HATNet: An End-to-End Holistic Attention Network for Diagnosis of Breast Biopsy Images

Sachin Mehta*, Ximing Lu*, Donald Weaver†, Joann G. Elmore‡, Hannaneh Hajishirzi*, Linda Shapiro*

*University of Washington, Seattle, WA
†University of Vermont College of Medicine, Burlington, VT
‡University of California, Los Angeles, CA

Abstract—Training end-to-end networks for classifying gigapixel size histopathological images is computationally intractable. Most approaches are patch-based and first learn local representations (patch-wise) before combining these local representations to produce image-level decisions. However, dividing large tissue structures into patches limits the context available to these networks, which may reduce their ability to learn representations from clinically relevant structures. In this paper, we introduce a novel attention-based network, the Holistic ATtention Network (HATNet) to classify breast biopsy images. We streamline the histopathological image classification pipeline and show how to learn representations from gigapixel size images end-to-end. HATNet extends the bag-of-words approach and uses self-attention to encode global information, allowing it to learn representations from clinically relevant tissue structures without any explicit supervision. It outperforms the previous best network Y-Net, which uses supervision in the form of tissue-level segmentation masks, by 8%. Importantly, our analysis reveals that HATNet learns representations from clinically relevant structures, and it matches the classification accuracy of human pathologists for this challenging test set.

Index Terms—Self-attention, Transformers, Histopathological images, Breast cancer, Image classification

I. INTRODUCTION

BR EAST cancer is the most common cancer in women accounting for approximately 25% of all cancer instances worldwide [1], [2]. Diagnostic classification errors among pathologists can have significant adverse consequences for patients. The “gold standard” for diagnosis of breast biopsy specimens relies on a pathologists visual assessment of tissue sections and cognitive processing of learned cytologic and morphological criteria, including architectural and cellular changes in the tissue, alterations of the tumor micro-environment, and immune-mediated host response. Assessment of these morphological criteria is subjective and pathologists, even expert pathologists, cannot always reach consensus on diagnostically challenging cases. Diagnostic disagreement occurs throughout the spectrum of benign to malignant lesions [3]. Diagnostic variability is a serious problem as misclassifying breast cancer as benign may lead to fatal progression, and diagnosing a benign lesion as malignant may result in significant morbidity including overtreatment, unnecessary emotional strain, anxiety, and cost. Additionally, misdiagnosis of breast cancer has been a leading cause for malpractice claims for decades, and misclassification undoubtedly undermines research quality and hinders medical progress [4], [5]. A computer-aided diagnostic system that could reduce classification uncertainties would have immediate positive clinical impact.

This paper introduces a self-attention-based network called Holistic ATtention Network (HATNet) for classifying breast biopsy images in an end-to-end manner. HATNet extends the self-attention network of Vaswani et al. [6]. The core principle is to factorize the input biopsy image into words (or patches) using a bag-of-words approach and then encode inter-word and inter-bag relationships in a hierarchical manner using self-attention. Self-attention enables interaction between inputs (bags or words), allowing the encoding of global information in an end-to-end fashion. This helps the network learn clinically relevant tissue structures without any supervision.

HATNet significantly outperforms state-of-the-art methods; it is 8% more accurate and about 2× faster than the previous best network, Y-Net [7], and also matches the classification performance of participant pathologists on the test set. Importantly, this network pays attention to ductal regions and stromal tissues, important bio-markers in breast cancer diagnosis, suggesting that there is clinical relevance in our method. To the best of our knowledge, this is the first work that (1) uses transformers to classify histopathological images in an end-to-end fashion and (2) correlates model decisions with clinically relevant structures. Our source code is available at https://github.com/sacmehta/HATNet.

II. RELATED WORK

Histopathological image classification: Convolutional neural networks (CNNs) are state-of-the-art networks for image classification (e.g., ResNet [8]), including histopathological image analysis [7], [9]–[13]. These methods follow a bag-of-words model for learning representations, wherein an image is treated as a bag while image patches are treated as words (or instances). The first line of research focuses on extracting word-level representations using CNNs, which are then aggregated to produce image-level decisions. Feature selection-based aggregation methods (e.g., [10], [11], [14]) allows identification of relevant features in these word representations.

This work was supported in part by the National Cancer Institute awards (R01 CA172343, R01 CA140560, and RO1 CA200690), NSF (IIS-1616112, IIS1252835), an Allen Distinguished Investigator award, and gifts from the Allen Institute for AI, Google, and Amazon.

S. Mehta and X. Lu contributed equally.

*Corresponding authors: Sachin Mehta, and Ximing Lu

—Training end-to-end networks for classifying gigapixel size histopathological images is computationally intractable. Most approaches are patch-based and first learn local representations (patch-wise) before combining these local representations to produce image-level decisions. However, dividing large tissue structures into patches limits the context available to these networks, which may reduce their ability to learn representations from clinically relevant structures. In this paper, we introduce a novel attention-based network, the Holistic ATtention Network (HATNet) to classify breast biopsy images. We streamline the histopathological image classification pipeline and show how to learn representations from gigapixel size images end-to-end. HATNet extends the bag-of-words approach and uses self-attention to encode global information, allowing it to learn representations from clinically relevant tissue structures without any explicit supervision. It outperforms the previous best network Y-Net, which uses supervision in the form of tissue-level segmentation masks, by 8%. Importantly, our analysis reveals that HATNet learns representations from clinically relevant structures, and it matches the classification accuracy of human pathologists for this challenging test set.

Index Terms—Self-attention, Transformers, Histopathological images, Breast cancer, Image classification

I. INTRODUCTION

BR EAST cancer is the most common cancer in women accounting for approximately 25% of all cancer instances worldwide [1], [2]. Diagnostic classification errors among pathologists can have significant adverse consequences for patients. The “gold standard” for diagnosis of breast biopsy specimens relies on a pathologists visual assessment of tissue sections and cognitive processing of learned cytologic and morphological criteria, including architectural and cellular changes in the tissue, alterations of the tumor micro-environment, and immune-mediated host response. Assessment of these morphological criteria is subjective and pathologists, even expert pathologists, cannot always reach consensus on diagnostically challenging cases. Diagnostic disagreement occurs throughout the spectrum of benign to malignant lesions [3]. Diagnostic variability is a serious problem as misclassifying breast cancer as benign may lead to fatal progression, and diagnosing a benign lesion as malignant may result in significant morbidity including overtreatment, unnecessary emotional strain, anxiety, and cost. Additionally, misdiagnosis of breast cancer has been a leading cause for malpractice claims for decades, and misclassification undoubtedly undermines research quality and hinders medical progress [4], [5]. A computer-aided diagnostic system that could reduce classification uncertainties would have immediate positive clinical impact.

This paper introduces a self-attention-based network called Holistic ATtention Network (HATNet) to classify breast biopsy images in an end-to-end manner. HATNet extends the self-attention network of Vaswani et al. [6]. The core principle is to factorize the input biopsy image into words (or patches) using a bag-of-words approach and then encode inter-word and inter-bag relationships in a hierarchical manner using self-attention. Self-attention enables interaction between inputs (bags or words), allowing the encoding of global information in an end-to-end fashion. This helps the network learn clinically relevant tissue structures without any supervision.

HATNet significantly outperforms state-of-the-art methods; it is 8% more accurate and about 2× faster than the previous best network, Y-Net [7], and also matches the classification performance of participant pathologists on the test set. Importantly, this network pays attention to ductal regions and stromal tissues, important bio-markers in breast cancer diagnosis, suggesting that there is clinical relevance in our method. To the best of our knowledge, this is the first work that (1) uses transformers to classify histopathological images in an end-to-end fashion and (2) correlates model decisions with clinically relevant structures. Our source code is available at https://github.com/sacmehta/HATNet.

II. RELATED WORK

Histopathological image classification: Convolutional neural networks (CNNs) are state-of-the-art networks for image classification (e.g., ResNet [8]), including histopathological image analysis [7], [9]–[13]. These methods follow a bag-of-words model for learning representations, wherein an image is treated as a bag while image patches are treated as words (or instances). The first line of research focuses on extracting word-level representations using CNNs, which are then aggregated to produce image-level decisions. Feature selection-based aggregation methods (e.g., [10], [11], [14]) allows identification of relevant features in these word representations.

This work was supported in part by the National Cancer Institute awards (R01 CA172343, R01 CA140560, and RO1 CA200690), NSF (IIS-1616112, IIS1252835), an Allen Distinguished Investigator award, and gifts from the Allen Institute for AI, Google, and Amazon.

S. Mehta and X. Lu contributed equally.

*Corresponding authors: Sachin Mehta, and Ximing Lu
However, such methods fail to capture the heterogeneity of diagnosis categories [12]. To address the limitations of these methods, multi-instance learning based methods have been proposed [12], [13], [15], [16]. These methods first identify salient instances (or words), which are then combined using different methods (e.g., thresholding, majority voting, and learned fusion) to produce image-level decisions.

The second line of research considers tissue type, size, and distribution to produce image-level decisions [17]–[19]. Instead of identifying salient words, these approaches produce word-level segmentation masks, which are then combined to produce image-level segmentation masks. Tissue-level structural information (e.g., size and distribution) extracted from these masks is then used to produce image-level decisions.

The third line of research integrates both saliency-based and segmentation-based approaches [7], [20]–[22]. These approaches simultaneously produce saliency maps and segmentation masks, which are then combined to extract structural information about tissues and to produce image-level decisions.

Though these methods are effective in classifying histopathological images, the context-capturing ability of saliency-based methods is limited to words and is not able to encode spatial relationships between words. Also, some of these methods (e.g., [12], [13]) require manual threshold selection to identify salient regions. The latter segmentation-based methods address these limitations; however, acquiring tissue-level segmentation labels at a large scale is difficult, because experts are required for annotating images. In this work, we introduce a transformer-based method, HATNet, to address the limitations of existing methods. Similar to previous work, HATNet is based on the bag-of-words model. However, unlike existing methods, it hierarchically aggregates information at different levels of the model using self-attention, which allows learning of spatial relationships between words and bags. HATNet outperforms existing methods (saliency-based or segmentation-based or their combination) by a significant margin (Section V-E). Moreover, this network pays attention to clinically relevant structures (Section VI).

**Spatial attention:** The most widely studied attention mechanism in visual recognition tasks (image classification, segmentation and object detection) is the spatial attention mechanism [23], [24], which weights the activation maps (or spatial planes) to identify regions of interest. Initially introduced to provide explanations for CNN outputs, variants of this mechanism (supervised [25], [26] and unsupervised [27], [28]) have been incorporated in CNNs to improve the performance across different visual recognition tasks (e.g., [29], [30]), including medical imaging (e.g., [31]–[35]). In general, these networks introduce a spatial attention module (e.g., Attention U-Net [31]) within a CNN. Identifying salient regions in histopathological images using spatial attention is difficult because of their large size (usually of the order of gigapixels). This paper introduces an end-to-end transformer-based network for classifying histopathological images.

**Transformers:** Recent work has extended transformers [6] (described in Section III for natural images (e.g., classification [36], segmentation [37], [38] and object detection [39]). Though these approaches are effective in learning local and global representations, extending these approaches to histopathological images is challenging primarily because of their large size (e.g., images in our dataset are $45 \times$ larger than the ImageNet dataset [40]). In this work, we extend transformers using bag-of-words models to classify breast biopsy images in an end-to-end fashion. Specifically, we introduce a bottom-up decoding method that allows us to hierarchically refine the information from words to bags to image and produce diagnostic categories. This hierarchical refinement allows clinically relevant tissue structures to be identified without any explicit supervision. To the best of our knowledge, this is the first work that 1) uses transformers for histopathological image analysis and 2) provides explanations for diagnostic decisions.

### III. BACKGROUND: TRANSFORMERS

Transformer-based attention networks [6] allow inputs to interact with each other, so that the model can automatically find important inputs on which to focus. The transformer module, shown in Figure 1 consists of two parts: (1) multi-head attention that models relationships between inputs, and (2) a feed forward network that learns wider representations.

The multi-head attention module has three projection branches that maps the input $X \in \mathbb{R}^{N \times d}$ to query (Q), key (K), and value (V), where $N$ is the number of inputs (words and bags in this case) and $d$ is the input dimensionality. In particular, each branch consists of $H$ linear layers (also called heads) that projects $X$ from $d$- to $d_h$-dimensional space, allowing us to learn multiple views of the input, where $d_h = d / H$. To allow the module to attend over different input positions, it computes $H$ dot-products between keys (K) and queries (Q) and produces $H$ attention weight matrices, each of size $N \times N$. These resultant $H$ attention weight matrices...
encode the relationships between \( N \) inputs and are combined with the values (V) after a softmax using another dot-product to produce weighted sum outputs. The independent \( H \) outputs are then concatenated and fused using another linear layer \( \beta_{mha} \in \mathbb{R}^{d \times d} \) to produce the output \( X_{mha} \in \mathbb{R}^{N \times d} \). The multi-head attention operation is defined as:

\[
X_{mha} = \text{MultiHead}(X_Q = X, X_K = X, X_V = X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)\beta_{mha}
\]

where \( \text{head}_i = A(X_Q\beta_Q, X_K\beta_K, X_V\beta_V) \) is the output of the \( i \)-th head and \( A = \text{softmax}(\frac{Q^{Ti}K^{T}}{\sqrt{d}}) \). \( V^i \) is the scalar dot-product attention.

The multi-head attention module learns narrower representations in \( d_H \)-dimensional space. To enable the transformer unit to learn wider representations, the output of multi-head attention \( X_{mha} \) is fed to the feed-forward network (FFN). The FFN is a stack of two linear layers. The first linear layer with weights \( \beta_E \in \mathbb{R}^{d' \times d} \) expands the input from \( d \)- to \( 4d \)-dimensional space, while the second linear layer with weights \( \beta_R \in \mathbb{R}^{4d \times d} \) projects back from \( 4d \)- to \( d \)-dimensional space. Mathematically, we can define this operation as:

\[
Y = \text{FFN}(X_{mha}) = \text{ReLU}(X_{mha} \beta_E)\beta_R
\]

This work extends the transformer unit using the bag-of-word model to encode inter-word and inter-bag relationships. This increases the context-capturing ability of the network and improves performance.

IV. HATNet: Holistic Attention Network

State-of-the-art CNN-based classification networks (e.g., ResNet [8]) stack convolutional layers and down-sampling layers to learn representations at multiple scales. These networks are difficult to apply to histopathological images, primarily because the resolution of these medical images (e.g., 10K \( \times \) 10K) are much larger than images used in standard image classification tasks (e.g., 224 \( \times \) 224). To address this resolution challenge, a standard approach is to learn patch-wise (or word-wise) representations using a sliding window method (e.g., [7], [12], [13]). Though these approaches have shown to be effective for histopathological image classification, the context-capturing ability of such approaches is still limited to patch-level. Also, such approaches are difficult to train in an end-to-end manner.

This paper introduces an end-to-end approach that allows diagnostic class prediction using the entire histopathological image at once. Our method extends the transformer architecture [6] using a bag-of-words approach and is shown in Figure 2. We call our model a Holistic Attention Network (HATNet) because of its ability to learn inter-word and inter-bag representations in an end-to-end fashion. With attention, we emphasize the progressive hierarchical refining from words to bags to image to produce the classification output.

Briefly, HATNet first encodes inter-word representations using self-attention (Section IV-A). These representations are then combined to produce bag-level representations (Section IV-B). We then encode inter-bag representations (Section IV-C), which are then combined to produce image-level representations (Section IV-D). These representations are classified to produce the diagnosis category (Section IV-E). Because of the bottom-up decoding (words \( \rightarrow \) bags \( \rightarrow \) image), representations learned using HATNet are expressive and allow the identification of important words and bags in an image. We believe that this will help us build tools to annotate clinically important words and explain diagnosis decisions.

A. Word-to-word attention

The word-to-word attention module (see Figure 2(b)) is comprised of a transformer unit (Section III) with multi-head

\[
\hat{B}_{w2w} = \text{Multi-head attention}(B_{w2w})
\]

\[
B_{w2w} = \text{feed forward network}(\text{FFN})(B_{w2w})
\]

\[
\text{Bag-to-bag attention}(B_{b2b})
\]

\[
B_{b2b} = \text{feed forward network}(\text{FFN})(B_{b2b})
\]

\[
\text{Function}\rightarrow \text{Linear}\rightarrow \text{Softmax}\rightarrow \text{Dot-product} \Rightarrow B_{w2w} \rightarrow \text{concatenated and fused using another linear layer} \beta_{mha} \rightarrow \text{Multi-head attention operation is defined as:}
\]

\[
X_{mha} = \text{MultiHead}(X_Q = X, X_K = X, X_V = X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)\beta_{mha}
\]

where \( \text{head}_i = A(X_Q\beta_Q, X_K\beta_K, X_V\beta_V) \) is the output of the \( i \)-th head and \( A = \text{softmax}(\frac{Q^{Ti}K^{T}}{\sqrt{d}}) \). \( V^i \) is the scalar dot-product attention.

The multi-head attention module learns narrower representations in \( d_H \)-dimensional space. To enable the transformer unit to learn wider representations, the output of multi-head attention \( X_{mha} \) is fed to the feed-forward network (FFN). The FFN is a stack of two linear layers. The first linear layer with weights \( \beta_E \in \mathbb{R}^{d' \times d} \) expands the input from \( d \)- to \( 4d \)-dimensional space, while the second linear layer with weights \( \beta_R \in \mathbb{R}^{4d \times d} \) projects back from \( 4d \)- to \( d \)-dimensional space. Mathematically, we can define this operation as:

\[
Y = \text{FFN}(X_{mha}) = \text{ReLU}(X_{mha} \beta_E)\beta_R
\]

This work extends the transformer unit using the bag-of-word model to encode inter-word and inter-bag relationships. This increases the context-capturing ability of the network and improves performance.
attention and a feed-forward network, allowing us to model the interactions between words and identify important words in the whole slide image.

The input image \( I \in \mathbb{R}^{w \times h} \) with width \( w \) and height \( h \) is first divided into \( n \) non-overlapping bags \( I = (B^1, \ldots, B^n) \in \mathbb{R}^{n \times \frac{w}{h}} \), where \( B^i \) represents the \( i \)-th bag. Each bag \( B^i \) is then divided into \( m \) non-overlapping words \( B^i = (W^{i1}, \ldots, W^{im}) \in \mathbb{R}^{m \times \frac{1}{h}} \), where \( W^{ij} \) represents the \( j \)-th word in the \( i \)-th bag. Similar to previous works (e.g. \cite{7}, \cite{12}), we feed words \( W^i \) inside each bag \( B^i \) to a CNN to produce word-level representations for each bag: \( \hat{B}^{icnn}_{2w} = (\hat{W}^{i1}, \ldots, \hat{W}^{im}) \in \mathbb{R}^d \). The representations from the CNN do not encode inter-word relationships. We encode inter-word relationships in each bag \( B^{icnn} \) using the transformer unit (Section III) to produce bag-level representations using a function \( \Psi \) and linear transformation \( \beta_{w2b} \in \mathbb{R}^{m \times m} \) to produce bag-level representations, \( \hat{B}^i_{w2b} \in \mathbb{R}^d \):

\[
\hat{B}^i_{w2b} = \text{softmax}(\Psi(\hat{B}^{icnn}_w) \beta_{w2b}) B^{icnn}_w, \quad 1 \leq i \leq n \quad (5)
\]

C. Bag-to-bag attention

The bag-level representations \( B_{w2b} \) and \( \hat{B}_{w2b} \) do not encode information about surrounding bags. To encode inter-bag relationships, we apply bag-to-bag attention (see Figure 2(d)). The bag-to-bag attention module is similar to the word-to-word attention (Section IV-A), except that we use \( \hat{B}_{w2b} \) (Eq. 5) as context to \( B_{w2b} \) (Eq. 4). This also mimics the typical skip-connection mechanism in neural networks \cite{8}, \cite{41}.

We first apply multi-head attention to \( \hat{B}_{w2b} \) to encode inter-bag representations and produce \( B_{b2b} \in \mathbb{R}^{n \times d} \) as:

\[
B_{b2b} = \text{Multihead}(X_Q = \hat{B}_{w2b}, X_K = \hat{B}_{w2b}, X_V = \hat{B}_{w2b}) \quad (6)
\]

To allow every bag obtained from CNN \( \hat{B}_{b2b} \) to attend over every bag \( \hat{B}_{w2b} \) obtained after word-level self-attention, we apply another multi-head attention in which \( \hat{B}_{b2b} \) serves as a query and \( \hat{B}_{w2b} \) serves as keys and values, allowing us to encode rich inter-bag representations and produce \( B_{b2b} \in \mathbb{R}^{n \times d} \). Mathematically, we can define the bag-to-bag attention operation as:

\[
B_{b2b} = \text{FFN}(\text{Multihead}(X_Q = \hat{B}_{b2b}, X_K = \hat{B}_{w2b}, X_V = \hat{B}_{w2b})) \quad (7)
\]

D. Bag-to-image attention

The inter-bag representations are encoded in \( B_{b2b} \in \mathbb{R}^{n \times d} \). We aggregate these representations to produce image-level representations. Similar to word-to-bag attention (Section IV-B), we combine these bag-level representations using a function \( \Psi \)
and linear transformation $\beta_{d2i} \in \mathbb{R}^{n \times n}$ to produce image-level representations $I_{d2i} \in \mathbb{R}^d$.

$$I_{d2i} = \text{softmax} (\Psi (B_{d2b}) \beta_{d2i}) B_{d2b} \quad (8)$$

Because of the bottom-up decoding (words to bags to image), these representations are expressive and allows us to identify important words and bags in an image (Figure 3).

E. Classification and Loss

We classify $I_{d2i} \in \mathbb{R}^d$ into $C$-diagnosis classes using a linear classifier with weights $\beta_{cls} \in \mathbb{R}^{d \times C}$ as:

$$\hat{y} = \text{softmax} (I_{d2i} \beta_{cls}) \quad (9)$$

We minimize the cross-entropy loss $\mathcal{L}$ between the ground truth $y$ and prediction $\hat{y}$ to train HATNet. During evaluation, we choose the index that has the highest confidence score in $\hat{y}$ as our predicted class label.

V. EXPERIMENTAL RESULTS

A. Dataset and evaluation

The breast biopsy dataset [3] consists of 240 whole slide images with haematoxylin and eosin (H&E) staining. The image dataset was designed to include a higher prevalence of cases from diagnostic categories that have lower prevalence in the general population. This provides a robust and more challenging image dataset. A total of 87 pathologists participated in a previous study and interpreted these cases. Each pathologist classified a random subset of 60 slides into three expert pathologists then interpreted these cases and provided a consensus label per slide. We treat these consensus labels as ground truth diagnostic labels. The pathologists also marked 420 regions of interest (ROIs) that best supported the diagnoses. Following previous studies on this dataset [7], [13], [15], [17], [18], we use these ROIs to train and evaluate our method using the following metrics: (1) classification (or top-1) accuracy, (2) F1-score, (3) sensitivity, (4) specificity, and (5) area under receiver operating characteristic curve (ROC-AUC). In particular, the dataset consists of 164 training, 42 validation, and 216 test ROIs (see Table I).

| Diagnostic Category | Number of ROIs | Average size (in pixels) |
|---------------------|----------------|-------------------------|
| Benign              | 48             | 13                      | 64                     | 125                     | 6733 \times 5839 |
| Atypia              | 40             | 8                       | 54                     | 102                     | 10668 \times 8967 |
| DCIS                | 60             | 17                      | 84                     | 161                     | 10778 \times 9547 |
| Invasive            | 16             | 4                       | 14                     | 34                      | 23866 \times 21402 |
| **Total**           | **164**        | **42**                  | **216**                | **422**                 | **10880 \times 9558** |

TABLE I: Statistics of breast biopsy whole slide image dataset.

B. Architecture

The size of the ROIs is variable. To use computational resources efficiently, each ROI is resized to spatial dimensions of about $12544 \times 12544$. The resized ROIs are then split into $n = 49$ non-overlapping bags, each bag with a spatial dimension of $1792 \times 1792$. Each bag is further split into $m = 49$ non-overlapping words, resulting in words with a spatial dimension of $256 \times 256$. Overall, each ROI has 2401 words. These words are fed to off-the-shelf CNNs to extract word-level representations. In our experiments, three state-of-the-art light-weight CNNs pretrained on the ImageNet dataset [42] are studied: (1) ESPNetv2 [43], (2) MobileNetv2 [44], and (3) MNASNet [45]. ESPNetv2 follows an Inception-style design [46] and uses four simultaneous $3 \times 3$ depth-wise convolutions with different dilation rates. MobileNetv2 follows ResNet-style design [8]. To improve the computational efficiency, MobileNetv2 uses $3 \times 3$ depth-wise convolutions instead of $3 \times 3$ standard convolutions. MNASNet uses the same basic building block as MobileNetv2; however, it uses neural architecture search [47] to identify the optimal model configuration, which provides best trade-off between different parameters. The proposed network is generic and any off-the-shelf heavy-weight CNNs (e.g., ResNet [8]) can be used to extract word-level representations. Heavy-weight networks are not explored because of computational constraints.

The dimensionality of word-level representations varies from CNN to CNN. Therefore, the output of a CNN is linearly projected to a 256 dimensional space ($d = 256$). To encode the inter-word and inter-bag representations, 4 heads ($H = 4$) are used in multi-head attention, resulting in a head dimension of $d_h = 64$. We use the function $\Psi$ to aggregate word-level representations into bag-level representations (Section IV-B) and bag-level representations into image-level representations (Section IV-D). In our experiments, we study three different functions: (1) Euclidean distance (or L2 norm), (2) Manhattan distance (or L1 norm), and (3) mean of a vector.

C. Training

Our models are trained end-to-end using the ADAM [48] optimizer with a learning rate warm-up strategy. We first warm-up the learning rate from $10^{-7}$ to $10^{-5}$ in 600 iterations and then train the model for the next 50 epochs with a learning rate of $10^{-4}$. After that, learning rate is decayed by half and then train for another 50 epochs. Our model takes about 36 hours for training on two NVIDIA GeForce GTX 1080 GPUs, each with a memory of 8 GB. We accumulate gradients for 8 iterations before updating the weights, yielding an effective batch size of 8 ROIs per update. Training data is augmented by randomly resizing ($192 \times 192$, $224 \times 224$, $256 \times 256$, $288 \times 288$, $320 \times 320$), flipping, and rotating (angle: $-10^\circ$ to $10^\circ$) the words. For evaluation, a single model is obtained by averaging the best 5 validation checkpoints. Compared to the best model, averaged models delivered 1-1.5 points higher accuracy.

D. Baseline networks

We compare our method with the following methods:
Network parameters are reported for single models only. We use majority voting for ensembling the models.

### TABLE II: Comparison with state-of-the-art networks

| Row | Model | Parameters | Evaluation metrics |
|-----|-------|------------|--------------------|
| R1  | Pathologists (avg. of 70 pathologists) [7] | | 0.70 0.71 0.70 0.90 |
| R2  | LAB & LBP hand-crafted features (w/o saliency) | | 0.28 |
| R3  | LAB & LBP hand-crafted features (w/ saliency) | | 0.45 |
| R4  | Bag-of-word (majority voting w/o saliency) | | 0.23 |
| R5  | Bag-of-word (majority voting w/ saliency) | | 0.55 |
| R6  | Bag-of-word (learned fusion w/o saliency) | | 0.38 |
| R7  | Bag-of-word (learned fusion w/ saliency) | | 0.55 |
| R8  | MRSegNet with histogram and co-occurrence features | 26.03 M NA | 0.55 0.56 0.55 0.85 |
| R9  | MRSegNet with structural features | 26.03 M NA | 0.56 0.57 0.56 0.85 |
| R10 | Y-Net | 3.91 M NA | 0.62 0.62 0.62 0.87 |
| R11 | HATNet (w/ ESPNetv2) | 2.21 M 2.37 M | 0.67 0.64 0.67 0.89 0.89 |
| R12 | HATNet (w/ MobileNetv2) | 2.22 M 2.37 M | 0.66 0.65 0.66 0.89 0.88 |
| R13 | HATNet (w/ MNASNet) | 3.10 M 2.37 M | 0.70 0.70 0.70 0.90 0.90 |
| R14 | HATNet (Ensemble) | NA NA | 0.71 0.70 0.71 0.90 0.90 |

**TABLE II: Comparison with state-of-the-art networks.** HATNet outperforms existing methods by a significant margin. Network parameters are reported for single models only. We use majority voting for ensembling the models.

1) **Bag-of-words model with hand-crafted features** [13]: An input image (bag) is split into words. Following [49], LAB histogram and LBP histogram features are extracted from these words. These word-level features are concatenated and then classified using logistic regression into diagnostic categories with and without saliency (see R2 and R3 in Table II). Similar to a standard practice in saliency-based approaches (e.g., [7], [12], [50]), the class with majority voting in saliency maps is selected as a diagnostic category.

2) **Bag-of-words model with deep features** [13]: Instead of using hand-crafted features, this approach uses a deep convolutional neural network, FCN [51], to obtain word-level representations. These representations are used to identify discriminative or salient regions. In addition to majority voting-based method, a learned fusion method [12] is also tried to model the relationships between words (see R4-R7 in Table II).

3) **Multi-resolution segmentation network (MRSegNet)** [17]: MRSegNet has two stages: (1) tissue-level segmentation and (2) diagnostic classification. The first stage is a multi-resolution encoder-decoder network which combines the outputs of many words (or patches) at different resolutions to reduce segmentation errors. In the second stage, histogram and co-occurrence features are extracted from tissue-level segmentation masks, which are then classified using a multi-layer perceptron into different diagnostic classes (see R8 in Table II).

4) **Structural features** [18]: This method extracts structural features from tissue-level segmentation masks produced using MRSegNet. These features allow capturing structural changes in ductal regions, an important biomarker for breast cancer diagnosis [52]–[55], in the breast (see R9 in Table II).

5) **Y-Net** [7]: Y-Net adds a classification branch to U-Net [41] and allows to jointly predict the tissue-level segmentation mask and the saliency map. The saliency map is then combined with the tissue-level segmentation mask to produce a discriminative segmentation mask. Similar to MRSegNet, Y-Net extracts histogram and co-occurrence features, which are then used for classify diagnostic classes (see R10 in Table II).

### E. Main Results

Table II shows that HATNet outperforms state-of-the-art methods significantly and delivers performance comparable to pathologists. For example, HATNet (R13) improves the performance of the best saliency-based models (R5, R7) by about 15%. When compared to approaches that uses tissue-level segmentation masks (R8-R10) to capture the structural changes in biopsies, HATNet delivers better performance. In particular, HATNet improves the F1-score of the previous best segmentation-based approach (R10) by 8%. Overall, these results shows that HATNet is effective. We note that ensembling these three models (R14) further improves the accuracy and sensitivity by 1%.

Furthermore, Table III shows that HATNet is fast. HATNet with MNASNet is about 1.8× faster than the previously best reported network, i.e., Y-Net.

### F. Model Ablations

**Effect of function Ψ:** Table IV compares the performance of three different Ψ functions with ESPNetv2 as a base feature.
Fig. 4: Class-wise performance comparison of HATNet with different CNN architectures. Overall, the models with MNASNet as a base feature extractor performs a little better than the other two networks across different metrics. However, MNASNet has a low sensitivity score for Invasive Cancer, while MobileNetv2 does much better in this regard.

Fig. 5: Receiver operating characteristic (ROC) curves of HATNet with different CNN architectures. The models with MNASNet as a base feature extractor has slightly higher area under curve (AUC) than the other two.

**TABLE IV**: Effect of different $\Psi$ functions.

| Function  | $\Psi$ | Accuracy | F1-score | Sensitivity | Specificity | ROC-AUC |
|-----------|--------|----------|----------|-------------|-------------|---------|
| Euclidean distance | 0.67 | 0.64 | 0.67 | 0.89 | 0.89 |
| Manhattan distance | 0.66 | 0.64 | 0.66 | 0.88 | 0.89 |

**TABLE V**: Effect of different bag and word sizes. Note the number of words in each configuration are the same (i.e. 49).

| Bag size | Word size | Accuracy | F1-score | Sensitivity | Specificity | ROC-AUC |
|----------|-----------|----------|----------|-------------|-------------|---------|
| 1792 x 1792 | 256 x 256 | 0.67 | 0.64 | 0.67 | 0.89 | 0.89 |
| 2016 x 2016 | 288 x 288 | 0.67 | 0.64 | 0.67 | 0.89 | 0.88 |
| 2240 x 2240 | 320 x 320 | 0.66 | 0.64 | 0.66 | 0.88 | 0.89 |

VI. ANALYSIS

Several clinical studies have shown that ductal regions and stromal tissue are important bio-markers for diagnosing breast cancer (e.g., [2], [52]–[57], [57]–[59]). Briefly, ducts are thin tubes in the breast and are responsible for carrying milk from lobules (milk glands) to the nipples. These regions are useful in identifying cancers that began in milk ducts (e.g., DCIS [53]–[55]). On the other hand, the stroma is part of the breast tissue with a structural and developmental role and may be involved in tumor promotion and progression. Many clinical studies have underlined the importance of stroma in tumor progression.
Fig. 6: HATNet identifies ducts of variable size and texture as an important structure. In (a-f), ductal regions (marked by pathologists) are shown in red, while the top-50% bags predicted by HATNet are shown in blue. In (g), the dice score is plotted between ductal regions and top-k bag predictions (k varies from 10% to 60%) for different diagnostic classes.

Fig. 7: HATNet identifies stroma as an important tissue. In (a-f), each sub-figure is organized from left to right as: breast biopsy image, stroma tissue labeled by pathologists, and the top-50% words (words that belong to stroma tissue are shown in pink while the remaining words are shown in blue) identified using our model. The remaining 50% words are shown in white. In (g), we plot the dice score between stromal tissue and top-k word predictions (k varies from 10% to 60%) for different diagnostic classes.

along with its contribution to risk factors that determines tumor formation [56], [57].

It was shown above that our model learns better representations, resulting in significant performance gains compared to existing methods (Table II). A closer analysis (see Figure 6 and 7) reveals that our model pays attention to these important bio-markers, which helps it to achieve these gains.

**Ductal regions:** To evaluate if our model pays attention to ductal regions or not, we compute the overlap between ductal regions (marked by experts) and top-k bag predictions of our model using dice score. Results are shown in Figure 6g. When considering top-50% bag predictions, HATNet achieves a dice score of 0.68. This shows that HATNet identifies ductal regions as an important structure. Furthermore, Figure 6g shows that HATNet is able to differentiate between ducts of variable size and texture. This shows that HATNet is effective in modeling inter-bag and inter-word relationships, which helps it improve performance.

**Stromal tissue:** We compute the overlap between stromal tissue (pixel-level annotations by pathologists) and top-k word predicted by our model to determine whether our model pays attention to stromal tissue. We use dice score to measure the overlap and vary k from 10% to 60%. Figure 7g shows that HATNet achieves a dice score of about 0.75 when top-50% word predictions are considered. This indicates that
HATNet also identifies stroma as an important tissue. This is further strengthened by visualizations in Figures 7a–7f which shows that majority of top-50% words lie in stromal tissue.

VII. CONCLUSION

This paper introduces an end-to-end attention-based network for classifying breast biopsy images. HATNet extends bag-of-words models using transformers 6 to learn global representations. Our approach effectively aggregates inter-word and inter-bag representations, which helps the model to learn representations from clinically relevant tissue structures. HATNet improves the state-of-the-art significantly and matches the classification performance of pathologists on the test set. Furthermore, HATNet learns representations from clinically relevant structures. We believe that top words and bags identified using our method will help us build tools to annotate cell-level structures (e.g., mitoses), which would help in providing detailed explanation of diagnosis decisions. In the future, we plan to build such tools and also apply HATNet to other histopathological images, including melanoma.

REFERENCES

[1] J. Makki, “Diversity of breast carcinoma: histological subtypes and clinical relevance,” *Clinical Medicine Insights: Pathology*, vol. 8, pp. 153–156, 2015.

[2] C. E. DeSantis, J. Ma, M. M. Gaudet, L. A. Newman, K. D. Miller, A. Goding Sauer, A. Jemal, and R. L. Siegel, “Breast cancer statistics, 2019,” *CA: a cancer journal for clinicians*, vol. 69, no. 6, pp. 438–451, 2019.

[3] Elmore et al., “Diagnostic concordance among pathologists interpreting breast biopsy specimens,” *JAMA*, 2015.

[4] A. K. Kern, “The delayed diagnosis of breast cancer: medicolegal implications and risk prevention for surgeons,” *Breast disease*, vol. 12, no. 1, pp. 145–158, 2001.

[5] L. M. Reisch, P. A. Carney, N. V. Oster, D. L. Weaver, H. D. Nelson, P. D. Frederick, and J. G. Elmore, “Medicolegal practice concerns and defensive medicine: a nationwide survey of breast pathologists,” *American journal of clinical pathology*, vol. 144, no. 6, pp. 916–922, 2015.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[7] S. Mehta, E. Merican, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro, “Y-net: joint segmentation and classification for diagnosis of breast biopsy images,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 893–901.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] D. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, “Mitosis detection in breast cancer histology images with deep neural networks,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2013, pp. 411–418.

[10] Y. Xu, Z. Jia, Y. Ai, F. Zhang, M. Lai, I. Eric, and C. Chang, “Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 947–951.

[11] A. Cruz-Roa, A. Basavanahally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, “Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks,” in *Medical Imaging 2014: Digital Pathology*, vol. 9041. International Society for Optics and Photonics, 2014, p. 904103.
