Title

Analysis of Regions of Interest and Distractor Regions in Breast Biopsy Images.

Authors

Lu, Ximing
Mehta, Sachin
Brunyé, Tad
et al.

Publication Date

2021-07-01

DOI

10.1109/bhi50953.2021.9508513

Peer reviewed

HHS Public Access

Author manuscript

IEEE EMBS Int Conf Biomed Health Inform. Author manuscript; available in PMC 2022 December 30.

Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2021 July ; 2021: . doi:10.1109/bhi50953.2021.9508513.

Analysis of Regions of Interest and Distractor Regions in Breast Biopsy Images

Ximing Lu[1], **Sachin Mehta**[1], **Tad T. Brunyé**[2], **Donald L. Weaver**[3], **Joann G. Elmore**[4], **Linda G. Shapiro**[1]

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle

[2]Center for Applied Brain and Cognitive Sciences, School of Engineering, Tufts University, Medford

[3]Department of Medicine, University of Vermont, Burlington

[4]David Geffen School of Medicine, University of California, Los Angeles

Abstract

This paper studies *why* pathologists can misdiagnose diagnostically challenging breast biopsy cases, using a data set of 240 whole slide images (WSIs). Three experienced pathologists agreed on a consensus reference ground-truth diagnosis for each slide and also a consensus region of interest (ROI) from which the diagnosis could best be made. A study group of 87 other pathologists then diagnosed test sets (60 slides each) and marked their own regions of interest. Diagnoses and ROIs were categorized such that if on a given slide, their ROI differed from the consensus ROI *and* their diagnosis was incorrect, that ROI was called a *distractor*. We used the HATNet transformer-based deep learning classifier to evaluate the visual similarities and differences between the true (consensus) ROIs and the distractors. Results showed high accuracy for both the similarity and difference networks, showcasing the challenging nature of feature classification with breast biopsy images. This study is important in the potential use of its results for teaching pathologists how to diagnose breast biopsy slides.

Index Terms—

biomedical image analysis; whole slide image; region of interest; machine learning

I. Introduction

According to the World Health Organization, cancer is the second leading cause of death globally and was responsible for an estimated 9.6 million deaths in 2018 [10]. Diagnostic classification errors among pathologists can have significant adverse consequences for patients. The “gold standard” for diagnosis of cancer relies on a pathologist’s visual assessment of tissue sections and perceptual and cognitive processing of learned cytological and morphological criteria. [2] Assessment of these criteria is subjective, and pathologists often can make mistakes in diagnostically challenging cases. To reduce diagnostic error, it is important to analyze and understand the circumstances under which pathologists fail to correctly diagnose a case.

Information processing frameworks propose that when pathologists diagnose a whole slide breast biopsy image, they proceed through three primary interpretive phases: visual search, recognition, and decision making [4], [9], [11]. The breadth-first visual search process identifies suspicious regions warranting further inspection. In-depth visual inspection results in recognizing critical histopathological features residing in a region of interest (ROI). Finally, recognized features are mapped onto one of four primary diagnostic classes: Benign, Atypia, Ducal Carcinoma in situ (DCIS), or Invasive Cancer.

Our study included 240 whole slide breast biopsy images, each with a single consensus diagnosis and ROI determined by an expert panel of three experienced pathologists. Consensus ROIs marked by the expert panel will be considered the correct ROIs in this study. The 240 whole slide images (WSI) were also diagnosed by a separate group of 87 pathologists, each of whom reviewed and marked ROIs on a test set of 60 slides. When diagnosing a WSI, if a pathologist chose the wrong area as ROI, and based on that ROI made the wrong diagnosis, we call this area that mislead the pathologist a distractor as illustrated in Figure 1.

Diagnostic mistakes can occur because of a visual search error (a missed cue) or a recognition error (a misinterpreted cue). By comparing correct ROIs to distractors, we can gain insights into why distractors are the misleading cues resulting in diagnostic errors. More specifically, what common clinical structures in distractors mislead pathologists, and what are their characteristics? On one hand, the distractors must have some similarity with the consensus ROIs that pathologists have studied or encountered before, increasing confusability. On the other hand, distractors must have some critical difference from consensus ROIs, which make them less suitable as diagnostic evidence. To examine these possibilities, our study will focus on analyses meant to reveal the surface similarity and underlying difference between consensus ROIs and distractors.

Our work aims to answer the following research questions: 1) What are the surface similarities between consensus ROIs and distractors? 2) What are their underlying differences?

II. Related Work

Detecting and recognizing critical histopathological features is fundamental to successful diagnosis of breast cancer biopsy slides. Our group and others have thus studied how to find such critical regions of interest and identify their features. Mercan et al. [7] developed a bag-of-words approach to finding critical ROIs that agreed with those selected by pathologists, using color and texture features to characterize the image patches and mouse tracking data to characterize pathologist image navigation behavior. Nagarkar et al. [8] studied the properties of ROIs and how well they correlated with the correct diagnosis of pathologists. Zheng et al. [13] developed a histopathological CBIR approach called CBHIR which is automatically processed throughout the WSI, based on which a probability map regarding the malignancy of breast tumors is calculated. In a small pilot study, Brunyè et al. [1] used eye tracking data to assess how eye movements were attracted to consensus ROIs versus visually salient image regions, using basic graph-based visual salience algorithms. Ersoy et al. [3] also worked

with eye tracking data and developed a platform that included eye tracking data analysis. To the best of our knowledge, ours is the most comprehensive study examining the effect of image distractors on diagnostic accuracy.

III. Background: HATnet

Holistic ATtention Network (HATNet) [5] is a transformer-based [12] network for classifying breast biopsy images in an end-to-end manner. It's the state-of-the-art method that was able to match the classification performance of participant pathologists on an independent test set [2].

HATNet factorizes the input biopsy image into words (or patches) using a bag-of-words approach and then encode their relationships in a hierarchical manner using self-attention. Moreover, the authors found that in diagnostic classification, this network paid high attention to ductal regions and stromal tissues, important bio-markers in breast cancer diagnosis, suggesting that there is clinical relevance in this method.

In our work, we will use HATNet as an efficient classifier to mimic the behavior of pathologists, and to determine on which areas of the image it focuses as it classifies distractor regions. Our work is not limited to HATNet and any other WSI classification network can be used.

IV. Method

Our methodology is to construct two different deep learning networks: one for learning differences between ROIs and distractors and the other for learning similarities (Figure 2). Unlike natural images where objects of interest (e.g., person and car) have different attributes and a single network can learn both similarities and differences [6], the distractor and consensus ROIs in our dataset can have similar attributes. In particular, they can be similar in appearance, size, and tissue distribution. This makes it potentially challenging to learn discriminative representations. Therefore, we chose to learn similarities and differences using two different networks.

A. Similarity Network

We aim to find out the visual similarity between ROIs and distractors that misleads pathologists via similarity network. To do that, we train a HATNet to mimic the behavior of pathologists who made mistakes, missing important regions and misinterpreting an erroneous ROI as critical, leading to an incorrect diagnosis.

We structure a classification problem to mimic such a situation. We put ROIs and distractors together in a set and label them as Category 1; we sample and label non-ROI/non-distractor areas as Category 2. The task to train similarity network is a binary classification problem between classes 1 and 2. After optimization for the task, the features extracted by the network should be similar intra-class and different inter-class. Thus, the features extracted from the ROIs and distractors in class 1 are what make them similar to each other and simultaneously different from the rest of regions.

B. Difference Network

We aim to find out the underlying differences between the consensus ROIs and the distractors via difference network.To do that, we train another HATNet network to mimic the behavior of the experienced pathologists.

We want difference network to have the ability to distinguish distractors from ROIs based on their subtle differences. We label the consensus ROIs as category 1A and the distractor ROIs as category 1B. The task for training difference network is to learn the ground-truth labels. In this classification, the features extracted from ROIs and distractors are hypothesized to be different, so that the network is able to assign different labels to them. Additionally, we analyze the important areas HATNet pay high attention to in this task, and present a case study in section VI.

V. Experiments

A. Dataset

We represent datatset statistics in Table I. For each type of the regions, we keep 80% for training, 10% for validation and 10% for testing.

1) Consensus ROI: The breast biopsy dataset consists of 240 whole slide images with haematoxylin and eosin (H&E) staining. [2] As described earlier, three experienced pathologists provide consensus ROIs and diagnostic label for each of the slide. We understand that consensus ROIs might not be the only areas that can help to diagnose a slide.

2) Distractor: A total of 87 pathologists participating in a previous study interpreted aforementioned breast biopsy dataset [2]. Each pathologist was randomized to classify a subset of 60 slides. If a participating pathologist chose an area that disagreed with the consensus ROI, and based on that choice, made the wrong diagnosis, we designate this area as a distractor ROI or just a distractor for short. There are 658 distractors in total for our data set. The percentages in Figure 1 illustrate the statistics of the pathologists' diagnoses. Notice that the participating pathologists often did not choose the consensus ROI, but still made the correct diagnosis on a different, but perhaps similar ROI. In this paper, we are only studying the distractors that were from different region of the consensus ROIs and seemed to thus cause them to make an incorrect diagnosis. Therefore, we enforce the overlapping ratio of a distractor with consensus ROIs must be smaller than 15% to be considered as a valid distractor. Moreover, we can see very few pathologists made an incorrect diagnosis when using the consensus ROI.

3) Non-ROI/distractor: We randomly sample regions with size $(10000 \pm 5000) \times (10000 \pm 5000)$ in pixels from WSIs which don't overlap with both ROI and distractor.

B. Training:

We use the default hyperparameters of the HATNet model [5] to train our binary classifiers.

C. Results

We report the classification result for the similarity network in Table IIa and for the difference network in Table IIb. The metric we used is standard accuracy (number of examples classified correctly / total number of examples). The difference network obtained an overall accuracy of 68.7%, while the similarity network obtained an overall accuracy of 77.2%. Similarity network achieves significantly higher accuracy for the DCIS and the Invasive classes, while difference network does better for the Atypia and the DCIS class. This reflects the difficulties of two tasks with respect to different diagnostic categories for neural notworks. Another interesting finding is that the similarity network has higher accuracy than the difference network, which indicates that the high-level similarity is easier for the network to realize than the subtle differences. The difference network is facing a more difficult task than the similarity network.

VI. Visual Analysis

HATNet is based on transformers [12] and can report which areas are of importance in its classifications. We studied some of the important areas identified by the difference network. These patches are areas that represent the differences between distractors and ROIs. We selected two correctly classified examples, one from the Atypia class and one from the DCIS class for visualization. Figure 3a shows the whole slide image and closeups of the ROI (in red) and a distractor (in green) of the Atypia class, while Figure 3b shows the whole slide image and close-ups of ROI and a distractor of the DCIS class. In both cases, the distractor was in a different part of the image from consensus ROI. In the Atypia image of Figure 3a, the important patches are mostly stroma and epithelial tissues, while in the distractor, many of the important patches include benign fibroglandular tissue. This is also true, but to a lesser extent in the DCIS image.

VII. Conclusions

This paper has analyzed the consensus regions of interest versus a set of regions we call *distractors* that some pathologists marked as the region of most interest when they incorrectly diagnose a case. In order to study the distractors, we trained two HATNet classifiers: one to discriminate between consensus ROIs and distractors and one to discriminate between a set containing both consensus ROIs and distractors and all other regions. The difference network obtained an overall accuracy of 68.7%, while the similarity network obtained an overall accuracy of 77.2%. Thus, we can conclude that there is some discernible difference between consensus ROIs and distractors, but there is also a lot of similarity, making it difficult for less experienced pathologists, especially those at the beginning of their training, to tell the difference. For example, similarity of features residing in Atypia and DCIS images may underlie a failure to detect, recognize, and/or correctly diagnose features when attempting to discriminate these challenging diagnostic classes. While many pathologists were able to correctly diagnose without using the exact consensus ROI, the large number who diagnosed incorrectly when choosing distractors makes this an important topic of study.

Acknowledgments

Research reported in this article was supported by grants R01 CA172343, R01 CA140560, U01 CA231782, and R01 CA200690 from the National Cancer Institute of the National Institutes of Health.

References

[1]. Brunyé TT, Carney PA, Allison KH, Shapiro LG, Weaver DL, and Elmore JG. Eye movements as an index of pathologist visual expertise: A pilot study. PLOS One, 2014.

[2]. Elmore Joann G, Longton Gary M, Carney Patricia A, Geller Berta M, Onega Tracy, Tosteson Anna NA, Nelson Heidi D, Pepe Margaret S, Allison Kimberly H, Schnitt Stuart J, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. Jama, 313(11):1122–1132, 2015. [PubMed: 25781441]

[3]. Ersoy I et al. Eye gaze pattern analysis of whole slide image viewing behavior in pathedex platform. Microscopy and Microanalysis, 23:248–249, 2017.

[4]. Kundel HL, Nodine CF, and Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Investigative radiology, 13(3):175–181, 1978. [PubMed: 711391]

[5]. Mehta S, Lu X, Elmore JG, Hajishirzi H, and Shapiro LG. Hatnet: An end-to-end holistic attention network for diagnosis of breast biopsy images, 2020. https://arxiv.org/pdf/2007.13007.pdf.

[6]. Melekhov I, Kannala J, and Rahtu E. Siamese network features for image matching. In 2016 23rd International Conference on Pattern Recognition (ICPR), pages 378–383. IEEE, 2016.

[7]. Mercan E et al. Localization of diagnostically relevant regions of interest in whole slide images: a comparative study. Journal of Digital Imaging, 29:496–506, 2016. [PubMed: 26961982]

[8]. Nagarkar Dilip B, Mercan E, Weaver D, Brunyé Tad T., Carney P, Rendi M, Beck A, Frederick P, Shapiro L, and Elmore J. Region of interest identification and diagnostic agreement in breast pathology. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc, 29:1004–1011, 2016. [PubMed: 27198567]

[9]. Nodine CF and Kundel HL. The cognitive side of visual search in radiology. In Eye movements from physiology to cognition, pages 573–582. Elsevier, 1987.

[10]. World Health Organization. Cancer. https://www.who.int/health-topics/cancer#tab=tab_1, 2021. Online; accessed 8 May 2021.

[11]. Patel VL, Kaufman DR, and Arocha JF. Emerging paradigms of cognition in medical decision-making. Journal of biomedical informatics, 35(1):52–75, 2002. [PubMed: 12415726]

[12]. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN,Kaiser L, and Polosukhin I. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[13]. Zheng Yushan, Jiang Z, Zhang Haopeng, Xie Feng ying, Ma Yibing, Shi H, and Zhao Y. Histopathological whole slide image analysis using context-based cbir. IEEE Transactions on Medical Imaging, 37:1641–1652, 2018. [PubMed: 29969415]

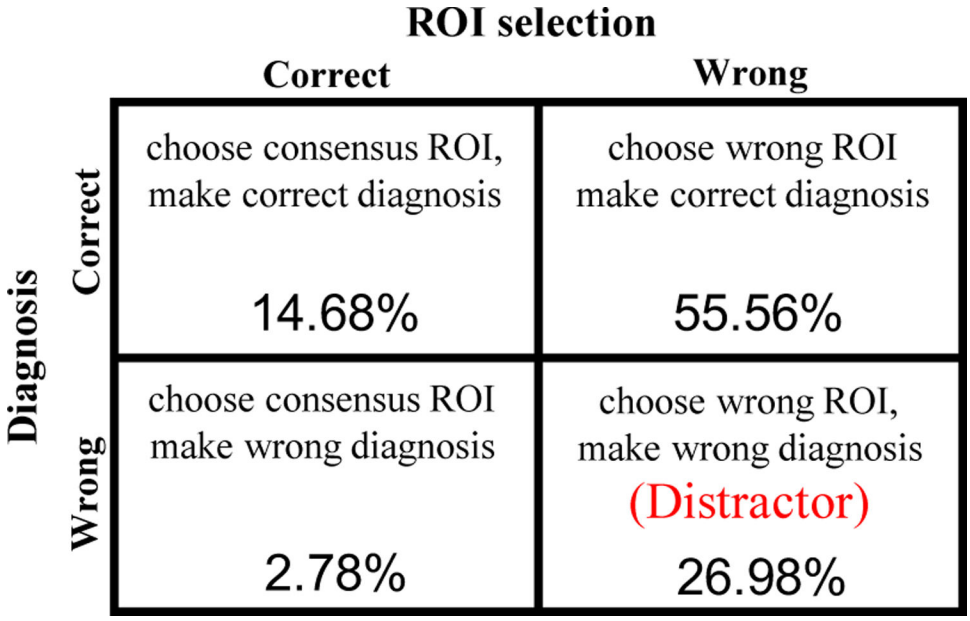


Fig. 1:

Confusion matrix for diagnostic procedure. The total diagnosises made by 87 participating pathologists is 5220.

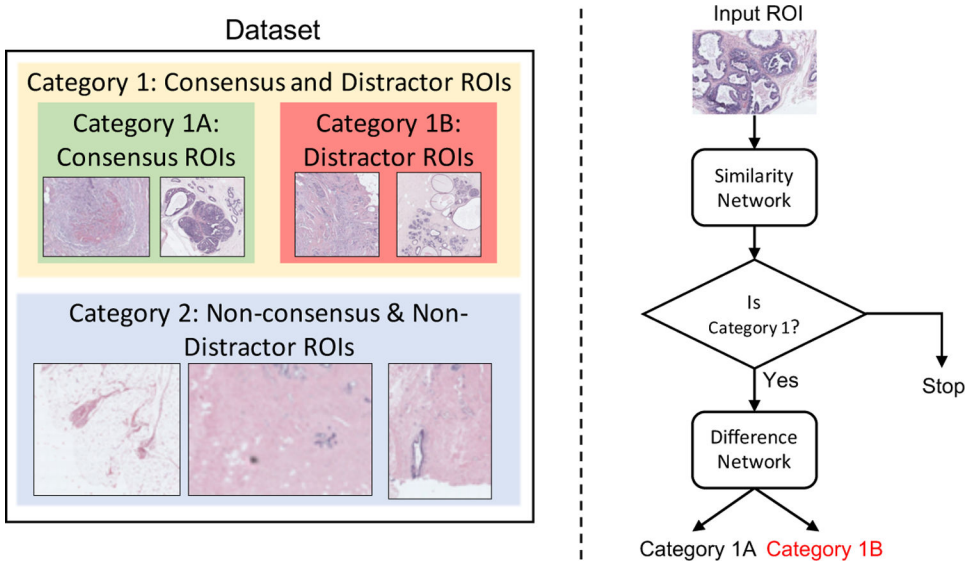
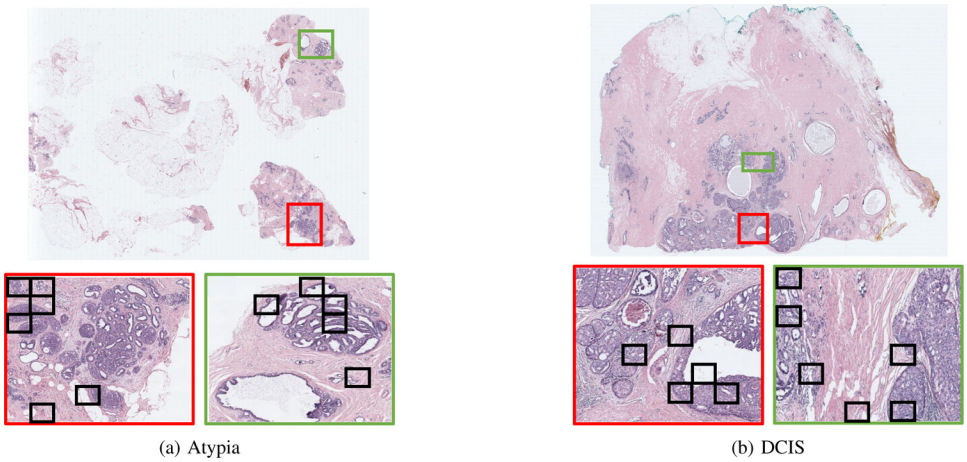


Fig. 2:

The overview of our method. On the left side, example regions from our dataset are shown, while the right side shows our network structure. The similarity network tries to differentiate between Category 1 (consensus ROIs and distractors) and Category 2 (other regions), while the difference network tries to discriminate between consensus ROIs (Category 1A) and distractors (Category 1B). Consensus and distractor ROIs can be similar, while Category 2 contains a mixture of appearances.



(a) Atypia (b) DCIS

Fig. 3:

Whole slide images for cases diagnosed as (a) Atypia and (b) DCIS with their consensus ROIs in red and example distractor ROIs in green. In the blowups of these regions (bottom panels), the small patches that the difference network found most discriminative are marked in black.

TABLE I:

Statistics of three types of region in our dataset.

Data Category	Number of examples					Average size (in pixels)
	Benign	Atypia	DCIS	Invasive	Total	
Consensus ROI	125	102	161	34	422	10880 × 9558
Distractor	329	169	126	24	658	10724 × 9213
Non-ROI/distractor	450	270	280	60	1060	9987 × 10021
Total	904	541	567	118	2130	10530 × 9597

TABLE II:

Diagnostic class-wise classification results

(a) Similarity network		(b) Difference network	
Diagnostic class	**Accuracy**	**Diagnostic class**	**Accuracy**
Atypia	71.9%	Atypia	75.3%
Benign	70.5%	Benign	61.1%
DCIS	79.8%	DCIS	70.9%
Invasive	86.4%	Invasive	67.3%
Overall	77.2%	Overall	68.7%