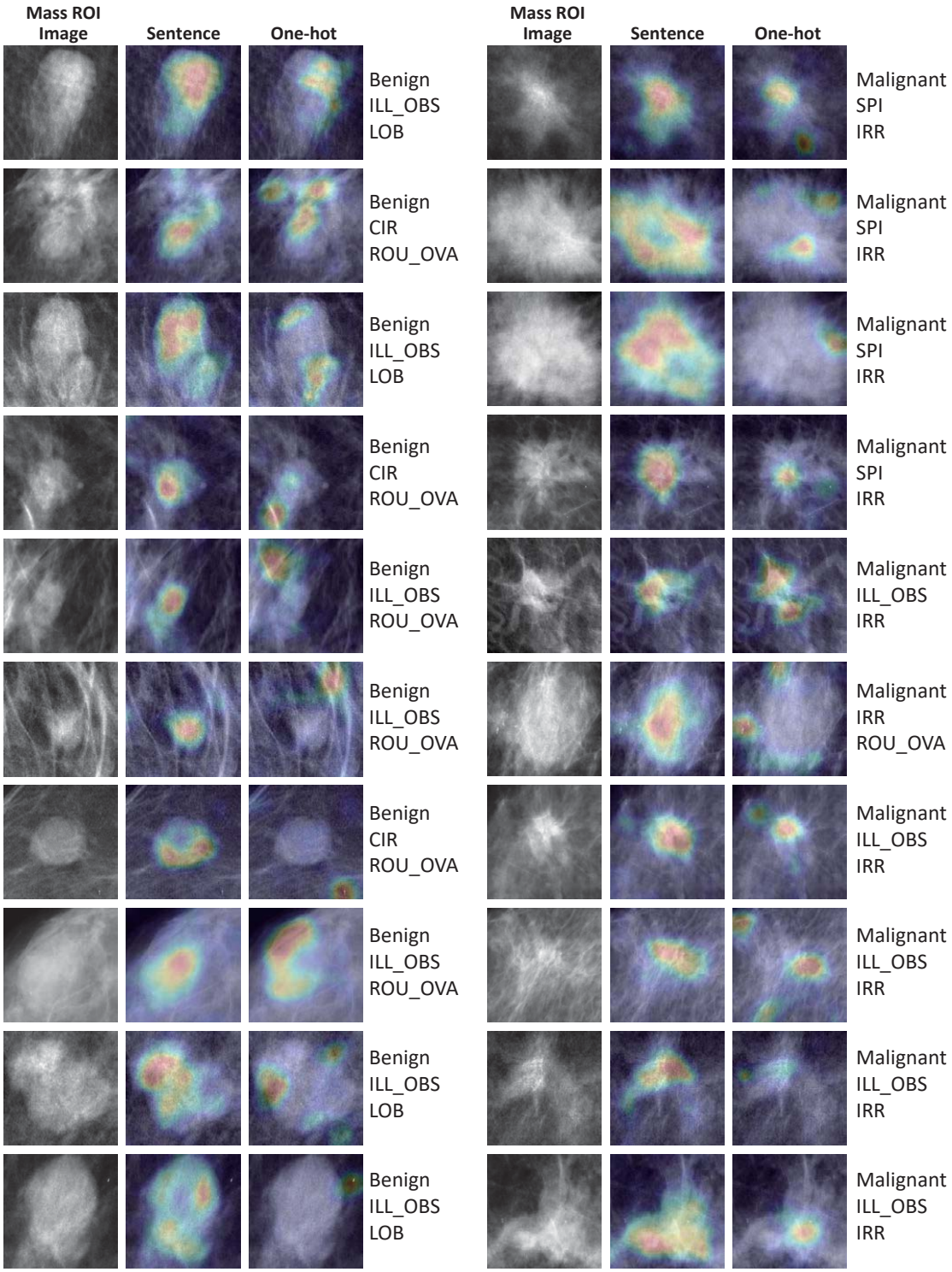


Figure 3. Comparison visual attentive map of the DDSM dataset guided by the sentence and the margin/shape one-hot vector (ILL_OBS: Ill-defined-obscured, CIR: circumscribed, SPI: spiculated, ROU_OVA: round-oval, IRR: irregular, LOB: lobulated)

[2] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[3] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

[4] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015.

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[7] B. M. Geller, W. E. Barlow, R. Ballard-Barbash, V. L. Ernster, B. C. Yankaskas, E. A. Sickles, P. A. Carney, M. B. Dignan, R. D. Rosenberg, N. Urban, et al. Use of the american college of radiology bi-rads to report on the mammographic evaluation of women with signs and symptoms of breast disease. *Radiology*, 222(2):536–542, 2002.

[8] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer. The digital database for screening mammography. In *Proceedings of the 5th international workshop on digital mammography*, pages 212–218. Medical Physics Publishing, 2000.

[9] D. Huk Park, L. Anne Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*, pages 8779–8788. IEEE, 2018.

[10] J. Lee. Practical and illustrated summary of updated bi-rads for ultrasonography. *Ultrasonography*, 36(1):71, 2017.

[11] K. Lee, N. Talati, R. Oudsema, S. Steinberger, and L. Margolies. Bi-rads 3: Current and future use of probably benign. *Current radiology reports*, 6(2):5, 2018.

[12] L. Liberman, A. F. Abramson, F. B. Squires, J. Glassman, E. Morris, and D. D. Dershaw. The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories. *AJR. American journal of roentgenology*, 171(1):35–40, 1998.

[13] L. Liberman and J. H. Menell. Breast imaging reporting and data system (bi-rads). *Radiologic Clinics*, 40(3):409–430, 2002.

[14] W. Moon, C. Lo, J. Chang, C. Huang, J. Chen, and R. Chang. Quantitative ultrasound analysis for classification of bi-rads category 3 breast masses. *Journal of digital imaging*, 26(6):1091–1098, 2013.

[15] A. C. of Radiology. *Breast Imaging Reporting and Data System® (BI-RADS®)*. American College of Radiology, Reston, Va, 4 edition, 2003.

[16] S. G. Orel, N. Kay, C. Reynolds, and D. C. Sullivan. Bi-rads categorization as a predictor of malignancy. *Radiology*, 211(3):845–850, 1999.

[17] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[18] T. Schlegl, S. M. Waldstein, W.-D. Vogl, U. Schmidt-Erfurth, and G. Langs. Predicting semantic descriptions from medical images with convolutional neural networks. In *International Conference on Information Processing in Medical Imaging*, pages 437–448. Springer, 2015.

[19] R. Selvi. *Breast Diseases Imaging and Clinical Management*. Springer India, 2015.

[20] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506, 2016.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[22] J. H. Suckling, J. Parker, S. M. Astley, I. W. Hutt, C. R. M. Boggis, I. W. Ricketts, E. A. Stamatakis, N. Cerneaz, S. Kok, P. S. Taylor, D. Betal, and J. Savage. The mammographic image analysis society digital mammogram database.

[23] B. Surendiran and A. Vadivel. Mammogram mass classification using various geometric shape and margin features for early detection of breast cancer. *International Journal of Medical Engineering and Informatics*, 4(1):36–54, 2012.

[24] S. H. Taplin, L. E. Ichikawa, K. Kerlikowske, V. L. Ernster, R. D. Rosenberg, B. C. Yankaskas, P. A. Carney, B. M. Geller, N. Urban, M. B. Dignan, et al. Concordance of breast imaging reporting and data system assessments and management recommendations in screening mammography. *Radiology*, 222(2):529–535, 2002.

[25] I. Thomassin-Naggara, A. Tardivon, and J. Chopier. Standardized diagnosis and reporting of breast cancer. *Diagnostic and interventional imaging*, 95(7-8):759–766, 2014.

[26] L. Tsochatzidis, K. Zagoris, N. Arikidis, A. Karahaliou, L. Costaridou, and I. Pratikakis. Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach. *Pattern Recognition*, 71:106–117, 2017.

[27] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.

[28] Y. Xue and X. Huang. Improved disease classification in chest x-rays with transferred features from report generation. In *International Conference on Information Processing in Medical Imaging*, pages 125–138. Springer, 2019.

[29] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017.

[30] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.