Double Attention for Pathology Image Diagnosis Network with Visual Interpretability

1st Hao Cheng
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China
jiaodachenghao@sjtu.edu.cn

2rd Kaijie Wu
Department of Automation
Shanghai Jiao Tong University
Shanghai, China
kaijiewu@sjtu.edu.cn

3rd Kai Ma
Department of Automation
Shanghai Jiao Tong University
Shanghai, China
makaay@sjtu.edu.cn

4th Jie Tian
Department of Automation
Shanghai Jiao Tong University
Shanghai, China
tianjie_13@sjtu.edu.cn

5th Rui Xu
Department of Automation
Shanghai Jiao Tong University
Shanghai, China
xuruihaha@sjtu.edu.cn

6th Chaochen Gu
Department of Automation
Shanghai Jiao Tong University
Shanghai, China
jacygu@sjtu.edu.cn

7th Xinping Guan
Department of Automation
Shanghai Jiao Tong University
Shanghai, China
xpguan@sjtu.edu.cn

***Abstract*—In recent years, cervical cancer has been one of the most common diseases in women's cancer. The advanced diagnosis of cervical precancerous lesions is essential for preventing cervical cancer. Its effectiveness and efficiency can be greatly improved by computer aided diagnosis, while challenged by the imprecise conclusions and uninterpretable process of diagnosis. To solve this problem, we propose a novel deep learning-based interpretable diagnosis system for pathology images, consisting of three interrelated models: an image model, an attention model and a conclusion model.**

Computer aided diagnosis improves the effectiveness and efficiency of the proposed image model uses a convolutional neural network (CNN) to ex-tract semantic features. Combining the model with the semantic attribute attention model, it aims to capture the discriminant relationship between se-mantic attributes by predicting the conclusion label through long-term and short-term memory (LSTM). The network is trained in an end-to-end manner, with different weights for each model. Experimental results on cervical intraepithelial neoplasia images, diagnostic reports and label datasets show that the proposed method achieves a significant improvement over traditional methods with a better interpretability.

Keywords—deep learning, visual interpretability, cervical precancerous lesions.

I. INTRODUCTION

Cervical cancer is one of the two most common malignant tumors in women, ac-counting for the majority in the female reproductive system. In recent years there has a resurgence, with the incidence of cervical cancer in developing countries being significantly higher than in developed countries. Every six minutes, on average, a woman dies of cervical cancer. China account, which accounts for nearly one sixth of the world's population, has a disproportionately high rate of cervical cancer, accounting for about one fourth of the global total. The five years survival rate amongst patients receiving early treatment stands at nearly 100%, while the five year survival rate for those receiving treatment in advanced stages is between 20 and 50%. It is the only malignant tumor type with clear etiology, where high-risk groups may be identified through cervical cancer screening. The lesion may be identified and treated early, in the precancerous stage, thus preventing transition into malignant tumors.

Screening results of cervical precancerous lesions are directly linked to the ability of trained pathologists. However, in China, trained pathologists number in the tens of thousands, which is far from adequate to allow diagnosis of such a large number of pathological sections. Moreover, in many cases, the workload and resulting stress on the pathologist may lead to misdiagnosis. Therefore, given these circumstances, it is very important to leverage technology, such as introducing convolutional neural net-works, to assist in automation of medical diagnostics.

With rapid development of deep learning, in recent years, significant progress has been made in the area of computer-aided medical diagnosis (CAD). Traditional deep learning methods treat processes as standard classification problems [1]. However, it is less efficient and less effective in diagnosing many diseases, such as Kidney renal clear cell carcinoma (KIRC) [2]. The reason is that the classification model simplifies the actual diagnosis process and lacks discriminant information supporting the conclusion. Doctors, even professional pathologists, often find it difficult to understand how models interpret features and make diagnostic conclusions. Therefore, a description is needed in order to address this challenge and support the clinical decision-making process.

In clinical practice, the process and output provided by a deep learning model is effective and important in the CAD process only after having been evaluated by a pathologist. Ideally, the model should capture the distinguishing features from the pathological image and generate written descriptions, as well as contextual and visual attention cues to assist the pathologist to reach their decision. For physicians, this method is more efficient and convenient than models which only produces conclusion labels.

This has prompted us to further investigate the premise for a model which automatically captures latent and discriminative features, then generates a corresponding report and visual attention map. Three key challenges remain to be addressed:

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

- The structured report has to contain different semantic attribute descriptions to support the conclusion.
- The output of the model must be clearly understood by medical professionals.
- The model needs to generate the discriminative attention map of semantic at-tribute label.

The above challenges are addressed through the implementation of an end-to-end network which consists of image, attention and conclusion models. The image model based upon CNNs extracts the discriminative features from the pathology image. A new method is proposed for the attention model, which generates the visible attention map and the structured report. The attention model is treated as a multi-label classification task so that the model generates full-structured context. In this approach, the conclusion model, combined with the attention model, possesses the ability to learn the contextual dependencies among the semantic attributes with the "memory" of LSTM for conclusion making.

Moreover, the main contributions of our method are:

- A new approach is proposed which can generate the structured report and support the conclusion of interpretable vision for pathology images.
- Construction of the pathology cervical intraepithelial neoplasia images, diagnostic report and attribute labels (CINDRAL) dataset (explained later in Section 3) with Shanghai International Peace Maternity and Child Health Hospital (IPMCH).
- Using two ways of attention method in our model.
- Undertaking extensive experimentation and evaluations on the CINDRAL dataset, and demonstrate accuracy and effectiveness of the approach.

The remainder of the paper is organized into several sections, beginning with a re-view of related work. This is followed by a description of the dataset, then by an explanation of the proposed model, which includes the extracted features and address-ing attention model's functionality. Finally, detailed performance studies and analysis on CINDRAL datasets are disclosed.

II. Related Work

A. Image and Conclusion Model

Recent advances in the performance of medical diagnostics [3,5,7,14,30] have achieved rapid progress as a result of the development of CNNs[4] and the memory mechanisms of recurrent neural networks (RNNs). Traditional methods treat the CAD as a classification problem, in cases such as skin lesions[6], lung squamous cell carcinoma [1], and conclusion of pathological images [8].

However, these methods for CAD typically aimed at finding one particular type of disease, concealing correlations between semantic attributes. Some researchers have considered latent dependent information from the report and pathology images by LSTM [11] in order to overcome this issue. The construction of CNN-RNN based framework model to predict the conclusion label of chest X-rays, is a prime example [9]. This method implements CNN to extract the feature of a disease and RNN to describe the attributes of the disease. Another methods uses CNN to obtain visual and semantic features from chest X-rays while using hierarchical LSTM to get a more intuitive report and conclusion [10]. The work most closely related to medical report and diagnosis generation was recently contributed by Zhang[12]. This group proposed the CNN-LSTM model to describe the semantic attribute report, which generates the conclusion. However, some words (e.g. "along", "the", and "is") present in their report (e.g. Polarity along the basement membrane is negligibly lost) contain no medically significant information supporting the conclusion. In this case, the most important word in the sentence is "negligibly," which provides the discriminative features to help the LSTM draw the conclusion.

Following the work of Wang et al.[13], this problem can be addressed by changing the conclusion problem into one of multi-label classification which allows the conclusion model to obtain accurate semantic attribute description labels. Our methods can generate the attribute report including the conclusion, which possess more accurate performance than the previously mentioned methods.

B. Attention Model

There has been a significant amount of research focused on the attention mechanism which achieves textual and visual interpretability in traditional natural image datasets, such as ImageNet [4], and other related methods [23,30]. For the attention model, our work is similar to several previous works [15-19]. Xu et al.[18] proposed the sequence to sequence model and attention model in the image captioning task. Within the context of that work, the attention map was determined using the CNN features and the previously hidden states of LSTM. Pedersoli et al. [20] provided associations between the attention map and caption words.

These above mentioned works have inspired further research in an effort to spur ongoing improvements to CAD, with several works focused on generation of reports for medical images. Additionally, these works focus on interpretable visual diagnosis [8,10,12,21,22], which can support the pathologist's decision-making process.

MDNET [12] introduced an attention model which focuses on the image region while every word is generated from the model. However, words such as, "like," "the," "a," and "along," amongst others, have no factual relationship to the attention map of the image.

In an attempt to address the problem, our method aimed at the visual interpretation of semantic attributes (e.g. in CINDRAL dataset, there are four types), rather than focusing on single words from the report.

III. Dateset

The cervical intraepithelial neoplasia images, diagnostic report and labels (CINDRAL) dataset was collected in collaboration with the Shanghai International Peace Maternity and Child Health Hospital (IPMCH). Whole-slide images (WSI) of stained tissue sections, captured under twenty times magnification, were obtained from fifty patients at risk of cervical neoplasm. One thousand, 600x600 RGB images were randomly selected from the dataset.

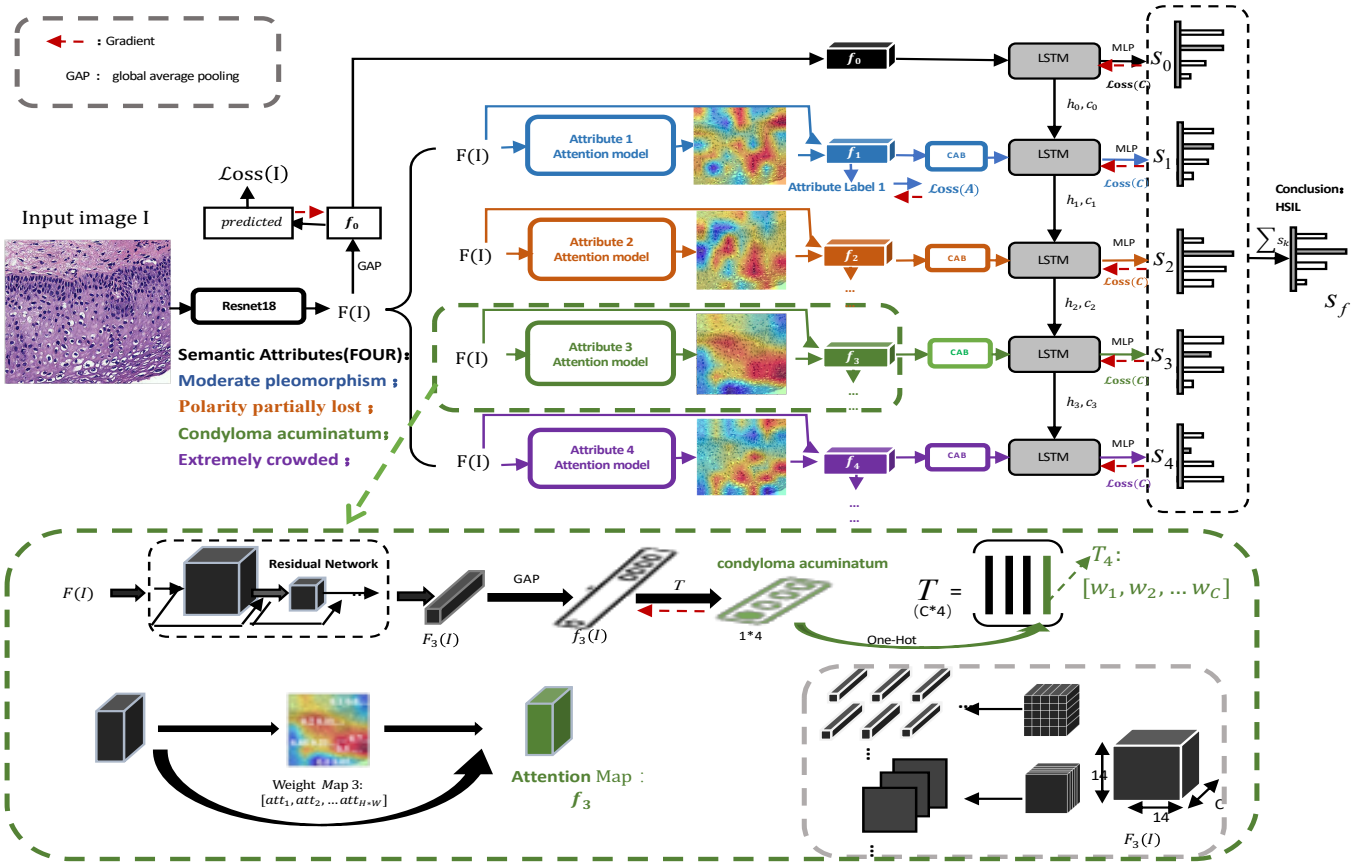

Fig. 2. The overall illustration of our model. A pathology image with its structured report and labels, presented as an example. Image model extracts the features, attention model demonstrates the attention map and the structured report, while conclusion semantic attributes, each with four state labels).

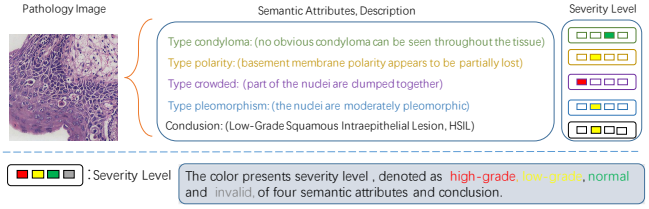

Fig. 1. One example in the CINDRAL dataset. It consists of a pathology image and a structured report (four semantic attributes, each with four state labels). No extra locations on images are needed in the dataset. (Color available online.)

The pathologists provided a paragraph describing four semantic attribute features (Fig. 1). These attributes included, the state of condyloma, cell polarity, cell crowding, and nuclear pleomorphism, followed by a diagnostic conclusion. The attributes and the conclusion are both grouped into one of four labels; normal, high-grade, low-grade, and insufficient information. Contained within different reports, each description of semantic attributes has two or three very similar descriptions (i.e. “portions of the nuclei are clumped together” and “extremely crowded nuclei can be seen”). Thus there are five sentences, four attributes and one conclusion, per image.

The dataset was pre-processed by cropping, rotating (90°,180° and 270°), and/or horizontal/vertical flipping for the purpose of data augmentation. We randomly selected 20% of the images as test data and the remaining 800 images were used for training and cross-validation. From the dataset, the four attributes and the conclusion are treated as five separate tasks for the structured report and LSTM, trained to support the conclusion model.

IV. Method

The method [25] (Fig. 2) uses a residual network (ResNet18) which is faster to train and achieves similar or better performance than most commonly used VGG16 [24] or AlexNet [31] models, on the training images. The network processes the image with convolutional layers to extract the feature map, denoted as F(I). Then the weighting is obtained from every attribute attention model and the conclusion model, comprised of an LSTM network, predicts the final result from the attention map. Finally, the scores from four attention maps are fused to achieve the final conclusion label distribution.

A. Attention Model with Structured Report

A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible

(for example, do not differentiate among departments of the same organization).

The attention mechanism, reported by Xu[18], is focused on learning the attention map for the whole image so that it can support the final prediction. However, in this work, we focus on the relationship between four semantic attributes to implicitly provide the discriminate information for LSTM to draw the conclusion. Therefore, we propose attention models which present the four semantic attributes of the CIN images. The attention model dynamically computes a weight map for every attribute, which presents the visual attention region corresponding to input image.

For one attribute (Fig.2), we duplicate the feature $F(I)$ four-fold, since every residual network in the attention model has the ability to learn the accurate features for each attribute with different parameters. The residual network extracts features, denoted as $F_k(I)$ with a dimension of $256 \times (14 \cdot 14)$, with k={1,2,3,4} representing the four attributes. After a global average pooling layer, a feature denoted as $f_k(I)$ with a dimension of 1×256, is obtained from $F_k(I)$.

Specifically, following [27], the weight map can be computed as follows:

$$O_k = \text{softmax}(f_k(I)T_k + b) \quad (1)$$

$$T_k^i = S_k T_k \quad (2)$$

$$W_k = (T_k^i)^T F_k(I) \quad (3)$$

Where T_k is a learned fully connection layer parameter with a dimension of 256×4. $S_k, k = \{1,2,3,4\}$ is the one-hot representation of the k-th image attribute generated by O_k. Then we may obtain one column of T_k, denoted as $T_k^i, i = \{1,2,3,4\}$ with a dimension of 256×1, and containing the discriminate information of the k-th semantic attribute. Finally, the weight map, denoted as W_k with dimension 14×14, which corresponds to the attribute's attentional regions in the input image is generated by bilinear interpolation.

The matrix W_k presents the visual interpretability to the pathologists, indicating which regions the attention model focuses on. The weight map W_k and the feature map $F_k(I)$ need to work in a cohesive manner so that the 256-dim feature f_k can contain the accurate k-th attribute feature selected by weight map (Fig. 2).

$$f_k = W_k(F_k(I))^T \quad (4)$$

Two loss functions are employed in order to better train the model. Initially, we use the semantic attributes label for the feature $f_k(I)$, and then apply the severity level label of each attribute for the feature f_k. The motivation behind this is two-fold. First, the feature generated by the ResNet can better extract the attribute information which is critical for next procedure. Second, a structured diagnostic report can be generated by the second loss function which can provide the information regarding the symptoms for every attribute. The two loss functions serve a supervisory role on the attention model, which can ensure the attention model training produces accurate semantic features for the next conclusion model.

We use the channel attention block (CAB) in our model (Fig. 3). Within the attention block, the weight channel can be extracted by the feature map1, which represents the weight of each channel. And then, we multiply the weights on each channel. The block "chooses" the discriminative feature for the conclusion module.

In the experiment, we compared the differences in the use of regional attention mechanisms and individual channel attention mechanism. Two different attention mechanisms can bring different enhancement effects to the model. The regional attention mechanism is more about telling us which area is more important. At the same time, the channel attention mechanism tells us more about which channels are the more important feature channels in the obtained features.

B. Conclusion Model

Feedback from the pathologists shows that symptom descriptions of semantic attributes as well as the latent relationship between attributes both support the conclusion. Considering relevance and dependence, we adopted the LSTM network to draw the conclusion.

There are numerous previous works on the generation of diagnostic reports using image captions as the input for LSTM training[8,12,18,21,22]. However, diagnostic reports often contain words which present no medically significant information. These words provide less accurate features from which to draw the conclusion. The proposed method treats the discriminate attribute features as the input to LSTM, in order to directly consider the critical feature as part of the conclusion.

Following [11], LSTM is defined by the following equations:

$$x_k = relu\big(W^{(x)} f_k + b_x\big), \text{k} \neq 0 \quad (5)$$

$$i_k = sigmoid\ (W^{(i)} x_k + U^{(i)} h_{k-1} + b_i) \quad (6)$$

$$f_k = sigmoid\ (W^{(f)} x_k + U^{(f)} h_{k-1} + b_f) \quad (7)$$

$$o_k = sigmoid\big(W^{(o)} x_k + U^{(o)} h_{k-1} + b_o\big) \quad (8)$$

$$\tilde{c}_k = tanh\ (W^{(c)} x_k + U^{(c)} h_{k-1} + b_c) \quad (9)$$

$$c_k = f_k * c_{k-1} + i_k * \tilde{c}_k \quad (10)$$

$$h_k = o_k * tan\, h(c_k) \quad (11)$$

$$z_k = relu(W^{(z)} h_k + b_z) \quad (12)$$

$$s_k = W^{(s)} z_k + b_s \quad (12)$$

Where the computation process represents k-th ($k \neq 0$) LSTM network, and f_k in equation (5) is a 256-dim vector containing the attribute information to support conclusion; h_{k-1} and c_{k-1} are the hidden states and memory cells of the previous LSTM; The k-th LSTM hidden state h_k is used to predict score distribution, denoted as s_k, of the conclusion by equation (12) and (13). Note that we initialize the hidden state with the feature f_0 extracted from an image, and the initial process for the LSTM network is set as:

$$h_0 = f_0 \quad (13)$$

$$z_0 = relu(W^{(z)} h_0 + b_z) \quad (14)$$

$$s_0 = W^{(s)} z_0 + b_s \quad (15)$$

We obtain the final predicted conclusion, denoted as s_f, which is a normalized exponential function, by adding the distributions of the five predicting scores, s_0, s_1, s_2, s_3, s_4 :

$$s_f = s_0 + s_1 + s_2 + s_3 + s_4 \qquad (17)$$

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

C. Network Optimization

The overall model has three sets of parameters: θ_I in the image model I, θ_A in the attention model A, θ_C in the conclusion model C. The overall optimization problem in our method is expressed as:

$$\max_{\theta_1,\ \theta_A,\ \theta_C} \begin{matrix} \mathcal{L}_I\big(l_c, I(I;\theta_1)\big) + \mathcal{L}_A(l_s, A[I(I;\theta_1);\theta_A]) \\ +\mathcal{L}_C(l_c, C\{A[I(I;\theta_1);\theta_A];\theta_C\}) \end{matrix} \qquad (18)$$

where (I, l_c, l_s) is a training tuple: I is a pathology image, l_c denotes the conclusion label, and l_s is the semantic attribute label. Modules I, A and C are supervised by three negative log likelihood loss functions $\mathcal{L}_I$, $\mathcal{L}_A$ and $\mathcal{L}_C$.

During the training stage, Adam and standard back-propagation are employed to optimize the joined model. For end-to-end training, we treat the loss function as two stages with different weights and learning rates. Thus, the training loss is computed as:

$$\mathcal{L}\text{oss}_{I,A,C} = \lambda(\mathcal{L}\text{oss}_I + \mathcal{L}\text{oss}_A) + (1-\lambda)\mathcal{L}\text{oss}_C \qquad (16)$$

where $\mathcal{L}\text{oss}(I, A, C)$ represents the combined loss function of the entire model. In the first 10 epochs, the parameter λ ($0 < \lambda < 1$) is larger so that the accurate feature can be extracted from input image. With the training process, λ is becoming smaller to support better predicted conclusion.

V. Experimental Results

In this section, we validate the proposed model on four aspects to demonstrate significant improvements compared with other methods. The experiments are implemented on the CINDRAL dataset as follows:

- The Implementation details In Training stage.
- Validation of the diagnosis conclusion accuracy (DCA) with the purpose of showing superior performance compared to that of other CNNs and image captioning methods.
- Then we conduct experiments on the semantic attributes prediction accuracy (SAPA) with MDNET to prove our method possesses the ability to generate the same diagnostic report using different methods.
- We use the method of MDNET method in CINDRAL dataset, which would show that the weakness of the method.
- This is followed by conducting the experiment on the different sequences of semantic attributes with the purpose of validation.
- We demonstrate that the attention map concurs with semantic attributes, and visually supports the conclusion.

A. Implementation Details

Our experiment is based upon the open source toolbox Pytorch, and implements the pretrained model for encoding the backbone Resnet18, which makes convergence faster in the training process. With a batch size of 32, we used a single GTX1080 GPU, and cross entropy was used as the loss function. We follow prior work to use the learning rate scheduling $\text{lr} = \text{baselr} * (1 - \frac{iter}{total\_iter})^{power}$. The base learning rate for Adam is set to 0.0001 for our dataset. The momentum is set to 0.9 and weight decay is set to 0.0001. The network is trained for 100 epochs, then we randomly shuffle the training samples, crop them to 500x500 pixels, and rotate them by between ten and fifteen degrees. During data augmentation, we apply 50% random probability to change brightness, contrast, and color channel.

B. Diagnosis conclusion accuracy on CINDRAL

In the computer-aided medical diagnosis process, the pathologist usually expects to get more accurate conclusions, therefore the DCA of evaluation metrics are critical aspects to consider. To validate the effectiveness of the proposed model, it was compared with results obtained by implementing AlexNet, VGG16, ResNet [28], DenseNet [26]. In this setting, the CNNs use the conclusion label. However, our model treats the semantic attributes and conclusion as the label in order to ensure that these attributes contribute significantly to improvements in decision-making. Factors, such as a pre-trained model on ImageNet, were also taken into consideration. Moreover, a comparison with MDNET in DCA, was also conducted, as it is the method with which ours is most similar.

TABLE 1. DIAGNOSIS CONCLUSION ACCURACY ON CINDRAL

Method	Image Classification							
Model	AlexNet		VGG16		Resnet18		Resnet34	
Pre-trained	✔	✘	✔	✘	✔	✘	✔	✘
DCA(%)±std	71.2 ±2.0	71.2 ±2.5	66.0 ±4.3	66.0 ±5.8	78.4 ±2.5	78.6 ±1.5	77.2 ±4.0	78.2± 3.5
Method	Image Classification	Image captioning					**Our**	
Model	Densenst40	MDNET		show and tell			**Our Method**	
Pre-trained	✔	✘	✔	✘	✔	✘	✔	✘
DCA(%)±std	75.0 ±2.4	75.4 ±1.9	78.0 ±2.3	78.2 ±2.2	75.2 ±4.5	74.9 ±3.6	**85.0 ±1.3**	**84.4± 1.1**

Pre-trained CNN model are not available in for this application in the medical im-age domain due to the fact that there is a huge difference in features between the natu-ral images from ImageNet and pathology images. With the increase of model depth, the CNN models achieve an accuracy which is at

best 78% in CINDRAL, which is a similar accuracy as that of MDNET.

Our method achieves a substantially higher accuracy rate of 84% in CINDRAL. The results demonstrate that by treating the semantic attributes as labels, substantially improvements may be realize in performance.

C. *MDNET method result on CINDRAL*

MDNET [12] is a model of image captioning that generates final diagnostic conclusions and obtains the attention maps corresponding to each word, was chosen to establish a baseline for comparison. In order to train MDNET, we modified CINDRAL sentences in into a similar format. The result of MDNET is show in the Figure5.

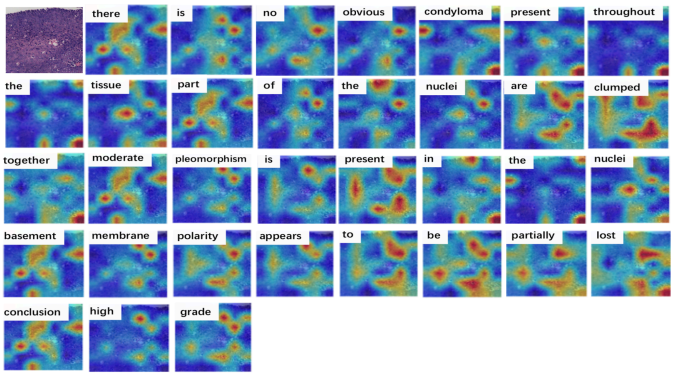

Fig. 5. MDNET attention result in CINDRAL dataset

As we can see from the diagram, MDNET reports contain a large number of words that are not related to diagnostic semantic properties because of the form of natural language. For example, in "there is no obvious condyloma present throughout the tissue", the words that have the most semantic information are "no obvious", and the other words like "throughout", "tissue", and "is" have no useful information. These adverbial phrases, such as "no obvious" are often only used to grammatically supplement other words in a sentence. However, these types of adverbial phrases often bear the most weight, with the remainder of the words making it more difficult for the model to learn its semantics and its corresponding attention region.

D. *Structured Diagnosis Report*

In clinical practice, not only natural language but also visual interpretation is necessary for pathologists to understand the specific symptoms of each semantic attribute. Therefore, this problem is addressed by generating structured diagnostic reports to help pathologists understand basic principles of the model's conclusions. In the experiment, we compare the discriminant information in the structured report in different ways, the semantic attribute prediction accuracy (SAPA) and MDNET.

TABLE 1. SAPA PERFORMANCE COMPARISON WITH MDNET

	Condyloma		Cell Polarity	
Model	MDNET	Our	MDNET	Our
SAPA(%) ±std	72.4±1.6	72.8±1.0	76.0±1.6	75.2±2.3
	Cell Crowding		Pleomorphism	
Model	MDNET	Our	MDNET	Our
SAPA(%) ±std	76.0±1.5	76.0±1.5	78.0±1.1	78.6±1.7

Table 2 shows an average score of more than 5 times. It can be seen that in the four semantic attributes, the accuracy of the results is almost the same as MDNET. In our approach, we implement predictions by treating the problem as a multi-label classification problem, rather than using natural language reports in MDNET to treat the problem as image subtitles. In general, we have proposed a new method for generating diagnostic reports that exhibits the same performance as MDNET.

E. *Different Sequence of Semantic Attributes*

We conducted a comparative experiment with CNN-RNN [29] with the purpose to investigate the influence of different sequences on LSTM. CNN-RNN model also considers the latent relationship between the attribute labels in natural images. Hence in our work, we change the attribute sequence. As an example, we denote the previous sequence, which is type condyloma, polarity, crowd and pleomorphism, as ABCD. Then we compare the DCA after permutation of the order of the sequence, as BCAD, DBCA and CADB.

TABLE 3. INFLUENCE OF SEMANTIC ATTRIBUTE INPUT SEQUENCE ON MODEL PERFORMANCE

Model	Sequence of attributes	DCA(%)±std
CNN-RNN	Previous order(ABCD)	76.6±4.2
	Order 1 (BCAD)	65.4±3.4
	Order 2 (DBCA)	70.2±3.8
	Order 3 (CADB)	70.0±3.1
Our method	Previous order(ABCD)	**85.0±1.7**
	Order 1 (BCAD)	**85.2±2.9**
	Order 2 (DBCA)	**84.4±1.3**
	Order 3 (CADB)	**84.8±1.2**

The results are shown in table 3. Our proposed method outperforms the baseline model by demonstrating significantly improved DCA with the different sequence.

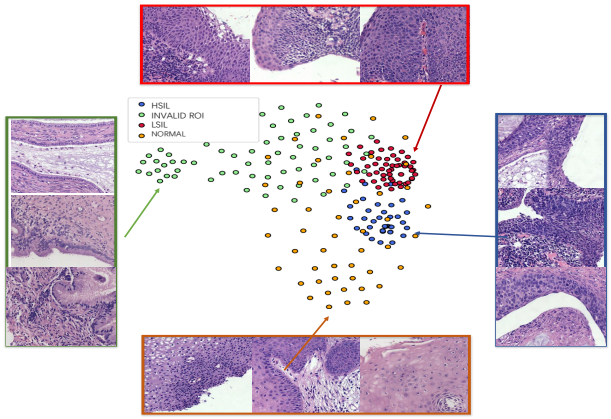

Fig.6. the t-SNE visualization of the features extracted by the Resnet18 module for all test images in our CINDRAL dataset.

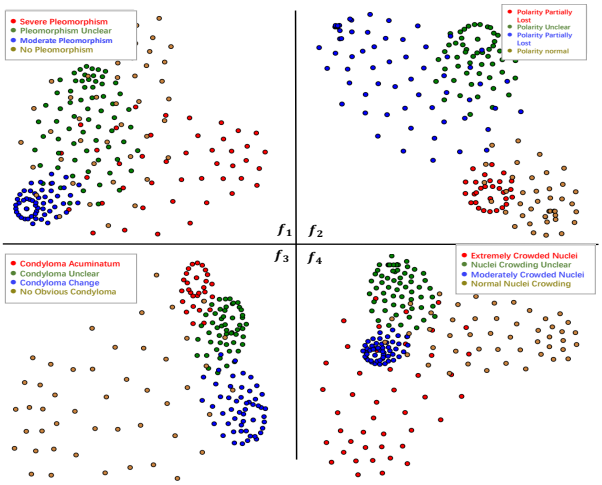

Fig.7. the t-SNE visualization of the features (f_1, f_2, f_3, f_4) extracted by attention module

In this experiment, we also tried to compare the t-SNE visualization results between the classification feature f_0 and the features f_1, f_2, f_3, f_4 which are extracted from our attention module. The first image (Figure 6) shows the t-SNE visualization of the features extracted by the Resnet18 module for all test images in our CINDRAL dataset. The second image (Figure 7) is the t-SNE visualization of the features (f_1, f_2, f_3, f_4) extracted by attention module. All images in the test set, have different distributions of test sample features, based upon different attributes and diagnostic conclusions. Therefore in our method, we map the final diagnostic conclusion space through the feature space under different attributes. Then we use LSTM to transform this mapping process into a form of sequence predictions, which can help us make better use of features to generate acceptable conclusions for clinicians.

F. The attention model with visually interpretation

In this experiment, we show the attention region for each semantic attribute (an example in Figure 8). The four attention models (as Figue.2) compute and show the attention map to interpret how the network supports the diagnostic conclusion. Rather than the attention region for a single word in , we generate four attention maps to support four semantic attributes. Our pathologist draws the region of interest (ROI) which significantly support the decision-making process. Thus we are able to observe the result which expresses strong correspondence between the pathologist annotations and our attention regions for four semantic attributes in the Figure 8. Note that there are no regional annotations in the training stage. Our work demonstrates the model has learned the critical information to support its conclusion.

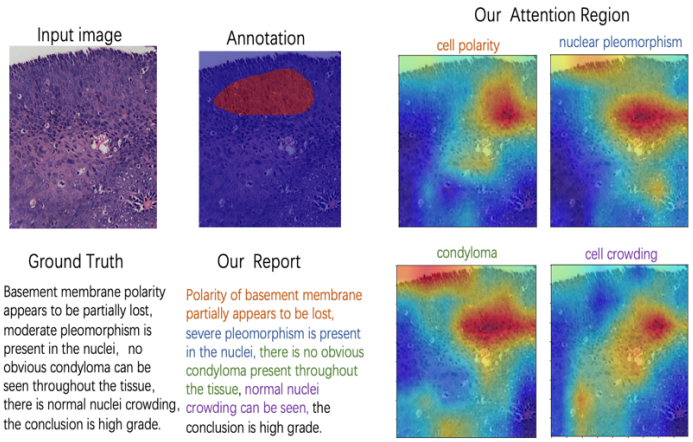

Fig. 8. The illustration of diagnosis report and four semantic attribute attention regions. Best viewed in color.

VI. Conclusion

This paper presents a new approach to interpreting pathology image date, through the generation of structured diagnostic reports with visual interpretation towards the attention map. Specifically, we have proved the efficacy of this method, which treats the four semantic feature as the input of LSTM, which then learns the accurate information to support the conclusion. Additionally, our pathologist also expresses appreciation for the utility of this work in the field of pathology image diagnosis. Experimental results on CINDRAL dataset demonstrate that our proposed deep model can significantly improve the performance in both accuracy and efficiency.

Acknowledgements

This work is supported by the National Key Scientific Instruments and Equipment Development Program of China (2013YQ03065101), the National Natural Science Foundation of China under Grant 61521063 and Grant 61503243.

References

[1] Zhang, X., Su, H., Yang, L., Zhang, S.: Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval. In: Computer Vision and Pattern Recognition. pp. 5361–5368

[2] Chang, H., Zhou, Y., Borowsky, A., Barner, K., Spellman, P., Parvin, B.: Stacked predictive sparse decomposition for classification of histology sections. Interna-tional Journal of Computer Vision 113(1), 3–18 (2015)

[3] Cirean, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: International Confer-ence on Medical Image Computing and Computer-Assisted Intervention. pp. 411–8 (2013)

[4] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.: Imagenet large scale visual recognition challenge. Inter-national Journal of Computer Vision 115(3), 211–252 (2014)

[5] Kisilev, P., Walach, E., Hashoul, S., Barkan, E., Ophir, B., Alpert, S.: Semantic descrip-tion of medical image findings: structured learning approach. In: British Machine Vi-sion Conference. pp. 171.1–171.11 (2015)

[6] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Corri-gendum: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639), 115–118 (2017)

[7] Chartrand, G., Cheng, P.M., Vorontsov, E., Drozdzal, M., Turcotte, S., Pal, C.J., Kadoury, S., Tang, A.: Deep learning: A primer for radiologists. Radiographics 37(7), 2113–2131 (2017)

[8] Zhang, Z., Chen, P., Sapkota, M., Yang, L.: TandemNet: Distilling Knowledge from Medical Images Using Diagnostic Reports as Optional Semantic References (2017)

[9] Kisilev, P., Walach, E., Barkan, E., Ophir, B.: From medical image to automatic medical report generation. Ibm Journal of Research Development 59(2/3), 2:1–2:7 (2015)

[10] Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports (2017)

[11] MURDOCK, B. B. (1970). SHORT- AND LONG-TERM MEMORY FOR ASSOCIATIONS. Biology of Memory, 11–13.

[12] Zhang, Z., Xie, Y., Xing, F., Mcgough, M., Yang, L.: Mdnet: A semantically and visually interpretable medical image diagnosis network pp. 3549–3557 (2017)

[13] Wang, Z., Chen, T., Li, G., Xu, R., Lin, L.: Multi-label image recognition by recurrently discovering attentional regions. In: IEEE International Conference on Computer Vi-sion. pp. 464–472 (2017)

[14] Shi, X., Xing, F., Xie, Y., Su, H., & Yang, L. Cell Encoding for Histopathology Im-age Classification (pp.30-38). (2017).

[15] Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching pp. 2156–2164 (2016)

[16] Pedersoli, M., Lucas, T., Schmid, C., Verbeek, J.: Areas of attention for image captioning pp. 1251–1259 (2017)

[17] Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. Computer Science (2015)

[18] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. Com-puter Science pp. 2048–2057 (2015)

[19] Yu, D., Fu, J., Mei, T., Rui, Y.: Multi-level attention networks for visual question an-swering. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4187–4195 (2017)

[20] Lu, J., Xiong, C., Parikh, D., & Socher, R. Knowing when to look: adaptive attention via a visual sentinel for image captioning. pp. 3242-3250 (2016)

[21] Shin, H.C., Roberts, K., Lu, L., Demnerfushman, D., Yao, J., Summers, R.M.: Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation pp. 2497–2506 (2016)

[22] Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding net-work for common thorax disease classification and reporting in chest x-rays (2018)

[23] Everingham, M., Winn, J.: The pascal visual object classes challenge 2010 (voc2010) development kit contents. In: International Conference on Machine Learning Chal-lenges: Evaluating Predictive Uncertainty Visual Object Classification. pp. 117–176 (2011)

[24] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computer Science (2014)

[25] Ma, K., Wu, K., Cheng, H., Gu, C., Xu, R., & Guan, X. (2018, December). A Pathology Image Diagnosis Network with Visual Interpretability and Structured Diagnostic Re-port. In International Conference on Neural Information Processing (pp. 282-293). Springer, Cham.

[26] Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)

[27] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)

[28] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition pp. 770–778 (2015)

[29] Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification pp. 2285–2294

[30] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Com-puter Vision. pp. 740–755 (2014)

[31] Krizhevsky, A., Sutskever, I., & Hinton, G. E. ImageNet classification with deep convo-lutional neural networks. pp.1097-1105 (2012)